GRAICE-DELFOS: Towards a GenAI-Powered Framework for Continuous Cybersecurity and Safety Compliance in High-Risk Healthcare AI Systems

María Dolores Acuña[1], Ivo Häring[2], Leticia Morales Trujillo[1], María José Escalona[1], David Lizcano[3]

[1]University of Seville, ETSII Avda. Reina Mercedes s/n. 41012, Seville, Spain
[2]Fraunhofer EMI, Ernst-Zermelo-Straße 4, 79104 Freiburg, Germany
[3]Madrid Distance Learning University, Collado Villalba A-6, 15, 28400, Madrid, Spain

{macuna, lmtrujillo, mjescalona}@us.es,ivo.haering@emi.fraunhofer.de, david.lizcano@udima.es

Abstract. *Ensuring AI safety and cybersecurity compliance for high-risk healthcare AI is complex under the EU AI Act, NIS2, and Cyber Resilience Act. This paper introduces GRAICE, a GenAI-based framework for automated remediation and continuous regulatory alignment, integrated into DELFOS, a clinical support tool for genetic diagnostics. By embedding GenAI agents into the AI lifecycle, the system replaces static audits with a continuous, evidence-driven compliance and cybersecurity continuum. Expected results include enhanced resilience, automated risk mitigation, and increased clinical trust.*

1. Introduction

While AI transforms genetic diagnostics [Abdelwanis et al. 2026], the EU AI Act's "high-risk" classification imposes strict governance requirements. Traditional manual audits are poorly suited for adaptive AI pipelines [Okonkwo et al. 2025]. This paper presents the GRAICE-DELFOS case study, demonstrating how GenAI-driven governance ensures continuous safety and cybersecurity compliance in medical settings.

2. The DELFOS Platform: An AI-Driven Clinical Decision Support System

DELFOS integrates genomic data, EHRs, and physiognomic data via FHIR servers, using ML models to detect genetic pathologies and analyze patient patterns for complex diagnostics [Enamorado-Díaz et al. 2025]. It addresses assisted reproduction bottlenecks by providing **geneticists** with streamlined workflows. **Patients** would be indirect beneficiaries because they would receive faster, data-driven decisions for better care quality. DELFOS's modular architecture utilizes PostgreSQL, FastAPI, and Apache Spark to manage the complete data lifecycle, now enhanced by GRAICE's AI safety and security layers.

3. The GRAICE Framework: Automating Compliance for DELFOS

For the high-risk DELFOS system, integration of the GRAICE framework provides several assessment layers designed to automate and manage this regulatory burden. Section 3 details the main objective to achieve continuous compliance.

3.1. Main Objective: From Static Audits to a Compliance Continuum

The primary objective of GRAICE-DELFOS integration is to transition from traditional, point-in-time compliance audits to a dynamic compliance continuum. Static audits are retrospective, expensive, and inadequate for continuously evolving AI systems, as widely noted in recent healthcare analyses. The integration between GRAICE and DELFOS ensures ongoing adherence to the EU AI Act, NIS2, and GDPR while enabling safe, iterative AI model refinement.

3.2. Technical Integration and Architecture

The project integrates GRAICE's GenAI-based compliance agents into the modular DELFOS architecture, where they monitor system artefacts and behavior, detect non-conformities, and generate automated compliance documentation aligned with European regulations. The integration centers on the FHIR communication layer, through which DELFOS exchanges data with external clinical and laboratory systems, and where GRAICE validates the security and trustworthiness of these connections. GenAI agents continuously oversee these interactions, detect anomalies, and issue adaptive risk alerts with compliance improvement recommendations (e.g., reinforcing patient identification workflows). The framework (Fig. 1) employs: (1) **Data ingestion** of technical artefacts (code, specs, feedback) and regulations; (2) **Risk-resilience panarchy** methodology for ongoing as-is evaluation; (3) **Layered structure** (system, development, resilience, legal, concept) with GenAI agents automating bottom-up and top-down analyses; (4) **Multi-metric outputs** classifying risks by frequency, susceptibility, robustness, recovery, and adaptive learning capacity, yielding efficient recommendations; and (5) **Transparency and Reliability mechanisms:** To ensure high-quality assessments and mitigate GenAI hallucinations, the framework employs a Retrieval-Augmented Generation (RAG) architecture. Agents are strictly grounded in the technical artefacts and specific European regulations (EU AI Act, NIS2, etc.) ingested in Layer 4, ensuring that generated human-readable audit reports are evidence-driven and verifiable against original legal texts.

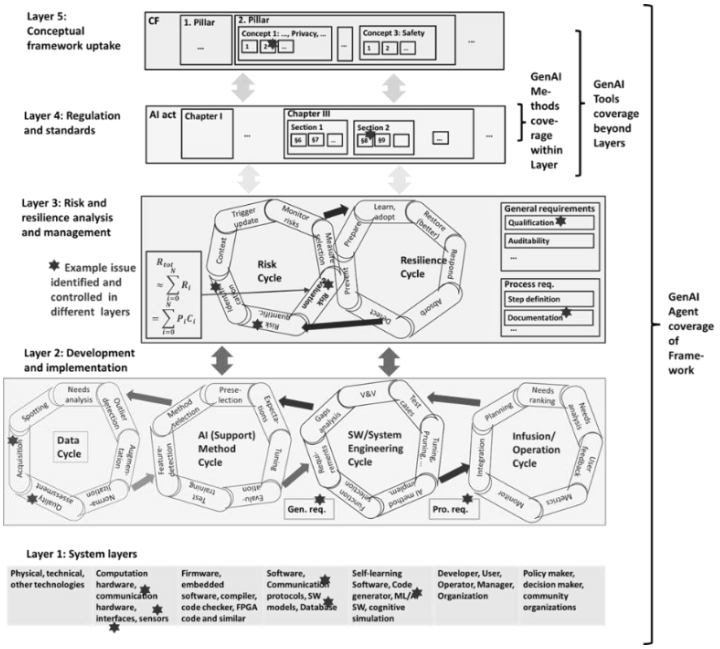

Figure 1. GRAICE cybersecurity and AI safety compliance framework

4. Expected Results and Impact

The success of the GRAICE-DELFOS integration will be measured through quantitative and qualitative indicators. Validation is designed to confirm the effectiveness of the GenAI-powered framework in a real-world clinical setting. This section presents the expected results and specific Key Performance Indicators (KPIs) established and analyses the anticipated impact on key stakeholders and clinical workflows.

4.1. Expected Results

The GRAICE-DELFOS project will deliver four main results. **R1 GenAI-powered governance:** Embedded GenAI agents on the FHIR communication layer continuously analyze behavior and detect non-conformities. By using a human-in-the-loop verification process, the agents generate compliance documentation that is both automated and expert-validated to ensure full alignment with the AI Act, NIS2, CRA, and GDPR. **R2 Dual-role intelligence:** These agents provide both defensive capabilities (real-time monitoring and automated remediation of policy violations) and anticipatory capabilities (predictive analytics and adversarial modelling of attacks such as data exfiltration, model inversion, and bias amplification). **R3 Legal and ethical rigor:** The project enforces privacy-preserving design and FHIR-based consent management in line with GDPR Article 9, while ensuring explainability, fairness, and lifecycle traceability through AI Model Cards and Data Sheets. **R4 Interoperability and scalability:** While validated on DELFOS, the framework's reliance on FHIR-compatible APIs and containerized microservices ensures high scalability. This design allows the GRAICE agents to be ported to any clinical AI system—such as those in oncology or cardiology—that utilizes standard healthcare communication protocols, facilitating a platform-agnostic compliance layer.

4.2. Validation through Key Performance Indicators (KPIs)

The effectiveness of the GRAICE-DELFOS integration is validated through a multi-phase pilot study within the DELFOS clinical environment, designed to link specific technical targets to measurable healthcare outcomes. To ensure a rigorous experimental design, the validation protocol follows three primary axes: (a) Controlled **adversarial simulations** to evaluate the framework's detection and response capabilities against simulated cyberattacks; (b) **Stakeholder surveys** utilizing Likert scales to quantify changes in clinical trust; and (c) A comparative **'Human vs. AI' audit**, where professional auditors provide a ground-truth baseline to verify the accuracy of the risks identified by the GenAI agents.

The success of the project is measured against the following five KPIs: (1) **Incident Detection & Response**: Achieve a ≥95% detection rate for security and compliance incidents, ensuring robust NIS2 and Cyber Resilience Act (CRA) resilience in all clinical data exchanges. (2) **Data Integrity & Traceability**: Maintain ≥99% cryptographically verified exchanges, providing robust, GDPR-compliant audit trails for sensitive genomic and patient data. (3) **AI Model Governance**: Ensure all AI models are equipped with standardized AI Model Cards and Data Sheets to meet the transparency and traceability requirements mandated by the EU AI Act. (4) **User Trust & Transparency**: Achieve a ≥30% improvement in trust—as measured by the pre and post pilot surveys—to drive clinical adoption and reduce operational errors. (5) Continuous Risk and Resilience Issue Ranking: Ensure ≥90% of security and compliance issues are

identified and correctly attributed to assessment layers 1 to 5 (as defined in Fig. 1) when compared against professional human audits.

4.3. Projected Impact on Clinical Workflows

By achieving these KPIs, DELFOS is expected to notably benefit its users. Continuous compliance and improved safety and security will allow IT specialists to speed up innovation cycles and to learn from near-real time operational data. It will reduce regulatory overhead for clinical geneticists and assisted reproduction specialists, enabling greater focus on clinical tasks rather than administrative work. Stronger model governance and transparency should increase clinicians' trust in AI-driven recommendations, so that in this high-stakes domain higher confidence and reliability translate into better diagnostic precision, clinical outcomes, and quality of patient care.

5. Conclusion and Future Work

A novel solution for the challenge of regulatory compliance of high-risk AI systems in healthcare was outlined. Integrating the GRAICE framework with the DELFOS clinical decision support platform is a significant step away from inadequate manual audits toward a GenAI-driven compliance continuum. By embedding intelligent agents for both defensive and anticipatory security, the framework ensures in an innovative fostering way that a complex, evolving AI system remains aligned with the multifaceted European regulations. Future work on the DELFOS platform will focus on expanding its datasets and refining its predictive models. Simultaneously, the GRAICE framework will undergo a cross-domain validation phase, where it will be integrated into diverse high-risk AI medical applications beyond genomics to empirically demonstrate its scalability and adaptability across the broader European digital health landscape.

Acknowledgments

Research supported by: (i) EQUAVEL project PID2022-137646OB-C31, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; (ii) DELFOS project 2021/C005/00151010 by Spanish organization red.es, 2021 call for proposals to fund projects related to AI and other digital technologies for value chains C005/21-ED.

References

Moustafa Abdelwanis, et al. (2026) "Artificial intelligence adoption challenges from healthcare providers' perspectives: A comprehensive review and future directions", Safety Science, Vol. 193, p107028, https://doi.org/10.1016/j.ssci.2025.10702828

Roy Okonkwo, et al. (2025) "A study on advanced AI-Driven continuous compliance monitoring for cybersecurity regulations in healthcare", WJARR, 26(03), 2249-2255 https://doi.org/10.30574/wjarr.2025.26.3.2424

Elena Enamorado-Díaz, et al. (2025) "A novel machine learning-based proposal for early prediction of endometriosis disease", Expert Systems with Applications, Vol. 271, p. 126621, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2025.126621