Ref. Ares(2026)4470225 - 30/04/2026

D9.2

Report initial pilot – Healthcare LSP

Related Work Package	WP9 – Implementation of healthcare LSP
Related Task	Task 9.2 – Initial Pilot
Lead Beneficiary	FORTH
Contributing Beneficiaries	PAGNH, UNIPI, CNR
Document version	v.1.0
Deliverable Type	R
Distribution level	PU
Contractual Date of Delivery	30.04.2026
Actual Date of Delivery	30.04.2026

Authors	K. Marias (FORTH); G. Kalliatakis (FORTH); K. Vardakas (FORTH); D. Zaridis (FORTH); T. Tsoumanis (FORTH); V. Pezoulas (FORTH); N. Tachos (FORTH); Francesca Pia Caputo (UNIPI); Dania Cioni (UNIPI); Sara Colantonio (CNR); Danila Germanese (CNR); Maria Antonietta Pascali (CNR).
Contributors	
Reviewers	SINTEF, AOA

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.

Version history

Version	Description	Date completed
v0.1	TOC preparation	26/02/2026
v0.2	Section 1 and Section 2	13/03/2026
v0.3	Early draft	02/04/2026
v0.4	Final draft to internal peer review	23/04/2026
V1.0	Ready for EC portal submission	30/04/2026

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer

This document contains material, which is the copyright of one or more FAITH consortium parties, and may not be reproduced or copied without permission.

All FAITH consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the FAITH consortium as a whole, nor individual FAITH consortium parties, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, accepting no liability for loss or damage suffered by any person using this information.

Executive summary

This report documents the execution, evaluation, and results of the initial pilot of the FAITH Large-Scale Pilot 6 (LSP6) — the Healthcare Large-Scale Pilot — conducted in the context of Task 9.2 (T9.2: Initial Pilot, M3–M26). Building upon the coordination and experimental infrastructure established in D9.1, this deliverable presents the outcomes of the prospective clinical deployment of the AI System Under Test (AI SUT) at the University Hospital of Heraklion, Greece (PAGNH), under the oversight of FORTH, and with contributions from PAGNH, UNIPI, and CNR.

Purpose and Scope of the Initial Pilot. The primary objective of the initial pilot is the real-world clinical evaluation of a deep learning-based prostate gland segmentation model deployed within the radiology workflow of PAGNH. The AI SUT is designed to replace the second reader of multiparametric MRI (mpMRI) images, targeting the delineation of the whole prostate gland, peripheral zone, and transition zone from T2-weighted MRI scans. In doing so, the pilot addresses the persistent clinical challenges of intra- and inter-observer variability in prostate segmentation, seeking to improve diagnostic reproducibility, reduce radiologist workload, and contribute to cost-efficient cancer imaging workflows.

AI System Under Test — Performance and Advancement. Entering the initial pilot at Technology Readiness Level (TRL) 5 with a whole-gland Dice Similarity Coefficient (DSC) of 0.93, the AI SUT has been further refined prospectively on cases acquired at PAGNH during the pilot period. Zonal segmentation performance stands at DSC 0.89 for the transition zone and DSC 0.83 for the peripheral zone, utilising a nnU-Net backbone architecture. Prospective model retraining on pilot-collected mpMRI cases — drawn from approximately 9.4 patient cases per month as per the project's target throughput — has incrementally strengthened generalisation across MRI vendors and field-strength variability (1.5 T and 3 T), directly targeting KPI2. The model is on track for advancement to TRL 6 upon completion of this phase.

Trustworthiness Assessment via the FAITH AI_TAF. The initial pilot operationalises the FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF) for the first time within a live clinical environment. Trustworthiness assessment activities have been structured across seven dimensions — safety, security and resilience, bias mitigation, explainability and interpretability, accountability and transparency, privacy enhancement, and validity and reliability — in direct alignment with the NIST AI Risk Management Framework and EU AI Act requirements applicable to Software as a Medical Device (SaMD). The completion of this trustworthiness evaluation cycle, generating a documented FAITH AI_TAF risk management profile for the healthcare domain, constitutes KPI3 of LSP6, with its outputs directly informing the replication pilot at UNIPI.

Clinical Validation and KPI Attainment. Evidence gathered through the structured clinical evaluation protocol at PAGNH — combining quantitative segmentation metrics, expert radiologist feedback via the embedded trustworthiness questionnaire, and qualitative usability assessments — is reported herein. Results indicate that more than 90% of AI-generated segmentations successfully exclude extra-prostatic tissues in line with KPI1. Stakeholder feedback has confirmed the clinical utility and interpretability of the confidence heatmap visualisations integrated into

the OHIF DICOM viewer. Regulatory and ethical compliance remains in force throughout, with the Greek Institutional Review Board (IRB) approval underpinning all data processing activities in accordance with GDPR and the EU Medical Device Regulation (MDR 2017/745).

Data and Infrastructure. The AI SUT has operated as a fully containerised, on-premise deployment at PAGNH throughout the pilot, maintaining a zero-data-transfer-outside-premises policy consistent with the project's data management framework. Retrospective data supplemented by continuous prospective acquisition has provided the clinical dataset for both model retraining and trustworthiness evaluation. Linkage to the EUCAIM European Cancer Imaging Data Space has been initiated in partial fulfilment of KPI5, with FORTH's dual role as EUCAIM partner facilitating data interoperability for secondary use.

Stakeholder Engagement. User and stakeholder engagement has been sustained throughout the pilot via fortnightly coordination meetings, dedicated training sessions for radiologists and clinical technical personnel, and structured focus groups to prioritise trustworthiness dimensions. Radiologist input has confirmed that transparency, robustness, and explainability are the most salient trustworthiness requirements in this clinical context, consistent with prior findings from the ProCancer-I project.

Next Steps. The results and trustworthiness risk profile generated in this deliverable form the definitive input for the replication pilot at the University Hospital of Pisa (UNIPI), led by UNIPI with contributions from FORTH, PAGNH, and CNR (T9.3, M20–M48). Key priorities for the replication phase include cross-site validation of generalisation performance, expansion of the EUCAIM data linkage, advancement to TRL 7, and finalisation of the FAITH AI_TAF profile towards CE Mark readiness as Software as a Medical Device.

Table of Contents

1. Introduction 14
1.1. Scope and objectives of the 1st pilot in the Healthcare LSP 14
1.2. Structure of the document 15
1.3. Evolution of the AI System Under Test (SUT) from initial setup 15
2. Progress on the AI system deployment(s) and execution of the Pilots 17
2.1. Status of the 1st Pilot execution 17
2.1.1 Deployment Context & Domain Specifics 18
2.1.2 Training 18
2.1.3 Users 18
2.1.4 Evaluation Process 19
2.2. Status of the Replication Pilot preparation/initialization 19
2.3. Deviations from the original deployment plan 20
2.4. Data collection throughout the 1st Pilot execution 21
2.5. Integration into LSP workflows: Lessons learned 23
2.5.1 Clinical Workflow Integration 23
2.5.2 Model Performance and Failure Mode Analysis 24
2.5.3 Stakeholder Engagement and Radiologist Feedback 24
2.5.4 Summary of Lessons Learned and Feed-Forward Actions 25
3. Application of FAITH Tools 26
3.1. Application of the FAITH AI TAF V1 methodology 26
3.1.1 Cartography 27
3.1.2 Threat Analysis 28
3.1.3 Impact Assessment 28
3.1.4 Vulnerability Assessment 29
3.1.5 Risk Analysis 29
3.1.6 Countermeasures 30

3.2 Tool Validation Planning & Training Processes 30
3.2.1 Dry-run 30
3.2.2 Plan development 31
3.2.3 Tool Training 33
3.3 FAITH Tool Validation Activities 34
3.3.1 TrustSense 34
3.3.2 Data & AI Model Passport in the FAITH AI Model Hub 36
3.3.2.1 AI Model Passport 37
3.3.2.1.1 Summary 37
3.3.2.1.2 Details 38
3.3.2.1.3 Pipeline 39
3.3.2.2 Data Passport 43
3.3.2.2.1 Summary 44
3.3.2.2.2 Authorship 44
3.3.2.2.3 Motivation 45
3.3.2.2.4 Data use 46
3.3.2.2.5 Distribution 46
3.3.2.2.6 Security 48
3.3.3 TrustGuard 48
3.4. Threat Analysis 50
3.4.1. Model 1 50
3.4.1.1 Threat Assessment - Model 1 50
3.4.1.2 Consequence Assessment - Model 1 52
3.4.1.3 Vulnerability Assessment - Model 1 54
3.4.2. Model 2 56
3.4.2.1 Threat Assessment - Model 2 56
3.4.2.2 Consequence Assessment - Model 2 57
3.4.2.3 Vulnerability Assessment - Model 2 59
3.5. Initial Risk Assessment results 61

3.5.1. Model 1 61
3.5.2. Model 2 62
3.6. Countermeasures 64
3.6.1. Model 1 64
3.6.2. Model 2 65
3.7 Formal Tool Feedback 67
3.7.1 TrustSense 67
3.7.2 AI Model Hub 67
3.7.3 TrustGuard 67
4. AI System Evaluation 68
4.1. Quantitative metrics on pilot data 68
4.2. Trustworthiness Assessment 70
4.2.1 Validity & Reliability 70
4.2.2 Security & Resilience 71
4.2.3 Transparency & Accountability 71
4.2.4 Explainability & Interpretability 71
4.2.5 Safety 72
4.2.5.1 Adversarial Attack Assessment Methodology 73
4.2.5.2 Key Findings 73
4.2.5.3 Clinical and Safety Implications 75
4.2.5.4 Planned Adversarial Defense: Phase 2 — Adversarial Training 75
4.2.6 Fairness 75
4.2.7 Privacy 76
4.3 Preliminary Impact Evaluation 76
5. Regulatory and Ethical Assessment 78
5.1. Pathway to CE Marking 78
5.1.1 SaMD Qualification and Classification 78
5.1.1.1 Qualification as Medical Device Software 79
5.1.1.2 MDR Risk-Based Classification Framework 79

5.1.1.3 Application of Rule 11 to the AI SUT....80
5.1.1.4 Indicative Classification for the Healthcare LSP AI System....80
5.1.1.5 Regulatory Implications of the Expected Classification....81
5.1.2 Pre-market Requirements....82
5.1.3 Post-Market Surveillance Considerations....83
5.2 Alignment with the EU AI Act....84
5.2.1 Risk Management and Quality Assurance....84
5.2.2 Data Governance and Dataset Quality....85
5.2.3 Transparency and Human Oversight....85
5.2.4 Technical Documentation and Record Keeping....86
5.3 Ethical Considerations for AI Deployment in Clinical Environments....86
5.3.1 Patient Privacy and Data Protection....87
5.3.2 Transparency and Explainability....87
5.3.3 Bias Mitigation and Fairness....88
5.4 Regulatory gap analysis: MDR vs EU AI Act vs current SUT status....88
6. Conclusions and Roadmap for the Final Pilot Phase....92
6.1. Summary of achievements in the second period....92
6.1.1 Key achievements....92
6.1.2 Clinicians feedback....93
6.2 Plan for Replication Phase....94
6.3. Identified risks for the Replication Phase....95
7. References....96

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
AI_TAF	Artificial Intelligence Trustworthiness Assessment Framework
AI SUT	Artificial Intelligence System Under Test
ASD	Average Surface Distance
bpMRI	bi-parametric Magnetic Resonance Imaging
CE	Conformité Européenne
CER	Clinical Evaluation Report
CI/CD	Continuous Integration / Continuous Deployment
CTC	Clinical Trial Centre
CTP	Clinical Trial Processor (anonymisation tool)
DAM	Medical Research Area Directorate (Direzione Area Medica)
DBDM	Data Bias Detection and Mitigation
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DPO	Data Protection Officer
DPIA	Data Protection Impact Assessment
DSC	Dice Similarity Coefficient
EUCAIM	European Cancer Imaging Data Space
FDA	Food and Drug Administration
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GSPR	General Safety and Performance Requirements
HCP	Health-Care Professional
HD	Hausdorff Distance
HIPAA	Health Insurance Portability and Accountability Act
IEC	International Electrotechnical Commission
IFU	Instructions For Use
IRB	Institutional Review Board
ISO	International Organization for Standardization
IT	Information Technology
KPI	Key Performance Indicator
LIME	Local Interpretable Model-agnostic Explanations
MDR	Medical Device Regulation (EU) 2017/745
MDCG	Medical Device Coordination Group
MDSW	Medical Device Software
MLOps	Machine Learning Operations
mpMRI	multi-parametric Magnetic Resonance Imaging

MRI	Magnetic Resonance Imaging
NIST	National Institute of Standards and Technology
OHIF	Open Health Imaging Foundation
PACS	Picture Archiving and Communication System
PI-CAI	Prostate Imaging: Cancer AI (dataset)
PI-QUAL	Prostate Imaging Quality
PII	Personally Identifiable Information
PMCF	Post-Market Clinical Follow-Up
PMS	Post-Market Surveillance
PSA	Prostate-Specific Antigen
QMS	Quality Management System
RBAC	Role-Based Access Control
RIS	Radiology Information System
RMF	Risk Management Framework
SaMD	Software as a Medical Device
SHAP	SHapley Additive exPlanations
SSL	Secure Sockets Layer
SUT	System Under Test
T2w	T2-weighted (MRI sequence)
TRL	Technology Readiness Level
TURP	Transurethral Resection of the Prostate
TWA	Trustworthy AI (characteristic/dimension)
UI	User Interface

List of Tables

Table 1. Expert qualitative evaluation of AI-generated prostate segmentations (N = 179). 22
Table 2. Clinical conditions associated with Good-rated segmentations (commented cases, n = 11). 22
Table 3. Summary of lessons learned from the WP9 1st pilot iteration (PAGNH) and feed-forward actions for the 2nd pilot iteration at UNIPI (T9.3). 25
Table 4. A summary of the data and AI model passports related to the AI SUT in the healthcare LSP. 36
Table 5. Metrics used for the validation of the AI SUT in the healthcare LSP. 68
Table 6. Regulatory gap analysis for the Healthcare LSP AI SUT. 89

List of Figures

Figure 1. An illustration of the AI SUT. 16
Figure 2. Detailed implementation plan for LSP6 – Healthcare pilot, including planning, training, deployment, and FAITH tool validation activities. 32
Figure 3. TrustSense assessment results for the radiologist end-user group (Domain AI User Team), indicating a Very High trustworthiness level with a score of 4.65. 35
Figure 4. TrustSense assessment results for the AI technical team, indicating a High trustworthiness level with a score of 4.10. 36
Figure 5. Summary section of the AI Model Passport for the prostate zonal segmentation model, illustrating identity and governance metadata including unique identifier, pilot context, publishing organization, verification status, and high-level model description within the FAITH ecosystem. 38
Figure 6. Details section of the AI Model Passport for the prostate zonal segmentation model, illustrating descriptive and technical metadata including algorithm, task type, learning method, model purpose, authorship, source repository, and governance-related information within the FAITH ecosystem. 39
Figure 7. Preprocessing stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, showing dataset preparation metadata, MRI preprocessing parameters, validation identifiers, and associated input/output artifact references. 40
Figure 8. Training stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, capturing learning and optimization metadata including hyperparameters, network architecture details, artifacts, and execution environment information. 41
Figure 9. Evaluation stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, showing validation metrics and associated artifacts and environment metadata. 42
Figure 10. Production stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, capturing the final deployable artifact with its URI, checksum, path location, implementation files, and environment metadata. 43
Figure 11. Summary section of the Data Passport for the ProCAncer-I UC1 dataset, capturing core dataset identifiers including title, acronym, version, unique ID, pilot context, publication dates, and a high-level description supporting governance and traceability within the FAITH ecosystem. 44
Figure 12. Authorship section of the Data Passport for the ProCAncer-I UC1 dataset, capturing governance and provenance metadata including publisher, dataset owner, and funding source, establishing stewardship and institutional responsibility. 45
Figure 13. Motivation section of the Data Passport for the ProCAncer-I UC1 dataset, describing intended use cases, including research and healthcare AI applications, and outlining usage limitations and constraints. 46
Figure 14. Data Use section of the Data Passport for the ProCAncer-I UC1 dataset, linking the dataset to AI workflows and indicating its use in training and supervised learning contexts. 46

Figure 15. Distribution section of the Data Passport for the ProCAncer-I UC1 dataset, detailing access conditions, licensing, provenance, data format, and key dataset characteristics such as size, annotations, and clinical origin. 47
Figure 16. Field-level dataset description within the Distribution section of the Data Passport for the ProCAncer-I UC1 dataset, providing detailed metadata for variables including names, descriptions, data types, units, allowed values, and semantic references. 47
Figure 17. Security section of the Data Passport for the ProCAncer-I UC1 dataset, capturing privacy and governance safeguards including access controls, authentication mechanisms, pseudonymisation, consent management, data protection oversight, and indications of sensitive personal data. 48
Figure 18. TrustGuard cartography stage for the zonal prostate segmentation model (Model 2), showing asset identification, lifecycle stage, and assessment setup. 49
Figure 19. TrustGuard criticality assessment applied to the healthcare AI system, resulting in a Very High criticality classification due to the medical domain context and potential impact on health-related decision-making. 49
Figure 20. TrustGuard assessment results interface, presenting trustworthiness characteristics evaluated for the zonal prostate segmentation model. 50
Figure 21. MDR SaMD Qualification and classification pathway (Healthcare LSP AI SUT). 82

1. Introduction

1.1. Scope and objectives of the 1st pilot in the Healthcare LSP

The present deliverable reports on the implementation and initial outcomes of the initial pilot of the FAITH Healthcare Large Scale Pilot (LSP6). The Healthcare LSP is one of the seven domain-specific pilots of the FAITH project and focuses on the application and validation of the FAITH Artificial Intelligence Trustworthiness Assessment Framework (FAITH AI_TAF) in a real healthcare environment. Within this context, LSP6 addresses the trustworthy deployment of artificial intelligence for medical image analysis, using prostate magnetic resonance imaging (MRI) as the representative clinical use case.

The pilot is coordinated by FORTH, with the active contribution of PAGNH, UNIPI, and CNR. The clinical validation activities involve the University Hospital of Heraklion (PAGNH) and the University Hospital of Pisa / AOUP, thereby enabling assessment of the AI solution across two representative European hospital environments. The AI System Under Test (SUT) concerns deep learning models for automated prostate segmentation from T2-weighted MRI, including whole-gland and zonal segmentation functionalities.

The Healthcare LSP provides a particularly relevant setting for the FAITH project, as healthcare is a high-sensitivity domain where the deployment of AI systems requires not only technical performance, but also strong guarantees in relation to safety, reliability, explainability, privacy, fairness, accountability, and regulatory compliance. The pilot therefore serves both as a technical validation exercise and as a real-world demonstration of how the FAITH methodology and tools can support trustworthy AI adoption in clinical practice.

This deliverable focuses on the initial pilot, during which the AI solution was deployed in a controlled clinical setting, pilot users were trained, trustworthiness assessment activities were initiated, and initial evidence was collected regarding usability, operational feasibility, and alignment with the FAITH framework. The results of this phase will inform the subsequent replication activities in the second clinical site and the continued maturation of the healthcare use case.

The scope of the initial pilot in the Healthcare LSP concerns the deployment and initial validation of an AI-based clinical support solution for automated prostate MRI segmentation within real hospital environments. More specifically, the pilot focuses on the use of deep learning models for segmenting the prostate gland and internal anatomical zones from T2-weighted MRI scans, with outputs reviewed by clinical experts through a dedicated visualisation environment.

Beyond technical model performance, the initial pilot also addresses the broader trustworthiness dimensions of the deployed AI system, in line with the objectives of the FAITH project. This includes the use of the FAITH AI_TAF v1 methodology and supporting tools to assess organisational readiness, system risks, governance maturity, transparency, and user trust.

The core objectives of the initial pilot are the following:

- Deployment of the SUT within a controlled clinical environment using local hospital infrastructure.
- Validation of operational feasibility of the solution in realistic radiology workflows.
- Assessment of clinical usability through interaction with radiologists and end users.
- Evaluation of trustworthiness characteristics including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy and fairness with harmful bias managed.
- Application of the FAITH AI_TAF v1 methodology and tools, specifically TrustSense, TrustGuard, and the AI Model Hub to support structured assessment and governance.
- Collection of user feedback and lessons learned to refine the solution and prepare the replication pilot.
- Preparation for cross-site deployment in the second clinical environment at the University Hospital of Pisa.
- Through these objectives, the initial pilot establishes the initial evidence base for the trustworthy use of AI-enabled prostate MRI analysis in healthcare settings.

1.2. Structure of the document

The remainder of this deliverable is organised into five main sections. **Section 2** presents the progress of the AI system deployment and execution of the initial pilot. It describes the deployment context, training activities, users involved, evaluation process, replication pilot preparation, data collection, and lessons learned from integration into clinical workflows. **Section 3** describes the application of the FAITH methodology and tools within the Healthcare LSP. It includes the use of the FAITH AI_TAF v1 methodology, planning and training activities, validation of TrustSense, TrustGuard, and the AI Model Hub, as well as the initial threat, risk, and countermeasure analysis. **Section 4** presents the evaluation of the AI SUT, including technical model performance, user-centred trustworthiness assessment, and preliminary impact observations from the initial pilot phase. **Section 5** addresses regulatory, ethical, and compliance considerations relevant to the healthcare deployment, including alignment with the EU AI Act, medical device requirements, privacy legislation, and related governance aspects. **Section 6** concludes the deliverable by summarising the main achievements of the initial pilot and outlining the roadmap for the replication phase at the second clinical site.

1.3. Evolution of the AI System Under Test (SUT) from initial setup

The AI SUT is a tool meant to support the activities of FAITH's WP9. Utilising T2-weighted magnetic resonance imaging (MRI) scans, it is designed to precisely define the prostate gland and its anatomical zones, including the complete gland, peripheral zone, and transition zone. The

algorithm generates segmentation masks that emphasise specific anatomical locations within the prostate by first anonymising the initial MRI scans. These segmented images are then superimposed on the original MRI scans, providing physicians with a visual representation of the prostate's anatomy that is enhanced by model-generated probability heatmaps that illustrate the AI's confidence in various areas. This feedback enables clinicians to immediately evaluate the accuracy and reliability of the segmentations. Initially, it offers real-time segmentation overlays on MRI images, thereby improving interpretability and facilitating interactive review during diagnosis evaluations. Secondly, healthcare professionals can easily upload MRI data, analyse segmentation results, and interact with the tool without the need for a high level of technical expertise, thanks to a user-friendly interface. Users have the ability to assess the results in the OHIF DICOM viewer and a Custom Heatmaps Visualiser endpoint, where they can examine the overlays and probability heatmaps after each segmentation is complete. This observer provides a user-friendly and adaptable implementation that facilitates the analysis while users are subsequently requested to complete a questionnaire that collects feedback on the trustworthiness of the tool. The AI SUT test can be securely implemented on-site to ensure compliance with healthcare data privacy and security standards due to its containerisation process and no need for internet connection. Fig. 1 presents the stakeholders (Clinical experts), the tech stack of the AI SUT and the outcomes along with the already generated passport.

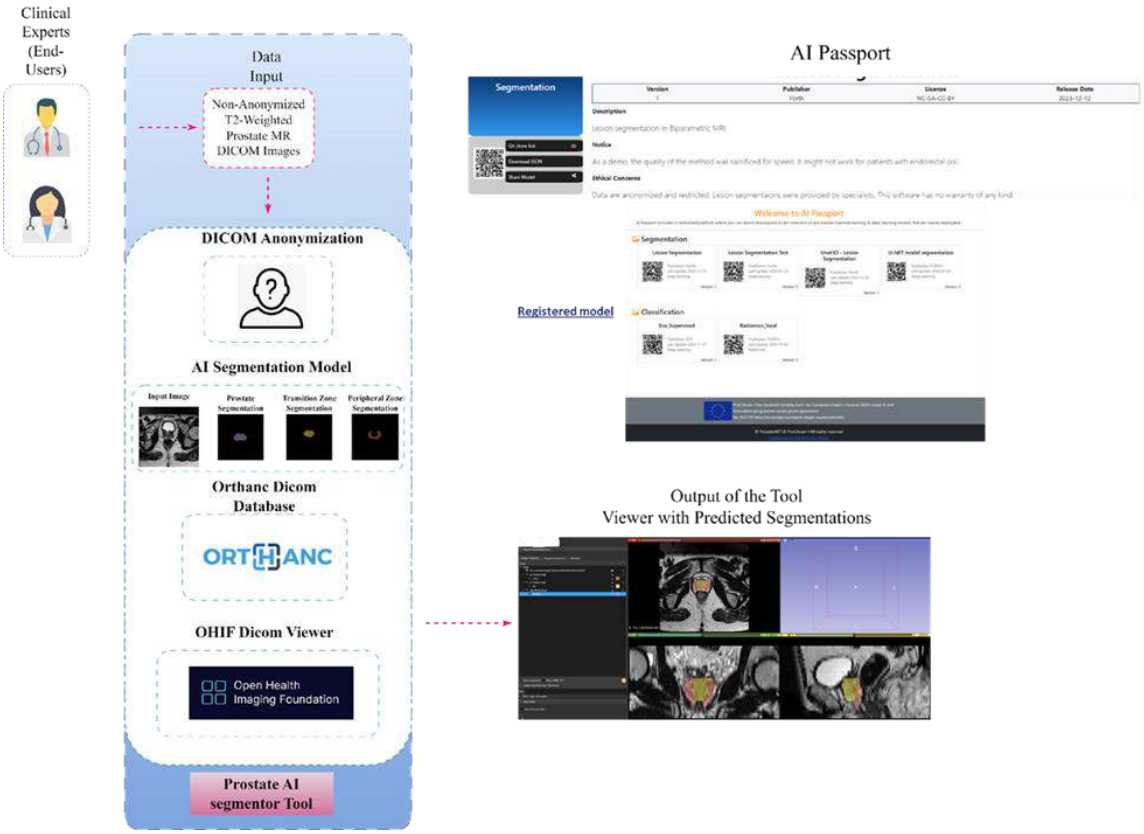


Figure 1. An illustration of the AI SUT.

Since the initial setup, the SUT has evolved in two main ways: (i) integration of a zonal segmentation model, and (ii) retraining of the whole-gland segmentation model with a revised dataset split to facilitate the generation of consistent and complete data passports.

The zonal segmentation model has been integrated into the AI SUT and made available through the FAITH Model Hub. This model performs segmentation of prostate anatomical zones from T2-weighted MRI scans, enabling the delineation of internal prostate regions that are relevant for clinical analysis. The model was trained on the PI-CAI Challenge: Public Training and Development Dataset v2.0, which provides annotated multiparametric prostate MRI scans for the development and benchmarking of AI models in prostate cancer research. Validation was performed using the ProCAncer-I UC1: Dataset for Prostate Cancer Detection, enabling an independent assessment of the model's segmentation performance. The zonal segmentation model achieved a Dice Score of 0.87, Jaccard Index of 0.78, Precision of 0.87, and Recall of 0.88, demonstrating strong agreement between predicted segmentation masks and reference annotations.

In parallel, the initial whole-gland segmentation model was retrained while preserving the same segmentation objective. In the revised configuration, the ProCAncer-I UC1 dataset was used for training, while validation was performed on the full Prostate158 dataset, including both the training and testing subsets. Following retraining, the model achieved a Jaccard Index of 0.84 and a Dice Score, Precision, and Recall values of 0.91, demonstrating high overlap and consistency between the predicted whole-gland segmentation masks and the corresponding ground-truth annotations. Furthermore, both models (Prostate WG and Zonal segmentation) have their respective AI Model and Data Passport in place.

2. Progress on the AI system deployment(s) and execution of the Pilots

2.1. Status of the 1st Pilot execution

The initial pilot of the Healthcare LSP has progressed from the planning phase to implementation and is currently being conducted in a real clinical environment for the assessment of the AI SUT in prostate mpMRI analysis. The pilot involves the clinical settings of the University Hospital of Heraklion (PAGNH) and the University Hospital of Pisa / AOUP, under the coordination of UNIPI, thereby supporting validation across two representative hospital sites.

At the present stage, the pilot is operationally active at PAGNH, where execution has been enabled through the completion of the main organizational, ethical, and technical prerequisites. Ethical approval for this first clinical site was granted on 27/01/2025, following submission of the relevant IRB application on 25/09/2024. In parallel, the Pisa site has advanced the preparation of the corresponding legal and organizational arrangements required for subsequent replication activities. From a system deployment perspective, the pilot concerns the local deployment of the AI SUT as a clinical support solution for prostate MRI segmentation, visualisation, and clinician

review. In this configuration, PAGNH serves as the validation site for the initial pilot, while Pisa/AOUP provides the setting for subsequent cross-site replication pilot and comparative assessment.

2.1.1 Deployment Context & Domain Specifics

The initial pilot is situated in the healthcare domain and, more specifically, within the radiology workflow for the analysis of prostate T2-Weighted MR images. In this context, the AI SUT is deployed as a clinical decision-support component for the automated segmentation of the prostate gland and relevant anatomical zones, while maintaining expert oversight over the interpretation and use of the generated outputs. The deployment takes place in a high-sensitivity environment characterised by the use of medical imaging data, strict ethical and regulatory requirements, and the need for secure and trustworthy system operation. For this reason, the SUT has been designed for local, on-premise deployment within controlled hospital infrastructures.

The domain-specific relevance of the pilot is linked to a well-recognised clinical challenge, namely that prostate segmentation in mpMRI is both time-consuming and subject to intra- and inter-observer variability. Within this setting, the pilot provides an appropriate environment for assessing not only the technical performance of the AI SUT, but also its practical suitability for integration into clinical workflows under real clinical, ethical, and organisational conditions.

2.1.2 Training

Training of end users was carried out to ensure that clinical participants could interact with the AI SUT in a consistent and informed manner during pilot execution. The training focused on familiarising users with the operational workflow of the system, including preparation of MRI input data, execution of the analysis, review of segmentation outputs in the visualisation environment, and completion of the trustworthiness assessment questionnaire.

Training activities also formed part of the broader stakeholder engagement process, through dedicated sessions aimed at presenting and testing the functionalities of the SUT while collecting feedback and requirements from clinical users and other relevant stakeholders. In this context, end-user training supported the effective use of the system and the collection of structured feedback relevant to trustworthiness assessment and future pilot refinement.

2.1.3 Users

The primary users of the AI SUT in the Healthcare LSP are clinical professionals involved in the radiological assessment of prostate T2-Weighted MRI, in particular radiologists. They interact with the system during image review, assess the AI-generated segmentation outputs within the visualisation environment and provide overall feedback on the AU SUT. Both the assessment of the outputs and the overall feedback was provided through a questionnaire. In addition to

supporting routine image assessment, the pilot allows expert users to contribute clinical judgement regarding the quality and usability of the generated segmentations.

Across the pilot activities, the user group spans the participating clinical sites in Heraklion and Pisa and includes radiologists with different levels of experience, ranging from senior clinical experts to radiology residents. This diversity supports evaluation of the AI SUT under realistic clinical conditions and enables feedback from users with different degrees of familiarity with prostate mpMRI and AI-supported image analysis.

2.1.4 Evaluation Process

The evaluation process in the Healthcare LSP is based on clinical interaction with the AI SUT during the review of prostate mpMRI cases. Following data preparation and execution of the analysis workflow, users examine the generated segmentation outputs in the dedicated visualisation environment, including overlays and probability heatmaps, and assess the results from a clinical perspective.

The evaluation is further supported through completion of a trustworthiness questionnaire, through which users provide structured feedback on aspects such as correctness, consistency, usability, and overall trustworthiness. Critically, this user-centred assessment directly supports the Description of Action (DoA) requirement for human expert involvement in model training and tuning: expert radiologists contribute anatomical ground truth annotations used for model fine-tuning, whilst their qualitative feedback on segmentation outputs — collected systematically during pilot use — informs iterative model refinement ahead of the replication pilot at UNIPI. In this way, the evaluation process combines system use with user-centred assessment, enabling the collection of evidence relevant both to validation in the first pilot and to the refinement of subsequent pilot activities across the participating clinical sites.

2.2. Status of the Replication Pilot preparation/initialization

The replication pilot has been planned and implemented at the University Hospital of Pisa (AOUP) as the replication piloting of the Healthcare LSP, with the objective of extending the validation of the FAITH AI SUT to a second clinical environment. In line with the Description of Action, this phase replicates the implementation and evaluation logic of the initial pilot to further assess the trustworthiness characteristics identified during the initial deployment and to examine the generalizability of the FAITH AI_TAF framework under a different operational setting. The replication activities are coordinated by UNIPI and form part of the broader iterative validation pathway foreseen for the healthcare use case.

At the current stage, UNIPI has started the preparation and initialization of the replication pilot by establishing the required organizational, legal, ethical and technical conditions for operational rollout at the University Hospital of Pisa

While defining the implementation of pilot, the workflow in UNIPI has been defined in UNIPI, covering patient selection and enrolment, image extraction from PACS, pseudonymization, secure transfer, and AI-based segmentation test and evaluation.

The available imaging base has been reviewed in order to confirm the feasibility of the study at scale. In parallel, preparatory work has addressed data-sharing arrangements, IRB-related documentation and data protection planning, as well as the organization of user participation, with 4 senior radiologists, 1 junior radiologist and 6 radiology residents, foreseen for the replication study depending on the outcomes of the initial pilot. Although the clinical activities cannot start without the local legal and ethical approval, the AI SUT tool has been installed in UNIPI, with the support of CNR and FORTH. Since the tool was originally optimized for the Windows operating system, and UNIPI uses a local machine running macOS, FORTH contributed to adapting the tool and implementing the necessary changes. With the SUT now fully installed on the clinical site, UNIPI personnel were trained in the use of the tool, and the UNIPI research team prepared training materials to guide radiologists in its use. An initial training of the radiologists has been completed to ensure they know exactly how to run the system, during which the tool has been explored by radiologists with test mpMRI images.

2.3. Deviations from the original deployment plan

The healthcare pilot remained overall aligned with the deployment rationale defined in the FAITH DoA, namely the deployment and evaluation of the AI-based prostate MRI segmentation solution within the radiology workflow of the participating hospital environments, with an initial pilot phase followed by a replication phase in a second clinical setting.

No substantial deviation was identified with respect to the scope of the use case, the participating clinical environments, or the intended pilot objectives. The main deviation concerned the operational deployment model adopted for the implementation of the solution. While the DoA describes deployment within the clinical workflow of the hospital setting, the implemented solution was delivered as a standalone local application based on Electron and Docker, rather than as a more tightly embedded component of the existing hospital IT environment. In practice, the application instantiated its own local technical environment, including dedicated services for DICOM management, image viewing, anonymisation, segmentation, and visualisation. This resulted in a more self-contained and portable deployment setup than initially envisaged. With respect to the planned TRL advancement, this architectural choice is considered compatible with the progression from the current TRL 5 towards TRL 6, as it enables rigorous validation of the AI model in an operationally relevant clinical environment without requiring full integration into hospital IT infrastructure at this stage. The containerised, on-site deployment preserves data privacy and security requirements and supports the collection of clinically meaningful evaluation evidence. However, it is acknowledged that the advancement to TRL 7 — which requires

demonstration of the system prototype in an operational environment with broader integration into clinical workflows — will necessitate closer coupling with existing hospital systems, such as the RIS/PACS infrastructure and radiology workstations. This integration will be specifically targeted during the replication pilot at UNIPI and through the EUCAIM data space connection, where a more deeply embedded deployment model is planned.

A further implementation-level adaptation concerned the processing workflow for imaging data. In the deployed setup, anonymisation was introduced as an explicit technical step prior to segmentation through a dedicated containerised component. The segmentation service was configured to process the anonymised data generated within the local workflow rather than the original input folder directly. This made privacy-preserving data handling an integral part of the deployed pilot configuration.

In addition, the deployed application incorporated the clinical assessment process directly into the operational interface. Beyond execution of the segmentation workflow and access to the resulting outputs, the application also provided direct access to the clinical expert trustworthiness questionnaire. Consequently, the deployed environment supported not only technical operation of the tool, but also the structured collection of stakeholder feedback required for pilot evaluation. This is consistent with the broader FAITH approach, which foresees stakeholder involvement and AI-based assessment with human experts during pilot execution.

Finally, due to the local deployment model, part of the technical setup effort was transferred to the pilot site level. The deployment documentation required installation of Node.js/npm and Docker Desktop, while the application also depended on local workstation resources meeting the specified minimum requirements. Compared to a more centrally managed deployment approach, this increased reliance on local setup and workstation readiness, while at the same time improving portability and reproducibility across deployment environments.

Overall, the deviations from the original deployment plan were limited and primarily implementation oriented. They reflect a pragmatic adaptation of the deployment approach in order to facilitate installation, support portability across pilot sites, integrate privacy-preserving processing into the execution workflow, and embed trustworthiness assessment within the same operational environment. These adaptations did not alter the core purpose of the pilot or the planned evaluation of usability, trustworthiness, and deployment feasibility in the healthcare domain.

2.4. Data collection throughout the 1st Pilot execution

During the initial pilot iteration at PAGNH, the AI-based prostate segmentation model was deployed prospectively within the clinical radiology workflow in line with the DoA T9.2 description and as specified in D9.1. Over a period of approximately 19 months (13 May 2024 – 22 December 2025), a total of 179 patient cases were processed by the AI SUT, and each

segmentation output was independently evaluated by expert radiologists using a three-tier qualitative rating scale: Very Good, Good, and Failed. Table 1 summarises the expert evaluation outcomes.

Table 1. Expert qualitative evaluation of AI-generated prostate segmentations (N = 179).

Rating	Count	Percentage (%)
Very Good	156	87.15
Good	22	12.29
Failed	1	0.56
Total	**179**	**100.00**

The AI SUT produced clinically acceptable segmentations (rated *Very Good* or *Good*) in **99.44% of cases** (178/179), whilst complete segmentation failure was observed in a single case.

Analysis of Reduced-Quality Segmentations

Among the 22 cases rated *Good*, expert annotators provided clinical comments in 11 instances, attributing the reduced rating to specific anatomical confounds. Table 2 presents the breakdown of identified clinical conditions.

Table 2. Clinical conditions associated with Good-rated segmentations (commented cases, n = 11).

Clinical Condition	Count	Description
Post-TURP	8	Surgically altered gland morphology following transurethral resection of the prostate
Transitional zone anomaly	3	Atypical zonal anatomy affecting boundary delineation accuracy

Post-TURP patients present with tissue defects, distorted boundaries, and signal heterogeneity on MRI — well-documented challenges for automated segmentation models. Similarly, transitional zone variability introduces ambiguity in the delineation of gland sub-structures. The remaining 11 *Good*-rated cases lacked explicit clinical commentary, although analogous morphological variations are plausible contributing factors. The single *Failed* case (case #55) was not accompanied by a clinical annotation; further investigation is warranted to determine the root cause and inform targeted model refinement.

Alignment with DoA KPIs.

The DoA stipulated a prospective deployment at PAGNH with an indicative throughput of 15 patient cases per month. Over the 19-month collection period, the pilot achieved an average throughput of **9.4 cases per month**. Whilst this falls below the indicative monthly target, the total volume of 179 evaluated cases provides a substantial evidence base for the trustworthiness assessment. Variability in throughput is attributable to clinical scheduling constraints, patient recruitment fluctuations, and scanner availability, factors inherent to prospective deployment in a live hospital environment. These outcomes contribute directly to **KPI3**, which mandates the demonstration of clinical trustworthiness in two clinical sites on prospective cases. The qualitative expert assessment constitutes part of the mixed-method evaluation strategy prescribed by the DoA for instantiating the NIST AI RMF MEASURE function, specifically addressing axis **A3** (performance measured under conditions similar to deployment settings) and axis **A5** (validity and reliability demonstration with documented limitations).

Identified Limitations and Feed-Forward Actions.

The data collected during this phase reveals a systematic performance degradation in anatomically atypical cases, particularly post-TURP patients. This finding will inform targeted data augmentation and domain adaptation strategies for the model refinement cycle preceding the 2nd pilot iteration at the University Hospital of Pisa (UNIPI), as detailed in D9.1 and in line with the DoA description of T9.3.

2.5. Integration into LSP workflows: Lessons learned

This section synthesises the principal lessons learned from the integration of the AI-based prostate segmentation model into the clinical workflow of the University Hospital of Heraklion (PAGNH) during the initial pilot iteration of the Healthcare LSP. Insights are drawn from three complementary evidence sources: (i) expert qualitative evaluation of 179 prospective cases (reported in §2.4), (ii) structured post-task questionnaires administered to participating radiologists, and (iii) operational observations from the MLOps monitoring infrastructure. Lessons are organised across five dimensions and are presented together with identified challenges and concrete feed-forward actions to inform the replication pilot at UNIPI.

2.5.1 Clinical Workflow Integration

The AI SUT was successfully embedded in the prospective mpMRI reporting workflow at PAGNH, operating as an AI-assisted decision-support tool within the radiology reporting pipeline. Segmentation outputs were reviewed inline by expert radiologists without disruption to standard reporting schedules, confirming the technical feasibility of deploying a DL-based segmentation model in an operational University Hospital environment. This validates the deployment strategy defined in D9.1 and demonstrates that the necessary ethical approvals, infrastructure

prerequisites, and interoperability requirements were adequately addressed during the preparatory phase. A notable operational challenge concerned case throughput. Over the 19-month collection period, the pilot achieved an average of approximately 9.4 cases per month, against the DoA indicative target of 15 patient cases per month. This shortfall is attributable to inherent constraints of prospective clinical deployment (variability in scanner availability, patient scheduling, and clinical workload fluctuations) rather than any technical failure of the AI SUT. For the replication pilot at UNIPI, a formalised case-scheduling protocol should be established in advance to stabilise monthly throughput and ensure that the evidence base required for KPI3 is accumulated in a timely and predictable manner.

2.5.2 Model Performance and Failure Mode Analysis

The initial pilot produced strong evidence of model reliability and robustness under real-world clinical conditions. With all but one (99.44%) of cases rated as clinically acceptable (Very Good or Good), the SUT demonstrated consistent segmentation performance across a heterogeneous patient cohort encompassing diverse prostate morphologies, sizes, and imaging characteristics. This outcome directly supports the validation of KPI3 and provides a positive empirical basis for the FAITH AI RMF profile under construction. Systematic performance degradation was, however, identified in two anatomically atypical case categories. Post-TURP patients — in whom surgical resection produces tissue defects, distorted gland boundaries, and altered MRI signal characteristics — accounted for 8 of the 11 annotated reduced-quality cases. A further 3 cases were associated with transitional zone anomalies, where atypical zonal anatomy introduces boundary ambiguity that challenges encoder-decoder segmentation architectures. These findings are consistent with known limitations in the prostate MRI segmentation literature and constitute well-characterised failure modes rather than novel or unexpected behaviours. The single Failed case (case #55) remains without a clinical annotation, precluding root-cause analysis at this stage. Determining whether this failure reflects an imaging artefact, an extreme morphological variant, or an edge case beyond the model's training distribution is essential for informing targeted model refinement prior to the UNIPI deployment. Additionally, the 11 Good-rated cases without explicit clinical commentary warrant follow-up to confirm or exclude analogous morphological confounders.

2.5.3 Stakeholder Engagement and Radiologist Feedback

Structured post-task questionnaires were administered to participating radiologists following each segmentation task. The questionnaire was designed to comprehensively evaluate the trustworthiness, accuracy, and clinical applicability of the SUT for prostate mpMRI analysis. It included structured items assessing qualitative segmentation accuracy, with emphasis on the

delineation of the peripheral and transitional zones and contour consistency, alongside items prompting identification of potential error sources such as over- and under-segmentation. The questionnaire further examined the model's consistency and robustness across heterogeneous patient datasets, accounting for variations in prostate morphology, size, and imaging characteristics, including structurally atypical cases such as hypertrophic transitional zones or post-operative anatomy. Usability was assessed through items addressing interface intuitiveness and the effectiveness of visualisation features, including overlay manipulation and transparency adjustment. Overall, the evidence gathered indicates high segmentation accuracy with clear delineation of prostate zones and consistent contour definition under routine conditions. Minor inaccuracies were observed in anatomically challenging cases, particularly involving hypertrophic prostates and post-operative anatomy, findings consistent with the quantitative evaluation data reported in §2.4. Despite variations in prostate size and imaging characteristics, the model maintained consistent performance across diverse patient datasets, reflecting acceptable robustness. Users reported high usability and expressed confidence in the system as a clinical decision-support tool, highlighting its potential to improve workflow efficiency, promote more standardised and reproducible prostate segmentation, and reduce inter-observer variability. These outcomes directly align with the core objective of T9.2: to replace the second reader of mpMRI images, ultimately reducing healthcare costs whilst maintaining the quality of the service provided. The positive user confidence findings are particularly significant for the human-centric dimension of the FAITH trustworthiness framework, as they provide initial qualitative evidence for end-user acceptance and perceived reliability under real-world clinical conditions.

2.5.4 Summary of Lessons Learned and Feed-Forward Actions

Table 3 consolidates the principal lessons learned, challenges identified, and feed-forward actions for each dimension addressed in the preceding sub-sections.

Table 3. Summary of lessons learned from the WP9 1st pilot iteration (PAGNH) and feed-forward actions for the 2nd pilot iteration at UNIPI (T9.3).

Dimension	Key Finding	Challenge Identified	Feed-Forward Action (T9.3)
Clinical Workflow Integration	AI SUT successfully embedded in prospective mpMRI reporting workflow at PAGNH without disruption to	Average throughput ~9.4 cases/month against DoA target of 15, attributable to clinical scheduling and scanner	Formalise a case-scheduling protocol with UNIPI ahead of replication pilot to stabilise monthly throughput.

	standard clinical schedules.	availability constraints.	
Model Performance & Failure Mode Analysis	99.44% clinically acceptable segmentation rate (178/179). Consistent performance across diverse prostate morphologies.	Systematic degradation in post-TURP (8 cases) and transitional zone anomaly (3 cases) presentations. Single unexplained failure (case #55). Annotation gaps in 12 non-Very Good cases limit failure-mode traceability.	Targeted data augmentation for post-TURP and TZ cases. Root-cause analysis for case #55. Mandatory clinical annotations for all non-Very Good cases in T9.3.
Stakeholder Engagement & Radiologist Feedback	High user confidence and usability. Radiologists recognise potential to improve workflow efficiency, promote reproducibility, and reduce inter-observer variability.	Minor inaccuracies in hypertrophic prostates and post-operative anatomy flagged as residual barriers to full second-reader replacement.	Structured radiologist feedback loops and cross-site usability assessments at UNIPI to validate generalisation of user confidence findings.

The predominantly positive performance and user confidence outcomes validate the core technical and clinical design decisions underpinning the FAITH healthcare LSP. The identified challenges centred on anatomically atypical case handling, failure-mode annotation completeness, and throughput consistency are well-characterised, actionable, and do not represent fundamental limitations to the generalisability of the approach. Their systematic resolution in the replication pilot will strengthen the overall evidence base for **KPI3**.

3. Application of FAITH Tools

3.1. Application of the FAITH AI TAF V1 methodology

The application of the AI_TAF v1 within the Healthcare Large Scale Pilot (LSP) follows a structured, phase-based approach for assessing the trustworthiness of AI systems and their associated human actors. At this stage of the FAITH project, the methodology is applied at the asset level,

meaning that individual components of the system, both AI assets (e.g., the segmentation models) and human assets (e.g., clinical users interacting with the system, AI developers, data analysts), are assessed in isolation, rather than as part of a fully integrated socio-technical system. This approach is consistent with the phased evolution of the FAITH methodology, where early-stage assessments focus on understanding the properties, risks, and behaviours of discrete assets before progressing to a holistic system-level evaluation in later project phases.

The FAITH AI TAF-v1 methodology structures the assessment across a sequence of interrelated phases, including cartography, threat analysis, impact assessment, vulnerability analysis, risk analysis, and countermeasure definition. Each phase contributes to building a progressively refined understanding of the risks and trustworthiness characteristics associated with the asset under evaluation. In the context of the Healthcare LSP, this structured workflow is operationalised through the FAITH toolset, including TrustGuard, TrustSense, and the FAITH AI Model Hub with Model and Data Passports, which support different aspects of the assessment lifecycle.

Importantly, the current application of FAITH AI TAF-v1 deliberately abstracts from the broader clinical ecosystem, focusing instead on the standalone behaviour and properties of the AI SUT and its direct users. For example, the prostate MRI segmentation model is evaluated as an independent AI asset operating in its production phase. Similarly, human users—primarily radiologists—are considered as individual actors interacting with the system, without yet modelling the full organisational, workflow, and socio-technical interdependencies present in real clinical environments. This abstraction enables a more controlled and systematic identification of risks at the component level, reducing complexity and supporting clearer attribution of identified issues.

Within the Healthcare LSP, this asset-level assessment approach is particularly relevant given the pilot nature of the deployment, where the AI system is implemented in a controlled, sandboxed environment and not yet fully integrated into hospital IT infrastructures.

3.1.1 Cartography

The cartography phase represents the initial step in the application of the FAITH AI_TAF v1 methodology and focuses on the structured identification and definition of the assets under assessment. In the Healthcare LSP, this phase involves the clear delineation of the AI SUT—specifically the prostate MRI segmentation model in its production phase—as well as the relevant human actors, primarily radiologists interacting with the system. This approach enables a clear and controlled understanding of each asset before progressing to more complex system-level interactions in later phases of the FAITH methodology.

In addition to asset identification, the cartography phase includes the characterization of key attributes such as the operational context, lifecycle stage, and criticality level of the domain. As part of this phase, a risk assessor was formally nominated for the Healthcare LSP, ensuring that

the assessment process is conducted by a qualified individual with relevant domain expertise. The selection of the risk assessor was based on the professional background, including experience in the medical device regulatory domain, with demonstrated knowledge of the Medical Device Regulation (MDR) and Quality Management Systems compliant with ISO 13485. The appointed assessor has previously served as a Quality/Compliance Engineer (QCE) within a medical device company, bringing practical experience in regulatory compliance, risk management, and quality assurance processes. This expertise supports the robustness and credibility of the asset-level assessment, ensuring alignment with established healthcare regulatory practices and standards.

3.1.2 Threat Analysis

The threat analysis phase within the Healthcare LSP was carried out through the structured use of the FAITH TrustGuard tool, which guided the risk assessor through a systematic identification of potential threats affecting the AI SUT in its production phase. The process involved the completion of a series of predefined questionnaires and assessment steps. These included threats related to model malfunction, adversarial manipulation (e.g., evasion attacks), unauthorized access, and compromise of AI components. Each identified threat was documented within the tool, together with contextual information reflecting the deployment conditions of the SUT in the Healthcare LSP pilot environment.

As part of the assessment activities, the risk assessor assigned qualitative likelihood levels and provided supporting justifications based on the operational setup of the system, including its deployment in a controlled and sandboxed environment. The outcome of this phase was a structured catalogue of threats relevant to the AI SUT, forming the basis for subsequent phases of the FAITH AI_TAF v1 methodology, including impact evaluation, vulnerability analysis, and risk calculation.

3.1.3 Impact Assessment

The impact assessment phase in the Healthcare LSP was conducted through a step-wise procedure, beginning with red teaming and desk-based analysis activities, followed by the structured evaluation within the TrustGuard tool. In the initial stage, the assessment team performed targeted red teaming exercises and internal discussions to explore how identified threats could materialize and what their potential consequences would be on the AI SUT during operation. This preliminary analysis enabled the team to establish a common understanding of their potential impact before formalising the results within the tool. Following this preparatory phase, the impact assessment was systematically implemented in TrustGuard, where each identified threat was assigned an impact level using a scale. This step was carried out collaboratively by the risk assessor, AI developers, and clinical experts, ensuring that both technical and clinical perspectives were incorporated into the evaluation. The structured input of impact levels and supporting rationale within the tool resulted in a consistent and traceable

mapping between threats and their potential consequences, providing a solid basis for the subsequent vulnerability assessment and overall risk analysis within the FAITH AI_TAF v1 methodology.

3.1.4 Vulnerability Assessment

The vulnerability assessment phase in the Healthcare LSP was initiated through a set of preparatory activities conducted within the assessment team, prior to the formal use of the TrustGuard tool. These activities aimed to establish a shared understanding of the potential weaknesses associated with the AI SUT in its production phase. The team engaged in structured internal discussions to review the system's operational characteristics, deployment setup, and interaction patterns with users, with the objective of identifying areas where the system could be exposed to previously identified threats. This preparatory step ensured that vulnerabilities were considered in a systematic and context-aware manner before being formally documented. As part of this preparation, the team aligned on the scope and perspective of the vulnerability assessment, focusing on the AI SUT as an isolated asset within a controlled pilot environment. This included reflecting on potential limitations in system behaviour, operational dependencies, and assumptions regarding user interaction and system monitoring. The outcome of these preparatory activities was a preliminary identification of vulnerability areas, which informed the subsequent structured assessment within TrustGuard. This step-wise approach ensured that the vulnerability analysis was grounded in a clear understanding of the system context and aligned with the overall methodology of the FAITH AI_TAF v1 methodology.

3.1.5 Risk Analysis

The risk analysis phase of the FAITH AI_TAF v1 methodology in the Healthcare LSP focused on the structured consolidation of the findings generated during the preceding threat analysis, impact assessment, and vulnerability assessment phases. The objective of this phase was to determine the relative significance of identified risks affecting the AI SUT and to support prioritisation of mitigation actions within the pilot environment. In practice, the risk analysis considered the combination of three core dimensions: the likelihood of threat occurrence, the severity of potential impact, and the degree of system exposure or vulnerability. By integrating these dimensions, the assessment team was able to distinguish between lower-priority risks requiring monitoring and higher-priority risks requiring direct mitigation or procedural controls. Given the healthcare context of the pilot, particular emphasis was placed on risks that could affect clinical reliability, data confidentiality, workflow continuity, and user trust. Even risks with moderate likelihood were treated carefully when associated with potentially significant clinical or organisational consequences. The outcome of this phase was a prioritised risk view supporting decision-making for the deployment and further refinement of the healthcare AI solution. Detailed pilot-specific risk analysis results are presented in Section 3.5.

3.1.6 Countermeasures

Following the identification and prioritisation of risks, the countermeasures phase of the FAITH AI_TAF v1 methodology focused on defining appropriate actions to reduce the likelihood, impact, or exposure associated with identified risks in the Healthcare LSP. Countermeasures were considered across multiple layers, including technical safeguards, organisational measures, operational procedures, and user-oriented actions. Examples include strengthening deployment security, improving documentation and traceability, enhancing user training, reinforcing monitoring processes, and maintaining expert clinical oversight during system use. In the healthcare setting, particular importance was assigned to countermeasures supporting patient data protection, robustness of AI outputs, transparency of model behaviour, and continuity of clinical workflows. Measures were selected proportionally to the identified risk level and the practical realities of pilot deployment within hospital environments. The countermeasure process was also iterative in nature. As pilot execution progressed and user feedback was collected, proposed controls could be refined or complemented by additional actions. The resulting pilot-specific mitigation measures and recommended controls are presented in detail in Section 3.6.

3.2 Tool Validation Planning & Training Processes

Robust validation of the AI SUT and the effective onboarding of clinical users are treated as interdependent processes within the Healthcare LSP. Meaningful evaluation evidence can only be collected if participating radiologists are sufficiently familiar with the SUT's operational workflow, visualisation outputs, and assessment instruments prior to commencing formal pilot activities. Accordingly, a structured approach to both tool validation planning and user training has been established, encompassing dedicated training sessions designed to present and test SUT functionalities, collect initial usability feedback, and ensure consistency of use across the two pilot sites (PAGNH and UNIPI).

3.2.1 Dry-run

Prior to the formal execution of the FAITH tool validation activities, a dry-run phase was conducted within the Healthcare LSP in order to verify the readiness of the assessment process, familiarise participants with the tools, and identify any practical issues requiring adjustment before full deployment. The dry-run activities focused primarily on the application of the FAITH risk assessment questionnaire and the preparatory workflows associated with TrustGuard. The exercise involved representatives from the Healthcare LSP consortium, including technical contributors and the designated risk assessor, who collaboratively reviewed the structure, terminology, and applicability of the questionnaire in relation to the prostate MRI segmentation use case. Emphasis was given to ensuring that the assessment items could be interpreted consistently within the healthcare context and that the questions were relevant to AI systems

operating as clinical decision-support tools. During the dry run, the team simulated the completion of the assessment workflow using the healthcare AI models as reference assets. This included preliminary identification of the AI system boundaries, lifecycle stage, criticality level, relevant trustworthiness characteristics, potential threats, expected impacts, and vulnerability areas. The exercise enabled the participants to test the logical sequence of the methodology and validate that the collected information was sufficient to support subsequent risk analysis activities. The dry run also served as a calibration activity among participants. Through structured discussion, the team aligned interpretations of likelihood, impact, and criticality levels, thereby improving consistency for the formal assessment phase. In addition, terminology requiring clarification was identified, particularly where general AI risk concepts needed adaptation to the realities of clinical imaging workflows and hospital deployment environments. From an operational perspective, the dry run confirmed the suitability of the selected assets for assessment, validated the readiness of the participating teams, and highlighted the need for concise supporting evidence related to deployment architecture, model documentation, and user interaction processes. Minor refinements were subsequently introduced in the preparation materials and internal guidance used during the formal TrustGuard assessment.

3.2.2 Plan development

The planning of the FAITH tool validation activities for the Healthcare LSP was further refined during the second FAITH consortium plenary meeting (Oslo, January 202), where consortium partners reviewed the status of pilot preparations and aligned the forthcoming implementation steps for the first clinical deployment. The plenary provided an important coordination point for translating the overall FAITH project methodology into a concrete and feasible validation roadmap adapted to the specific operational context of LSP6. During the planning sessions, FORTH, together with the contributing partners CNR, UNIPI, and PAGNH, reviewed the detailed implementation schedule for the healthcare pilot, covering the period from the preparation phase through the initial pilot deployment and the first iterations of FAITH tool usage. Particular attention was given to sequencing the activities in a manner compatible with the clinical environment, local ethical approvals, availability of radiologists, and the technical readiness of the AI SUT.

A dedicated implementation plan was consolidated for the Healthcare LSP, defining responsibilities, milestones, and dependencies for the deployment and assessment process. The plan included activities related to overall pilot coordination, training of users, deployment of the AI SUT, onboarding of stakeholders to the FAITH tools, execution of trustworthiness assessments, and collection of structured feedback. In this context, FORTH assumed the role of technical coordinator for the deployment of the AI models and Model Hub assets, while UNIPI and PAGNH

focused on local clinical execution and end-user engagement. CNR supported technical integration and tool-related activities.

Detailed Implementation Plan LSP6 - Healthcare		Phase:	Deployment		SANDBOX DEPLOYMENT							
FORTH / CNR / UNIPI / PAGNH												
History (last upd): 19.12.2026 (init)	Project Start:		M23									
	Display Month:		1		M23	M24	M25	M26	M27	M28	M29	M30
SWB	Assigned to	PROGRESS	START	END								
Planning												
Define plan	FORTH / CNR / UNIPI / PAGNH	100%	M23	M24								
Agree plan	FORTH / CNR / UNIPI / PAGNH	100%	M24	M24								
Training and support												
Attend training for AI Model Passport (python library)	hospitals 1 & 2	100%	M26	M28								
Attend training for AI Data Passport	hospitals 1 & 2	100%	M26	M28								
Attend training for trustBENSE	hospitals 1 & 2	100%	M23	M23								
Attend training for trustGUARD	hospitals 1 & 2	50%	M26	M28				24/02/2026 at 09:00 CET				
Attend training for SSM	hospitals 1 & 2	50%	M32	M33								
Attend introduction to FAITH STM Platform registration and workflows	hospitals 1 & 2	0%	M30	M30								
FAITH Tools implementation and iteration 2026												
Deploy SUT in hospital 1	hospital 1 & FORTH	100%	M10	M10								
Use trustBENSE for team maturity assessment after 6 months use of SUT (Domain User Team - early deployment phase)	Risk Assessor + Domain (Radiologists)	100%	M20	M20								
Provide formal feedback on Usability and Usefulness of trustBENSE	Risk Assessor	100%	M26	M26								
Implement trustBENSE mitigations	hospital 1	50%	M27	M33								
Create assessories in AI Model Hub	Risk Assessor	100%	M23	M24								
Provide formal standardised feedback on Usability and Usefulness of AI Model Hub	Risk Assessor	100%	M26	M26								
Iterate AI Model Hub v2 based on LSP feedback	FORTH / CNR	0%	M28	M32								
Use trustGUARD v1 for in-isolation AI Model assessment @ deployment phase	Risk Assessor	100%	M26	M28								
Use trustGUARD v1 for in-isolation ad-hoc on-going Input Dataset assessment @ deployment phase	Risk Assessor	0%	M28	M28								
Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during deployment phase	Risk Assessor	100%	M29	M29								Provided feedback in Survey
Apply 1 or more Controls to AI Model or Input Data during deployment phase based on trustGUARD recommendations	hospital 1 & FORTH	15%	M29	M32								Activity 1 assessment of D9.2
Update in-isolation AI Model assessment + new Controls @ deployment phase using trustGUARD v1	Risk Assessor	0%	M30	M32								
Re-train SUT with extended image datasets from hospital 1 (design phase)	FORTH	0%	M30	M33								
Use trustGUARD v1 for in-isolation Training Dataset assessment @ design phase	Risk Assessor	0%	M31	M31								
Address trustGUARD training dataset risks (bias/robustness trade-offs)	FORTH	0%	M32	M33								
Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during design phase	Risk Assessor	0%	M31	M31								
Use trustGUARD v1 for in-isolation Testing Dataset assessment @ development phase	Risk Assessor	0%	M34	M34								
Address trustGUARD testing dataset risks (bias/robustness trade-offs)	FORTH	0%	M35	M38								
Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during development phase	Risk Assessor	0%	M34	M34								
On-going iteration of trustGUARD v2 based on LSP feedback @ all 3 lifecycle phases (inc. on-going (re-)validation upon updates)	ATC	0%	M30	M36								
Use FAITH STM Platform	hospital 1 & 2 - Doctors & Risk Assessor	0%	M30	M39								
Provide formal standardised feedback on Usability and Usefulness of FAITH STM Platform	hospital 1 & 2 - Doctors & Risk Assessor	0%	M30	M39								
Revisit FAITH STM Platform	UCS	0%	M31	M34								
Create SSM system model	FORTH / UCS	50%	M33	M33			TBAL					
Use SSM for in-context SUT assessment	Risk Assessor	0%	M33	M38								
Provide formal standardised feedback on Usability and Usefulness of SSM	Risk Assessor	0%	M35	M35								
On-going iteration of SSM v2 based on LSP feedback (inc. on-going Knowledge-based updates)	UCS	0%	M36	M48								
Deploy SUT in hospital 2	hospital 2 & FORTH / CNR	100%	M33	M36								Deployed

Figure 2. Detailed implementation plan for LSP6 – Healthcare pilot, including planning, training, deployment, and FAITH tool validation activities.

As illustrated in Fig. 2, the implementation roadmap was organised into three main workstreams: planning, training and support, and FAITH tools implementation and iteration. The planning stream focused on coordination across partners and agreement of the operational pilot plan. The training stream covered attendance in consortium training sessions for the AI Model Hub including Model and Data Passport, TrustSense, TrustGuard, and supporting modules. The implementation stream focused on the deployment of the AI SUT, execution of team maturity and usability assessments, creation of model assets within the AI Model Hub, and iterative improvement of tools based on pilot feedback.

Within the FAITH consortium plenary meeting (Oslo, January 2026), the consortium also agreed on the staged introduction of the FAITH tools in the Healthcare LSP. In the first phase, emphasis was placed on establishing the required Data and AI Model Passports, ensuring that the healthcare AI assets were properly documented before broader trustworthiness assessments commenced. Subsequently, TrustSense and TrustGuard activities were scheduled to follow once pilot users had gained practical experience with the deployed SUT. This phased approach ensured that assessments would be informed by real user interaction rather than purely theoretical assumptions. The planning activities also addressed resource allocation and expected participant involvement. It was agreed that radiologists and clinical users from the participating hospitals would contribute to usability and trustworthiness assessments after an initial familiarisation period with the platform. In parallel, the designated risk assessor would lead the formal TrustGuard assessment activities, supported by technical experts and clinical stakeholders where required. Finally, the Oslo plenary was used to align timelines between the initial pilot at PAGNH at the replication pilot at UNIPI. This ensured that lessons learned from the initial pilot could be transferred efficiently to UNIPI/AOUP, thereby improving readiness for the second clinical site and supporting the iterative validation logic of the FAITH project.

3.2.3 Tool Training

The training process was organised through consortium-level sessions delivered within the FAITH project and complemented by internal follow-up activities among the Healthcare LSP partners. Participants from FORTH, CNR, UNIPI, and PAGNH attended the relevant training sessions according to their roles in the pilot. The training programme covered both technical and methodological aspects of the tools, including data and model documentation, maturity assessment, trustworthiness evaluation, and risk analysis. As reflected in the implementation plan (Fig. 3), training activities for the Healthcare LSP were scheduled progressively in alignment with the deployment timeline of the pilot. Early sessions focused on the governance-oriented tools, namely the AI Model Passport and Data Passport, allowing the pilot partners to document datasets and AI models prior to broader validation activities. Subsequent sessions addressed the

use of TrustSense and TrustGuard, enabling participants to prepare for team maturity assessments, usability evaluations, and structured risk assessment processes.

Training on the **AI Model Passport** focused on the creation and management of AI model documentation within the FAITH Model Hub. Participants were introduced to the registration of models, metadata completion, linkage to datasets, lifecycle stage documentation, and publication workflows. This was particularly relevant for the Healthcare LSP, where multiple segmentation models required formal documentation and publication within the platform. Training on the **Data Passport** concentrated on dataset governance and documentation procedures. Participants were instructed on how to capture key metadata related to dataset provenance, ownership, intended use, structure, privacy safeguards, and access conditions. This supported the preparation of passports for the clinical datasets used in training and validation of the healthcare AI models.

TrustSense training focused on the execution of organisational and team maturity assessments, including the interpretation of maturity dimensions, questionnaire completion processes, and the use of outputs for identifying areas of improvement. In the Healthcare LSP, this training was relevant for assessing the readiness of the involved teams and the broader conditions for trustworthy AI adoption in clinical environments.

TrustGuard training addressed the structured methodology for asset identification, threat analysis, impact assessment, vulnerability evaluation, risk calculation, and countermeasure definition in accordance with the FAITH AI_TAF v1 methodology. Particular emphasis was placed on how to perform assessments for production-stage AI assets in high-criticality domains such as healthcare. The designated risk assessor and technical contributors received focused guidance on applying the tool in the context of the deployed prostate MRI segmentation solution.

In addition to formal tool sessions, internal knowledge-transfer activities were conducted within the Healthcare LSP consortium. These included coordination meetings, walkthrough demonstrations, clarification sessions, and practical exchanges among technical and clinical partners to ensure common understanding of responsibilities and expected outputs. This was particularly important given the multidisciplinary nature of the pilot, which involved AI developers, engineers, clinicians, and regulatory-oriented stakeholders.

3.3 FAITH Tool Validation Activities

3.3.1 TrustSense

Group Radiologists end users

TrustSense was applied within the Healthcare LSP to assess the organisational and human readiness dimensions associated with the trustworthy adoption of the AI SUT. The tool was used with two key stakeholder groups involved in the pilot: the radiologist end users, representing the clinical domain users of the deployed solution, and the AI technical team, representing the developers and technical maintainers of the system. Through structured questionnaires,

TrustSense provided an evidence-based overview of team maturity, awareness, preparedness, and organisational behaviours relevant to trustworthy AI deployment.

For the radiologist end-user group (Domain AI User Team), the TrustSense assessment resulted in a Very High trustworthiness level, with a calculated score of 4.65 (Fig. 3). This outcome indicates strong readiness of the clinical user group to engage with the AI solution in a responsible and effective manner. The result suggests high levels of professionalism, adherence to operational procedures, critical judgement during AI-supported image review, and capacity to integrate the system within clinical decision-support workflows while maintaining expert oversight. From a pilot perspective, this finding is particularly positive, as clinician trust, competence, and informed use are central prerequisites for successful adoption of AI systems in healthcare environments.



Figure 3. TrustSense assessment results for the radiologist end-user group (Domain AI User Team), indicating a Very High trustworthiness level with a score of 4.65.

For the AI technical team (AI Technical Team), the TrustSense assessment resulted in a High trustworthiness level, with a calculated score of 4.10 (Fig. 4). This indicates a mature and capable technical group with strong foundations in AI development, deployment, and maintenance practices, while also identifying opportunities for further improvement. The result reflects solid competence in model development, engineering collaboration, and responsible AI implementation, together with the need for continued enhancement in selected trustworthiness dimensions such as structured governance processes, cross-disciplinary knowledge exchange, and continuous improvement practices. Within the Healthcare LSP, this demonstrates that the technical team is well positioned to support pilot execution and iterative refinement of the AI SUT.



Figure 4. TrustSense assessment results for the AI technical team, indicating a High trustworthiness level with a score of 4.10.

3.3.2 Data & AI Model Passport in the FAITH AI Model Hub

FAITH AI Model Hub provides two complementary governance tools: the Data Passport and the AI Model Passport. Together, they support transparent documentation of both datasets and AI models, enabling traceability, compliance, reproducibility, and responsible use across the FAITH ecosystem. The Data Passport serves as the digital identity of each dataset, while the AI Model Passport serves as the digital identity of each AI model.

Within the Healthcare LSP, these governance mechanisms have been applied to the AI models developed for prostate cancer segmentation on T2-weighted MRI. More specifically, two AI Model Passports were created to document the models used in this context: the LSP6_ProstateWholeGlandSegmentationModel, which supports whole-gland segmentation, and the LSP6_ProstateZonalSegmentationModel, which supports zonal segmentation. The corresponding model documentation captures the datasets used for training and validation, thereby strengthening traceability and transparency with respect to model development and evaluation.

Table 4. A summary of the data and AI model passports related to the AI SUT in the healthcare LSP.

Model	Data Passport	UID	Stage
Whole-Gland	The PI-CAI Challenge: Public Training and Development Dataset v2.0	cmk45mq5f005ewx0rvfqdftta	Training

	ProCAncer-I UC1: Dataset for Prostate Cancer Detection (Validation)	cmk5bnlgt00a3wx0r03o3y9lk	Validation
Zonal	ProCAncer-I UC1: Dataset for Prostate Cancer Detection	cmjikog2w0001wx0r8i1mvog2	Training
	Prostate158 - Training Data (Validation): An expert-annotated 3T MRI dataset	cmk5kl4ao00hhwx0rzhrpylm0	Validation
	Prostate158 - Testing Data (Validation): An expert-annotated 3T MRI dataset	cmk5qwow100kawx0rs1b30r6e	Validation

3.3.2.1 AI Model Passport

The AI Model Passport is used as a structured governance mechanism for documenting the lifecycle, provenance, and operational context of an AI model. Rather than serving only as a static description, it provides a traceable record of how a model is defined, linked to datasets and pilots, and documented across the main lifecycle stages. In the Healthcare LSP, this functionality can be illustrated through the zonal prostate segmentation model passport, corresponding to the model LSP6_ProstateZonalSegmentationModel. The screenshots below (Fig. 4-10) show an example of a successfully published passport and demonstrate the type of metadata that can be captured through the platform.

The overall workflow consists of three different sections as follows:

- Summary
- Details
- Pipeline

3.3.2.1.1 Summary

The first section, Summary, provides the summary-level identity of the AI model passport, establishing the model as a governed digital asset. It links the model to its unique identifier, the relevant pilot, and the publishing organization, acting as the entry point to the passport and allowing users to quickly verify existence, ownership, and publication status before exploring detailed technical and lifecycle information.

In the zonal segmentation example, the summary page records core metadata such as the unique ID, pilot context (Healthcare LSP), publisher and authoring organization (FORTH), and the connection to the corresponding documentation entity. It also indicates that the passport has been successfully published and verified within the platform. This layer supports discoverability, governance, and high-level traceability within the FAITH ecosystem, while providing general model information, including its description as a prostate zonal segmentation model in T2-

weighted MRI, task type (segmentation), and governance-related fields such as risk assessment and ethical considerations.



Figure 5. Summary section of the AI Model Passport for the prostate zonal segmentation model, illustrating identity and governance metadata including unique identifier, pilot context, publishing organization, verification status, and high-level model description within the FAITH ecosystem.

3.3.2.1.2 Details

The second section, Details, captures the descriptive and technical identity of the model itself. At this stage, the passport records the metadata needed to explain what the model is, what it is intended to do, who created it, and under what technical and governance context it should be understood. This layer is important because it translates the model from a mere technical artifact into a documented AI asset with a clear purpose, provenance, and accountability context.

For the zonal segmentation example, the details section records metadata such as the algorithm used, the model/task type (segmentation), the learning method (deep learning), the model name, the description of the model purpose, the author and publisher information, the source code repository, and governance-related fields such as risk assessment and ethical concerns. Together, these fields provide a compact but meaningful description of the model's intended function and origin, while also linking its technical characterisation to governance-oriented information relevant for transparency and responsible use.



Figure 6. Details section of the AI Model Passport for the prostate zonal segmentation model, illustrating descriptive and technical metadata including algorithm, task type, learning method, model purpose, authorship, source repository, and governance-related information within the FAITH ecosystem.

3.3.2.1.3 Pipeline

The last section, Pipeline, captures the metadata recorded for each stage of the AI model development pipeline, providing the most granular layer of the passport. For every stage, it documents core execution details (stage name, type, status, timestamps), references to input and output artifacts with their checksums, URIs, and path locations, implementation scripts, software environment, and relevant configuration parameters, including hyperparameters or evaluation metrics where applicable. Preserving this information supports technical traceability, validation of outputs, monitoring of model performance, and reproducibility throughout the model lifecycle, with the differences between stages arising only from the specific content of these artifacts and parameters.

In the zonal segmentation example, the passport includes an AI Model Development Pipeline consisting of four stages: preprocessing, training, evaluation, and production. This workflow representation helps users understand the overall development process and locate the detailed metadata for each stage. Importantly, this configuration is an example of one successfully completed passport; the number and type of stages may differ across models and are not rigidly prescribed by the platform.

Preprocessing: Metadata documents the preparation of the dataset and generation of preprocessing artifacts. Key parameters captured include MRI channels, normalization method, voxel spacing, and validation patient identifiers.



Figure 7. Preprocessing stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, showing dataset preparation metadata, MRI preprocessing parameters, validation identifiers, and associated input/output artifact references.

Training: Metadata captures model learning and parameter optimization. Hyperparameters such as batch size, epochs, optimizer settings, learning rate, and network architecture (blocks, layers, strides, kernel sizes) are preserved, along with input/output artifacts and execution environment details.



Figure 8. Training stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, capturing learning and optimization metadata including hyperparameters, network architecture details, artifacts, and execution environment information.

Evaluation: Metadata records model validation and performance assessment. Evaluation metrics, such as Dice score, Jaccard index, precision, recall, and 95 Hausdorff distance, are documented alongside the usual artifact and environment metadata.



Figure 9. Evaluation stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, showing validation metrics and associated artifacts and environment metadata.

Production: Metadata captures the final operational output, including the resulting artifact, its URI, checksum, path location, implementation files, and software environment. Additional technical or configuration metadata may also be included, preserving the deployable form of the workflow and its context.



Figure 10. Production stage of the AI Model Development Pipeline in the AI Model Passport for the prostate zonal segmentation model, capturing the final deployable artifact with its URI, checksum, path location, implementation files, and environment metadata.

3.3.2.2 Data Passport

Complementing the AI model passport presented above, the data passport documents the dataset as a governed digital asset within the FAITH ecosystem. Whereas the model passport focuses on the characteristics and lifecycle of the AI model, the data passport captures the provenance, stewardship, access conditions, privacy safeguards, structure, and intended use of the underlying data resource. Taken together, the two passports strengthen transparency, traceability, and governance across both the data and model dimensions of the AI system lifecycle.

Data Passport consists of six different sections as follows

- Summary
- Authorship
- Motivation

- Data Use
- Distribution
- Security0

3.3.2.2.1 Summary

In the present example, the data passport refers to the ProCAncer-I UC1 dataset for Prostate Cancer Detection, which serves as the training dataset for the zonal segmentation model described above. The summary page records the core identifying metadata of the dataset, including its title, acronym, version, unique identifier, pilot context (Healthcare LSP), and publication dates, together with a short description of its purpose. In this way, the summary layer supports discoverability, governance, and high-level traceability of the dataset within the FAITH ecosystem.

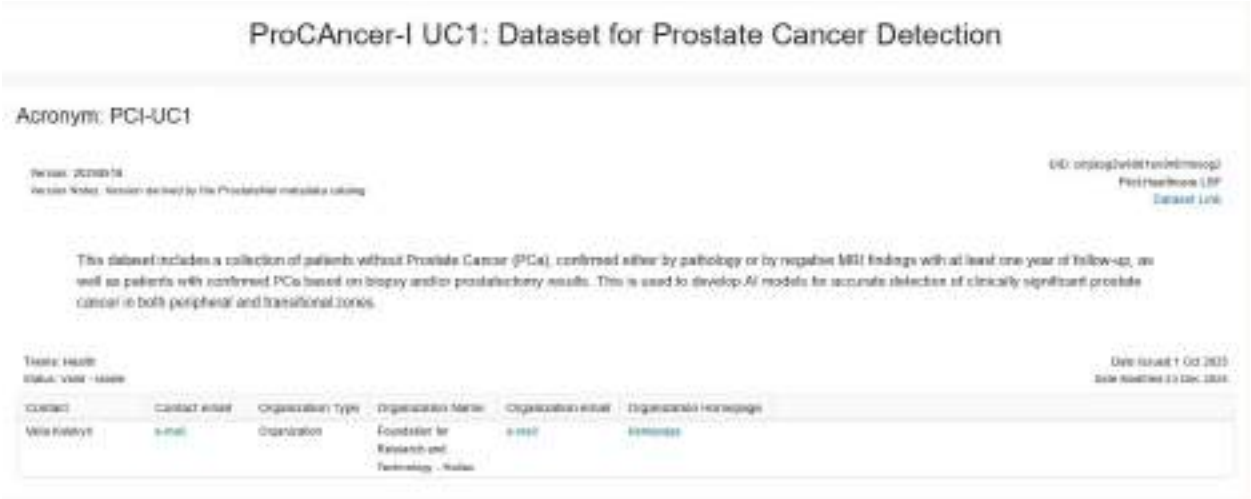


Figure 11. Summary section of the Data Passport for the ProCAncer-I UC1 dataset, capturing core dataset identifiers including title, acronym, version, unique ID, pilot context, publication dates, and a high-level description supporting governance and traceability within the FAITH ecosystem.

3.3.2.2.2 Authorship

The second section captures the authorship and governance context of the dataset. It records the metadata needed to identify the organisations responsible for publishing, owning, and supporting the dataset, thereby establishing its provenance and accountability framework. In the ProCAncer-I UC1 example, this section includes the publisher (FORTH), the dataset owner (the ProCAncer-I consortium), and the funding source linked to the European Commission grant. Together, these elements provide a concise account of dataset stewardship, institutional responsibility, and funding context.

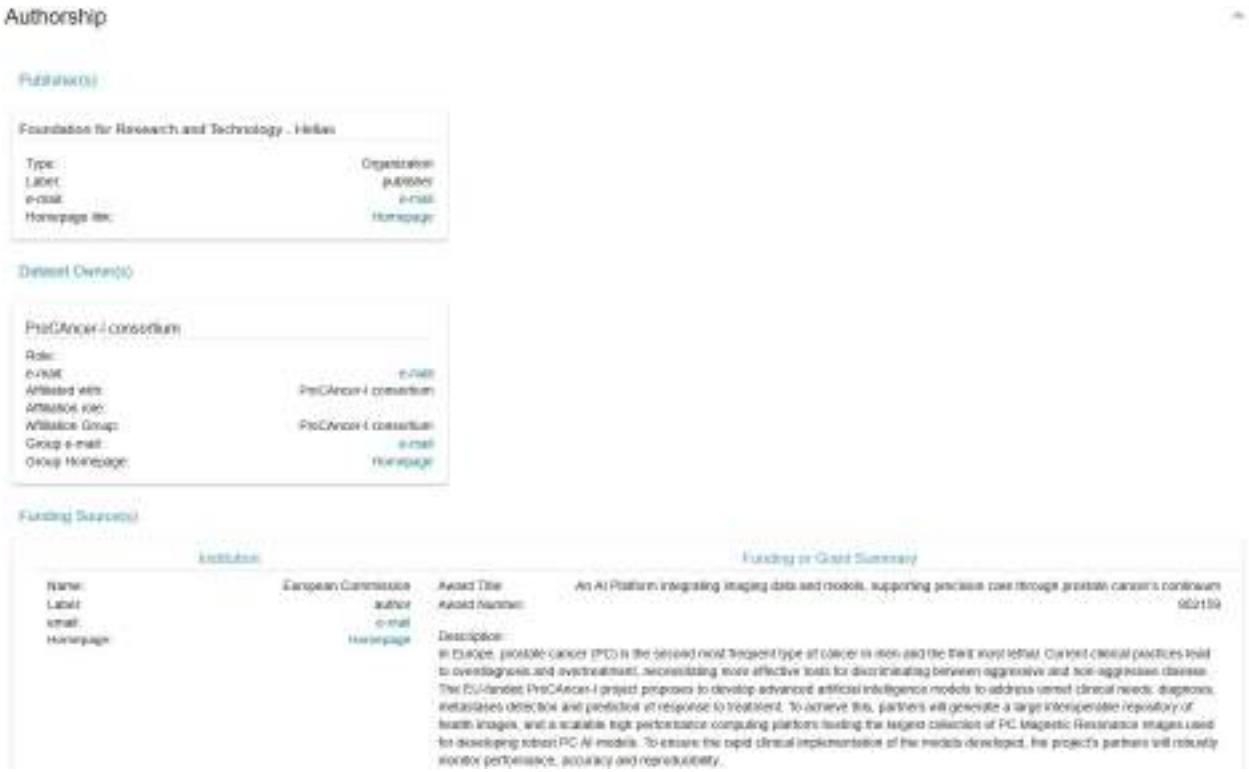


Figure 12. Authorship section of the Data Passport for the ProCAncer-I UC1 dataset, capturing governance and provenance metadata including publisher, dataset owner, and funding source, establishing stewardship and institutional responsibility.

3.3.2.2.3 Motivation

The third section addresses the motivation and intended use of the dataset. Its purpose is to clarify why the dataset exists, for which purposes it is suitable, and where its use would be inappropriate. For the ProCAncer-I UC1 dataset, the passport identifies purposes such as academic research, public benefit, and research and development, and specifies its suitability for health-related AI uses, including organ and lesion segmentation tasks and prostate cancer detection. At the same time, it records limitations on unsuitable use, thereby supporting transparency regarding both the intended scope and the boundaries of application.



Figure 13. Motivation section of the Data Passport for the ProCAncer-I UC1 dataset, describing intended use cases, including research and healthcare AI applications, and outlining usage limitations and constraints.

3.3.2.2.4 Data use

The fourth section links the dataset to its role within AI or machine learning workflows. In this case, the passport indicates that the dataset is used in the training phase and is associated with supervised learning. Although concise, this information is important because it situates the dataset within the broader AI development lifecycle and connects the data passport to the model-related documentation presented elsewhere in the platform.



Figure 14. Data Use section of the Data Passport for the ProCAncer-I UC1 dataset, linking the dataset to AI workflows and indicating its use in training and supervised learning contexts.

3.3.2.2.5 Distribution

The fifth section documents the distribution characteristics of the dataset, including how it is made available, under which access conditions it may be used, which provenance statement accompanies it, and which licensing terms govern its reuse. In the present example, the passport records information such as dataset subject, size, number of instances, access rights, annotation type, format, relevant links, provenance statement, and license. More specifically, it indicates that the dataset contains patient-level data in DICOM format, is not publicly accessible, and originates from the aggregation and harmonisation of data contributed by multiple European clinical centres. This section therefore strengthens legal clarity, controlled reuse, and traceability.

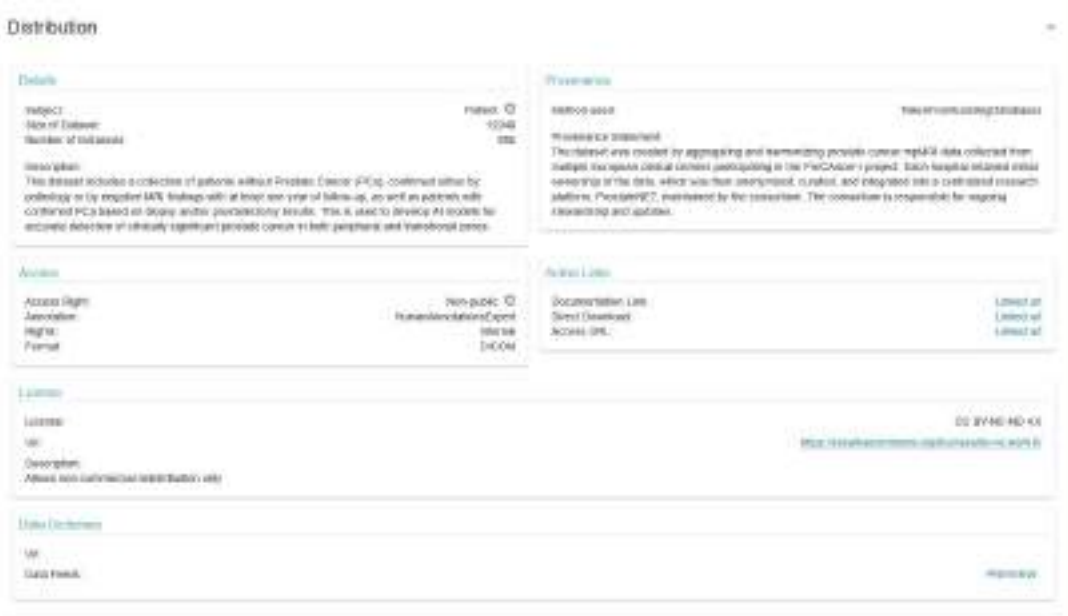


Figure 15. Distribution section of the Data Passport for the ProCAncer-I UC1 dataset, detailing access conditions, licensing, provenance, data format, and key dataset characteristics such as size, annotations, and clinical origin.

This section also provides the most granular description of the dataset by documenting each data field along with its label, description, data type, allowed values, units, and, where available, semantic references. In the example, the passport records variables such as age, digital rectal examination findings, biopsy details, PSA measurements, MRI status, and examination dates. This structured representation supports interpretability, schema-level understanding, and consistent reuse across users and systems.



Figure 16. Field-level dataset description within the Distribution section of the Data Passport for the ProCAncer-I UC1 dataset, providing detailed metadata for variables including names, descriptions, data types, units, allowed values, and semantic references.

3.3.2.2.6 Security

The sixth and last section captures the security and privacy context of the dataset by documenting the legal, technical, and organisational safeguards in place, together with the relevant categories of personal or sensitive data. In the ProCAncer-I UC1 example, the passport records measures such as data processing agreements, password authentication, single sign-on, pseudonymisation, consultation with the data protection officer, consent recording, and data interoperability assessment. It also indicates the presence of sensitive personal data related to age, demographics, gender, and health. This section provides an essential governance layer for understanding how the dataset is protected and managed.



Figure 17. Security section of the Data Passport for the ProCAncer-I UC1 dataset, capturing privacy and governance safeguards including access controls, authentication mechanisms, pseudonymisation, consent management, data protection oversight, and indications of sensitive personal data.

3.3.3 TrustGuard

TrustGuard was applied within the Healthcare LSP as the principal tool supporting the structured risk and trustworthiness assessment of the AI System Under Test in accordance with the FAITH AI_TAF v1 methodology. The tool was used to assess both segmentation models deployed in the pilot environment, namely the whole-gland prostate segmentation model (Model 1) and the zonal prostate segmentation model (Model 2). Through its guided workflow, TrustGuard enabled the systematic execution of the main assessment phases, including cartography, criticality assessment, threat analysis, impact assessment, vulnerability assessment, risk analysis, and countermeasure definition.

The first stage of the assessment involved the registration and characterisation of the AI asset under evaluation (Fig. 18). For each model, the assessment scope was defined by specifying the model identity, lifecycle stage, and operational context. In the present case, the AI systems were assessed in their production phase, reflecting their deployment as operational tools within the pilot environment for clinical image analysis. This step ensured clear traceability between the deployed models and the corresponding trustworthiness assessment records.

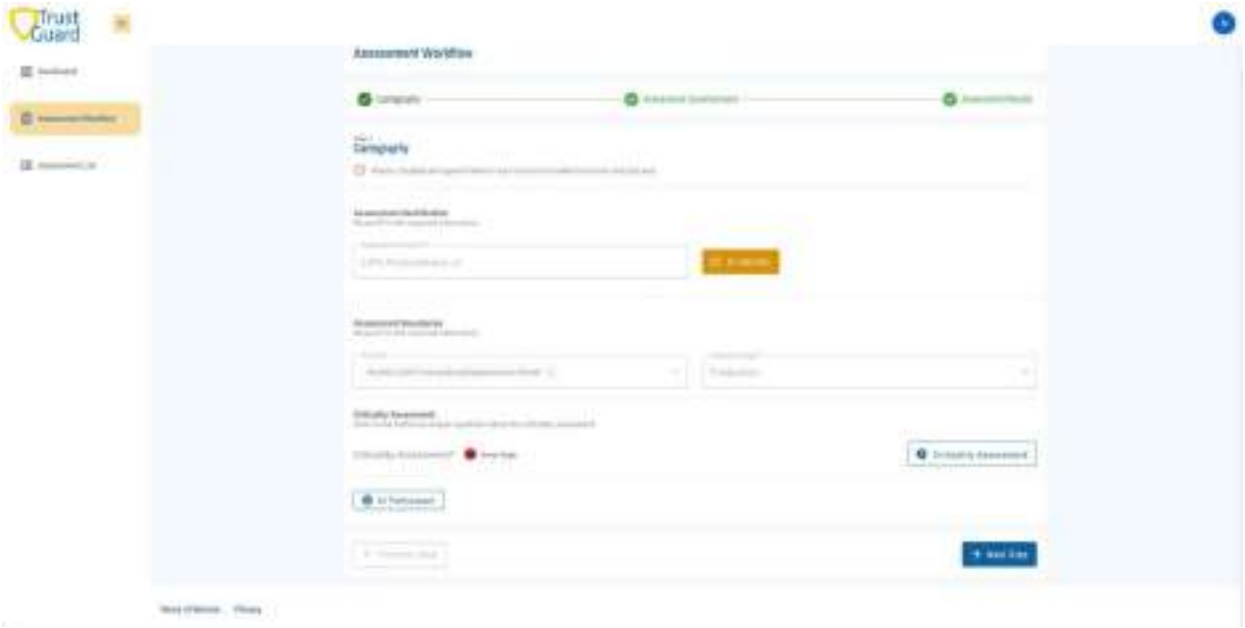


Figure 18. TrustGuard cartography stage for the zonal prostate segmentation model (Model 2), showing asset identification, lifecycle stage, and assessment setup.

A key component of the TrustGuard process was the criticality assessment (Fig. 19). This stage examined whether the AI system falls under regulatory or operational contexts associated with increased risk and evaluated the potential impact of the system on health, safety, and decision-making. Given that the assessed models operate in the healthcare domain and support medical imaging analysis, the resulting classification indicated a Very High criticality level. This outcome is consistent with the sensitivity of clinical environments, the importance of reliability in medical workflows, and the potential consequences of degraded performance or misuse. The same criticality rationale was considered applicable to both Model 1 and Model 2, as both systems serve comparable clinical functions within prostate MRI segmentation.



Figure 19. TrustGuard criticality assessment applied to the healthcare AI system, resulting in a Very High criticality classification due to the medical domain context and potential impact on health-related decision-making.

Following the criticality assessment, TrustGuard supported the structured evaluation of trustworthiness risks across the FAITH dimensions (Fig. 20). For the healthcare pilot, the assessment considered characteristics such as safety, security and resilience, explainability and interpretability, privacy enhancement, fairness with harmful bias managed, accountability and transparency, and validity and reliability. This allowed the assessment team to identify the dimensions most relevant to the deployed clinical AI solution and to prioritise subsequent mitigation actions.

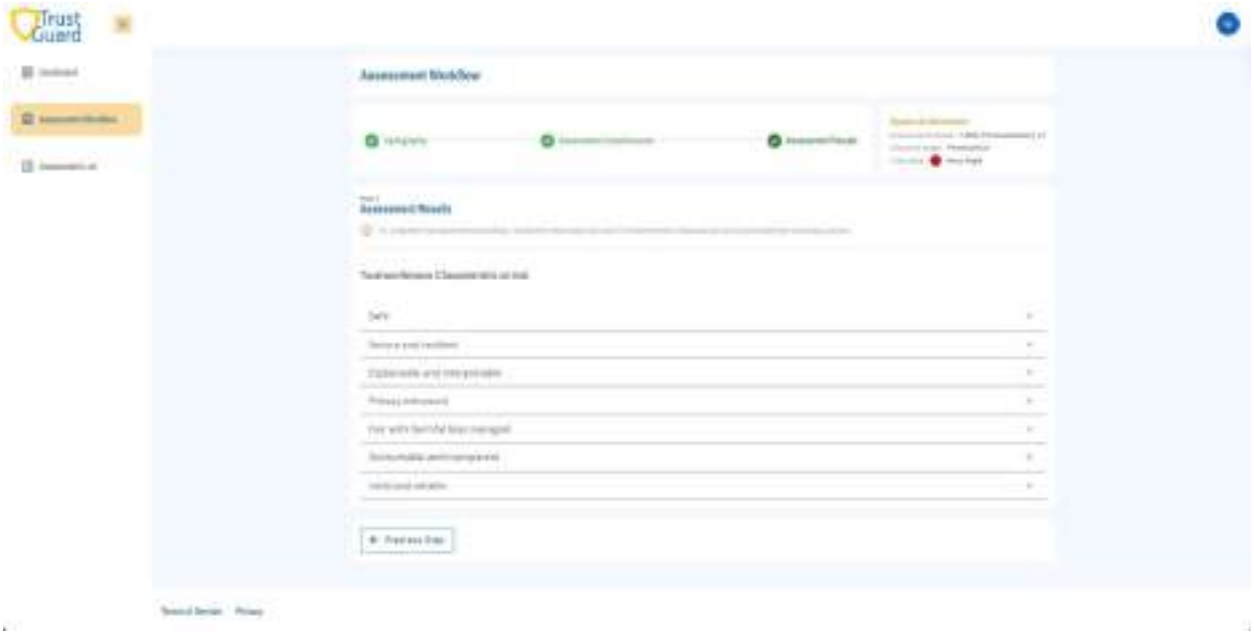


Figure 20. TrustGuard assessment results interface, presenting trustworthiness characteristics evaluated for the zonal prostate segmentation model.

Overall, the use of TrustGuard provided a consistent and traceable mechanism for conducting risk-oriented trustworthiness assessments within the Healthcare LSP. It enabled the translation of methodological FAITH AI_TAF principles into practical pilot-level assessment activities and supported evidence generation for both regulatory readiness and responsible deployment of AI systems in healthcare environments. Detailed threat, risk, and countermeasure outcomes derived from the TrustGuard process are presented in Sections 3.4, 3.5, and 3.6.

3.4. Threat Analysis

3.4.1. Model 1

3.4.1.1 Threat Assessment - Model 1

The threat assessment for Model 1 (LSP6_ProstateWholeGland_v1), conducted within the FAITH AI_TAF v1, reflects a high-criticality AI system operating in a healthcare production environment, where risks to patient safety and clinical decision-making are inherently significant. Despite this high criticality, the assessment indicates that the system is deployed in a highly regulated and

controlled clinical context, which substantially influences both the likelihood and nature of the identified threats.

From the perspective of the security and resilience trustworthiness dimension, the analysis highlights that threats such as evasion attacks and adversarial examples (white/grey-box attacks) are considered plausible but relatively infrequent, typically expected to occur every few years. These attacks are particularly relevant in medical imaging, where subtle perturbations (e.g., imperceptible noise in MRI scans) could potentially alter segmentation outputs without being detectable by clinicians. Similarly, threats related to oracle-based vulnerabilities and system-level exploits are acknowledged, especially in light of real-world vulnerabilities in widely used healthcare software systems, indicating that dependencies on external platforms may introduce indirect risks. In contrast, data poisoning attacks are considered less likely due to strict data governance, controlled pipelines, and regulatory safeguards governing medical datasets.

Regarding threats pertaining to the privacy enhancement trustworthiness dimension, the assessment identifies risks such as label modification, data disclosure, and model disclosure, though these are generally assessed as low-frequency events. Label modification is particularly notable, as it can occur through compromised annotation processes or hybrid labeling pipelines, potentially degrading model performance or introducing hidden biases. However, the absence of frequent occurrences reflects the structured data curation processes and validation mechanisms typically applied in clinical AI development. Similarly, risks related to unauthorized access or leakage of sensitive data are mitigated by compliance with healthcare regulations and secure data handling practices.

From the perspective of the security and resilience and validity and reliability trustworthiness dimensions, threats such as compromise of ML components, system malfunction, and human error are recognized as more practically relevant. These risks are associated with software dependencies, API vulnerabilities, and user interaction with the system. Human error remains an important factor despite training efforts, highlighting the importance of user education and workflow integration. Additionally, emerging threats, such as, *sponge attacks* (a form of computational denial-of-service) are identified, where system performance degradation rather than outright failure could impact clinical usability, potentially leading to delays or misinterpretation in time-sensitive environments.

The assessment also considers risks pertaining to the accountability and transparency trustworthiness dimension, including unreported cybersecurity incidents, regulatory non-compliance, and misinformation propagation. However, these are generally assessed as low likelihood due to the presence of quality management systems (QMS), certification processes, and strict regulatory oversight in the healthcare domain. This reflects the maturity of governance structures in clinical environments, which play a critical role in mitigating systemic risks beyond the technical AI components.

In terms of the explainability and interpretability and validity and reliability trustworthiness dimensions, several threats relate to interpretability, robustness, and generalizability. Issues such as reduced interpretability, model drift, and lack of generalization to new clinical settings are acknowledged as potential long-term risks rather than frequent events. These are particularly important in clinical AI, where lack of transparency may reduce clinician trust, while model drift could lead to degraded performance over time if not properly monitored. Similarly, risks associated with insufficient documentation, limited explainability, or inadequate robustness checks are mitigated by regulatory requirements and auditing practices but remain relevant for continuous monitoring.

Finally, the assessment addresses risks pertaining to fairness, including biased training data, discriminatory behavior, and limitations in fairness metrics. These risks are recognized as inherent to AI systems trained on historical clinical data, where demographic or institutional biases may be embedded. While their occurrence is not considered frequent, their impact can be significant, especially in clinical decision-support contexts. Additional concerns include inaccuracies introduced by bias mitigation techniques (e.g., reweighing), highlighting the complexity of achieving fairness without compromising model performance.

Overall, the threat assessment for Model 1 presents a balanced risk landscape, where high-impact threats exist but are generally low in frequency due to the controlled, regulated nature of the healthcare domain. The analysis underscores that the most relevant risks are not necessarily frequent cyberattacks, but rather operational, data-related, and trustworthiness challenges, including human factors, system reliability, and long-term model behavior. This structured identification of threats provides a solid foundation for subsequent impact, vulnerability, and risk analysis within the FAITH AI_TAF.

3.4.1.2 Consequence Assessment - Model 1

The consequence assessment for Model 1 (LSP6_ProstateWholeGlandSegmentationModel) highlights a high-impact risk profile, reflecting the critical role of the AI system within a clinical decision-support environment. Overall, the analysis indicates that if the safe operation of the AI system is compromised, the resulting business and clinical consequences are classified as *very high*. This classification is consistent with the system's deployment in healthcare, where failures may directly or indirectly affect diagnostic workflows, clinical decision-making, and ultimately patient outcomes.

From the perspective of the safety and security and resilience trustworthiness dimensions, the consequences of compromised system security or resilience are assessed as high. This suggests that disruptions such as cyberattacks, system malfunctions, or reduced robustness could significantly impact clinical operations, potentially leading to delays in image analysis, reduced availability of decision-support tools, or degraded performance in segmentation tasks. While the system is not fully autonomous and operates under human oversight, its role in supporting

radiological interpretation means that any degradation in reliability can affect workflow efficiency and clinician confidence.

From the perspective of the explainability and interpretability trustworthiness dimension, the assessment assigns a high consequence level to scenarios where the AI system lacks transparency. In clinical contexts, the inability to interpret or justify model outputs can reduce trust among radiologists and hinder the effective use of the system in practice. This is particularly important for segmentation tasks, where clinicians must validate AI-generated contours and integrate them into diagnostic reasoning. A lack of explainability may therefore not only limit adoption but also increase the risk of unnoticed errors in model outputs.

From the perspective of the privacy enhancement trustworthiness dimension, with high consequences associated with any compromise of sensitive medical data. Given that the system processes patient-level MRI data, breaches of confidentiality could lead to significant ethical, legal, and reputational implications, including violations of GDPR and medical data protection standards. Even though the system operates within controlled environments, the potential impact of privacy failures remains substantial due to the sensitivity of healthcare data.

From the perspective of the fairness dimension, risks related to fairness and bias are assessed as having high consequences. If the AI system produces biased or unfair outputs, potentially due to imbalances in training data or insufficient fairness controls, this could lead to unequal performance across patient groups. In a clinical setting, such disparities may translate into differences in diagnostic support quality, raising concerns about equity, ethical compliance, and adherence to emerging regulatory frameworks such as the AI Act.

From the perspective of the accountability and transparency and validity, and reliability trustworthiness dimensions, risks related to transparency, robustness, and accuracy are also associated with high consequence levels. Failures in these areas could result in incorrect or inconsistent segmentation outputs, which may require additional manual correction by clinicians or, in worst-case scenarios, contribute to misinterpretation of imaging data. Although human oversight mitigates the direct clinical risk, the cumulative impact on workflow efficiency, trust, and system usability remains significant.

From the perspective of the accountability and transparency trustworthiness dimension, the assessment also addresses consequences relating to broader governance and lifecycle considerations, including documentation, fairness processes, and data representation practices. While threats such as poorly defined fairness thresholds or insufficient documentation are not considered highly frequent, their consequences can still be substantial if they materialize, particularly in terms of regulatory compliance, auditability, and long-term system trustworthiness.

Overall, the consequence assessment demonstrates that Model 1 is characterized by low-to-moderate likelihood risks but consistently high to very high impact across all major

trustworthiness dimensions. This imbalance underscores the importance of robust risk mitigation, continuous monitoring, and strict adherence to regulatory and ethical standards. In the context of the FAITH framework, these findings reinforce the need to prioritize safeguards not only against frequent threats but also against low-probability, high-impact failures, which are particularly critical in healthcare AI systems.

3.4.1.3 Vulnerability Assessment - Model 1

The vulnerability assessment for Model 1 (LSP6_ProstateWholeGlandSegmentationModel) reveals a mixed maturity profile, where strong foundations in data governance and regulatory compliance coexist with notable gaps in advanced ML security, monitoring, and robustness mechanisms. The assessment, conducted through a structured questionnaire within the FAITH framework, highlights that while the system benefits from operating in a controlled healthcare environment, several technical and organizational safeguards are either partially implemented or absent.

From the perspective of the security and resilience trustworthiness dimension, the assessment indicates limited evidence of systematic defenses against adversarial attacks. Key techniques such as adversarial training, input sanitisation, or dedicated adversarial input detection mechanisms are not clearly established. Similarly, there is no strong indication that the model architecture has been explicitly designed with robustness against evasion attacks as a primary requirement. This suggests that, although the likelihood of such attacks may be low in a controlled clinical environment, the system could remain vulnerable to sophisticated adversarial manipulation if exposed to malicious inputs.

From the perspective of the security and resilience trustworthiness dimension, the assessment points to gaps in continuous security monitoring and lifecycle-based protection. While standard organizational controls may exist, there is limited confirmation of dedicated processes for continuously assessing the security of ML components, including models, pipelines, and dependencies. Additionally, integration of the AI system into incident detection and response frameworks appears incomplete, potentially delaying the identification and mitigation of security incidents. The lack of explicit evaluation of the model's exposure surface (e.g., APIs, deployment configuration) further suggests that attack surface management could be improved.

From the perspective of the privacy enhancement trustworthiness dimension, the assessment presents a more robust picture with respect to data protection and privacy-preserving mechanisms. The use of well-established, anonymized datasets (e.g., PI-CAI and ProstateNet) significantly reduces risks related to sensitive data exposure, and the absence of personally identifiable information is explicitly acknowledged. However, more advanced privacy-preserving techniques, such as differential privacy or federated learning, are not implemented. This indicates that privacy is primarily ensured through data anonymization and governance practices rather than through algorithmic safeguards embedded within the model itself.

From the perspective of the accountability and transparency trustworthiness dimension, the evaluation of data and model governance practices shows partial strengths. External datasets are validated and originate from reputable, benchmark sources, which enhances trustworthiness and data integrity. However, there is limited evidence of comprehensive mechanisms for detecting data poisoning, monitoring dataset integrity throughout the lifecycle, or systematically controlling information leakage from model outputs. Similarly, restrictions on model transparency (e.g., limiting output detail or protecting architectural information) are not strongly enforced, potentially increasing the risk of reverse engineering or adversarial probing.

From the perspective of the validity and reliability trustworthiness dimension, the assessment highlights important vulnerabilities in monitoring, maintenance, and lifecycle management. Continuous monitoring of model performance, drift detection, and anomaly detection mechanisms are not clearly implemented. Although periodic retraining is planned, the absence of real-time monitoring and early warning indicators may limit the system's ability to detect performance degradation or unexpected behavior in production. This is particularly relevant in clinical settings, where even gradual performance decline can have significant implications.

From the perspective of the fairness, and explainability and interpretability trustworthiness dimensions, the assessment indicates that fairness, bias, and explainability are not yet systematically embedded into the development and deployment lifecycle. While some fairness evaluation practices exist (e.g., fairness testing tools and embedding invariance checks), there is no evidence of continuous bias monitoring, integration of fairness criteria into CI/CD pipelines, or comprehensive bias regression testing. Similarly, explainability and traceability mechanisms are not explicitly enforced as design requirements, which may affect both regulatory compliance and clinician trust.

From the perspective of the accountability and transparency dimension, the assessment suggests that ML-specific considerations relating to organizational governance are not yet fully integrated into broader security-by-design and cyber-resilience strategies. While the system operates within a regulated healthcare framework, dedicated policies addressing AI-specific risks, such as model integrity, adversarial threats, and lifecycle vulnerabilities, are not comprehensively defined. This creates a gap between traditional IT/security governance and the emerging requirements of AI system management.

Overall, the vulnerability assessment demonstrates that Model 1 benefits from strong data quality, controlled deployment, and regulatory alignment, but exhibits moderate vulnerabilities in technical robustness, continuous monitoring, and ML-specific security practices. The findings highlight that the primary weaknesses are not related to data integrity or compliance, but rather to the lack of advanced, proactive mechanisms for detecting, preventing, and responding to AI-specific threats. Addressing these gaps will be essential to strengthen the system's resilience and ensure alignment with emerging standards for trustworthy and secure AI in healthcare.

3.4.2. Model 2

3.4.2.1 Threat Assessment - Model 2

The threat assessment for Model 2 (LSP6_ProstateZonalSegmentationModel), conducted within the FAITH AI_TAF framework, characterizes a very high-criticality AI system deployed in a clinical production environment, where risks to patient safety, diagnostic workflows, and healthcare operations are inherently significant. Similar to Model 1, the system operates under strict regulatory conditions (e.g., MDR, GDPR, ISO 27001), which influence both the frequency and nature of potential threats, generally reducing their likelihood while maintaining high potential impact.

From the perspective of the security and resilience trustworthiness dimension, the assessment indicates that risks such as evasion attacks and white/grey-box adversarial examples are present but relatively infrequent, typically occurring once every few years. These threats are particularly relevant in medical imaging, where adversarial perturbations, such as imperceptible noise in MRI scans, can alter segmentation outputs without being detectable by clinicians. Similarly, oracle-based attacks and data poisoning are considered low-frequency threats, largely due to the regulated nature of healthcare data pipelines and strict governance mechanisms. However, their inclusion in the threat landscape highlights the system's theoretical exposure to advanced adversarial techniques, especially in scenarios involving increased system connectivity or broader deployment.

In terms of the FAITH AI_TAF's privacy enhancement trustworthiness dimension, threats such as label modification, model disclosure, and data disclosure are identified as low-frequency but non-negligible risks. Label modification is particularly relevant in modern healthcare settings where hybrid annotation processes (e.g., involving external teams or AI-assisted labeling) are used to manage large datasets. While regulatory safeguards and controlled environments mitigate these risks, the assessment acknowledges that compromised labeling pipelines could introduce hidden biases or degrade model performance over time. Similarly, risks related to unauthorized access to sensitive data or model internals remain low due to anonymization practices and restricted access controls.

From the perspective of the security and resilience and validity and reliability trustworthiness dimensions, the assessment highlights that threats such as compromise of ML application components, failure or malfunction of the ML system, and model drift are among the most practically relevant. While their expected frequency is moderate (e.g., every few years), these threats are critical due to their potential impact on system performance and clinical usability. The emphasis on rigorous testing, certification processes, and regulatory compliance reduces their likelihood, but does not eliminate the risk of failures arising from software dependencies, integration issues, or evolving clinical data distributions.

In terms of the safety trustworthiness dimension, human-related risks including human error and over-reliance on model predictions, are assessed as relatively infrequent due to structured training and regulatory requirements for clinical users. However, these risks remain important, particularly in scenarios where clinicians may implicitly trust AI outputs without sufficient validation. The presence of human-in-the-loop operation mitigates these risks, but also introduces dependencies on user behavior and workflow integration.

The assessment also identifies threats pertaining to the fairness dimension. Notably, risks such as biased training data, discriminatory behavior, insufficient safeguards against bias, and lack of effective fair data representations are considered to occur periodically (e.g., every few years), reflecting the inherent challenges of learning from historical clinical data. Additional concerns include reweighing inaccuracies, where bias mitigation techniques may inadvertently introduce new errors, and failure to quantify bias, which may limit the system's ability to ensure equitable performance across patient groups. These findings highlight that fairness-related risks are not only present but also complex and dynamic, requiring continuous monitoring and refinement.

In terms of the FAITH AI_TAF's validity and reliability trustworthiness dimension, threats such as lack of generalizability, model drift, poor data quality, and limited diversity in training data are identified as recurring challenges. These risks emphasize the importance of ensuring that the model performs consistently across different clinical settings and patient populations. The absence of continuous adaptation mechanisms or real-time monitoring could exacerbate these issues, particularly as the system is exposed to new data distributions in production.

Finally, the assessment considers risks pertaining to the accountability and transparency trustworthiness dimension, including lack of documentation, insufficient explainability methods, reduced interpretability, and accountability gaps. While these threats are generally assessed as low frequency, their impact can be significant in terms of regulatory compliance, auditability, and user trust. The structured use of documentation tools and regulatory frameworks mitigates some of these risks, but further integration into organizational processes is necessary to ensure long-term accountability.

Overall, the threat assessment for Model 2 presents a comprehensive and balanced risk landscape, where most threats are characterized by low-to-moderate frequency but high potential impact. The most relevant risks are concentrated in system reliability, adversarial robustness, fairness, and long-term model performance, rather than in traditional cybersecurity or governance domains alone. This reinforces the need for continuous monitoring, robustness enhancement, and lifecycle management to ensure the system's safe and trustworthy operation in clinical practice.

3.4.2.2 Consequence Assessment - Model 2

The consequence assessment for Model 2 (LSP6_ProstateZonalSegmentationModel) highlights a consistently high to very high impact profile across all major trustworthiness dimensions,

reflecting the critical role of the system within a clinical diagnostic support environment. According to the FAITH assessment, the overall business consequence of compromised safe operation is classified as very high, indicating that failures in the system could have significant implications for clinical workflows, patient outcomes, and organizational operations.

From the perspective of the safety and security and resilience trustworthiness dimensions, the consequences associated with compromised security or resilience are assessed as high. This suggests that disruptions such as cyberattacks, adversarial inputs, or system malfunctions could lead to reduced availability, degraded performance, or incorrect segmentation outputs, ultimately affecting the efficiency and reliability of radiological processes. While the system is deployed within a controlled healthcare environment, the potential impact of such disruptions remains substantial due to the dependency of clinical workflows on accurate and timely image analysis.

From the perspective of the explainability and interpretability trustworthiness dimension, the consequence level for interpretability and explainability is also assessed as high. This reflects the importance of transparency in clinical AI systems, where radiologists must be able to understand and validate model outputs. A lack of interpretability could reduce clinician trust, hinder adoption, and increase the risk of undetected errors in segmentation results. Although the system provides inherently interpretable outputs (e.g., segmentation masks), the assessment acknowledges that insufficient explainability mechanisms could still have significant operational and clinical consequences.

From the perspective of the privacy enhancement trustworthiness dimension, consequences within this dimension are similarly characterized by high consequences. Any compromise of patient data confidentiality, despite the use of anonymization techniques, would have serious ethical, legal, and reputational implications, particularly in the context of strict regulatory frameworks such as GDPR. Even in controlled environments, the sensitivity of medical imaging data elevates the potential impact of privacy breaches.

From the perspective of the fairness dimension, the assessment assigns a high consequence level to scenarios where the system produces biased or unfair outputs. Given that the model is trained on clinical datasets, the presence of demographic or institutional biases could lead to unequal performance across patient groups, raising concerns related to equity, ethical compliance, and regulatory adherence. These consequences are particularly important in healthcare, where fairness directly affects the quality of care provided to different populations.

From the perspective of the accountability and transparency and validity and reliability trustworthiness dimensions, risks related to transparency and robustness/accuracy are also associated with high consequences. A lack of transparency may lead to accountability issues and reduced auditability, while insufficient robustness or accuracy could result in incorrect segmentation outputs, increased clinician workload, or potential misinterpretation of imaging

data. Although human oversight mitigates the risk of direct clinical harm, the cumulative impact on workflow efficiency, system trust, and decision-making remains significant.

Overall, the consequence assessment for Model 2 demonstrates a risk profile dominated by high-impact outcomes across all evaluated dimensions, even when the likelihood of threats is relatively low. This reflects the inherent nature of clinical AI systems, where even infrequent failures can lead to substantial operational, ethical, and regulatory consequences. The findings emphasize the importance of robust mitigation strategies, continuous monitoring, and strict compliance with trustworthiness principles to ensure the safe and effective deployment of the system in real-world healthcare settings.

3.4.2.3 Vulnerability Assessment - Model 2

The vulnerability assessment for Model 2 (LSP6_ProstateZonalSegmentationModel) reveals a moderate-to-high vulnerability profile, where strong practices in data governance and clinical deployment are counterbalanced by notable gaps in advanced ML security, monitoring, and lifecycle management. While the system benefits from operating in a controlled hospital sandbox environment, the assessment highlights that several AI-specific security and robustness mechanisms are either not implemented or only partially addressed.

From the perspective of the security and resilience trustworthiness dimension, the system exhibits significant vulnerabilities from a security and adversarial robustness standpoint, due to the absence of key defensive mechanisms. There is no clear evidence of adversarial input detection, adversarial training, input sanitisation, or architecture-level robustness design, which are essential for mitigating evasion and adversarial attacks. Additionally, the exposure level of the model, such as its accessibility and potential attack surface, has not been formally evaluated. While basic access control mechanisms (e.g., RBAC for repositories) are in place, these do not fully address the broader spectrum of ML-specific security risks. Furthermore, the system is not fully integrated into incident detection and response frameworks, limiting its ability to detect and react to security breaches in a timely manner.

From the perspective of the validity and reliability trustworthiness dimension, the assessment identifies critical gaps in monitoring, reliability, and lifecycle management. Continuous monitoring mechanisms, such as drift detection, anomaly detection, and real-time performance tracking, are not implemented, and there are no clearly defined indicators for model health that are actively monitored. Although user feedback is collected during system operation, this manual approach does not provide sufficient coverage for early detection of performance degradation or unexpected behavior. Similarly, scheduled retraining, adaptive learning mechanisms, and automated performance validation pipelines are absent, increasing the risk of undetected model drift and reduced generalizability over time.

From the perspective of the privacy enhancement trustworthiness dimension, the assessment presents a relatively stronger profile with respect to data protection and privacy, primarily due to

the use of trusted and curated datasets (e.g., ProstateNet) and operation within a controlled clinical environment. However, advanced privacy-preserving techniques, such as differential privacy or federated learning, are not implemented, indicating that privacy is ensured mainly through governance and anonymization rather than through embedded algorithmic safeguards. Additionally, there is limited evidence of mechanisms to prevent information leakage from model outputs, which could expose sensitive data or enable model inversion attacks under certain conditions.

From the perspective of the accountability and transparency trustworthiness dimension, the system demonstrates partial strengths from a data and model governance perspective. Data sources are validated and originate from high-quality repositories, and documentation practices are supported by regulatory requirements and tools such as the AI model passport. However, there are no comprehensive mechanisms for detecting data poisoning, validating data integrity throughout the lifecycle, or systematically restricting access to model internals and outputs. Moreover, a formal risk analysis of data/model disclosure scenarios across the lifecycle is not fully established, which may leave certain attack vectors unaddressed.

From the perspective of the accountability and transparency trustworthiness dimension, the assessment highlights significant gaps in organizational and policy-level integration relating to this dimension. There are no clearly defined security-by-design policies specific to ML systems, and existing organizational security frameworks have not been updated to address AI-specific risks such as adversarial attacks, model integrity, or data provenance. The system is not fully embedded into cyber-resilience strategies or asset management processes, and security policies addressing ML lifecycle risks are either absent or not yet implemented. This creates a disconnect between traditional IT governance and the requirements of AI system management.

From the perspective of the fairness, and explainability and interpretability trustworthiness dimensions, the system exhibits limited maturity with respect to fairness, bias, and explainability. Although some explainability support exists through saliency maps, there is no systematic implementation of bias detection, fairness monitoring, or mitigation pipelines. Automated fairness testing, bias audits, and CI/CD integration of fairness criteria are not present, and fairness risks are not continuously evaluated during system updates. Additionally, while the system includes human-in-the-loop validation, the absence of formal fairness controls increases the risk of undetected disparities in model performance across patient populations.

From the perspective of the safety trustworthiness dimension, the assessment notes partial strengths in human oversight and operational validation, as clinicians are actively involved in evaluating system outputs during deployment. However, this reliance on manual validation highlights the lack of automated safeguards and decision-support mechanisms for detecting abnormal system behavior. The absence of advanced assessment tools, such as uncertainty quantification or outlier detection, further limits the system's ability to proactively identify risks.

Overall, the vulnerability assessment indicates that Model 2 benefits from strong data quality, controlled deployment, and regulatory alignment, but exhibits substantial vulnerabilities in technical robustness, continuous monitoring, and ML-specific security practices. The most critical weaknesses are related to the lack of proactive, automated mechanisms for detecting, preventing, and responding to AI-specific threats, as well as the limited integration of ML considerations into organizational governance. Addressing these gaps will be essential to enhance the system's resilience and ensure compliance with emerging standards for trustworthy AI in healthcare.

3.5. Initial Risk Assessment results

3.5.1. Model 1

The overall risk assessment for Model 1 (LSP6_ProstateWholeGlandSegmentationModel), as derived from the FAITH TrustGuard tool, reveals a risk landscape dominated by high and very high risks concentrated around system reliability, security, and fairness-related dimensions, despite the presence of several low-risk areas. The most critical risks are consistently associated with *compromise of ML application components*, *failure or malfunction of the ML application*, *evasion attacks*, *data poisoning*, *lack of generalizability*, and *insufficient safeguards against bias*, all of which are assessed as very high risk across multiple trustworthiness properties.

A key finding is that system-level failures and component compromise represent the highest and most recurrent risks. These risks appear across multiple assets, including the model itself and the associated datasets (PI-CAI and ProCAncer-I), indicating that vulnerabilities are not isolated but propagate across the entire AI lifecycle. The high criticality of these risks reflects the dependency of the system on complex ML pipelines, software components, and data flows, where failures may impact not only technical performance but also clinical usability and trust. Closely related to this, *model malfunction* is consistently rated as very high risk, emphasizing the importance of reliability in a clinical decision-support context where segmentation errors can affect downstream diagnostic interpretation.

From a security and robustness perspective, *evasion attacks* and *data poisoning* are identified as major high-risk threats. These risks highlight the system's exposure to adversarial manipulation, even within a controlled healthcare environment. While the likelihood of such attacks may be limited, their potential impact on model outputs and clinical workflows justifies their high-risk classification. Additionally, *denial-of-service scenarios* (e.g., sponge attacks) are assessed as very high risk at the dataset level, further underlining the importance of ensuring robustness not only at the model level but also in data handling processes.

The assessment also reveals significant risks related to fairness and bias, particularly *lack of effective fair data representations* and *insufficient safeguards against bias*, both rated as very high. These findings indicate that, although bias detection activities are already being performed

(e.g., via the DBDM tool), fairness considerations are not yet fully integrated into the development and deployment lifecycle. This creates a risk of unintended disparities in model performance, which is particularly critical in healthcare applications where equitable outcomes are essential.

Another important category of high-risk concerns model lifecycle and performance sustainability, including *model drift leading to inaccurate assessments* and *lack of generalizability in production settings*. These risks highlight the potential degradation of model performance over time or across different clinical environments. The absence of continuous monitoring and drift detection mechanisms further amplifies these risks, as performance issues may remain undetected until they significantly affect system outputs.

In contrast, several risks are assessed as low, including *human error*, *model or data disclosure*, *oracle attacks*, and *regulatory non-compliance*. This reflects the mitigating effect of the controlled deployment environment, strong regulatory frameworks (MDR, GDPR, AI Act), and continuous training of clinical users. However, even though human-related risks are rated as low, they remain relevant due to their potential interaction with high-risk technical failures, particularly in scenarios where clinicians rely on AI outputs.

Overall, the risk profile of Model 1 is characterized by a clear asymmetry: low likelihood but very high impact risks dominate the assessment. The most critical risks are concentrated in technical robustness, system reliability, and fairness integration, rather than in traditional cybersecurity or governance domains. This highlights that the primary challenge for the system is not external threats alone, but the internal resilience and long-term trustworthiness of the AI model and its supporting infrastructure.

3.5.2. Model 2

The overall risk assessment for Model 2 (LSP6_ProstateZonalSegmentationModel), as derived from the FAITH TrustGuard results, reveals a risk landscape dominated by very high risks concentrated in system reliability, robustness, and fairness-related dimensions, while traditional risks such as data disclosure or human error remain comparatively low. The analysis clearly shows that the most critical risks are not isolated but recur across multiple trustworthiness properties, indicating systemic vulnerabilities affecting the model, data pipeline, and operational context.

The highest-risk category relates to system reliability and operational performance, with threats such as *failure or malfunction of the ML application*, *compromise of ML application components*, and *model drift leading to inaccurate assessments* consistently classified as very high risk. These risks reflect the dependence of the clinical workflow on the correct functioning of the segmentation model, where any degradation in performance may directly impact diagnostic support. In particular, the lack of automated monitoring and drift detection mechanisms amplifies these risks, as performance degradation may remain undetected over time. Closely related is the

lack of generalizability in production settings, also rated as very high, highlighting the challenge of maintaining consistent performance across diverse clinical environments and imaging conditions.

From a security and adversarial robustness perspective, threats such as *evasion attacks* and *data poisoning* are also assessed as very high risk, despite their relatively low frequency. This reflects a typical healthcare AI risk pattern, where low-probability but high-impact adversarial events can significantly affect model outputs. The absence of strong adversarial defenses further increases the system's exposure. Additionally, risks related to *compromise of ML components* extend beyond the model itself, affecting the broader infrastructure and data pipeline, reinforcing the systemic nature of these vulnerabilities.

A second major cluster of very high risks has to do with fairness and bias-related issues. Threats such as *insufficient safeguards against bias*, *lack of effective fair data representations*, and *bias in learning processes* are consistently classified as very high. These findings indicate that fairness is a critical concern for Model 2, particularly given the reliance on clinical datasets that may encode demographic or institutional biases. The assessment also highlights more nuanced risks, such as *reweighing inaccuracies* and *failure to quantify bias*, emphasizing that even mitigation strategies may introduce unintended consequences if not carefully managed.

In addition, data-related threats, including *poisoning* and *data quality issues*, contribute significantly to the overall risk profile. Although the use of trusted datasets reduces the likelihood of such risks, their potential impact on model behavior remains substantial. The assessment also identifies *lack of robustness checks* and *poor data diversity* as contributing factors, which may affect both fairness and generalizability.

In contrast, several risks are assessed as low, including *human error*, *model or data disclosure*, *oracle attacks*, and *regulatory non-compliance*. This reflects the mitigating effect of the controlled clinical environment, strict regulatory frameworks, and structured user training. Human teams, particularly clinicians, play a stabilizing role by validating model outputs and reducing the likelihood of harmful consequences. However, risks such as *over-reliance on model predictions* and *reduced ability to detect errors* remain relevant in the long term, especially as trust in AI systems increases.

Overall, the risk profile of Model 2 is characterized by a strong asymmetry between likelihood and impact, where most threats occur infrequently but carry very high consequences when they materialize. The most critical risks are concentrated in technical robustness, system reliability, and fairness integration, rather than in governance or compliance domains. This indicates that the primary challenge for Model 2 lies in ensuring continuous, resilient, and unbiased performance throughout its lifecycle, rather than simply preventing external attacks.

3.6. Countermeasures

3.6.1. Model 1

The FAITH TrustGuard assessment identifies a comprehensive set of recommended controls targeting the highest-risk areas of Model 1, with varying levels of implementation maturity. These controls can be broadly categorized into monitoring and reliability enhancements, adversarial robustness measures, fairness and bias mitigation, data security improvements, and governance integration. While some controls are already being implemented ("treated"), a significant number are postponed for future deployment, and others are considered already addressed or not applicable due to system design choices.

A central group of controls focuses on improving system monitoring and reliability, particularly in response to the high risk of model malfunction and component compromise. The most critical control, *defining and monitoring indicators for proper model functioning,* has been marked as treated, with a clear implementation plan to introduce monitoring mechanisms within the next 12 months. In parallel, additional controls such as *drift detection, anomaly detection, and performance monitoring* are currently postponed, with the development team planning to integrate these capabilities as part of future system enhancements. These controls are expected to be implemented by the AI development team (FORTH and technical partners), in collaboration with clinical users who will provide feedback during operation.

Another key area of intervention has to do with adversarial robustness and security hardening. Controls such as *adversarial training techniques* are already marked as treated, with planned integration in the upcoming development cycle. Additional measures, including *detection of adversarial inputs*, *poisoning mitigation pipelines*, and *secure lifecycle management of ML components*, are currently postponed, indicating that they will be implemented progressively as the system moves closer to certification and broader deployment. These actions are primarily the responsibility of the AI developers, with potential support from cybersecurity experts where needed.

In the domain of fairness and bias mitigation, the recommended controls emphasize the need for stronger integration of fairness into the development lifecycle. Controls such as *embedding fairness criteria into CI/CD pipelines* and *regular bias regression testing* are currently postponed, although preliminary bias assessment is already performed using the DBDM tool. Future implementation will involve extending these practices into automated pipelines and continuous evaluation processes. These actions are expected to be carried out by the data science and AI development teams, with oversight from project stakeholders responsible for ethical and regulatory compliance.

With respect to data security and quality, the assessment recommends controls such as *data cleaning pipelines for adversarial samples*, *data quality assessment tools*, and *secure handling of*

datasets. Many of these controls are currently postponed, with the intention to integrate them during system maturation and potential MDR certification. Some controls, such as ensuring reliable data sources, are considered already satisfied due to the use of trusted clinical datasets and hospital PACS systems. Implementation in this area will involve both data engineers and clinical partners, ensuring alignment with real-world data workflows.

At the organizational and governance level, several controls relate to the integration of ML-specific requirements into existing security and risk management processes. These include *integration into asset management systems*, *alignment with security policies*, and *inclusion in cyber-resilience strategies*. Most of these controls are currently postponed, as they are planned to be implemented during later stages of system deployment, particularly when transitioning from a sandboxed pilot environment to a certified medical device context. These actions will require coordination between technical teams, hospital IT departments, and regulatory/compliance stakeholders.

Finally, some controls are marked as ignored, primarily because they are either already covered by existing mechanisms (e.g., anonymization for privacy protection) or not applicable to the specific model architecture (e.g., LLM-related controls such as prompt attack mitigation). This indicates a selective and context-aware application of the FAITH framework, ensuring that only relevant controls are prioritized. In summary, the control strategy for Model 1 reflects a progressive implementation approach, where critical monitoring and robustness measures are being actively addressed, while more advanced controls, particularly in fairness, security integration, and lifecycle governance, are planned for future deployment. The implementation responsibility is primarily assigned to the AI development team, with clinical users contributing through feedback and validation, and organizational stakeholders supporting governance and compliance alignment. This phased approach is consistent with the current pilot-stage deployment and provides a clear pathway toward achieving a fully robust, secure, and trustworthy AI system in subsequent project phases.

3.6.2. Model 2

The FAITH TrustGuard results for Model 2 identify a comprehensive set of recommended controls targeting the highest-risk areas, with a clear distinction between controls that are already implemented (treated), those planned for future deployment (postponed), and those deemed not applicable or already covered (ignored). The overall strategy reflects a progressive, maturity-driven approach, where core monitoring and compliance mechanisms are in place, while more advanced AI-specific safeguards are scheduled for future integration.

A central group of controls focuses on system reliability and performance monitoring, addressing the very high risks associated with model malfunction, drift, and component compromise. The most critical control, *defining and monitoring indicators for proper model functioning*, is already

treated, with the system currently tracking performance through user feedback and planning to integrate automated monitoring mechanisms in future versions. Similarly, controls related to *outlier detection and malfunction identification* are partially implemented, with ongoing development efforts aimed at enhancing automated detection capabilities. These actions are primarily the responsibility of the AI development team (FORTH and technical partners), with clinical users contributing through continuous evaluation and feedback during system operation.

In the domain of adversarial robustness and security, several controls are identified but remain postponed for future implementation. These include *adversarial training*, *input modification techniques*, and *post-evaluation poisoning detection mechanisms*. The team has already conducted adversarial assessments and plans to incorporate adversarial samples into future training cycles, indicating a phased approach to robustness enhancement. However, some controls—such as changing model architecture or applying differential privacy, are marked as ignored, either because they are not compatible with the nnU-Net architecture or are not considered necessary in the current controlled deployment context.

A significant set of controls αφορά fairness and bias mitigation, reflecting the high-risk profile in this area. While bias detection and fairness audits are already treated and performed as ongoing processes, more advanced controls, such as *automated fairness testing*, *bias regression testing*, and *integration of fairness criteria into CI/CD pipelines*, are currently postponed. These are planned as future enhancements, with the development team aiming to introduce fairness thresholds as part of release validation processes. Additional actions, such as *testing embedding invariance across demographic groups*, are also planned, indicating a move toward more systematic fairness evaluation.

Regarding data security and governance, many controls are marked as ignored because they are already implicitly addressed through existing practices, such as the use of trusted datasets (e.g., ProstateNet), data and model passports, and controlled access mechanisms. For example, controls related to data provenance, dataset versioning, and access control are considered already implemented within the current framework. However, some advanced controls, such as automated data cleaning for adversarial samples or formal lifecycle risk analysis, are not yet fully developed, suggesting potential areas for future improvement.

At the organizational and policy level, several controls remain postponed, particularly those related to integrating ML-specific requirements into security policies, asset management processes, and cyber-resilience strategies. The assessment explicitly notes the absence of formal policies in these areas, indicating that implementation will depend on organizational maturity and alignment with institutional frameworks. These actions will require collaboration between AI developers, hospital IT departments, and governance stakeholders, especially as the system moves toward certification and broader deployment.

Finally, a number of controls are marked as ignored due to non-applicability, particularly those related to generative AI (e.g., prompt attacks, RAG, token-level probability estimation), which are not relevant to the nnU-Net-based segmentation model. This demonstrates a context-aware application of the FAITH framework, ensuring that only relevant controls are considered.

In summary, the control strategy for Model 2 reflects a balanced and pragmatic approach, where essential controls for monitoring, compliance, and basic robustness are already in place, while more advanced mechanisms, particularly in adversarial robustness, fairness integration, and lifecycle governance, are planned for future implementation. The responsibility for implementation lies primarily with the AI development team, supported by clinical users for validation and organizational stakeholders for policy integration, ensuring a coordinated pathway toward a more robust and trustworthy AI system.

3.7 Formal Tool Feedback

Formal feedback on the FAITH tools was collected following the completion of the trustworthiness assessment activities within the Healthcare LSP. Feedback was provided through structured questionnaires designed to capture an initial high-level assessment of each tool's usefulness and usability. Detailed feedback will be reported in the dedicated WP2 deliverable; only a brief summary is provided here.

3.7.1 TrustSense

The formal feedback questionnaire for TrustSense was completed by the radiologists involved in the pilot. No specific comments or improvement suggestions were recorded, reflecting a straightforward and satisfactory assessment experience.

3.7.2 AI Model Hub

The formal feedback questionnaire for the AI Model Hub was completed by the risk assessor. At a high level, the tool was assessed as highly useful, particularly in supporting transparency and explainability of the AI SUT. One improvement suggestion was recorded: establishing a direct linkage between the Model Hub and Data Passport entries and the corresponding TrustGuard assessment reports, so that asset documentation and risk assessment outputs are navigable as an integrated record.

3.7.3 TrustGuard

The formal feedback questionnaire for TrustGuard was completed by the risk assessor following the conclusion of the structured risk assessment activities. The tool was considered effective in supporting the systematic execution of the FAITH AI_TAF methodology. Three improvement suggestions were recorded: (i) the tool should display the version of TrustGuard in use, (ii) it should indicate the version of the Knowledge Base it relies upon, and (iii) it should provide explicit

traceability of how each risk score is calculated, so that assessors can verify and audit the underlying reasoning.

4. AI System Evaluation

4.1. Quantitative metrics on pilot data

The AI SUT is validated against well-known segmentation metrics, summarised in Table 5 below.

Table 5. Metrics used for the validation of the AI SUT in the healthcare LSP.

Metric	Abbreviation	Description
Dice Coefficient	DS	Measures volumetric overlap between predicted and ground-truth segmentation. Ranges from 0 (no overlap) to 1 (perfect agreement).
Hausdorff Distance	HD	Quantifies worst-case boundary error. A lower value indicates the predicted contour closely follows the reference everywhere; a high value reveals localised deviations.
Average Surface Distance	ASD	Measures mean point-to-point distance between predicted and ground-truth boundaries. Reflects overall boundary alignment and is less sensitive to outlier points than HD.
Recall	—	Proportion of true positive pixels in the ground truth correctly identified by the model; captures how well the model avoids missing relevant regions.
Precision	—	Proportion of pixels predicted as positive that belong to the ground truth; reflects the model's ability to avoid false positive predictions.

The two prostate segmentation models deployed in the Healthcare LSP demonstrate strong quantitative performance across all evaluated metrics. The Whole Gland Segmentation Model (LSP6) achieves a Dice Score of 91.14% and a Jaccard Index of 84.03%, reflecting a high degree of volumetric overlap between AI-generated and ground-truth segmentations, while Precision (91.16%) and Recall (91.41%) are closely aligned, indicating a well-balanced model with minimal over- or under-segmentation. The Zonal Segmentation Model, which addresses the more anatomically complex task of delineating distinct prostate zones, achieves a Dice Score of 86.99% and a Jaccard Index of 77.58%, with similarly balanced Precision (87.19%) and Recall (88.14%). Both models attain a 95th percentile Hausdorff Distance of 3mm, indicating that even the worst-case boundary deviations remain within a clinically acceptable range. The slightly lower scores of the Zonal Model relative to the Whole Gland Model are consistent with the inherently greater

segmentation difficulty posed by zonal boundaries, which are less anatomically distinct and more subject to inter-observer variability. Taken together, these results confirm that both models meet a high standard of segmentation accuracy suitable for clinical decision-support in prostate MRI analysis.

In addition to these quantitative metrics, the Healthcare LSP incorporated a real-world clinical usability and quality assessment based on direct radiologist interaction with the deployed system. In addition to quantitative segmentation metrics, the Healthcare LSP incorporated clinician-based qualitative assessment of AI outputs across 179 real clinical cases. As reported in detail in Section 2.4, aggregated clinician ratings demonstrated highly positive acceptance of the deployed solution, with 99.4% of cases rated as clinically acceptable. These real-world results are particularly important because they complement purely technical segmentation metrics with evidence from operational clinical use. While metrics such as Dice Score and Hausdorff Distance quantify geometric agreement against reference annotations, clinician ratings reflect practical usefulness, visual adequacy, trust, and suitability for workflow integration. The combination of technical benchmarking and user-centred assessment provides a robust evaluation framework for the Healthcare LSP and supports the continued progression toward broader clinical adoption and replication pilot activities.

To complement the quantitative segmentation metrics, a structured qualitative assessment was conducted through direct clinician interaction with the deployed AI system. The measurement instrument was a standardised post-task questionnaire covering five criteria: (i) segmentation accuracy, (ii) consistency, (iii) usability, (iv) suggested improvements, and (v) overall recommendation for clinical use. A total of 171 questionnaire responses were collected across the deployment period. Detailed findings across all five criteria, including results on usability and model robustness, are reported in Section 2.5.3. In summary, clinicians reported high segmentation accuracy, acceptable robustness across heterogeneous patient datasets, and high usability, expressing unanimous confidence in the system as a clinical decision-support tool.

Clinicians consistently rated the segmentation accuracy of the AI tool at a high level. Across 166 valid responses, the dominant assessment (approximately 85% of cases) described the accuracy as "generally very good", with clear and reliable delineation of both the peripheral and transitional zone boundaries. In the remaining cases, accuracy was rated as "generally good", with minor inaccuracies localised to anatomically challenging regions, most notably the prostate apex, base, and hypertrophic transitional zone. These suboptimal outcomes were predominantly associated with post-TURP anatomy and enlarged prostates presenting with heterogeneous signal intensity, consistent with the known challenges these anatomical contexts pose for automated segmentation. A single case documented a complete detection failure, representing an isolated outlier across the entire evaluation dataset.

Clinician feedback on consistency was highly uniform across all 171 responses. The tool was described as demonstrating consistent performance across patients with typical prostate anatomy, with robustness maintained across varying prostate sizes and morphologies. Minor inaccuracies were acknowledged as occasionally or sporadically occurring at the base and apex in cases of prostatic enlargement with heterogeneous signal, and in a subset of post-operative cases. Importantly, all responses affirmed that these limitations did not undermine the model's overall acceptable robustness in clinical deployment conditions.

Usability received the strongest and most uniform positive signal across the entire questionnaire dataset. All 171 clinicians provided functionally identical feedback, describing the interface as intuitive and easy to navigate even for first-time users. The ability to visualise segmentation outputs efficiently was highlighted, with particular appreciation expressed for the overlay toggle and transparency adjustment features, which were noted as especially helpful during clinical review. The complete unanimity of this response across all cases and sessions reflects a consistently positive user experience throughout the deployment period.

4.2. Trustworthiness Assessment

The application of the FAITH tools, primarily TrustGuard, supported by the Model and Data Passports, of the AI Model Hub has enabled a systematic, asset-level evaluation of the AI SUT across key Trustworthy AI characteristics. The assessment integrates evidence from threat analysis, consequence and vulnerability assessments, and risk evaluation results, providing a multi-dimensional view of how the model, data, and human actors perform against the core trustworthiness characteristics. Given the nature of the healthcare use case, particular emphasis is placed on Validation & Reliability, Security & Resilience, Transparency & Accountability, Explainability, Safety, Fairness, and Privacy, as these dimensions are most critical for clinical AI systems operating under regulatory constraints and human oversight.

4.2.1 Validity & Reliability

The FAITH tools indicate that the AI SUT demonstrates strong baseline performance in terms of validation metrics, with high segmentation accuracy (e.g., Dice score and Jaccard index) and consistent behavior across evaluation datasets, as documented in the Model Passport. This supports a positive assessment of the system’s initial validity and technical reliability. However, TrustGuard highlights very high risks related to model malfunction and performance degradation, particularly in relation to *model drift* and *lack of generalizability in production settings*. These risks stem from the absence of fully implemented continuous monitoring, drift detection, and anomaly detection mechanisms. Although user feedback is collected during pilot execution, the lack of automated monitoring reduces the system’s ability to detect performance issues in real time. Overall, while the system is valid and reliable under controlled evaluation conditions, its long-

term reliability in dynamic clinical environments remains an area for improvement, requiring the integration of continuous validation and monitoring processes.

4.2.2 Security & Resilience

The FAITH assessment identifies Security & Robustness as one of the most critical dimensions for the AI SUT. TrustGuard results show very high risks associated with adversarial threats, including evasion attacks, data poisoning, and compromise of ML components. On the positive side, the system benefits from deployment in a controlled, sandboxed hospital environment, limited external exposure, and the use of trusted datasets. These factors significantly reduce the likelihood of external attacks. However, the vulnerability assessment reveals that advanced robustness mechanisms, such as adversarial training, input sanitisation, and adversarial detection, are only partially implemented or planned for future integration. Additionally, the lack of continuous security monitoring and lifecycle-based protection mechanisms introduces further vulnerabilities. While some controls (e.g., adversarial training) are planned within the next development cycle, the current setup does not fully align with best practices for robust AI systems. In summary, the system exhibits moderate robustness due to environmental controls, but technical robustness against AI-specific threats remains a high-priority improvement area.

4.2.3 Transparency & Accountability

Transparency and accountability are partially supported through the FAITH Model and Data Passports, which provide structured documentation of datasets, model development processes, and evaluation metrics. These tools enhance traceability, reproducibility, and governance, aligning with regulatory expectations (e.g., MDR, AI Act). However, TrustGuard identifies very high risks related to the compromise of ML application components, as well as gaps in the integration of ML systems into organizational security policies, asset management processes, and lifecycle governance frameworks. Many controls recommended in TrustGuard in this area are currently postponed, particularly those related to formal integration into hospital IT governance and cybersecurity strategies. Accountability is further supported through human-in-the-loop operation, where radiologists remain responsible for final clinical decisions. Nevertheless, the lack of full integration into institutional processes limits the system's organizational accountability maturity. Overall, the system demonstrates good transparency at the documentation level, but accountability mechanisms at the organizational and lifecycle level require further strengthening.

4.2.4 Explainability & Interpretability

The AI SUT performs relatively well in this dimension due to its inherently interpretable outputs, such as segmentation masks and probability heatmaps, which are directly visualized and assessed by clinicians. This aligns well with the requirements of radiological workflows, where visual

interpretability is critical. Despite this, the FAITH tools identify residual risks related to limited explainability mechanisms beyond visual outputs, including the lack of advanced techniques for explaining model behavior or uncertainty. While the system does not rely on black-box classification outputs, the absence of formal explainability frameworks or uncertainty quantification methods may limit deeper understanding and auditability. In practice, the presence of clinical oversight mitigates many explainability-related risks, as radiologists can validate and correct outputs. Therefore, the system achieves a moderate-to-high level of interpretability, though further enhancements could improve transparency and user trust.

4.2.5 Safety

The safety dimension of the FAITH AI_TAF for the Healthcare LSP is hard linked with the impact of the AI SUT has in human life or/and health deterioration. Thus, the resilience of the AI SUT against perturbations that may compromise its clinical outputs is of high concern. A key concern identified during the initial pilot relates to adversarial vulnerability: the susceptibility of deep learning-based segmentation models to carefully crafted input perturbations that, while imperceptible to the human eye, can induce significant degradation in model performance. In the context of the Healthcare LSP, where the AI SUT performs prostate gland and zone segmentation from multiparametric MRI, such vulnerabilities carry direct clinical implications, including erroneous delineation of anatomical boundaries that could affect biopsy targeting, volume estimation, and treatment planning.

As part of the security and resilience, and subsequently the safety dimension, a systematic adversarial attack assessment was conducted on both nnU-Net models comprising the AI SUT.

While the controlled clinical deployment environment reduces the likelihood of such attacks occurring in practice, their potential consequences, including erroneous segmentation outputs affecting biopsy targeting, volume estimation, or treatment planning, are sufficiently severe to warrant systematic empirical characterisation.

Furthermore, deep learning-based medical image segmentation models are known to be inherently susceptible to imperceptible input perturbations, a vulnerability that is not captured by standard performance metrics evaluated under clean conditions.

The specific mitigation measure was described by the Risk Assessor during the TrustGuard usage and specifically the high-risk evasion threat in the dimensions of “Safe and “Secure and Resilience”.

This assessment constitutes the first phase (Phase 1) of a two-phase adversarial robustness plan. The second phase (Phase 2) will incorporate adversarial training and defense strategies to

increase the adversarial robustness of the AI models. The findings of the adversarial attack assessment are summarised below and inform the countermeasures identified in Section 3.6.

4.2.5.1 Adversarial Attack Assessment Methodology

The adversarial attack assessment was performed on the two nnU-Net v2 models that constitute the AI SUT: (i) the Whole Gland (WG) model, which performs binary segmentation of the whole prostate from T2-weighted MRI, and (ii) the Prostate Zones model, which performs multi-class segmentation of the transition and central zones (TZ+CZ) and the peripheral zone (PZ). Both models were evaluated under the Fast Gradient Sign Method (FGSM)[1], a white-box single-step adversarial attack in which the adversary has complete access to model architecture, trained weights, and gradient information. This white-box scenario represents the strongest possible adversary and provides an upper bound on model vulnerability. The FGSM perturbation is bounded in the L-infinity norm, constraining each voxel intensity modification to at most ε from its original value. Experiments were conducted across eleven attack strengths ($\varepsilon \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$), with $\varepsilon = 0.0$ serving as the clean-inference baseline. Robustness was evaluated using three complementary segmentation metrics: the Dice Similarity Coefficient (DSC), measuring volumetric overlap; the 95th percentile Hausdorff Distance (HD95), capturing boundary outlier displacement; and the Average Surface Distance (ASD), quantifying mean boundary-level deviation. The WG model was evaluated on a validation set of 258 cases, and the Prostate Zones model on 158 cases, both drawn from the ProstateNet dataset of T2-weighted MRI scans with expert-annotated ground-truth segmentation masks. The spatial distribution of the intensity modification is defined by the vector δ (which is the sign of the gradient vector of the loss function w.r.t. the input x, scaled by the attack strength ε). The gradient vector of the loss function for a given input indicates the direction in the input space where the AI model is most sensitive to input x. Due to the sign operator, the magnitude of the intensity modification is uniformly equal to the attack strength ε throughout the input.

4.2.5.2 Key Findings

Under clean conditions, the WG model achieves a DSC of approximately 0.89, reflecting strong baseline segmentation performance. However, adversarial vulnerability is pronounced even at low attack strengths: a perturbation of $\varepsilon = 0.1$ reduces the DSC to approximately 0.73, representing a drop of 0.16 points from a single-step, low-magnitude attack. Beyond $\varepsilon = 0.3$, the Dice curve exhibits partial saturation, plateauing around 0.66–0.70. In contrast, distance-based metrics (HD95 and ASD) continue to degrade monotonically: HD95 increases from approximately 3 mm at baseline to approximately 16 mm at $\varepsilon = 1.0$, and ASD rises from 0.8 mm to approximately

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

5.5 mm. This divergence between volumetric and boundary-level metrics underscores the importance of using multiple complementary evaluation criteria: while Dice saturation may suggest partial robustness, the progressive boundary displacement revealed by HD95 and ASD indicates continuing structural degradation at higher attack strengths.

The Prostate Zones model exhibits differential vulnerability between the two segmentation targets. At baseline, TZ+CZ achieves a DSC of approximately 0.90 and PZ achieves approximately 0.83. Under adversarial perturbation, PZ is consistently more vulnerable across all metrics: its DSC declines to approximately 0.72 at $\varepsilon = 0.1$ and continues falling to approximately 0.45 at $\varepsilon = 1.0$, compared to TZ+CZ which declines more gradually from 0.90 to approximately 0.63. This zone-specific vulnerability is attributable to PZ's thinner, more irregular morphology, which provides less geometric redundancy against adversarial perturbations. HD95 values at $\varepsilon = 1.0$ reach approximately 21 mm for PZ and 19 mm for TZ+CZ, both representing clinically significant boundary displacement. The elevated vulnerability of the PZ is particularly salient given that this zone is the primary site of prostate cancer detection.

A fine-grained slice-level analysis revealed that adversarial degradation is non-uniformly distributed along the superior–inferior axis of the prostate. Across all three segmentation targets (WG, TZ+CZ, PZ), the basal and apical extremities consistently exhibit higher adversarial vulnerability than the mid-gland region. For the WG model, the base shows the highest mean Dice drop (approximately 0.20–0.25), followed by the apex (approximately 0.17–0.23), while the mid-gland remains the most resilient (approximately 0.13–0.18). A statistically significant negative correlation was observed between ground-truth foreground area and Dice drop (Spearman $\rho \approx -0.18$ to -0.28, $p < 10^{-34}$): slices with smaller prostate cross-sections are disproportionately vulnerable to adversarial attacks, reflecting the reduced geometric redundancy of the gland at its anatomical extremities. The PZ, due to its uniformly thin morphology, does not exhibit a clear mid-gland resilience advantage, showing elevated adversarial vulnerability across all anatomical positions.

The worst-case failure analysis at $\varepsilon = 1.0$ reveals catastrophic segmentation collapse in the most adversarially affected patients. Patients achieving clean Dice scores above 0.80 under natural conditions experience drops to below 0.10 under maximum-strength attack. The dominant failure mode is spatial displacement: FGSM attack causes predicted masks to shift far from their ground-truth positions, producing massive over-segmentation into surrounding anatomical tissue and complete loss of the characteristic morphological shapes of the prostate zones. These extreme failure cases underscore the clinical risk associated with the absence of adversarial defense mechanisms.

4.2.5.3 Clinical and Safety Implications

The results of this adversarial attack assessment confirm that both nnU-Net models comprising the Healthcare LSP AI SUT are not robust to adversarial perturbations in their current state. From a clinical safety perspective, the identified vulnerabilities translate to a tangible deployment risk: minor intensity modifications to an MRI scan — whether introduced maliciously or arising from data corruption, scanner artefacts, or domain shift — could cause the segmentation model to produce erroneous boundaries. Incorrect prostate delineation could lead to suboptimal biopsy targeting, inaccurate volume estimation, or flawed radiotherapy treatment planning. These failure modes are directly relevant to the AI SUT's intended purpose as a second reader in a clinical radiology workflow.

4.2.5.4 Planned Adversarial Defense: Phase 2 — Adversarial Training

This assessment constitutes Phase 1 of a two-phase adversarial robustness programme for the Healthcare LSP. Building directly on the vulnerability profile characterised above, Phase 2 will introduce adversarial training as the primary mitigation strategy. Adversarial training augments the model training process with adversarially perturbed examples, enabling the model to learn representations that are robust to such perturbations. Specifically, future model versions of the AI SUT will be retrained incorporating FGSM-generated adversarial examples across representative attack strengths identified in this assessment as most impactful (particularly ε = 0.1–0.3, where the steepest performance degradation occurs). The efficacy of adversarial training will be evaluated using the same three-metric framework (DSC, HD95, ASD) and validated across the anatomical regions identified as most vulnerable (base, apex, and PZ). This plan is reflected in the countermeasure status reported under the Safety trustworthiness dimension of the FAITH AI_TAF risk management profile (Section 3.6). The adversarial training mitigation is currently classified as Postponed pending the Phase 1 assessment outcomes, which have now been documented in this deliverable, and will be actioned in the replication pilot phase at UNIPI. The completion of Phase 2 adversarial training is expected to advance the AI SUT's adversarial robustness profile and contribute to its progression towards TRL 7 and CE Mark readiness as Software as a Medical Device.

4.2.6 Fairness

Fairness is identified as a critical (fairness dimension is very critical in the healthcare domain) for which the team pays attention and has already performed relevant assessment and activities. TrustGuard results highlight very high risks related to insufficient safeguards against bias and lack of effective fair data representations. While the development team has already conducted bias assessments using the DBDM (in-house) tool, fairness evaluation is not yet fully integrated as an automatic step into the development lifecycle. In particular, there is a lack of: (i) continuous bias

monitoring, and (ii) integration of fairness criteria into CI/CD pipelines. The use of large, multi-center datasets partially mitigates bias risks but does not eliminate them. The absence of automated and continuous fairness controls increases the risk of undetected disparities in model performance across patient groups. Therefore, fairness remains an area where the system shows initial awareness but reduced operational maturity, requiring further development to meet the high expectations of trustworthy AI in healthcare.

4.2.7 Privacy

Privacy is one of the strongest dimensions of the AI SUT, supported by both technical and organizational measures. The system relies on fully anonymized datasets, and additional anonymization processes are applied to input data during operation. These practices significantly reduce the risk of personal data exposure. Furthermore, the controlled deployment environment and compliance with GDPR and healthcare data protection standards reinforce the system's privacy posture. As a result, most privacy-related risks, including data disclosure and model leakage, are assessed as low. However, the FAITH tools highlight that advanced privacy-preserving techniques, such as differential privacy or federated learning, are not implemented. While not strictly necessary given the current setup, their absence may limit scalability or future deployment in more distributed environments. In summary, the system demonstrates a high level of privacy protection through data governance and anonymization, with potential for further enhancement through advanced privacy-preserving methods.

4.3 Preliminary Impact Evaluation

A preliminary impact evaluation was conducted to assess the effects of the AI SUT within the operational context of the Healthcare LSP. At this stage of the pilot, the evaluation is primarily qualitative and focuses on both observed and anticipated impacts associated with the deployment of the prostate MRI segmentation solution in clinical environments. The assessment considers positive and negative dimensions, as well as actual impacts already evidenced during pilot execution and potential impacts expected through broader operational use.

From a positive perspective, the most immediate observed impact concerns the improvement of workflow efficiency during radiological image review. Automated generation of whole-gland and zonal prostate segmentations has the potential to reduce the manual effort required for contouring anatomical structures, particularly in time-constrained clinical settings. Even when final clinician review remains necessary, the availability of pre-generated segmentations can accelerate case preparation and support more standardised image interpretation. A second positive impact relates to consistency and reproducibility. Manual prostate segmentation is known to be affected by inter-observer and intra-observer variability. By providing systematic AI-generated contours based on consistent inference logic, the SUT may contribute to more harmonised delineation practices across users and institutions. This is particularly relevant for

multicentre settings, training environments, and future downstream applications such as quantitative imaging analysis or radiotherapy planning support. The pilot has also demonstrated positive effects in relation to user engagement with trustworthy AI practices. Through the use of TrustSense, TrustGuard, and the AI Model Hub, clinical and technical stakeholders were actively involved in structured trustworthiness assessment activities. This has increased awareness regarding explainability, governance, risk management, and responsible AI deployment, thereby generating organisational value beyond the AI model itself. Another anticipated positive impact concerns scalability and transferability. The deployment of the solution across two clinical sites (Heraklion and Pisa) creates the basis for future multi-site adoption of similar trustworthy AI workflows in radiology. Lessons learned during the initial pilot may therefore extend beyond the immediate use case and inform future implementations in other hospitals or imaging domains.

These expected benefits are also consistent with recent evidence on AI adoption in radiology. A 2026 umbrella review of systematic reviews and meta-analyses reported that AI systems show strong potential as simultaneous second readers across multiple radiological tasks, supporting clinicians in image interpretation, case prioritisation, and workflow efficiency. The review also highlighted the potential of AI to reduce fatigue-related diagnostic variability and assist radiologists during high workload or off-hour conditions. Although the present Healthcare LSP focuses on segmentation rather than direct diagnosis, the same human-AI collaboration principles apply, whereby AI can reduce repetitive workload, improve consistency, and allow clinicians to focus on higher-value interpretative tasks. At the same time, several potential negative or limiting impacts were identified. One important consideration is the risk of overreliance on AI outputs if users were to accept segmentation results without adequate clinical verification. Although the pilot was explicitly designed with expert oversight, this remains an important factor for future operational scaling and underlines the continued need for human-in-the-loop use models. A second potential negative impact concerns workflow disruption during early adoption. Introducing a new standalone AI application into established hospital routines may initially require additional time for training, case preparation, and familiarisation. If integration is not sufficiently smooth, short-term efficiency gains may be offset by temporary operational burden. Another relevant challenge is performance variability across data sources. As with many medical imaging AI systems, segmentation performance may vary according to scanner vendor, acquisition protocol, image quality, or population characteristics. While the pilot includes multicentre validation elements, broader deployment would require continued monitoring to ensure robust generalisation. Potential organisational impacts were also considered. These include the need for ongoing technical maintenance, version control, user support, cybersecurity oversight, and alignment with evolving regulatory requirements such as the EU AI Act and medical device obligations. Such requirements may increase operational complexity but are necessary for responsible clinical deployment.

Overall, the preliminary impact evaluation indicates that the Healthcare LSP AI solution has a strong positive potential, particularly in improving efficiency, consistency, and trust-aware AI adoption in radiology workflows. The identified risks and limitations are manageable provided that human oversight, technical monitoring, governance controls, and user training remain integral parts of deployment. As pilot execution continues and replication activities commence, further evidence will be collected to support a more mature quantitative and longitudinal impact assessment.

5. Regulatory and Ethical Assessment

The deployment and evaluation of the AI SUT within the Healthcare LSP require careful consideration of both regulatory and ethical frameworks governing the development and use of Artificial Intelligence in clinical environments. The AI SUT performs automated segmentation of the prostate gland and its anatomical zones from T2-weighted MRI scans and is therefore positioned within the regulatory category of Software as a Medical Device (SaMD), as defined under the European Union Medical Device Regulation (EU MDR 2017/745).

The regulatory and ethical assessment performed in the context of the Healthcare LSP focuses on three complementary dimensions:

- Compliance with medical device regulatory frameworks, particularly the EU MDR requirements applicable to software-based diagnostic tools.
- Alignment with the emerging EU AI Act requirements, especially those related to high-risk AI systems in healthcare.
- Ethical and governance considerations, including patient privacy, transparency, clinical oversight, and responsible AI deployment.

The objective of this section is not to represent a formal regulatory submission but to assess the extent to which the AI SUT and its deployment within the LSP environment are aligned with the regulatory expectations that would apply to its potential future clinical deployment.

5.1. Pathway to CE Marking

The AI-based prostate segmentation software developed within the Healthcare LSP is expected to fall under the regulatory scope of Software as a Medical Device (SaMD) under the EU Medical Device Regulation (EU MDR 2017/745). According to the MDR definition, software intended for diagnostic or therapeutic purposes is considered a medical device and must comply with the applicable regulatory requirements before being placed on the European market.

5.1.1 SaMD Qualification and Classification

The qualification of the AI SUT as medical device software is primarily determined by its intended clinical purpose and its role within the diagnostic workflow. In the Healthcare LSP, the SUT is designed to assist radiologists in the segmentation of the prostate gland and anatomical zones from T2-weighted MRI images, supporting the interpretation of prostate mpMRI examinations within the radiology workflow. The system provides automated segmentation outputs that are subsequently reviewed and interpreted by clinical experts, enabling improved efficiency and consistency in prostate imaging analysis.

5.1.1.1 Qualification as Medical Device Software

Under the EU Medical Device Regulation (EU MDR 2017/745), software qualifies as a medical device when it is **intended by the manufacturer for medical purposes**, including diagnosis, prevention, monitoring, prediction, prognosis, treatment, or alleviation of disease. Software that performs such functions independently of hardware devices is typically referred to as **Software as a Medical Device (SaMD)** or **Medical Device Software (MDSW)** within European regulatory guidance.

According to the MDR interpretation guidelines (MDCG 2019-11), the qualification of software as medical device software typically follows a structured decision process that includes the following steps:

Determining whether the software has a medical purpose.

The AI SUT processes medical imaging data (MRI) and produces segmentation outputs that are used during diagnostic image interpretation. Because these outputs directly contribute to the clinical assessment of prostate anatomy and potential pathological findings, the software clearly fulfils a medical purpose.

Assessing whether the software performs an action beyond simple data storage or communication.

Software that merely stores, archives, or transfers medical images is generally not considered medical device software. However, the SUT performs **algorithmic analysis of imaging data**, generating segmentation masks and visualization outputs that assist in clinical interpretation. This analytical functionality places the system within the scope of regulated medical software.

Evaluating whether the software output influences clinical decision-making.

The segmentation results produced by the AI system support radiologists in identifying anatomical boundaries and potentially relevant regions in prostate MRI scans. While clinicians retain full responsibility for diagnostic interpretation, the software provides **information used to support diagnostic decision-making**, which is a key criterion for qualification as medical device software.

Based on these considerations, the AI SUT implemented in the Healthcare LSP can be reasonably considered **Medical Device Software under the MDR framework**.

5.1.1.2 MDR Risk-Based Classification Framework

Once software is qualified as medical device software, the next step is determining its risk class under the MDR classification rules. The MDR uses a risk-based classification system, where the

regulatory requirements increase with the potential risk to patient safety associated with incorrect outputs or software malfunction.

Medical devices in the MDR are categorized into four classes:

- Class I – Low risk
- Class IIa – Moderate risk
- Class IIb – Higher risk
- Class III – Highest risk

For software, the most relevant classification rule is Rule 11 (Annex VIII of the MDR), which specifically addresses software used for medical purposes. Rule 11 states that software intended to provide information used for diagnostic or therapeutic decisions is generally classified as Class IIa, unless the consequences of incorrect decisions could result in more serious harm.

5.1.1.3 Application of Rule 11 to the AI SUT

Under Rule 11, the classification of medical device software depends primarily on the potential impact of incorrect outputs on patient health outcomes. The classification thresholds are typically interpreted as follows:

Class IIa

Software intended to provide information used to support diagnostic or therapeutic decisions, where incorrect outputs would not directly result in serious deterioration of health.

Examples include clinical decision-support tools or software that assists clinicians in interpreting medical data without determining treatment autonomously.

Class IIb

Software where incorrect information could lead to serious deterioration of a patient's health or require surgical intervention. This may include software used in critical diagnostic processes or systems whose outputs significantly influence treatment decisions.

Class III

Software where incorrect outputs could directly lead to death or irreversible deterioration of health, such as systems used in acute diagnostic decision-making for life-threatening conditions.

This rule reflects a core regulatory principle: the classification is driven by the severity of harm that could result if the software output is wrong.

5.1.1.4 Indicative Classification for the Healthcare LSP AI System

Considering the intended role of the AI SUT within the Healthcare LSP, several factors influence its expected classification:

Clinical function of the software

The system performs automated segmentation of prostate MRI images, supporting the identification of anatomical regions during radiological interpretation.

Human oversight

The AI outputs are reviewed by radiologists, who remain responsible for the final diagnostic interpretation and clinical decision-making.

Nature of clinical impact

While segmentation results may influence image interpretation and diagnostic workflow efficiency, the system does not autonomously determine diagnoses or treatment decisions.

Potential consequences of incorrect outputs

Errors in segmentation could affect workflow efficiency or require manual correction by the radiologist, but they would not independently lead to clinical intervention without human validation.

Given these characteristics, the expected classification of the AI SUT would most likely fall within Class IIa, as it provides information that supports diagnostic decision-making without directly determining patient management. However, if the software were to evolve toward more advanced diagnostic functionality, such as automated lesion detection or direct diagnostic classification, the classification could potentially increase to Class IIb, depending on the level of clinical reliance placed on the system outputs.

5.1.1.5 Regulatory Implications of the Expected Classification

The expected classification has several important implications for the future regulatory pathway of the software. Under MDR, Class IIa and IIb medical devices require:

- Assessment by a Notified Body before CE marking,
- Implementation of a Quality Management System (ISO 13485),
- A risk management process compliant with ISO 14971,
- Software lifecycle documentation aligned with IEC 62304, and
- Demonstration of clinical performance through clinical evaluation and validation studies.

These requirements ensure that the safety, performance, and reliability of the software are rigorously evaluated before it can be placed on the European market. They also highlight the importance of early regulatory planning during the development of AI-based healthcare technologies.

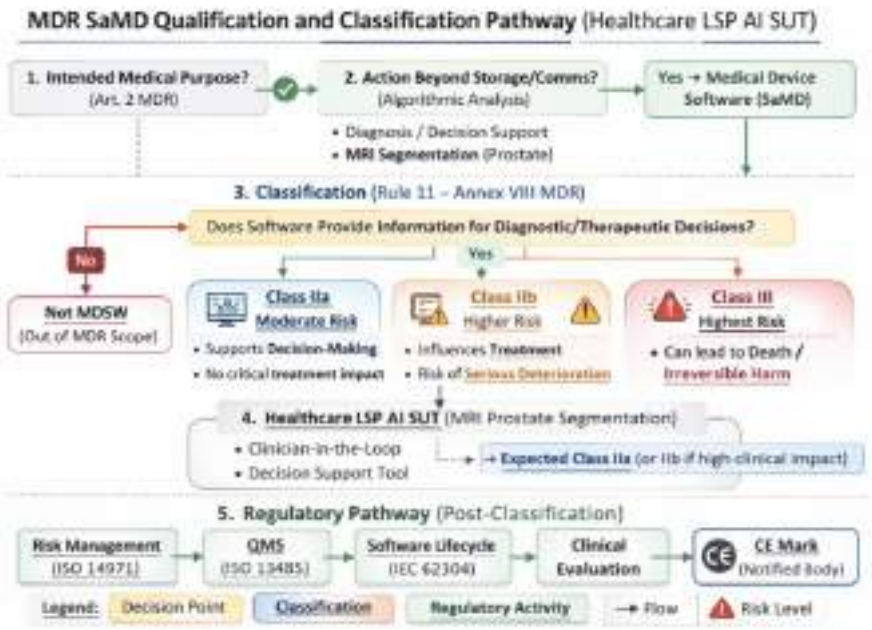


Figure 21. MDR SaMD Qualification and classification pathway (Healthcare LSP AI SUT).

5.1.2 Pre-market Requirements

To obtain CE marking under the EU Medical Device Regulation (EU MDR 2017/745), the manufacturer of a Software as a Medical Device (SaMD) must demonstrate compliance with the General Safety and Performance Requirements (GSPR) defined in Annex I of the regulation. These requirements ensure that medical devices placed on the European market meet rigorous standards of safety, reliability, and clinical effectiveness. Compliance with the GSPR requires manufacturers to establish a comprehensive regulatory framework covering the full lifecycle of the software, from design and development to validation and post-market monitoring. Key elements of this framework include the implementation of a Quality Management System (QMS) compliant with ISO 13485, which governs the organizational processes related to medical device design, development, and production. In addition, manufacturers must apply a structured risk management process in accordance with ISO 14971, ensuring that potential hazards associated with the software are systematically identified, evaluated, and mitigated.

For medical device software specifically, regulatory expectations also require documentation of the software lifecycle and development processes, typically aligned with the international standard IEC 62304, which defines best practices for software design, verification, maintenance, and configuration management in medical devices. Another critical component of the pre-market regulatory pathway is the clinical evaluation, which must demonstrate that the software achieves its intended purpose and delivers clinically meaningful performance and safety outcomes. This evaluation typically combines analytical validation, clinical performance data, and evidence from scientific literature or clinical investigations. Furthermore, manufacturers must establish

appropriate cybersecurity and data protection mechanisms, ensuring that the software operates securely within healthcare environments and that sensitive patient data are protected in accordance with applicable legal frameworks.

Within the Healthcare LSP of the FAITH project, several of these regulatory elements are already partially addressed through the experimental infrastructure and evaluation activities described in D9.1. The pilot implementation provides a structured environment for assessing the AI System Under Test (SUT) within real clinical settings, thereby generating early evidence relevant to regulatory evaluation. In particular, the pilot activities support clinical validation of the prostate MRI segmentation algorithms, allowing their performance to be assessed using real-world clinical data and workflows. Quantitative evaluation metrics, including segmentation performance indicators such as the Dice similarity coefficient and related accuracy measures, provide objective evidence of the technical performance of the AI models. In addition to these technical metrics, the pilot also incorporates qualitative assessments from clinical users, including radiologists who interact with the system during routine imaging analysis and provide feedback on usability, reliability, and trustworthiness.

Although the Healthcare LSP does not represent a formal regulatory approval process, the evidence generated through these pilot activities constitutes an important step toward future regulatory readiness. The performance metrics, clinical feedback, and operational insights collected during the pilot can contribute to the development of a Clinical Evaluation Report (CER), which is a core component of the technical documentation required for CE marking under the MDR. As such, the pilot serves not only as a technical validation exercise but also as an early-stage mechanism for generating evidence that could support a future regulatory submission.

5.1.3 Post-Market Surveillance Considerations

Under the EU Medical Device Regulation (EU MDR 2017/745), manufacturers are required to establish and maintain a Post-Market Surveillance (PMS) system to continuously monitor the safety and performance of medical devices after they are placed on the market. The purpose of PMS is to ensure that any emerging risks, performance limitations, or usability issues are identified and addressed throughout the lifecycle of the device. For software-based medical devices, this requirement is particularly important because software systems may evolve over time through updates, algorithm improvements, or integration into new clinical environments. The PMS system must therefore include mechanisms for collecting and analyzing real-world performance data, managing user feedback, detecting potential safety signals, and implementing corrective or preventive actions when necessary. For higher-risk devices, including Class IIa and above, the MDR also requires the implementation of Post-Market Clinical Follow-Up (PMCF) activities to systematically gather additional clinical data that confirm the device’s safety and performance in routine clinical practice.

Although the AI SUT in the Healthcare LSP is currently deployed within a research and pilot environment rather than a commercial setting, the pilot activities provide an early approximation of several PMS-related processes. The evaluation conducted across participating clinical sites allows the continuous observation of system performance in realistic operational conditions, including the monitoring of segmentation accuracy, system reliability, and clinician interaction with the AI outputs. Feedback from radiologists using the system contributes to identifying potential usability issues, limitations of the segmentation results, or contexts in which the AI output may require additional human verification. These observations can inform future improvements to the AI models and system interfaces while also generating insights relevant for a potential future PMS and PMCF framework. In this way, the Healthcare LSP not only supports the technical validation of the AI system but also contributes to establishing the types of real-world evidence and monitoring practices that would be expected in a post-market regulatory context.

5.2 Alignment with the EU AI Act

The EU Artificial Intelligence Act (AI Act) establishes a comprehensive regulatory framework for artificial intelligence systems deployed within the European Union, adopting a risk-based approach to ensure that AI technologies are developed and used in a safe, transparent, and trustworthy manner. Under this framework, AI systems are categorized into different risk levels depending on their potential impact on fundamental rights, safety, and societal outcomes. AI systems used in healthcare, particularly those that support clinical decision-making or diagnostic processes, are generally classified as high-risk AI systems. As a result, they must comply with a range of regulatory requirements related to risk management, transparency, data governance, human oversight, and technical documentation.

Within the Healthcare LSP, the AI SUT supports radiological image analysis by performing automated segmentation of prostate MRI images. Because the system contributes information used by clinicians during diagnostic interpretation, it falls within the high-risk AI category defined in Annex III of the AI Act, which includes AI systems intended for use as safety components in medical devices or as part of regulated healthcare technologies. Although the SUT is currently deployed within a research and evaluation context rather than as a commercial product, assessing its alignment with the AI Act provides an important foundation for ensuring that future versions of the system can meet emerging regulatory expectations for trustworthy healthcare AI.

5.2.1 Risk Management and Quality Assurance

A central requirement of the AI Act for high-risk AI systems is the implementation of a comprehensive risk management system that covers the entire lifecycle of the AI solution. This includes the identification, analysis, evaluation, and mitigation of potential risks arising during system design, model training, deployment, and operational use. The risk management process

must be continuously maintained and updated as new data become available or as the system evolves. In addition, organizations deploying high-risk AI systems must ensure that appropriate quality assurance processes are in place to verify the reliability, robustness, and safety of the AI system before and during its operational use.

Within the Healthcare LSP, these requirements are partially addressed through the FAITH AI_TAF, which provides a structured methodology for evaluating and managing risks associated with AI systems. The framework supports systematic threat identification, vulnerability assessment, and trustworthiness evaluation, allowing developers and stakeholders to identify potential weaknesses in the system and implement appropriate mitigation strategies. By integrating these risk assessment activities into the development and evaluation process of the AI SUT, the project contributes to establishing a lifecycle-oriented approach to AI governance that aligns with the AI Act's expectations for continuous monitoring and risk management of high-risk AI systems.

5.2.2 Data Governance and Dataset Quality

The AI Act places strong emphasis on data governance and dataset quality, recognizing that the reliability and fairness of AI systems depend heavily on the quality of the data used during model development. For high-risk AI systems, training, validation, and testing datasets must be relevant, representative of the intended target population, and free from systematic biases that could negatively affect system performance. In addition, dataset characteristics and preprocessing procedures must be properly documented in order to ensure transparency and traceability throughout the AI lifecycle.

In the Healthcare LSP, the development and evaluation of the prostate MRI segmentation models rely on several large-scale datasets derived from publicly available and clinically validated repositories, including PI-CAI, Prostate158, and ProCancer-I. These datasets provide diverse imaging data collected from multiple institutions and imaging environments, supporting the development of models that generalize across different clinical settings. Furthermore, the FAITH project introduces structured documentation mechanisms through Data Passports and Model Passports implemented within the FAITH Model Hub. These artifacts record key information such as dataset provenance, model development history, training and validation configurations, and model performance metrics. Such documentation contributes directly to the traceability and transparency requirements defined by the AI Act and supports responsible management of data throughout the AI development lifecycle.

5.2.3 Transparency and Human Oversight

The AI Act also requires that high-risk AI systems incorporate mechanisms that enable appropriate human oversight, ensuring that AI-generated outputs do not replace professional judgment in contexts where safety or critical decision-making is involved. Human oversight mechanisms must allow qualified individuals to interpret AI outputs, detect potential errors, and

intervene when necessary. These requirements aim to ensure that AI systems operate as supportive tools rather than autonomous decision-makers in high-stakes environments such as healthcare.

The AI SUT evaluated in the Healthcare LSP is explicitly designed as a clinical decision-support tool, meaning that the final responsibility for clinical interpretation remains with the radiologist. The system produces segmentation outputs that are visualized directly on the MRI images, often through overlays or confidence representations that allow clinicians to assess the model's output in context. Radiologists review these outputs during the imaging analysis process and can accept, modify, or disregard the AI-generated segmentation as part of their diagnostic workflow. This human-in-the-loop design ensures that the AI system enhances the efficiency and consistency of radiological analysis while preserving full clinical control and accountability.

5.2.4 Technical Documentation and Record Keeping

Another key requirement of the AI Act for high-risk AI systems concerns the maintenance of comprehensive technical documentation and record-keeping mechanisms. These materials must provide sufficient information for regulatory authorities to assess whether the system complies with the applicable regulatory obligations. Technical documentation typically includes detailed descriptions of the system architecture, training procedures, dataset characteristics, model evaluation results, and risk management measures implemented during development and deployment.

Within the FAITH project, several mechanisms support the generation and maintenance of such documentation. The AI Model Passports developed within the FAITH framework provide structured descriptions of model characteristics, development stages, and performance metrics across different evaluation contexts. Complementing these, Data Passports document dataset provenance, governance practices, and the conditions under which training and validation data are collected and processed. In addition, project deliverables and internal documentation contribute to establishing a traceable record of system development, evaluation results, and experimental configurations. Together, these elements contribute to the level of traceability, transparency, and auditability expected under the AI Act for high-risk AI systems.

5.3 Ethical Considerations for AI Deployment in Clinical Environments

Beyond regulatory compliance, the deployment of artificial intelligence systems in healthcare also raises important ethical considerations related to patient safety, fairness, transparency, and trust. AI technologies that support clinical decision-making must be developed and evaluated with particular attention to how they affect patient outcomes, clinician responsibilities, and the broader healthcare ecosystem. Ethical governance in healthcare AI therefore extends beyond technical performance metrics to include responsible data handling, transparency in algorithmic behavior, and the prevention of unintended bias that could affect patient groups differently.

Within the Healthcare Large Scale Pilot (LSP), these ethical dimensions are considered as part of the broader evaluation of the AI System Under Test (SUT). The pilot activities emphasize responsible data management, clinician oversight, and transparent system behavior to ensure that the use of AI in the radiology workflow remains aligned with ethical principles in healthcare innovation. By incorporating ethical considerations alongside technical validation and regulatory alignment, the Healthcare LSP aims to promote the development of trustworthy AI solutions that support clinicians while maintaining patient safety and public confidence in AI-enabled medical technologies.

5.3.1 Patient Privacy and Data Protection

Healthcare AI systems typically rely on the processing of highly sensitive patient data, including medical imaging and associated clinical information. As a result, compliance with the General Data Protection Regulation (GDPR) is essential to ensure that patient privacy and data protection rights are respected throughout the development and deployment of AI systems. In clinical environments, strict safeguards must be implemented to ensure that personal health information is processed lawfully, securely, and only for legitimate medical purposes. These requirements are particularly relevant for projects involving large datasets used to train or validate machine learning models.

Within the Healthcare LSP, several privacy-preserving measures have been implemented to reduce the risk of unauthorized access or disclosure of patient information. Medical imaging data used by the AI SUT undergo multi-stage anonymization and pseudonymization procedures, ensuring that personal identifiers are removed before data are used for analysis or model evaluation. Additionally, patient identifiers embedded in DICOM metadata are systematically removed or replaced with pseudonymous identifiers. The AI system is also deployed within on-premise hospital infrastructures, minimizing the need for external data transfers and reducing potential cybersecurity risks. Together, these safeguards support compliance with European data protection regulations and contribute to maintaining patient trust in the use of AI within healthcare environments.

5.3.2 Transparency and Explainability

Transparency and explainability are key ethical principles in the development of AI systems used in clinical contexts. Clinicians must be able to understand how AI-generated outputs relate to the underlying medical data in order to interpret the results responsibly and integrate them into clinical decision-making processes. AI systems that operate as opaque “black boxes” may reduce clinician confidence and make it difficult to detect potential errors or limitations in algorithmic outputs. Therefore, explainability mechanisms are essential for supporting safe and trustworthy interactions between clinicians and AI systems.

In the Healthcare LSP, the AI SUT incorporates several features that enhance the interpretability of its outputs. The segmentation results generated by the model are visualized directly on the MRI images through segmentation overlays, allowing clinicians to compare AI-generated contours with the underlying anatomical structures. In addition, the system may provide probability or confidence visualizations, such as heatmaps that indicate the model's certainty in specific regions of the image. Interactive visualization tools enable clinicians to review and adjust segmentation results when necessary, ensuring that the AI outputs remain interpretable and clinically meaningful. These capabilities help clinicians maintain control over the diagnostic process while benefiting from the efficiency gains provided by automated image analysis.

5.3.3 Bias Mitigation and Fairness

One of the major ethical challenges associated with AI in healthcare is the potential introduction of algorithmic bias, which may occur when training datasets do not adequately represent the diversity of the patient populations on which the system will be used. If an AI system is trained on limited or unbalanced datasets, its performance may vary across different demographic groups, imaging devices, or clinical settings. Such disparities could result in unequal diagnostic accuracy or reduced reliability for certain patient groups, raising concerns about fairness and equitable access to AI-supported healthcare.

To address these risks, the Healthcare LSP adopts several strategies aimed at promoting fairness and minimizing bias in the development and evaluation of the AI SUT. The segmentation models are trained and validated using multi-institutional datasets that include imaging data from diverse clinical environments and patient populations. This diversity helps improve the generalizability of the models across different imaging systems and healthcare settings. In addition, model performance is evaluated across multiple datasets and clinical contexts to identify potential variations in accuracy or reliability. Continuous monitoring and evaluation of system performance across different subgroups further contribute to ensuring that the AI system supports equitable and responsible healthcare delivery, consistent with broader ethical principles guiding the development of trustworthy AI.

5.4 Regulatory gap analysis: MDR vs EU AI Act vs current SUT status

Table 6 provides a high-level regulatory gap analysis comparing the main expectations of the EU MDR, the EU AI Act high-risk AI framework, and the current status of the AI SUT as implemented in the Healthcare LSP. The purpose of this table is not to claim full compliance, but to identify areas already partially addressed within the pilot and areas requiring additional work before any real-world market placement.

Table 6. Regulatory gap analysis for the Healthcare LSP AI SUT.

Regulatory domain	Key requirement	Current SUT status in Healthcare LSP	Gap level	Recommended next step
EU MDR	Clear intended use, indications, user profile, and clinical purpose	The SUT is described as a radiology decision-support tool for prostate gland and zonal segmentation in MRI, with radiologist oversight clearly emphasized.	**Partial**	Freeze a market-facing intended use statement and formal indications for use suitable for regulatory submission.
EU MDR	Medical device software qualification and classification under MDR/Rule 11	D9.1 recognizes the SUT as SaMD/MDSW-relevant, but no formal notified body classification decision is documented.	**Partial**	Conduct formal classification rationale under Rule 11 and document borderline/qualification analysis.
EU MDR	Quality Management System (QMS)	Project documentation evidences structured development, but not a certified ISO 13485 QMS.	**High**	Establish or align development under an ISO 13485-compliant QMS.
EU MDR	Software lifecycle process	The SUT has documented development, testing, deployment, and refinement, but not yet in a formal IEC 62304 compliance structure.	**High**	Map development lifecycle, maintenance, anomaly handling, and version control to IEC 62304 requirements.
EU MDR	Risk management	FAITH AI_TAF supports structured threat, impact, vulnerability, and risk analysis.	**Partial**	Convert FAITH trustworthiness and risk outputs into a formal ISO 14971 medical-device risk management file.
EU MDR	Clinical evaluation	Pilot validation, user feedback, and performance metrics are being generated in clinical environments.	**Partial**	Consolidate evidence into a Clinical Evaluation Report with claims-linked endpoints and state-of-the-art comparison.
EU MDR	Usability / human factors	End-user interaction, workflow use, and questionnaire-based feedback are embedded in the pilot.	**Partial**	Expand into structured usability engineering and summative validation, aligned with intended users and use environment.

EU MDR	Cybersecurity	Local/on-premise deployment and limited external communication reduce exposure, but formal cybersecurity submission evidence is not yet described.	**Partial**	Develop formal cybersecurity documentation, SBOM, patching policy, threat modelling, and secure update procedures.
EU MDR	Post-market surveillance and PMCF	Pilot monitoring and clinician feedback provide an early analogue, but not a formal PMS/PMCF system.	**High**	Define PMS and PMCF plans for a post-CE lifecycle.
EU AI Act	High-risk AI qualification	Because the software is intended for medical imaging decision support, it is likely to fall within the high-risk scope when marketed as an AI medical device in the EU. This follows the AI Act's high-risk treatment of AI safety components/products covered by sectoral legislation.	**Partial**	Explicitly map the final product to AI Act scope and conformity route once intended use is frozen.
EU AI Act	Risk management system for AI	The FAITH AI_TAF provides a strong project-specific basis for structured AI risk analysis.	**Low to Partial**	Extend FAITH outputs to cover AI Act lifecycle risk governance, including residual risk acceptance and update governance.
EU AI Act	Data and data governance	The project already uses Data Passports, Model Passports, and documented dataset provenance.	**Low to Partial**	Add a formal representativeness, bias, and data-quality justification specifically tied to intended use and target population.
EU AI Act	Technical documentation and record keeping	D9.1, D9.2 draft materials, passports, and pilot documentation provide a strong base for traceability.	**Partial**	Consolidate into an AI Act-ready technical documentation package with versioned evidence and audit trail.
EU AI Act	Transparency to deployers and users	The SUT includes explainability-oriented visual outputs and instructions for use; human oversight is explicit.	**Low to Partial**	Formalize user-facing transparency statements on limitations, expected performance, and misuse boundaries.
EU AI Act	Human oversight	The pilot architecture is explicitly clinician-in-the-loop, with	**Low**	Retain and formally document human oversight procedures,

		radiologists reviewing outputs before use.		override capability, and training obligations.
EU AI Act	Accuracy, robustness, and cybersecurity	Pilot metrics, cross-site validation plans, and trustworthiness assessment provide partial evidence.	**Partial**	Define acceptance thresholds, drift monitoring, robustness testing protocol, and formal cybersecurity evidence.
Current deployment reality	Research/pilot vs market-ready status	The current SUT is a pilot-stage research deployment, locally installed, clinically assessed, and not yet positioned as a CE-marked or FDA-authorized product.	**N/A**	Keep language explicit that current activities support readiness-building, not regulatory approval.

6. Conclusions and Roadmap for the Final Pilot Phase

6.1. Summary of achievements in the second period

6.1.1 Key achievements

During the second reporting period, the Healthcare LSP achieved substantial progress in the deployment, validation, and trustworthiness assessment of the FAITH healthcare use case. The activities performed during this period significantly advanced the maturity of the pilot from a preparatory stage toward real-world operational validation, while also generating evidence to support the broader objectives of the FAITH project. A major achievement was the successful deployment of the AI SUT in the first clinical environment. The deployed solution, consisting of deep learning models for whole-gland and zonal prostate MRI segmentation, was operationalised through a standalone application adapted for hospital use. This enabled practical interaction between clinicians and the AI system under realistic workflow conditions.

A second key achievement was the completion of the initial pilot execution with real clinical users. Radiologists and domain experts engaged with the system using real-world cases and provided structured feedback regarding segmentation quality, usability, and practical utility. The results demonstrated strong clinical acceptance, with the vast majority of evaluated cases considered clinically acceptable, thereby confirming the operational relevance of the solution as an AI-based mpMRI second reader. Another important milestone was the successful application of FAITH trustworthiness tools within the Healthcare LSP. TrustSense was used to assess organisational and team readiness among both clinical users and the technical development team, yielding high to very high trustworthiness maturity levels. In parallel, TrustGuard was applied to conduct a structured assessment of criticality, risks, vulnerabilities, and mitigation needs for the deployed AI models.

The second period also saw progress in AI governance and transparency mechanisms. Relevant healthcare AI assets were documented through the AI Model Hub and associated passporting processes, supporting traceability, lifecycle management, and alignment with trustworthy AI principles. These activities strengthened the governance dimension of the pilot beyond purely technical model evaluation. A further achievement was the generation of multidimensional evidence on system performance. As described in Section 4.1, the Healthcare LSP combined conventional segmentation metrics such as Dice Score, Hausdorff Distance, Average Surface Distance, Recall, and Precision with clinician-based real-world qualitative assessment. Overall, the Whole Gland Model achieved a Dice Score of 91.14%, Jaccard Index of 84.03%, Precision of 91.16%, Recall of 91.41%, and 95th percentile Hausdorff Distance of 3mm, while the Zonal Model achieved corresponding values of 86.99%, 77.58%, 87.19%, 88.14%, and 3mm respectively. This hybrid evaluation approach provided a richer understanding of system value than technical

metrics alone, achieving 157 out of 179 (87.15%) very good responses, thus reaching very good rating from experienced radiologists.

In addition, meaningful progress was made toward cross-site replication readiness. Lessons learned from the first deployment were consolidated to support the next pilot phase at the second clinical site in Pisa. This includes technical deployment experience, user onboarding insights, and trustworthiness assessment know-how that can be transferred to the replication phase. Overall, the second reporting period can be considered highly productive for the Healthcare LSP. It transformed the pilot from concept and preparation into a validated real-world demonstrator, while simultaneously showcasing how FAITH tools can support the responsible deployment of AI systems in sensitive healthcare environments. These achievements provide a strong basis for the final pilot phase, broader validation activities, and future sustainability of the healthcare use case.

6.1.2 Clinicians feedback

Clinician suggestions for improvement were highly consistent across all 171 responses, converging on four recurring themes: (1) enhanced segmentation performance at the prostate apex and base; (2) improved handling of cases with marked anatomical distortion, including post-operative and hypertrophic presentations; (3) the introduction of a manual correction option to allow clinician-driven refinement of segmentation outputs; and (4) the provision of uncertainty maps to increase transparency and clinical trust in the AI-generated results. These suggestions are broadly aligned with the known performance boundaries of the system and reflect constructive, clinically grounded feedback from experienced radiologists.

The recommendation criterion yielded a unanimous outcome across all 171 responses. Every clinician indicated that the tool could be recommended for clinical use as an assistive technology, describing it as providing reliable and time-efficient zonal segmentation to support radiologists in prostate MRI interpretation. Minor phrasing variation was observed across responses but carried no substantive difference in meaning. One response additionally noted that final review and clinical validation should remain clinician-driven, which is consistent with the intended assistive — rather than autonomous — role of the deployed system. The 100% positive recommendation rate, observed across a heterogeneous case mix including anatomically challenging presentations, constitutes strong evidence of clinical acceptance and operational readiness.

6.2 Plan for Replication Phase

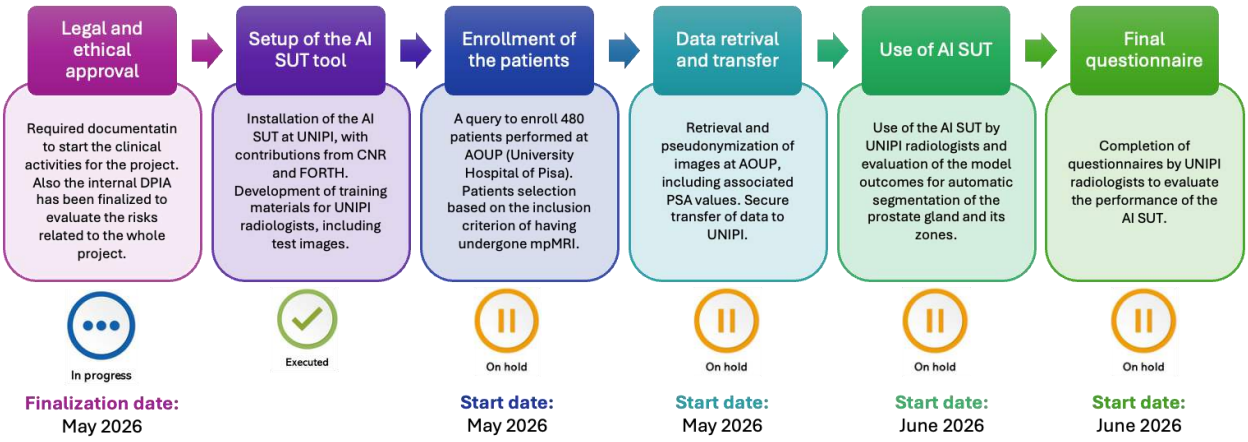


Figure X. UNIPI timeframe and workflow.

The UNIPI clinical activities for the project started in June 2025 with the preparation of the study protocol, patient information documents, risk assessment, and Data Protection Impact Assessment (DPIA). To date, UNIPI has obtained approval from the UNIPI Data Protection Officer (DPO) and the Medical Research Area Directorate (DAM). The documentation is currently under evaluation by the AOUP Clinical Trial Center (CTC) and AOUP DPO, prior to submission to the Ethics Committee. At the local level, the AI SUT tool has been installed on a dedicated UNIPI physical machine with the support of CNR and FORTH. In particular, FORTH contributed to adapting the tool to ensure compatibility with a macOS environment. UNIPI personnel involved in the installation received appropriate training and developed dedicated training materials, including test images, to support radiologists in the use of the tool.

The evaluation team will consist of 4 senior radiologists, 1 junior radiologist, and 6 radiology residents. At present, one senior and one junior radiologist have been trained using the developed materials and have conducted preliminary testing with sample images to gain familiarity with the system and ensure independent use.

Following Ethics Committee approval, AOUP will perform a query to enroll 480 consecutive patients who have undergone prostate mpMRI, without further stratification criteria. Subsequently, imaging data will be retrieved and pseudonymized at AOUP, including associated PSA values, and securely transferred to UNIPI. The AI SUT will then be used by trained UNIPI radiologists to perform automatic segmentation of the prostate gland and its zones, and to evaluate the model's performance. Finally, the structured questionnaires will be administered to collect feedback from the radiologists and assess the overall performance and usability of the tool.

6.3. Identified risks for the Replication Phase

The main critical aspect concerns possible delays in obtaining approval from the Ethics Committee, which could postpone patient enrollment and subsequent study activities. Additional minor delays may arise from further requirements requested by regulatory bodies, such as the DPO or CTC, which could require updates to the protocol or DPIA; however, the likelihood of this occurrence is considered low. Regarding the evaluation phase, some variability in questionnaire responses may be expected due to the current structure of the questionnaire, which includes open-text fields. Similarly, there is a low probability of incomplete or missing responses, which could slightly limit the comprehensiveness of the usability assessment. Finally, as the AI SUT currently operates on a per-patient basis, the processing of the full cohort of 480 patients may result in longer analysis times and a potential extension of the evaluation phase; however, this risk is also considered to have a low likelihood of significantly impacting the overall project timeline.

7. References

The references in the current document are presented as footnotes.