Check for updates

OPEN LETTER

REVISED A sustainable platform for federated health data access, AI innovation, and regulatory acceptance in alignment with the European Health Data Space principles

[version 4; peer review: 1 approved, 1 approved with reservations]

Erwin Boutsma[1*], Katja Herzog[2*], Rebecca C. Rancourt[3], Torsten G. Blum[3,4*], Alberto Moreno[5*], Francisco J. Núñez-Benjumea[5*], Hanna Ćwiek-Kupczyńska[6*], Rada Hussein[7*], Holger Fröhlich[8*], Lotte Groth Jensen[9*], Anne Funck Hansen[10*], Nadja Kartschmit[11*], Anja Burmann[12], Mira Grättinger[2], Sumit Madan[8], Rita Peeters[13], Venkata P. Satagopam[6], Tieneke B.M. Schaaij-Visser[1], Philip Gribbon[2], Christian Muehlendyck[14], the IDERHA Consortium

[1]Lygature Foundation, Utrecht, Netherlands Antilles
[2]Discovery Research Screening Port, Fraunhofer Institute for Translational Medicine and Pharmacology, Hamburg DE, Germany
[3]Institute of Clinical Research, MSB Medical School Berlin GmbH, Berlin, Berlin, Germany
[4]HELIOS Klinikum Emil von Behring GmbH, Berlin, Berlin, Germany
[5]Digital Health & Artificial Intelligence, Institute of Biomedicine of Seville, IBiS/Virgen Macarena University Hospital/CSIC/University of Seville, Seville, Spain
[6]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg
[7]Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria
[8]Fraunhofer SCAI, Sankt Augustin, Germany
[9]DEFACTUM, Aarhus N, Denmark
[10]Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., München, France
[11]Medical University of Vienna, Center for Medical Data Science, Vienna, Austria
[12]Fraunhofer ISST, Dortmund, Germany
[13]Johnson & Johnson MedTech, Machelen, Belgium
[14]Johnson & Johnson Medical GmbH, Norderstedt, Germany

* Equal contributors

v4 **First published:** 28 May 2026, **6**:181
https://doi.org/10.12688/openreseurope.23459.1
Second version: 26 Aug 2026, **6**:181
https://doi.org/10.12688/openreseurope.23459.2
Third version: 21 Sep 2026, **6**:181
https://doi.org/10.12688/openreseurope.23459.3
Latest published: 06 Oct 2026, **6**:181
https://doi.org/10.12688/openreseurope.23459.4

Abstract

The IDERHA (**I**ntegration of Heterogeneous **D**ata and **E**vidence towards **R**egulatory and **H**TA **A**cceptance) project aims to enhance

Open Peer Review

Approval Status ? ✓

	1	2
version 4 (revision) 06 Oct 2026		✓ view
version 3		?

medical research by establishing one of Europe's first pan-European, disease-agnostic health data spaces. Aligned with the European Health Data Space (EHDS) principles, IDERHA addresses critical challenges in data quality, standardization, and governance, ensuring compliance with GDPR, the AI Act, and emerging EHDS regulations. Its ambition is to enable secure, federated access and analysis of health data, fostering data-driven collaboration and innovation in healthcare.

IDERHA's technical infrastructure employs a 'privacy-by-design' approach, leveraging federated analytics and learning to maintain data sovereignty and reduce privacy risks. The project focuses on lung cancer as a high-impact use case, utilizing AI and machine learning to improve early detection, diagnosis, and personalized care. It also aims to develop policy recommendations for the acceptance of real-world evidence (RWE) for regulatory decision-making through multi-stakeholder engagement and public consultations.

However, challenges remain, including semantic interoperability, and scaling federated AI methods across borders. IDERHA's modular, standards-based architecture and emphasis on ethical, legal, and FAIR compliance provide a robust framework for addressing these issues. By collaborating with other initiatives in the health data domain to drive compatibility, IDERHA seeks to accelerate innovation by creating a sustainable, scalable model for health data access and thereby a positive impact for patients across Europe.

Keywords
European Health Data Space, Real-World Evidence, AI, lung cancer, governance frameworks, federated learning, health data, regulatory acceptance

HE This article is included in the Horizon Europe gateway.

	1	2
		view
(revision) 21 Sep 2026		
version 2 (revision) 26 Aug 2026	? view	
version 1 28 May 2026	? view	

1. **Daniel Happy Putra**, Universitas Esa Unggul, West Jakarta, Indonesia
2. **Yash Mehandiratta**, All India Institute of Medical Sciences (AIIMS), Guwahati, India

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Erwin Boutsma (erwin.boutsma@lygature.org)

Author roles: Boutsma E: Conceptualization, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Herzog K**: Conceptualization, Visualization, Writing – Review & Editing; **Rancourt RC**: Writing – Review & Editing; **Blum TG**: Writing – Original Draft Preparation, Writing – Review & Editing; **Moreno A**: Writing – Original Draft Preparation, Writing – Review & Editing; **J. Núñez-Benjumea F**: Writing – Original Draft Preparation, Writing – Review & Editing; **Ćwiek-Kupczyńska H**: Writing – Original Draft Preparation, Writing – Review & Editing; **Hussein R**: Writing – Original Draft Preparation, Writing – Review & Editing; **Fröhlich H**: Writing – Original Draft Preparation, Writing – Review & Editing; **Groth Jensen L**: Writing – Original Draft Preparation, Writing – Review & Editing; **Funck Hansen A**: Writing – Original Draft Preparation, Writing – Review & Editing; **Kartschmit N**: Writing – Original Draft Preparation, Writing – Review & Editing; **Burmann A**: Writing – Review & Editing; **Grättinger M**: Writing – Review & Editing; **Madan S**: Writing – Review & Editing; **Peeters R**: Writing – Review & Editing; **Satagopam VP**: Writing – Review & Editing; **Schaaij-Visser TBM**: Writing – Review & Editing; **Gribbon P**: Conceptualization, Supervision, Writing – Review & Editing; **Muehlendyck C**: Conceptualization, Supervision, Writing – Review & Editing;

Competing interests: Rita Peeters and Christian Muehlendyck are employees of Johnson & Johnson; the paper expresses the personal opinion of the authors, but not of the company. All other authors declare no competing interests.

Grant information: This work was conducted in the IDERHA project which is supported by the Innovative Health Initiative Joint Undertaking (JU) under grant agreement No 101112135. The JU receives support from the European Union's Horizon Europe research and innovation program and life science industries represented by COCIR, EFPIA / Vaccines Europe, EuropaBio and MedTech Europe. In addition, there are financial and/or in-kind contributions from our Swiss and UK partners. The UK participant is supported by UKRI grant No 10079453 (National Institute for Health and Care Excellence). IDERHA is funded by the European Union, the private members, and those contributing partners of the IHI JU.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Boutsma E *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Boutsma E, Herzog K, Rancourt RC *et al.* **A sustainable platform for federated health data access, AI innovation, and regulatory acceptance in alignment with the European Health Data Space principles [version 4; peer review: 1 approved, 1 approved with reservations]** Open Research Europe 2026, **6**:181 https://doi.org/10.12688/openreseurope.23459.4

First published: 28 May 2026, **6**:181 https://doi.org/10.12688/openreseurope.23459.1

REVISED **Amendments from Version 3**

We added text to best address the reviewer's points which we agree will improve describing clinical relevance and help demonstrate how the IDERHA platform can move beyond technical infrastructure toward safe and meaningful use in real-world healthcare.

Any further responses from the reviewers can be found at the end of the article

Introduction

IDERHA (**I**ntegration of Heterogeneous **D**ata and **E**vidence towards **R**egulatory and **H**TA **A**cceptance) is a European public-private partnership launched in April 2023, under the Innovative Health Initiative (IHI) program, with funding from the European Union and Swiss and UK partners, and the European life sciences industries.

This project aims to enhance medical research by overcoming challenges in accessing, integrating, and analysing health data. To achieve this, IDERHA is building one of the first pan-European open, disease-agnostic health data spaces, aligned with European Health Data Space (EHDS) principles. The project addresses critical challenges in data quality and standardization, as well as developing an innovative data governance framework, which includes a tailored Data Management Plan (DMP), ensures adherence to key technical standards and interoperability, in compliance with European Union (EU) legal regulations like the General Data Protection Regulation (GDPR), the AI Act, and emerging EHDS regulations. IDERHA's ambition is to enable connected, secure, efficient federated data access and analysis of health data, ultimately supporting better clinical decision-making and patient care.

As a functional demonstration and alignment with the Europe Beating Cancer Plan, IDERHA will develop artificial intelligence (AI) solutions to improve early detection and diagnosis of lung cancer, and in doing so illustrate the added value of a data-driven approach for improving clinical decision making and personalized disease management.

Furthermore, IDERHA will benefit patients and healthcare providers by developing policy recommendations on health data access and heterogeneous health research for regulatory and HTA decision-making. Those are developed through literature review, multi-stakeholder engagement, expert panels, and public consultations.

The project will not only create a technically robust and legally and ethically compliant solution but also establish a sustainable model that will ensure ongoing use, future growth, and further development of the IDERHA platform. This includes welcoming new partners and expanding into additional indication and application areas domains. In doing so, we will enable data-driven collaboration, interoperability, and innovation within the healthcare sector.

Value of lung cancer screening and early diagnosis: The medical benefit

The European Cancer Information System (ECIS) estimated that in 2022 nearly 320,000 patients were diagnosed with LC within the EU, while some 250,000 patients died of LC in same year. Among all cancers, LC ranked fourth and first in incidence and mortality rates, respectively.[1] Within European countries, in the majority of patients LC continues to be first diagnosed at advanced and consequently incurable stages.[2] In addition, LC places the highest burden on national health systems in the EU among all malignancies.[3] Furthermore, people with LC often have substantial co-morbidities, making treatment more challenging and contributing to a higher symptom load when compared to many other solid cancers.[4,5]

Besides smoking cessation, early detection of LC by low dose computed tomography (LDCT) screening in high-risk populations is one of the most promising options to reduce both mortality and morbidity. Two large randomized controlled trials (RCT), the US National Lung Screening Trial (NLST; NCT00047385; 53,454 participants) and the Dutch-Belgium Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON; ISRCTN63545820; 15,789 participants)[6] explored the efficacy of chest LDCT screening in individuals at high risk for LC versus chest radiography and no screening, respectively. Both demonstrated a significant reduction in LC-specific mortality rates by LDCT compared to the control arms. In addition, earlier diagnosis and resection rates were significantly enhanced in the screening cohorts in both RCTs. NLST was also large enough to show a significant reduction in all-cause mortality rate, something not observed in breast or colorectal cancer screening trials.

Based on a systematic review of the evidence, the Science Advice for Policy by European (SAPEA) concluded in March 2022 that there exists a strong scientific basis for introducing life-saving screening programmes in EU Member States.[7] Subsequently in December 2022, the EU Council suggested in their updated recommendation on strengthening

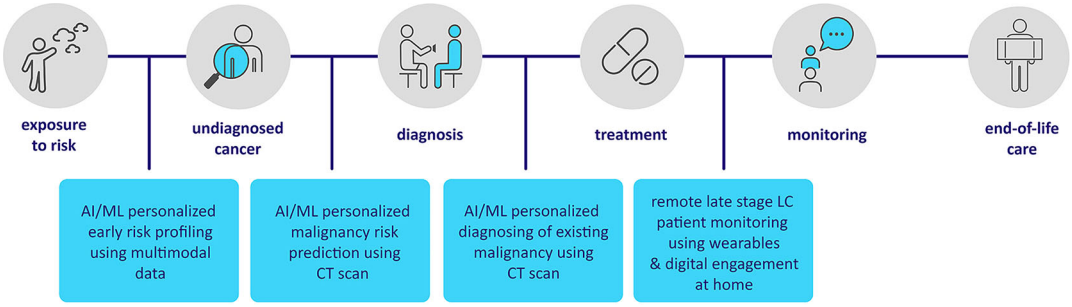

Figure 1. IDERHA's patient journey: aiming to improve overall cancer screening, risk prediction and remote monitoring.

prevention through early detection to Member States to explore the feasibility and effectiveness of LDCT-based LC screening programmes.[8]

While population-based LDCT screening programs were already implemented in the US following NSLT, Croatia was the first EU-country successfully initiating its own national screening programme for LC in October 2020.[9] Meanwhile three other national LC screening programs were launched in the Czech Republic, Poland and the United Kingdom, along with several implementation pilots in other European countries.[10] Recent evidence from on-going national programs in the US, Taiwan and Croatia as well as the German HANSE study could confirm the efficacy of LDCT screening in real-world settings, thus highlighting a stage-shift towards early, curable LC stages achieved via higher surgical resection rates.[11–15]

The three IDERHA use case studies (Figure 1) will address the still unmet needs to:

1) better define individuals at high risk for LC beyond the currently used age- and tobacco-exposure-based risk criteria:

This use case is intended to complement rather than replace established screening eligibility criteria. Candidate predictors will be derived from routinely available longitudinal EHR and claims data and may include demographic and clinical characteristics, comorbidities and healthcare history, in addition to smoking exposure where available. To minimize the risk of merely reproducing smoking-driven risk models, model performance will be evaluated across relevant smoking strata, including never-smokers where sufficient data are available. Any potential extension of LDCT screening beyond currently established eligibility groups would require separate prospective clinical validation.

2) extend the utilization of chest CT image features beyond pulmonary nodule detection towards improved LC risk prediction:

In contrast to nodule-specific malignancy prediction, this approach will investigate information contained in the overall CT examination beyond the presence or absence of a suspicious pulmonary nodule and may combine image-derived features with available clinical risk information. The objective is to determine whether such multimodal information can improve individualized longitudinal LC risk estimation.

3) improve overall malignancy risk prediction for screen-detected nodules from chest LDCTs:

This use case will investigate quantitative imaging features derived from LDCT and, where available, combine these with relevant clinical and patient-level characteristics to improve individualized malignancy-risk estimation for detected nodules. The objective is to support clinically meaningful risk stratification and subsequent nodule management while reducing unnecessary diagnostic procedures for benign findings.

Real-world data and standard models in healthcare: A healthcare provider perspective

EU countries in general have solidaric health care systems with mostly equal access to health care, leading to a relatively inclusive and complete patient cohort covering much of the population. The large amount of Real-World Data (RWD), collected in routine care, provide significant opportunities for deriving data-driven insights that can facilitate decision making in both patient care (i.e., enabling preventive approaches) and clinical management (i.e., predicting demand and optimizing resources utilization). However, while data quantity is not a limitation and theoretically advantageous,

healthcare providers often face challenges related to quality assurance and standardization, as well as managing, accessing, and analysing the data effectively. This inhibits full utilization of RWD. Also, the large-scale nature of RWD requires considerable resources and investments to host, manage and process them, especially for the purpose of enabling secondary use of the data. Healthcare providers frequently lack the necessary resources, including advanced data analytics infrastructures, skilled personnel, and governance frameworks. This limited access to skilled data scientist personnel, interoperability solutions, and Secure Processing Environments (SPE) hinders RWD-based research.[16]

Data quality

High-quality RWD is critical for clinical decision-making and secondary uses like observational studies, yet its unstructured or semi-structured nature poses major challenges.[17] Real-world data are often drawn from both automatically collected sources – such as imaging or device-generated measurements – and manually entered records, typically from electronic health records (EHRs), claims data, and Patient Reported Outcome Measures (PROMs). Automatically collected data tend to be more complete and more standardized, while manually entered data, recorded to support routine clinical care rather than research needs, are more prone to incompleteness, inconsistencies, or errors.[18] This variability in data collection can compromise reliability for research purposes.[19] Issues such as incomplete or inaccurate entries, misclassification, and coding inconsistencies particularly affect clinical validity of manually registered data.

Data standardization

The lack of standardization in clinical data models across different healthcare systems as well as within the same healthcare system also poses a significant challenge for providers. Data collection is often highly heterogeneous, using different terminologies, formats, and structures across institutions or even individual clinicians.[20] There is a big variability in how clinical events, diagnoses, and interventions are monitored and documented. This lack of uniformity complicates the aggregation, analysis, and interpretation of data and reduces the generalizability of insights derived from this data. For healthcare providers, reconciling these disparate data models (in this context: an agreed combination of structure and standard vocabularies to represent the information) often necessitates extensive data cleaning, mapping, and harmonization efforts, which are both time-consuming and resource-intensive.

To address data heterogeneity, the scientific community is increasingly adopting the Observational Medical Outcomes Partnership (OMOP) common data model (CDM)[21,22] which offers a standardized approach to structuring and organizing healthcare data, thus facilitating the integration of data from diverse sources.[23]

Within IDERHA, a key step in network studies is achieving data interoperability through data standardisation and harmonisation using the OMOP Common Data Model (OMOP-CDM). To ensure semantic interoperability without compromising clinical validity, IDERHA applies a structured Semantic Mapping Quality Assurance Framework that combines automated and semi-automated semantic matching, terminology-based mapping, expert review, structural mapping, ETL transformation, and quality controls. Mapping activities are supported by OHDSI tools including WhiteRabbit, Usagi, Athena, Achilles, and the Data Quality Dashboard (DQD), together with provenance-aware ETL processes and terminology services. Semantic mapping can be performed either automatically when the source data relies on OMOP-supported ontologies such as ICD10, SNOMED, LOINC, etc.; in a semi-automated manner using Usagi, where suggested mappings are reviewed manually on a concept-by-concept basis; or manually using Athena. This process is performed in parallel with structural mapping and is followed by ETL transformation to convert the source data into the target OMOP-CDM representation. An initial quality check compares general counts and other basic data-volume indicators between the transformed and source data to identify potential discrepancies. This is followed by systematic data quality assessment using the DQD, including conformance, completeness, and plausibility checks, as well as further validation through cohort concordance analyses and comparison with source-system representations.

An OMOP readiness assessment then verifies the completeness, semantic consistency, quality, and analytical suitability of each dataset for the specific study, confirming mapping, ETL, validation, and alignment with network requirements. Discrepancies and remaining harmonisation needs are addressed iteratively through interactions among OMOP mapping experts, clinicians, and database/data-engineering experts, supported by documented decision rules and version-controlled mappings. IDERHA manages this process with a specific focus on the study protocol and the parameters of the relevant IDERHA use case, ensuring that data quality, interoperability, and analytical suitability meet the study-specific requirements across all federated nodes and that the resulting data are “fit for purpose”.

Healthcare providers are also recognizing the potential of OMOP to enable federated observational studies across institutions and geographic regions, without needing to centrally harmonise sensitive patient data. By mapping local clinical data to the OMOP CDM, data holders can contribute to large-scale research efforts while maintaining data privacy and security. Furthermore, the European Medicines Agency (EMA) is leveraging this OMOP CDM in order to

facilitate post-market pharmacovigilance services through the DARWIN initiative.[24] However, the process of OMOP-ing also holds its own challenges, such as ambiguous mappings due to local practices, and transitioning to this model requires significant upfront investment in data transformation and ongoing maintenance, which can be burdensome for resource-limited healthcare institutions.

Data regulations

Healthcare providers also face considerable hurdles in obtaining clearance for RWD use in research. Ethical approvals and data-sharing agreements are often slow and cumbersome, further complicated by varying regulations across regions and institutions. The lack of harmonization and/or shared interpretation in these clearance processes creates inefficiencies and delays, preventing timely use of RWD for research purposes.

Moreover, current processes are not optimized for the scale and complexity of modern data-driven studies, particularly those involving federated data networks or multi-institutional collaborations. For healthcare providers, navigating these clearance processes remains a significant bottleneck in using RWD.

Connecting health data in a decentralized manner: A technical perspective

RWD is collected and stored at all healthcare institutions across Europe in a highly decentralized manner. Making these data available via Secure Processing Environments (SPE) according to the Data Governance Act (COM/2020/767), is one of the visions of the EHDS. An SPE is an environment which enables a safe use of data through third parties, representing interoperable, secure and regulated data processing infrastructures to enable data users (e.g., healthcare providers and researchers) to access sensitive health data, while data security and integrity remain intact. To realise this ambition in practice, IDERHA has developed a data privacy preserving research infrastructure, which supports federated discovery, access, and analysis of patient-level RWD including federated learning capabilities (Figure 2).

Driven by security and data sovereignty, IDERHA's architecture embodies a 'privacy-by-design' approach:

- All personal data remains at all times under the control of data holders at their Federated Nodes. Algorithms for the analysis are moved to the data and are executed within SPEs at data holders' IT infrastructure. Only non-personal aggregated results are allowed to leave the data holders' site and be shared for centralized joint analyses or returned to the data user (Figure 3).

- The platform entry point at a Central Node ensures user authentication, and allows users to discover data, request federated access and initialise analysis tasks throughout the environment.

- A metadata catalogue within the Central Node, describes available datasets using standardized DCAT-AP metadata model to ensure interoperable descriptions of datasets, facilitating compliance with the EHDS Regulation, and allowing researchers and policymakers to find relevant data.

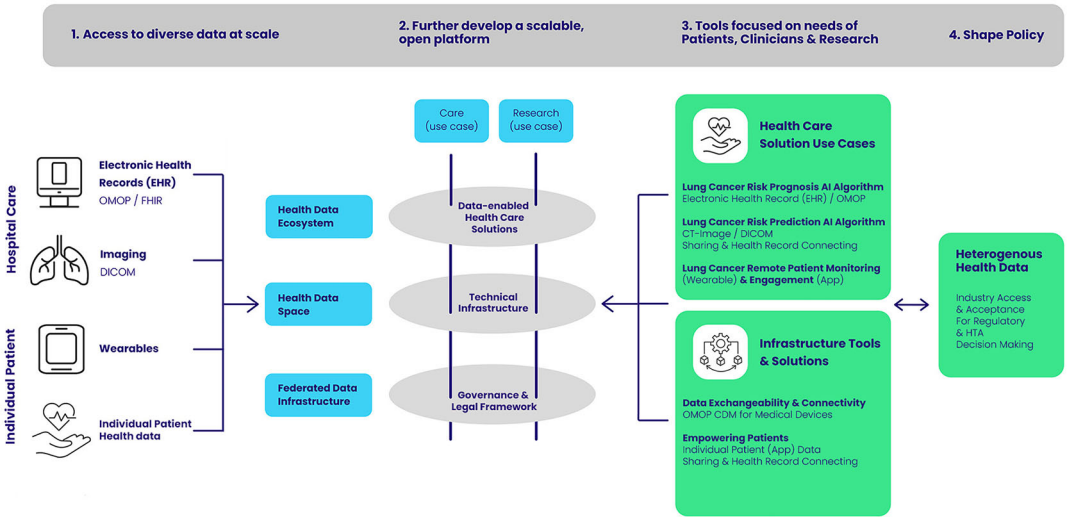

Figure 2. IDERHA's approach to 'privacy-by-design': implementing a platform (or way) to support research, federated discovery, access, and analysis of patient-level RWD while preserving data privacy.

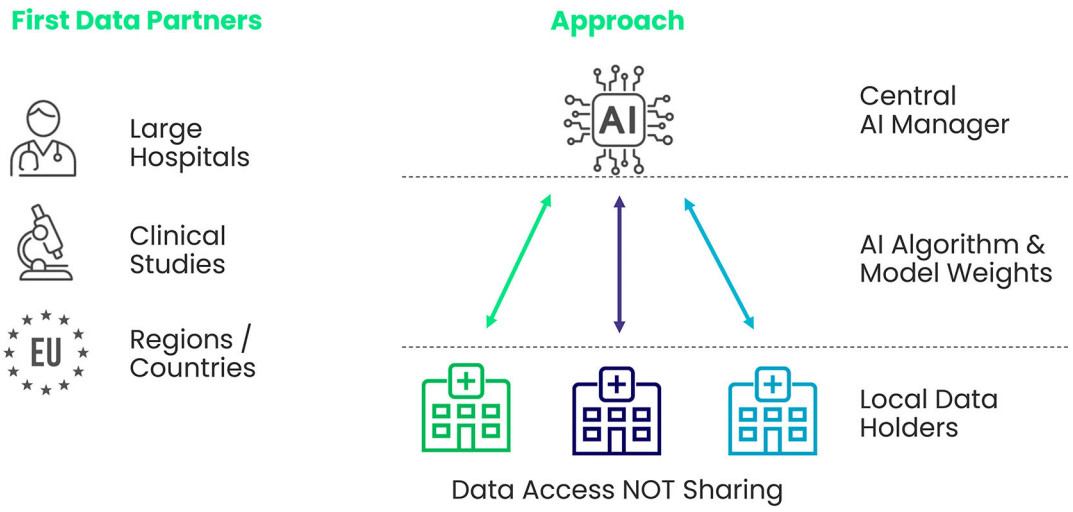

Figure 3. IDERHA's federated approach: all personal data remains under the control of data holders at local Federated Nodes.

- Data holders evaluate incoming data access requests and issue and manage data use permits (aligned to specific access policies). Domain-specific standards provide a flexible, standardized way to define and communicate user permissions.

- Access management components, located within Federated Nodes of data holders, enforce holder-defined data use conditions and access policies, ensuring data is only released for approved users and analyses within approved scopes and timelines.

In both federated analysis and federated learning a compute-to-data paradigm is applied, with computational tasks delivered to the data rather than data being centralized for pooled evaluation. This approach allows for powerful, multi-site analyses while respecting data sovereignty and reducing privacy risks,[25] hence helping to meet GDPR requirements. Whilst not removing the responsibility of data holders to ensure that a suitable data usage basis exists, including a proper legal basis etc., it offers by design lower risks of data breaches and thus promotes trust building with this group.[26]

To ensure interoperability and adherence to best practices, IDERHA will implement several established standards and good practices, such as:

- Global Alliance for Genomics and Health (GA4GH) standards, for managing complex workflows and executing computational tasks across different environments, and for abstracting data access and standardised communication of use permissions.

- International Data Spaces Association's (IDSA) Dataspace Protocol, implemented through Eclipse Dataspace Components (EDC), for sovereign and secure data exchange between participants.[27]

- Open industry standards for secure communication, authentication, and authorization, enabling user identity management and access control across the platform[28] and for cross-border recognition of electronic identification (eIDAS).

- An SPE, supported through technical and organizational measures for ensuring data privacy, as well as documentation and supervision of data processing.

These standards combine to create a robust, interoperable framework that supports secure, efficient data access and analysis while respecting data sovereignty and privacy. Moreover, IDERHA's modular, standard-based design aims to ensure technical interoperability and extensibility through:

- Integration with existing health IT systems and emerging data space initiatives.
- Inherent flexibility to incorporate new data types, analysis methods, and regulatory requirements as they evolve.
- Potential for reuse of components and learnings in other health data projects across Europe.

While the compute-to-data paradigm reduces privacy risks by keeping patient-level data within local Federated Nodes, IDERHA also recognizes that federated learning infrastructures introduce several technical challenges. In federated learning, the central node aggregates model updates computed by the client nodes. These updates may vary because of non-IID data or variations in data quality. Robust aggregation methods such as FedProx,[29] introducing a penalty term, or FedMeanStd,[30] focusing on the mean and standard deviation of client updates, can be utilized to reduce the influence of noisy or unreliable contributions. Furthermore, personalized federated learning (PFL) has emerged to address the distribution shifts originating from non-IID client data. PFL aims to learn client-specific models that better reflect local data characteristics.[31] The selection of suitable robustness mechanisms for IDERHA's federated learning platform remains part of the ongoing technical development.

IDERHA also addresses the privacy and governance challenges posed by large multimodal datasets, particularly imaging data such as DICOM files, by keeping raw data with their respective Federated Node and avoiding any cross-border transfer of any raw data completely. Instead, federated analysis and learning workflows are executed close to the data, and only aggregated summary statistics and model parameters or updates are exchanged. Although this approach avoids the transfer of raw files, it does not eliminate the communication overhead caused by repeated exchanges of large parameter updates. The implementation of communication-efficient measures such as update compression or sparsification, asynchronous communication, optimization of only a subset of model layers (for e.g., in fine-tuning), and less frequent aggregation is subject to ongoing assessment.

This approach, combined with data standards (such as OMOP and DICOM) and workflow orchestration, is intended to facilitate federated data analysis and improve scalability across geographically distributed servers while preserving data sovereignty and privacy. In addition, IDERHA introduces technical safeguards such as manual code approval before execution and monitoring for anomalous outputs to reduce the influence of unreliable participants, thereby strengthening the trustworthiness of the platform. However, the human-in-the-loop review process can introduce significant time delays, which may slow deployment and execution of new model iterations by increasing operational overhead.

By leveraging these technical approaches, IDERHA aims to create a secure, scalable 'privacy-by-design' infrastructure for connecting health data across Europe.

IDERHA interoperability framework

In alignment with the current EHDS technical and governance requirements for the secondary use of data, the IDERHA interoperability framework has been conceptualized. An agile plan was created using the World Health Organization (WHO) planning process and TEHDAS Joint Action[32] results to adopt any future refined updates of the EHDS.[33]

The plan explored the current projects and initiatives shaping the EHDS framework while focusing on the lessons learned in implementing the main three categories of data standardization, namely: 1) data discoverability through the Data Catalog Vocabulary (DCAT-AP); 2) enabling semantic interoperability through OMOP); and 3) application of health data exchange standards (DICOM and FHIR). The plan also identified the minimum set of requirements to align the IDERHA platform with the EHDS, including data governance, user journeys, standards and interoperability, and data quality framework as outlined by the TEHDAS2 project.

To actively participate in shaping the EHDS interoperability framework for the secondary use of data, synergy was created with other EU-funded data-driven projects in collaboration with the HSbooster.eu initiative,[34] namely, EUCAIM, HealthData@EU pilot, INCISIVE, BigPicture, iHelp and ASCAPE projects.[35]

Initially, the health-standardized models, ontologies, and terminologies implemented in these projects, including FHIR, OMOP, DICOM, ISO health informatics standards DCAT, IoT, and eHealth coding standards, were compared. Afterwards, other standards related to data quality, security, AI reasoning, and wearables were explored.[36] Based on that, an interactive toolkit was developed, based on the new European Interoperability Framework (EIF), covering the main four layers of interoperability: legal, organizational, semantic and technical, to support existing and future projects in developing the interoperability framework of these projects in alignment with EHDS.[37] The toolkit is used to create the

IDERHA interoperability framework, linking the other activities and deliverables of IDERHA, such as, DMP, standards manual, data quality framework, risk management, etc. in a harmonized and interoperable way.

The need for using artificial intelligence

Health data is typically highly complex, heterogenous in content and format, and comes in large volumes. Manual analysis of such data is not feasible. Algorithmic solutions from the field of AI and more specifically machine learning (ML) are needed to identify information in extended data sets, which may be otherwise imperceptible for humans. Within IDERHA, these AI components primarily comprise predictive ML, including federated approaches; large language models are not used to generate clinical diagnoses or autonomous treatment recommendations. Generative AI is considered separately for synthetic-data generation and research and does not replace source clinical information in patient-level decision-making. IDERHA will consolidate a substantial collection of electronic health records (EHRs), LC screening, and claims datasets, which will include comprehensive clinical and imaging information from four internal consortium partners and additional external data providers. ML algorithms will be trained on these data with the following aims:

- Predicting the probability of a LC diagnosis using EHR's, potentially prior to a chest computed tomography (CT) scan. The intent is to identify individuals (including never-smokers), who would benefit from participating in a low-dose CT (LDCT) based LC screening program.

- Predicting the risk of LC in individuals where no malignancy has been identified in chest CT scans.

- Predicting the risk of malignancy in screening-detected pulmonary nodules in individuals within LC screening programs based on an automated analysis of their LDCT scans.

Importantly, ML algorithms allow for predictions on an individual level, considering their personal characteristics thus improving medical diagnoses and leading to better patient treatment and care.

For models progressing towards clinical application, predictive accuracy must be complemented by clinically interpretable output. Where methodologically applicable, this will include calibrated risk estimates, information on prediction uncertainty, relevant contributing features (such as through SHAP analysis), and attention maps for imaging-based models, together with assessment of performance across clinically relevant subgroups. The aim is to provide healthcare professionals not only with a risk estimate but also sufficient contextual information to assess its reliability and clinical applicability.

The final IDERHA case study aims to develop a mobile application (app), which allows for remote monitoring of the well-being of patients after treatment induction by tracking their physical activity, heart rate, etc., with wearable devices (such as smart watches) and asking the patient questions. These patient-reported outcomes can provide valuable feedback to the treating physicians e.g., for early complication identification or drug dosage adjustment. AI/ML algorithms can play a relevant role in processing such data, e.g., to detect potential disease- or treatment-related harms even earlier.

Challenges for regulatory adoption

While the merits of RWD and RWE are recognized, policies for their use and adoption in regulatory and HTA decision making are not universally available. Even in areas where there is a commitment from regulators to use it, concerns around scientific rigor exist.[38,39] One of IDERHA's goals is to develop policy recommendations to accelerate RWE policy development to meet the current and future advances in the digitization of health data, research methods, and health care innovations.[39] The long-term goal is to develop consensus recommendations to firstly enable appropriate and compliant data sharing, access, and governance policies, and, secondly, to support policies for the use and acceptance of RWD and RWE in public health decision-making (i.e., regulatory and HTA). Where possible, the final recommendations will seek to advance the wider global harmonization or RWE and RWD policies (Figure 2).

Clinical implementation of AI requires explicit safeguards and human oversight. IDERHA-derived predictions are intended to support, rather than autonomously replace, clinical decision-making. Models intended for clinical use will require validation for their defined intended purpose, appropriate regulatory assessment, and integration into institutional clinical-governance processes. AI-derived output should be considered together with established clinical guidelines, source data and the individual clinical context; discordant model predictions should not automatically supersede guideline-based care. Responsibility for clinical use and potential liability will depend on the regulatory status and intended use of the respective system as well as applicable national and institutional frameworks.

Policy recommendations in IDERHA are developed through a multi methodology approach, including stakeholder workshops with different relevant stakeholders, interviews, document analysis and desk research. Multiple stakeholder involvement is sought through public consultations, public forums, and the involvement of the IDERHA Patient Advisory Board, the Clinical Advisory Board, the Ethics Advisory Board, and the Integrated Data Access Governance Council (IDAGC). In this way, policy recommendations in IDERHA are developed within an iterative process of draft releases and feedback from relevant stakeholders. A priority is the transparent development of policy recommendations through involving all relevant stakeholders, including e.g., patients and patient advocates, researchers, regulators, industry representatives, data protection officers, and diversity and inclusion advocates. To ensure that the eventual policy recommendations will have the support of the broader health care ecosystem, it is essential to obtain ongoing multi-stakeholder input. In order to support IDERHA's objectives, IDERHA has established the IDAGC as a multi-stakeholder advisory board bringing together legal, ethical, technical, regulatory, and data-holder perspectives. During the project's lifetime, the IDAGC provides strategic guidance on facilitating the secondary use of health data and strengthening the generation and acceptance of RWE for HTA and regulatory decision-making. However, the IDAGC does not assess or decide on individual data access requests. Such decisions remain the responsibility of the respective data holders, in accordance with applicable legal and governance requirements. Importantly, the federated and decentralized architecture of IDERHA enables data holders to retain local control over their datasets and participate only in activities that are compatible with their institutional and national legal obligations. Where full alignment cannot be achieved, analyses can be adapted to reflect local constraints while maintaining transparency regarding differences in participation and data availability.

IDERHA will also establish contacts and work together with other projects with overlapping interest in policy development to avoid duplication of work and dividing resources in this fast evolving and complex area. This will include other European initiatives in the health data sharing space like, but not exclusively, the TEHDAS2 project (Second Joint Action Towards European Health Data Space) and the H_2O project (Health Outcomes Observatory).[40]

Ethical and legal commitment

IDERHA is focusing on easier use and reuse of digital health data, adhering to GDPR, AI Act and ethical standards, while navigating a complex and dynamic legal landscape. An important objective of IDERHA is to identify and verify resilient legal solutions for modular and large-scale data sharing within a cross-sectoral and cross-jurisdictional setting.

Focused on data-driven and patient-centered health innovation, IDERHA is pursuing its research aims within the statutory acts stemming from the EU strategies for data[41] and digitalization.[41,42] Corresponding to a fundamental principle of accountability and the risk-based approach within EU data and digitalization law, sustainability and flexibility are ensured within IDERHA through a consistent emphasis on a complete documentation of all data-driven project activities. A pivotal means to this end is the Data Management Plan (DMP) which, next to its traditional scope orientated towards scientific data management standards and ensuring data FAIRness[43,44] has been recalibrated to meet the legal requirements of documentation. In the meantime, documentation standards of e.g. the AI Act must be included and based upon the DMP. As the central element within the IDERHA data sharing framework, the DMP ensures that all risk assessments, privacy contracts, and transparency documents refer to a single factual basis, documenting all data sources and data flows in a legally, ethically and FAIR compliant manner.

This way IDERHA aims to ensure coherence and transparency towards data subjects and data holders, contributing to trust. By checking the relevant parts of the DMP, institutional stakeholders can familiarize themselves with the data processing taking place within IDERHA and verify that all legal and ethical requirements are met. Through its innovative use of the DMP as a central piece of real-time documentation, IDERHA aims to be well-positioned to adapt as necessary to changing statutory demands.

To this end, IDERHA's data sharing framework aims to deliver a proof-of-concept on how to enable modular and large-scale data sharing by enhancing compliance.

Alignment with the European health data space regulation

The EHDS will be an ecosystem for health-specific data access, sharing, and use. It aims to establish shared practices, standards, rules, a digital infrastructure, and a governance framework[45] and entered into force on 26 March 2025.[46] Depending on the use cases, data types, and data use categories, it will become applicable in various stages, ranging from two years after its entry into force to four or six years, with certain exemptions up to ten years.[45]

The EHDS will give citizens from the EU the right to access and manage their health data within the EU as well as to opt out from: 1) Primary data use (MyHealth@EU) for the provision of their individual healthcare,[47] and 2) Secondary data

use (HealthData@EU) for regulatory, research, and innovative purposes as well as for policymaking. The EHDS thus specifies the framework for connecting and sharing health data in the EU.[48]

Both the EHDS and IDERHA are fully aligned with the Europe Beating Cancer Plan,[49] and both aim to improve public health by making better use of health data for research, innovation, and regulatory purposes. Hence, there are various synergy opportunities. The creation of Health Data Spaces by public-private initiatives such as IDERHA is important for sharing expertise and informing the EHDS implementing acts of practical lessons learnt within those initiatives. By March 2027, the Commission is expected to adopt implementing acts that will set out the detailed rules necessary for the operationalisation of the regulation.

Public consultations on key projects, such as TEHDAS2, provide IDERHA with a valuable approach to review ongoing work, deliver feedback, and issue recommendations, particularly in relation to federated infrastructure. The planned Stakeholder Forum of the EHDS, scheduled for establishment by 2029, could serve as a pivotal platform for showcasing IDERHA's insights, though its launch will fall outside the project's active timeframe. On the other hand, the IDERHA project is aiming to align with the EHDS principles, e.g., concerning interoperability, legal compliance, and ethical standards, and to be ready for EHDS once the regulation is applicable. For this reason, IDERHA has developed a comprehensive plan on how to align with the EHDS requirements for the secondary use of health data, including services, architecture, user journey, and standards.[33]

While EHDS alignment is central to IDERHA, it is essential to emphasize that the EHDS will function as an additional framework for data access and will not replace existing or ongoing data sharing networks such as IDERHA.

Conclusion & discussion

The IDERHA project is a pioneering initiative designed to overcome critical barriers to accessing, integrating, and analyzing health data, with the aim of maximizing its value for patient care and medical research. By developing a scalable, privacy-preserving platform embedded in a coherent data governance framework and accompanying policy recommendations, IDERHA aims to support healthcare professionals, patients, and researchers in improving clinical outcomes and driving medical innovation (Figure 2).

IDERHA monitors and incorporates insights and developments from other European initiatives such as TEHDAS2, EHDS2 Pilot,[50] DSSC, DARWIN EU, OPTIMA and HealthyCloud and ensures alignment wherever possible. Together, we are laying the groundwork for the EHDS. This interconnectedness allows IDERHA to ensure that its solutions are compatible, scalable, and reusable across future health data infrastructures, within Europe and beyond. A key strength of IDERHA lies in its alignment with the EHDS principles, underscoring its commitment to interoperability, ethical integrity, and legal compliance.

Technically, IDERHA's privacy-by-design approach ensures that data remains with its holders, mitigating privacy risks while maintaining data sovereignty. By implementing federated analytics/learning, where analyses are brought to the data, IDERHA ensures compliance with the GDPR and enhances trustworthiness.[25] The adoption of internationally recognized standards, such as those issued from Global Alliance for Genomics and Health (GA4GH) and International Data Spaces Association (IDSA), further ensures the platform's scalability and security, making it fit-for-purpose for a European-wide data infrastructure (Figure 2).

While ultimately disease-agnostic, IDERHA leverages lung cancer as a high-impact use case – reflecting both the clinical urgency and the rich potential of improved health data integration for early detection and treatment.

The project also explores the transformative potential of artificial intelligence and machine learning. These tools offer the ability to detect subtle patterns in complex health data – imperceivable to humans and crucial for applications such as lung cancer screening. Through its use cases, IDERHA aims to significantly enhance early diagnosis and personalized care for citizens and patients.

Nevertheless, achieving the project's vision involves significant challenges. Regulatory adoption remains a major hurdle – particularly the integration of real-world data (RWD) and real-world evidence (RWE) into clinical and policy contexts. IDERHA's multi-method approach, including stakeholder workshops and public consultation, is a constructive step toward policy alignment. However, broader harmonization across EU Member States and international partners will be essential. In parallel, IDERHA collaborates with initiatives such as the SYNTHIA Project,[51] which develops tools and methods for synthetic health data generation. These synthetic datasets enable faster experimentation and analysis in a privacy-preserving manner,[52] without the need to transfer sensitive personal data or having legal procedures in place– helping to accelerate innovation while aligning with strict data protection standards.

Ethical, legal, and FAIR compliance is another major challenge to which IDERHA responds with a modular and scalable data governance framework accommodating all data-driven activities within IDERHA to the extent relevant for legal, ethical or FAIR purposes. IDERHA's DMP documents all data related activities and transparency measures – offering a living framework to ensure alignment between progression of the project, evolving legal standards and public expectations.

In addition to the answers to these immediate challenges, IDERHA also anticipates and builds solutions for emerging issues, such as scaling federated AI methods to handle complex data modalities across borders, integrating patient-reported data into clinical workflows, and addressing persistent gaps in semantic interoperability across research and care domains. By incorporating responsive compliance mechanisms – able to adapt to evolving legal and regulatory contexts – IDERHA aims towards alignment with data protection requirements across jurisdictions. Its flexible, standards-based architecture and emphasis on data sovereignty will make IDERHA's tools and frameworks reusable by other national data hubs, cancer missions, and EU digital health programs.

In conclusion, IDERHA represents a bold and necessary step towards building a secure, interoperable, and legally and ethically grounded European health data ecosystem. By tackling technical, regulatory, and ethical challenges in parallel, the project is laying a foundation for scalable solutions that improve public health. Success will depend on continued collaboration with stakeholders – including healthcare professionals, patients, researchers, and policymakers. With its forward-looking vision and strong alignment with EHDS goals, IDERHA is well-positioned to contribute meaningfully to the responsible and impactful use of health data across Europe. As such there are strong ambitions to not only establish the technical and legally/ethically compliant solution, but also to develop a sustainable framework that will allow the use, further expansion and development of this EU funded infrastructure for the long term with new partners and indications.

Disclaimer

Funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them. Publication in Open Research Europe does not imply endorsement of the European Commission.

Ethics and consent

Ethics and consent were not required.

Data availability

Underlying data

There are no new data associated with this article.

Acknowledgements

This publication was built on the work of many people involved in the IDERHA consortium. It consists of the following partners and members:

ASK: Nele Gessler, Dagmar Pilz; **CDR**: Lotte Groth Jensen, Camilla Palmhøj Nielsen, Morten Deleuran Terkildsen, Kathrine Carstensen, Stina Lou, Jakob Feldtfos Christensen, Rasmus Mølgaard Hansen, Maria Heilskou Pedersen; **DKMA**: Niels Henrik Meedom, Jakob Wested, Stine Hasling Mogensen, Jakob Wested; **EPF**: Flavia Topan, Gözde S. Briggs, Valentina Strammiello, Bianca Pop, Claudia Louati, Arlet Viader; **ERS**: Brian Ward, Celine Genton, Valérie Vaccaro; **FHI**: Philip Gribbon, Lena Hegel, Farid Halabi, Marcel Klötgen, Mostafa Kamal Mallick, Simon Geller, Timo Renzelmann, Joel Weihrauch, Quang Tinh Le, Holger Fröhlich, Sumit Madan, Erik Ostrowski, Anja Burmann, Mira Grättinger, Carsten Claussen, Katja Herzog, Katerina Tschausovska, Anne Funck Hansen, Eduard Wolf, Yukai Sun, Daniel Buschman, Reagon Karki, Felix Ritter, Leonie von Berlin; **FHTW**: Lilly Treml, Birgit Pohn, Matthias Frohner, Stefan Sauermann, Lukas Rohatsch; SAS: David Vicente-Baz, Luis G. Luque-Romero, Alberto Moreno-Conde; **FISEVI:** Francisco J. Nuñez-Benjumea, Miguel Giraldez-Álvarez, Cristina Simarro-Castellanos, Lourdes Martinez-Martinez; **FORTREA:** Elke Jahn, Hannah Simonds, Melissa Harris, Pernilla Aarskog, Malgorzata Mirowska, Denitza Muller, Piqueras Benavides, Maria Jose; **HULAFE**: Carlos López Gómez, Eugenia Gas López, Francisco de Asis Aparisi, Leonor Cerda Puchades, Luis Martí Bonmatí, Matías Fernandez Patón, Oscar Juan Vidal, Gloria Ribas Despuig; **HYG:** Dominik Geller, Ivan Nesic, Josef Sevcik, Niamh O'Shea; **i-HD:** Jens Declerck, Nathan Lea, Dipak Kalra, Peter Singleton, Petra Wilson, Nathan Gatpandan; **ITTM:** Andreas Kremer, Erno Lindfors, Gabriel Sieglerschmidt, Maria Quaranta, Romain Tching, Chi Yen, Sanna Palmqvist; **J-C Neuss**: Dorothea Perings, Joern Sindern; **J-C UK:** Maria Bardaji-Cruz, Susan M. Sandler, Pablo Del Yesso; **JJM:** Christian Muehlendyck, Jose-Felipe Golib-Dzib, Rita Peeters, Sanziana Negreanu Arboreanu, Ulrich Juknat, Emanuela Mauro, Monika Brand, Alan Curley, Mariana Constantinovici,

Angel Martin, Elisabetta Gatti, Dominika Domanska, Delia Strunz, Francisco Talamas, Nyree Louise Kelly, Sharad Shukla, Zhennan Zhu, João Pedro Linheiro De Sá Gomes Pereira, **JPNV:** Hans Verstraete, Hans Vergauwen; **JPV:** Andrée Willemse; **LBG:** Rada Hussein, Somayeh Abedian **LH**: Stanley Letovsky, Helio Gil, Paul Mesange, Eric Thomas; **LUCE:** Anne-Marie Baird, Maeve O'Sullivan, Anniina Hentinen; **LYG:** Jan-Willem Boiten, Tieneke B. M. Schaaij-Visser, Robin Navest, Erwin Boutsma, Herman Verheij, Wim-Jan Koot, Nebo Savic, Leone Bock, Judy Löwik, Edwine Beya, Mark Pieters; **MDBSI:** Heather Colvin, Kathy Chen, Hanane Omichessan, Stacey Bentley; **MSB:** Rebecca C. Graebig-Rancourt, Torsten Gerriet Blum; **MUW:** Tanja Stamm, Margaret Andrews, Erika Mosor, Nadja Kartschmit, Kathrin Morasek, Maryam Jafarpour; **NICE:** Stephen Duffield, Omnia Bilal, Bethany Shinkins; **PMS:** Saurav Baidya, Srivatsa RV, Raf Cuypers, Catharina Ibrahim, Yusuf Sayıta; **Q1.6**: Filip Vissers, Guillaume de Decker, Cristina Cumpanasoiu; **ROCHE**: Eric Boernert, Lesley Maloney, Kimberly Noel, Milan Obradovic, Marlene Thomas, Gracy Crane, Katalin Ersek, Afua van Haasteren, Andre Trottier, Brad Spring, Lukasz Kaczmarek, Andy Nguyen, Melanie Huntley, Carly McWilliams, Matthias Antonin, Erika Schirghuber, Frederik Buijs, Anna Greco; **SARD:** Phillip Dennis, Zhaoling Meng, Rebecca Lumsden, Philippe Bordes, Marc Dietrich Voss, Gerasimos Konidari, Medha Sasane, Paulo Carita, Charlie Nicholls, Nathalie Largeron, Yves Jacob, Philip Dennis, Audry L'Homme, Lionel Bascle, Eric Génévois-Marlin, Amaya Clemente; **SAS:** Sophie Monteau; **SERGAS:** Portal Ramos Vieiro, Carlos Tilve, Irene Martinez Martin, Daniel Llamas, Pablo Feijoo-Vilanova, Yolanda Vazquez; **TPIZ:** Alain Anyouza, Sunil Dravida, Saurin Jani, Jugal Sharma, Julia Zhu, Nick Bott, Missy Heidelberg, Rose Purcell, Meni Styliadou, Nataliya Kemenyash, Eda Niland, Hrvojka Kostelac, Marco Palladino, Ling Li, Liang Feng, Meredith Trabinska; **TUM-MED**: Daniel Rueckert, Maulik Chevli, Johannes Brandt; **UNILU:** Linda Ebermann, Soumyabrata Ghosh, Sascha Herzinger, Venkata Pardhasaradhi Satagopam, Hanna Ćwiek-Kupczyńska, Jaleh Shoshtarian Malak, Vilem Ded, Jenny Thuy Dung Tran, Artur Jorge Oliveira de Carvalho, Sandrine Medves; **VTT**: Jaakko Lähteenmäki, Liedes Hilkka, Mika Hilvo, Ilona Ruotsalainen, Claus Møldrup.

Mistral Le Chat (version Small 3) was used to assist in summarizing the article in the abstract.

References

1. **ECIS - European Cancer Information System | ECIS - European Cancer Information System.** Accessed: Jun. 23, 2025.
Reference Source
2. **Cancer in Germany 2019/2020.**
Reference Source
3. Luengo-Fernandez R, Leal J, Gray A, *et al.*: **Economic burden of cancer across the European Union: a population-based cost analysis.** *Lancet Oncol.* Nov. 2013; **14**(12): 1165–1174.
PubMed Abstract | Publisher Full Text
4. Alt-Epping B, Stäritz AE, Simon ST, *et al.*: **What is special about patients with lung cancer and pulmonary metastases in palliative care? Results from a nationwide survey.** *J. Palliat. Med.* Sep. 2012; **15**(9): 971–977.
Publisher Full Text
5. Ellis J: **The impact of lung cancer on patients and carers.** *Chron. Respir. Dis.* Feb. 2012; **9**(1): 39–47.
Publisher Full Text
6. van der Aalst CM, van den Bergh KAM, Willemsen MC, *et al.*: **Lung cancer screening and smoking abstinence: 2 year follow-up data from the Dutch–Belgian randomised controlled lung cancer screening trial.** *Thorax.* Jul. 2010; **65**(7): 600–605.
Publisher Full Text
7. **Improving cancer screening in the European Union – Scientific Advice Mechanism.** Accessed: Jun. 23, 2025.
Reference Source
8. **Council updates its recommendation to screen for cancer, Consilium.** Accessed: Jun. 23, 2025.
Reference Source
9. van Meerbeeck JP, Franck C: **Lung cancer screening in Europe: where are we in 2021?** *Transl. Lung Cancer Res.* May 2021; **10**(5): 2407–2417.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Catarata MJP, *et al.*: **ERS Congress 2024: highlights from the Thoracic Oncology Assembly.** *ERJ Open Res.* Apr. 2025; **11**(2).
Publisher Full Text
11. Potter AL, Rosenstein AL, Kiang MV, *et al.*: **Association of computed tomography screening with lung cancer stage shift and survival in the United States: quasi-experimental study.** *BMJ.* Mar. 2022; **376**: e069008.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Yang C-Y, *et al.*: **Stage Shift Improves Lung Cancer Survival: Real-World Evidence.** *J. Thorac. Oncol.* Jan. 2023; **18**(1): 47–56.
Publisher Full Text
13. Balata H, *et al.*: **Analysis of the baseline performance of five UK lung cancer screening programmes.** *Lung Cancer Amst. Neth.* Nov. 2021; **161**: 136–140.
Publisher Full Text
14. Samaržija M, *et al.*: **Unpublished data.**
15. **HANSE-Studie mit ersten Ergebnissen.** Accessed: Jun. 23, 2025.
Reference Source
16. Abu-Odah H, Said NB, Nair SC, *et al.*: **Identifying barriers and facilitators of translating research evidence into clinical practice: A systematic review of reviews.** *Health Soc. Care Community.* Nov. 2022; **30**(6): e3265–e3276.
PubMed Abstract | Publisher Full Text
17. Liu F, Panagiotakos D: **Real-world data: a brief review of the methods, applications, challenges and opportunities.** *BMC Med. Res. Methodol.* Nov. 2022; **22**(1): 287.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Bell SK, Delbanco T, Elmore JG, *et al.*: **Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes.** *JAMA Netw. Open.* Jun. 2020; **3**(6): e205867.
Publisher Full Text
19. Botsis T, Hartvigsen G, Chen F, *et al.*: **Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.** *Summit Transl. Bioinforma.* Mar. 2010; **2010**: 1–5.
20. de Mello BH, Rigo SJ, da Costa CA, *et al.*: **Semantic interoperability in health records standards: a systematic literature review.** *Health Technol.* 2022; **12**(2): 255–272.
PubMed Abstract | Publisher Full Text | Free Full Text
21. FitzHenry F, *et al.*: **Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership.** *Appl Clin Inform.* Aug. 2015; **6**(3): 536–547.
Publisher Full Text
22. Kent S, Burn E, Dawoud D, *et al.*: **Common Problems, Common Data Model Solutions: Evidence Generation for Health**

Technology Assessment. *Pharmacoeconomics.* 2021; **39**(3): 275–285.
Publisher Full Text

23. Ahmadi N, Peng Y, Wolfien M, *et al.*: **OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review.** *Int. J. Mol. Sci.* Oct. 2022; **23**(19): 11834.
PubMed Abstract | Publisher Full Text | Free Full Text

24. **Real-world evidence for regulatory purposes: The example of DARWIN EU® - ScienceDirect.** Accessed: Jul. 23, 2025.
Reference Source

25. **Journal of Medical Internet Research - Reality Check: The Aspirations of the European Health Data Space Amidst Challenges in Decentralized Data Analysis.** Accessed: Sep. 24, 2025.
Reference Source

26. Bak M, Madai VI, Celi LA, *et al.*: **Federated learning is not a cure-all for data ethics.** *Nat. Mach. Intell.* Apr. 2024; **6**(4): 370–372.
Publisher Full Text

27. **Dataspace Protocol 2024–1|IDS Knowledge Base.** Accessed: Jun. 23, 2025.
Reference Source

28. **How OpenID Connect Works - OpenID Foundation.** Accessed: Sep. 24, 2025.
Reference Source

29. Li T, Sahu AK, Zaheer M, *et al.*: **Federated optimization in heterogeneous networks.** *Proceedings of Machine learning and systems.* 2020 Mar 15; **2**: 429–450.

30. Singh T, Kundroo M, Kim T, *et al.*: **FedMeanStd: Federated Aggregation With Outliers Filtering.** *IEEE Access.* 2026; **14**: 6194–6206.
Publisher Full Text

31. T Dinh C, Tran N, Nguyen J: **Personalized federated learning with moreau envelopes.** *Advances in Neural Information Processing Systems.* 2020; **33**: 21394–21405.

32. **Recommendations on a Data Quality Framework for the European Health Data Space for secondary use.** Accessed: Sep. 24, 2025.
Reference Source

33. Hussein R, Balaur I, Burmann A, *et al.*: **Getting ready for the European Health Data Space (EHDS): IDERHA's plan to align with the latest EHDS requirements for the secondary use of health data.** *Open Res. Eur.* 2024; **4**: 160.
PubMed Abstract | Publisher Full Text | Free Full Text

34. **HSBooster.eu.** Accessed: Jun. 23, 2025.
Reference Source

35. Gyrard A, *et al.*: **Synergies Among Health Data Projects with Cancer Use Cases Based on Health Standards.** *Stud. Health Technol. Inform.* Aug. 2024; **316**: 1292–1296.
Publisher Full Text

36. Gyrard A, Abedian S, Gribbon P, *et al.*: **Lessons Learned From European Health Data Projects With Cancer Use Cases: Implementation of Health Standards and Internet of Things Semantic Interoperability.** *J. Med. Internet Res.* Mar. 2025; **27**: e66273.
Publisher Full Text

37. Hussein R, Gyrard A, Abedian S, *et al.*: **Interoperability Framework of the European Health Data Space for the Secondary Use of Data: Interactive European Interoperability Framework–Based Standards Compliance Toolkit for AI-Driven Projects.** *J. Med Internet Res.* Apr. 2025; **27**: e69813.
Publisher Full Text

38. Wieseler B, Neyt M, Kaiser T, *et al.*: **Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy?** *BMJ.* Mar. 2023; **380**: e073100.
PubMed Abstract | Publisher Full Text

39. Cresswell K, *et al.*: **Report on Global Regulatory Best Practices Analysis (D6.2).** Apr. 2024. Accessed: Jun. 23, 2025.
Reference Source

40. Stamm T, *et al.*: **Building a Value-Based Care Infrastructure in Europe: The Health Outcomes Observatory.** *Catal. Non-Issue Content.* Jun. 2021; **2**(3).
Publisher Full Text

41. **A European strategy for data|Shaping Europe's digital future.** Accessed: Jun. 23, 2025.
Reference Source

42. **A Europe fit for the digital age - European Commission.** Accessed: Jun. 23, 2025.
Reference Source

43. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci. Data.* Mar. 2016; **3**(1): 160018.
Publisher Full Text

44. Rocca-Serra P, Gu W, Ioannidis V, *et al.*: **The FAIR Cookbook - the essential resource for and by FAIR doers.** *Sci. Data.* May 2023; **10**(1): 292.
PubMed Abstract | Publisher Full Text | Free Full Text

45. **Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space.** 2022. Accessed: Jun. 23, 2025.
Reference Source

46. **Regulation - EU - 2025/327 - EN - EUR-Lex.** Accessed: Jul. 23, 2025.
Reference Source

47. **Electronic cross-border health services - European Commission.** Accessed: Jun. 23, 2025.
Reference Source

48. **Data sharing through eDelivery in the HealthData@EU.** Accessed: Jun. 23, 2025.
Reference Source

49. **Europe's Beating Cancer Plan - Communication from the commission to the European Parliament and the Council.**
Publisher Full Text

50. Bacry E, Jendrossek M: **EHDS2 pilot infrastructure for the secondary use of health data.** *Eur. J. Public Health.* Oct. 2022; **32**(Suppl 3): ckac129.368.
Publisher Full Text

51. **Home - Synthia.** Accessed: Sep. 24, 2025.
Reference Source

52. Kühnel L, Schneider J, Perrar I, *et al.*: **Synthetic data generation for a longitudinal cohort study – evaluation, method extension and reproduction of published data analysis results.** *Sci. Rep.* Jun. 2024; **14**(1): 14412.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓

Version 4

Reviewer Report 12 October 2026

https://doi.org/10.21956/openreseurope.27534.r82631

© **2026 Mehandiratta Y.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ **Yash Mehandiratta**
All India Institute of Medical Sciences (AIIMS), Guwahati, Assam, India

Authors have addressed the suggestions provided in the last review report. the manuscript now clearly defines, how IDERHA will use Artificial intelligence. The manuscript better describes the healthcare side of the framework and how the platform derived predictions are intended to support, rather than autonomously replace, clinical decision making.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence in Healthcare, Surgical Oncology, Robotic and minimal invasive surgery, Implementation of AI in Surgical Practice, Global implementation of AI in Healthcare

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 3

Reviewer Report 29 September 2026

https://doi.org/10.21956/openreseurope.27357.r81590

© **2026 Mehandiratta Y.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Yash Mehandiratta**

All India Institute of Medical Sciences (AIIMS), Guwahati, Assam, India

Thank you to the authors for the opportunity to review their work. I must praise them for the timely topic and interesting use of the upcoming technology for detection of lung cancer. It addresses an important clinical domain and may help design future of cancer screening and prevention.
The authors have very well described the collection of quality data and appropriate use of technology to turn this heterogenous high quality data into standardized information that healthcare workers are able to use.
The authors have very well described the decentralization of information and privacy first approach to the technology, that is an area of concern in the modern times. The article also highlights the practical challenges to implementation of this technology and commitment to adhere with European regulatory guidelines.
I believe article will benefit by adding more details regarding medical and healthcare side of the working technology, as it will be the main interface of the practical use and implementation. The authors describe *"better define individuals at high risk for LC beyond the currently used age and tobacco-exposure-based risk criteria"* as a use case, however there is no mention of any alternative plans or criteria for the same. Since the current screening methods are for high risk individuals (mostly smokers), it will be difficult to optimize the available data that authors are using to train ML to screen for non smokers as authors clarify the intent to use the technology for screening selection before LDCT (including never-smokers). Put simply , how will IDERHA avoid simply reproducing existing smoking-dominated lung-cancer prediction models?
Similarly, a brief description of parameters for predicting the risk of LC in individuals where no malignancy has been identified in the chest CT may also help interest the real user (Physicians and hospitals) of the proposed technology.
The authors have promptly described the use of patient feedback and use of smartwatches to collect treatment data.
It will be beneficial to clarify whether the proposed AI components are conventional predictive ML/federated learning models, generative AI/LLMs, or a combination of both. If generative AI is used for data synthesis or clinical summarization, the authors should describe mechanisms for preventing unsupported or fabricated information from entering the clinical workflow, including source attribution, verification mechanisms and human oversight.
For predictive ML models, the authors should describe the proposed approach to explainability, including whether clinicians will receive feature level explanations, uncertainty estimates, calibration information or other clinically interpretable outputs. As authors have acknowledged, the heterogenous nature of health data and OMOP-CDM as a solution to this, it will be worthy to describe how authors plan to make this ML output interpretable in simpler terms for frontline healthcare professionals to understand and finally use in clinical practice. How the resulting model outputs will be translated into clinically interpretable information for healthcare professionals.
Adding a paragraph on safety in implementation of this technology for final clinical decision making will be a well worth addition, such as, in case where ML model provides interpretation based on regional data that may cause physicians to deviate from standard guidelines then what will be the regulatory and legal implications of that, what are the safety nets that may be imposed for patient protection in such cases and how the final decision becomes the responsibility and may become legal obligation of the physician.
Addressing these points would improve the manuscript's clinical relevance and help demonstrate how the IDERHA platform can move beyond technical infrastructure toward safe and meaningful use in real-world healthcare.

Is the rationale for the Open Letter provided in sufficient detail? (Please consider whether existing challenges in the field are outlined clearly and whether the purpose of the letter is explained)
Yes

Does the article adequately reference differing views and opinions?
Yes

Are all factual statements correct, and are statements and arguments made adequately supported by citations?
Yes

Is the Open Letter written in accessible language? (Please consider whether all subject-specific terms, concepts and abbreviations are explained)
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow? (Please consider whether others in the research community would be able to implement guidelines or recommendations and/or constructively engage in the debate)
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence in Healthcare, Surgical Oncology, Robotic and minimal invasive surgery, Implementation of AI in Surgical Practice, Global implementation of AI in Healthcare

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 29 Sep 2026

Rebecca Graebig-Rancourt

Dear reviewer, We thank you for taking the time to provide feedback on our article. We have made the suggested changes (see in track changes in the manuscript). Please find our point-by-point responses below.
Review: Thank you to the authors for the opportunity to review their work. I must praise them for the timely topic and interesting use of the upcoming technology for detection of lung cancer. It addresses an important clinical domain and may help design future of cancer screening and prevention.
The authors have very well described the collection of quality data and appropriate use of technology to turn this heterogenous high quality data into standardized information that healthcare workers are able to use.

The authors have very well described the decentralization of information and privacy first approach to the technology, that is an area of concern in the modern times. The article also highlights the practical challenges to implementation of this technology and commitment to adhere with European regulatory guidelines. I believe article will benefit by adding more details regarding medical and healthcare side of the working technology, as it will be the main interface of the practical use and implementation. The authors describe "better define individuals at high risk for LC beyond the currently used age and tobacco-exposure-based risk criteria" as a use case, however there is no mention of any alternative plans or criteria for the same. Since the current screening methods are for high risk individuals (mostly smokers), it will be difficult to optimize the available data that authors are using to train ML to screen for non smokers as authors clarify the intent to use the technology for screening selection before LDCT (including never-smokers). Put simply, how will IDERHA avoid simply reproducing existing smoking-dominated lung-cancer prediction models?
Response to reviewer: Thank you for this important point. We agree that a model trained predominantly on populations selected by conventional smoking-based criteria could simply reproduce existing risk stratification. We have therefore clarified that the first IDERHA use case is intended to complement, rather than replace, established screening eligibility criteria and to investigate whether routinely available multimodal health data add predictive information beyond age and tobacco exposure. Model performance will be assessed across relevant subgroups, including smoking strata and, where sample size permits, never-smokers. Importantly, the identification of risk in currently non-eligible populations would require dedicated prospective validation before any extension of screening eligibility could be considered. Text was added in manuscript as seen with track changes. Similarly, a brief description of parameters for predicting the risk of LC in individuals where no malignancy has been identified in the chest CT may also help interest the real user (Physicians and hospitals) of the proposed technology.

Response to reviewer: We agree and have added a brief description of the intended input information. This use case differs from pulmonary-nodule malignancy classification: it aims to extract predictive information from the complete chest CT even when no suspicious malignant lesion is identified and, where appropriate, combine this with available clinical risk information. The exact feature representation will depend on the modelling approach and data available across participating federated nodes. We have added also a brief description for use case 3 Text was added in manuscript as seen with track changes. The authors have promptly described the use of patient feedback and use of smartwatches to collect treatment data.
It will be beneficial to clarify whether the proposed AI components are conventional predictive ML/federated learning models, generative AI/LLMs, or a combination of both. If generative AI is used for data synthesis or clinical summarization, the authors should describe mechanisms for preventing unsupported or fabricated information from entering the clinical workflow, including source attribution, verification mechanisms and human oversight.

Response to reviewer: Thank you for requesting this clarification. The clinical AI use cases described in IDERHA are based primarily on predictive machine-learning approaches, including their implementation within federated analytics/federated

learning environments. Generative AI or large language models are not currently used to generate clinical diagnoses, risk predictions, or autonomous clinical recommendations in these use cases. Synthetic-data generation, mentioned separately in relation to the SYNTHIA project, serves methodological and research purposes and should not be confused with generation of patient-level clinical information. Text was added in manuscript as seen with track changes. For predictive ML models, the authors should describe the proposed approach to explainability, including whether clinicians will receive feature level explanations, uncertainty estimates, calibration information or other clinically interpretable outputs. As authors have acknowledged, the heterogenous nature of health data and OMOP-CDM as a solution to this, it will be worthy to describe how authors plan to make this ML output interpretable in simpler terms for frontline healthcare professionals to understand and finally use in clinical practice. How the resulting model outputs will be translated into clinically interpretable information for healthcare professionals.

Response to reviewer: We agree that predictive performance alone is insufficient for clinical implementation. We have therefore clarified that models progressing toward clinical use should provide clinically interpretable outputs in addition to the numerical prediction itself. Depending on the underlying model, these may include information on relevant contributing features following SHAP, information gain, or similar methodologies, attention maps for imaging-based predictive models, calibrated risk estimates, uncertainty, and performance across clinically relevant subgroups. Such information is intended to support clinicians in assessing whether a prediction is plausible and applicable to an individual patient. Text was added in manuscript as seen with track changes. Adding a paragraph on safety in implementation of this technology for final clinical decision making will be a well worth addition, such as, in case where ML model provides interpretation based on regional data that may cause physicians to deviate from standard guidelines then what will be the regulatory and legal implications of that, what are the safety nets that may be imposed for patient protection in such cases and how the final decision becomes the responsibility and may become legal obligation of the physician.
Addressing these points would improve the manuscript's clinical relevance and help demonstrate how the IDERHA platform can move beyond technical infrastructure toward safe and meaningful use in real-world healthcare.

Response to reviewer: We fully agree that clinical safety and human oversight are essential. We have added a paragraph emphasizing that IDERHA-derived AI outputs are intended as clinical decision support rather than autonomous decision-making tools. Before clinical implementation, models require validation for their defined intended use, appropriate regulatory assessment, and integration into local clinical governance. AI output should be interpreted together with established clinical guidelines and individual patient information. Where an AI-derived prediction differs from guideline-based management, the model output should not automatically override established standards of care. We have deliberately avoided assigning legal liability generically to an individual physician, as responsibilities depend on the intended use of the technology, its regulatory status and the applicable national and institutional legal framework. Text was added in manuscript as seen with track

changes.

Competing Interests: No competing interests were disclosed.

Version 2

Reviewer Report 02 September 2026

https://doi.org/10.21956/openreseurope.27054.r79743

© **2026 Putra D.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Daniel Happy Putra
Universitas Esa Unggul, West Jakarta, Jakarta, Indonesia

Thank you for the detailed point-by-point response and for revising the manuscript. I compared Version 2 with Version 1 and checked each response against the corresponding manuscript location. The revision improves the discussion of federated learning, but several issues remain before I can recommend full approval.

1. Federated-learning limitations and mitigation. The new paragraphs on page 8 appropriately acknowledge non-IID data, unreliable client updates, local handling of DICOM data, and human-in-the-loop safeguards. However, the technical distinctions should be sharpened. FedProx primarily addresses statistical and systems heterogeneity; it should not be presented as a safeguard against malicious client updates. In addition, keeping raw DICOM data local reduces raw-data transfer but does not, by itself, resolve communication overhead from repeated exchange of large model parameters or updates. Please briefly state which optimization or orchestration measures IDERHA plans to use, or clearly identify this as ongoing work.

2. OMOP semantic-mapping quality assurance. The response provides useful and concrete information about WhiteRabbit, Usagi, Athena, Achilles, the Data Quality Dashboard, expert review, confidence classification, provenance-aware ETL, and reconciliation procedures. However, this information does not appear in the revised manuscript; the relevant “Data standardization” text remains essentially unchanged. Please add a concise version of this material to the manuscript so that the Open Letter functions as an actionable blueprint for external users.

3. Conflicting national legal interpretations and IDAGC resolution. The response states that the manuscript will be revised to explain the IDAGC's role and the flexibility of the federated model, but I could not locate this addition in Version 2. Please add the promised explanation, including the decision process, relevant criteria, documentation of unresolved disagreement, and how participation or analysis is adapted when national requirements cannot be aligned.

Please also correct two typographical errors in the new page 8 text: “statisics” and “signficant.”

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research areas include medical record and health information management, health informatics, and applied public health and epidemiology. My work particularly focuses on health data quality, interoperability, governance, and the secondary use of routine and administrative health data for research, service evaluation, and evidence-informed decision-making.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Sep 2026

Rebecca Graebig-Rancourt

We thank the reviewer for taking the time and reviewing our article. Version 3 of this article now includes the suggested text which will be seen with track changes and highlight (as detailed below in our responses).
Thank you. Reviewer: 1. Federated-learning limitations and mitigation. The new paragraphs on page 8 appropriately acknowledge non-IID data, unreliable client updates, local handling of DICOM data, and human-in-the-loop safeguards. However, the technical distinctions should be sharpened. FedProx primarily addresses statistical and systems heterogeneity; it should not be presented as a safeguard against malicious client updates. In addition, keeping raw DICOM data local reduces raw-data transfer but does not, by itself, resolve communication overhead from repeated exchange of large model parameters or updates. Please briefly state which optimization or orchestration measures IDERHA plans to use, or clearly identify this as ongoing work.
Response: We thank the reviewer for highlighting these important distinctions. As the reviewer suggested, we revised the manuscript to clarify the intended usage of FedProx and FedMeanStd as robust aggregation methods that help to reduce the influence of noisy contributions. Furthermore, as the reviewer suggested, we clarified the discussion of raw DICOM data that stays at their respective federated nodes by introducing and discussing the need of effective communication overhead measures. The following changes have been introduced: Modified paragraphs and highlight changes in green are below: While the compute-to-data paradigm reduces privacy risks by keeping patient-level data within local Federated Nodes, IDERHA also recognizes that federated learning infrastructures introduce several technical challenges. In federated learning, the central node aggregates gradients model updates computed by trained on the client nodes. These updates may vary because of non-IID data or variations in data quality.Some clients may send noisy, low-quality, or even malicious updates that can degrade the global model. To mitigate this challenge rR obust aggregation methods such as FedProx, [29] introducing a penalty term, or FedMeanStd, [30] focusing on the mean and standard deviation of client updates, can be utilized to reduce the influence of noisy or unreliable contributions. Furthermore, personalized federated learning (PFL) has emerged to address the distribution drifts shifts originating from non-IID

client data. PFL aims to learn client-specific models, which that better reflect local data characteristics. [31] The selection of suitable robustness mechanisms for IDERHA's federated learning platform remains part of the ongoing technical development. IDERHA also addresses the the privacy and governance challenges posed by communication overhead associated with large multimodal datasets, particularly imaging data such as DICOM files, by keeping raw data with their respective Federated Node and avoiding any cross-border transfer of any raw data completely. Instead, in federated analysis and learning computational workflows are executed close to the data, and only aggregated summary statistics and model parameters or updates are exchanged. Although this approach avoids the transfer of raw files, it does not eliminate the communication overhead caused by repeated exchanges of large parameter updates. The implementation of communication-efficient measures such as update compression or sparsification, asynchronous communication, optimization of only a subset of model layers (for e.g., in fine-tuning), and less frequent aggregation is subject to ongoing assessment. This approach, combined with data standards (such as OMOP and DICOM) and workflow orchestration, is intended to allow facilitate federated data analysis and improve scalability across geographically distributed servers while preserving data sovereignty and privacy.

Reviewer: 2. OMOP semantic-mapping quality assurance. The response provides useful and concrete information about WhiteRabbit, Usagi, Athena, Achilles, the Data Quality Dashboard, expert review, confidence classification, provenance-aware ETL, and reconciliation procedures. However, this information does not appear in the revised manuscript; the relevant “Data standardization” text remains essentially unchanged. Please add a concise version of this material to the manuscript so that the Open Letter functions as an actionable blueprint for external users.

Response: Thank you for this feedback. The manuscript has been updated accordingly (in track changes and highlighted). Added text is here below (blue). Within IDERHA, a key step in network studies is achieving data interoperability through data standardisation and harmonisation using the OMOP Common Data Model (OMOP-CDM). To ensure semantic interoperability without compromising clinical validity, IDERHA applies a structured Semantic Mapping Quality Assurance Framework that combines automated and semi-automated semantic matching, terminology-based mapping, expert review, structural mapping, ETL transformation, and quality controls. Mapping activities are supported by OHDSI tools including WhiteRabbit, Usagi, Athena, Achilles, and the Data Quality Dashboard (DQD), together with provenance-aware ETL processes and terminology services. Semantic mapping can be performed either automatically when the source data relies on OMOP-supported ontologies such as ICD10, SNOMED, LOINC, etc.; in a semi-automated manner using Usagi, where suggested mappings are reviewed manually on a concept-by-concept basis; or manually using Athena. This process is performed in parallel with structural mapping and is followed by ETL transformation to convert the source data into the target OMOP-CDM representation. An initial quality check compares general counts and other basic data-volume indicators between the transformed and source data to identify potential discrepancies. This is followed by systematic data quality assessment using the DQD, including conformance, completeness, and plausibility checks, as well as further validation through cohort concordance analyses and comparison with source-system representations.

An OMOP readiness assessment then verifies the completeness, semantic consistency,

quality, and analytical suitability of each dataset for the specific study, confirming mapping, ETL, validation, and alignment with network requirements. Discrepancies and remaining harmonisation needs are addressed iteratively through interactions among OMOP mapping experts, clinicians, and database/data-engineering experts, supported by documented decision rules and version-controlled mappings. IDERHA manages this process with a specific focus on the study protocol and the parameters of the relevant IDERHA use case, ensuring that data quality, interoperability, and analytical suitability meet the study-specific requirements across all federated nodes and that the resulting data are “fit for purpose".

Reviewer: 3. Conflicting national legal interpretations and IDAGC resolution. The response states that the manuscript will be revised to explain the IDAGC's role and the flexibility of the federated model, but I could not locate this addition in Version 2. Please add the promised explanation, including the decision process, relevant criteria, documentation of unresolved disagreement, and how participation or analysis is adapted when national requirements cannot be aligned.
Response: Thank you for highlighting this point, and apologies for the lack of clarity in the original wording. We have revised the paragraph to clarify the role of the IDAGC and, importantly, that it does not assess or decide on individual data access requests. Such decisions remain the responsibility of the respective data holders, in accordance with applicable legal and governance requirements. The revised text (added in track changes and yellow) now more clearly reflects the IDAGC’s broader advisory role within the project.
In order to support IDERHA's objectives, IDERHA has established the IDAGC as a multi-stakeholder advisory board bringing together legal, ethical, technical, regulatory, and data-holder perspectives. During the project's lifetime, the IDAGC provides strategic guidance on facilitating the secondary use of health data and strengthening the generation and acceptance of RWE for HTA and regulatory decision-making. However, the IDAGC does not assess or decide on individual data access requests. Such decisions remain the responsibility of the respective data holders, in accordance with applicable legal and governance requirements. Importantly, the federated and decentralized architecture of IDERHA enables data holders to retain local control over their datasets and participate only in activities that are compatible with their institutional and national legal obligations. Where full alignment cannot be achieved, analyses can be adapted to reflect local constraints while maintaining transparency regarding differences in participation and data availability.

Reviewer: Please also correct two typographical errors in the new page 8 text: “statisics” and “signficant.”
Response: Thank you and we have made these corrections (as seen with track changes).

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 24 June 2026

https://doi.org/10.21956/openreseurope.25388.r75888

© 2026 Putra D. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Daniel Happy Putra
Universitas Esa Unggul, West Jakarta, Jakarta, Indonesia

This Open Letter presents a comprehensive conceptual blueprint for the IDERHA (Integration of Heterogeneous Data and Evidence towards Regulatory and HTA Acceptance) project, an Innovative Health Initiative public-private partnership launched in April 2023. The manuscript outlines the framework for a disease-agnostic, pan-European digital health data space designed to align cleanly with the European Health Data Space (EHDS) regulations, GDPR, and the EU AI Act.
To overcome the challenges of severe data fragmentation and administrative silos across European healthcare systems, IDERHA details a decentralized, "privacy-by-design" infrastructure. This infrastructure operates on a compute-to-data paradigm using federated analytics and learning, ensuring that patient-level Real-World Data (RWD) remains strictly at the local node under the control of individual data holders. Semantic interoperability is achieved by mapping disparate clinical datasets to the standard Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).
The paper uses lung cancer as a high-impact clinical use case to demonstrate how multi-modal machine learning models can be trained across this federated architecture to improve early screening risk profiling, malignancy prediction in chest CT scans, and post-treatment patient monitoring via wearables. Finally, it outlines strategies to cultivate policy consensus among regulators, Health Technology Assessment (HTA) bodies, and clinical stakeholders to accelerate the official acceptance of Real-World Evidence (RWE) in public health decision-making.

The article does a good job acknowledging operational and institutional barriers, such as the initial upfront resource investments required for "OMOP-ing" historical hospital datasets and the historical skepticism from HTA bodies regarding non-RCT data. However, it operates heavily on the assumption that federated learning architectures are an absolute panacea for data security. The paper would benefit from explicitly referencing active scientific debates regarding the limitations of federated analysis—such as statistical heterogeneity (non-IID data) across nodes, the potential for privacy leakage through model inversion attacks, or communication overhead friction across cross-border networks.

The manuscript successfully tracks internal project next steps, such as its ongoing multi-methodology framework involving the Integrated Data Access Governance Council (IDAGC) and policy coordination with the TEHDAS2 project. However, because this is an Open Letter meant to act as an ecosystem blueprint, the current recommendations remain highly specific to the internal workflows of the IDERHA consortium. The paper does not yet offer a clearly formalized, universally applicable operational template or step-by-step guideline that external research communities or emerging national data hubs can easily implement to replicate their governance success.

Revision or Input to Author:
While the "compute-to-data" framework protects patient privacy by design, it introduces unique technical vulnerabilities. The authors should briefly add text addressing how the IDERHA technical infrastructure intends to mitigate known federated challenges, such as ensuring model integrity

against malicious/low-quality local data nodes and managing the high communication overhead when handling heavy multimodal image features (like DICOM files) across distant cross-border servers.

Section 3 accurately mentions that transitioning to the OMOP Common Data Model (CDM) can lead to "ambiguous mappings due to local practices". To make this blueprint highly valuable to external researchers, the authors must briefly expand on the specific validation tools or standard operating procedures (SOPs) IDERHA utilizes to identify, measure, and reconcile these local semantic mapping discrepancies without compromising clinical data validity.

While the text highlights using Data Management Plans (DMPs) as a living documentation tool to satisfy the GDPR and the AI Act, it glosses over the fragmented manner in which individual EU Member States interpret data protection laws. The authors should explicitly state how their Integrated Data Access Governance Council (IDAGC) plans to navigate situations where two participating national health nodes hold conflicting legal interpretations regarding the secondary use of specific patient-reported data

Is the rationale for the Open Letter provided in sufficient detail? (Please consider whether existing challenges in the field are outlined clearly and whether the purpose of the letter is explained)
Yes

Does the article adequately reference differing views and opinions?
Partly

Are all factual statements correct, and are statements and arguments made adequately supported by citations?
Yes

Is the Open Letter written in accessible language? (Please consider whether all subject-specific terms, concepts and abbreviations are explained)
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow? (Please consider whether others in the research community would be able to implement guidelines or recommendations and/or constructively engage in the debate)
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My primary areas of expertise and research encompass public health, epidemiology, and medical record and health information management. Specifically, my work focuses on health data systems, data governance frameworks, and the strategic application of real-world evidence (RWE) within epidemiological contexts.Given this background, I am well-equipped to evaluate the public health rationale of the lung cancer screening use case, the ethical

and legal governance frameworks, and the structural aspects of data standardization using the OMOP Common Data Model (CDM). However, I have restricted my assessment to these domains; I have not evaluated the underlying raw machine learning source code, algorithmic optimization mathematics, or the specialized software engineering configurations of the federated architecture, as these fall outside my primary academic focus.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Aug 2026

Rebecca Graebig-Rancourt

The authors' point-by-point responses to reviewer **Putra D.**

COMMENT 1 While the "compute-to-data" framework protects patient privacy by design, it introduces unique technical vulnerabilities. The authors should briefly add text addressing how the IDERHA technical infrastructure intends to mitigate known federated challenges, such as ensuring model integrity against malicious/low-quality local data nodes and managing the high communication overhead when handling heavy multimodal image features (like DICOM files) across distant cross-border servers.

Response: We thank the reviewer for this important comment. We agree that federated analytics and learning are not a complete solution to all privacy, security, and performance challenges. To address this point, we expanded the manuscript to acknowledge key limitations and mitigation strategies associated with federated approaches. Specifically, we noted that IDERHA is investigating mechanisms for quality assurance and model governance, including data quality assessment, monitoring of local node performance, harmonized analytical pipelines, and validation procedures designed to identify inconsistent or potentially problematic local contributions. Furthermore, we acknowledge known privacy risks, such as model inversion and information leakage attacks, and describe the role of technical safeguards and organizational controls within the IDERHA framework. Regarding multimodal data and medical imaging, IDERHA follows a “compute-to-data” paradigm in which large imaging datasets remain at the originating institutions. Rather than transferring DICOM images across borders, model training and inference are performed locally, and only aggregated model updates, parameters, or summary statistics are exchanged. This approach substantially reduces communication burdens while preserving local control of sensitive data. Appropriate optimization and orchestration strategies are being explored to address the communication overhead associated with large-scale federated analyses. The following corresponding paragraphs have been added to the manuscript to explicitly discuss these limitations and mitigation approaches: *“While the compute-to-data paradigm reduces privacy risks by keeping patient-level data within local Federated Nodes, IDERHA also recognizes that federated learning infrastructures introduce several technical challenges. In federated learning, the central node aggregates gradients trained on the client nodes. Some clients may send noisy, low-quality, or even malicious updates that can degrade the global model. To mitigate this challenge robust aggregation methods such as FedProx (Ref),*

introducing a penalty term, or FedMeanStd (Ref), focusing on the mean and standard deviation of client updates, can be utilized. Furthermore, personalized federated learning (PFL) has emerged to address the distribution drifts originating from non-IID client data. PFL aims to learn client-specific models, which better reflect local data characteristics (Ref). IDERHA also addresses the communication overhead associated with large multimodal data, particularly imaging data such as DICOM files, by avoiding cross-border transfer of any raw data completely. Instead, in federated analysis and learning computational workflows are executed close to the data and only aggregated summary statisics and model parameters are exchanged. This approach, combined with data standards (such as OMOP and DICOM) and workflow orchestration, is intended to allow data analysis and improve scalability across geographically distributed servers while preserving data sovereignty and privacy. In addition, IDERHA introduces technical safeguards such as manual code approvals before execution and monitoring for anomalous outputs to reduce the influence of unreliable participants, thereby strengthening the trustworthiness of the platform. However, the human-in-the-loop review process can introduce signficant time delays, which may slow deployment and execution of new model iterations by increasing operational overhead."

COMMENT 2 Section 3 accurately mentions that transitioning to the OMOP Common Data Model (CDM) can lead to "ambiguous mappings due to local practices". To make this blueprint highly valuable to external researchers, the authors must briefly expand on the specific validation tools or standard operating procedures (SOPs) IDERHA utilizes to identify, measure, and reconcile these local semantic mapping discrepancies without compromising clinical data validity.

Response: We thank the reviewer for highlighting the importance of semantic-quality assurance in OMOP CDM transformations. To ensure semantic interoperability without compromising clinical validity, IDERHA applies a structured Semantic Mapping Quality Assurance Framework that combines automated semantic matching, terminology-based mapping, expert review, and federated quality controls. Mapping activities are supported by OHDSI tools including WhiteRabbit, Usagi, Athena, Achilles, and the Data Quality Dashboard, together with provenance-aware ETL processes and terminology services. Candidate mappings are generated and ranked using semantic similarity approaches, reviewed by clinical and informatics experts, and classified according to mapping confidence and semantic equivalence. Validation is performed through data quality assessment (conformance, completeness, and plausibility), cohort concordance analyses, and comparison against source-system representations to ensure preservation of clinical meaning. Discrepancies identified across partner nodes are reconciled through iterative review, version-controlled mapping repositories, and documented decision rules. IDERHA manages these discrepancies, focusing on the study protocol and the specific IDERHA use case parameters, and by ensuring data quality with respect to those use case parameters across all federated nodes. In this way, the data should be "fit for purpose."

COMMENT 3 While the text highlights using Data Management Plans (DMPs) as a living documentation tool to satisfy the GDPR and the AI Act, it glosses over the fragmented manner in which individual EU Member States interpret data protection laws. The authors should explicitly state how their Integrated Data Access Governance Council (IDAGC) plans to navigate situations where two participating national health nodes hold conflicting legal interpretations regarding the secondary use of specific patient-reported data.

Response: We thank the reviewer for raising this important governance consideration. We fully recognize that, despite the harmonizing objectives of the GDPR, national legal interpretations and implementation practices may differ across Member States, creating challenges for cross-border secondary use of health data. To address such situations, IDERHA has established the Integrated Data Access Governance Council (IDAGC) as a multi-stakeholder governance body bringing together legal, ethical, technical, regulatory, and data-holder perspectives. In cases where differing legal interpretations arise between participating nodes, the IDAGC serves as a forum for structured assessment and consensus-building. The Council evaluates the specific use case, applicable national requirements, ethical considerations, and project objectives to identify a compliant and practical path forward. Importantly, the federated and decentralized architecture of IDERHA enables data holders to retain local control over their datasets and participate only in activities that are compatible with their institutional and national legal obligations. Where full alignment cannot be achieved, analyses can be adapted to reflect local constraints while maintaining transparency regarding differences in participation and data availability.

We will revise the manuscript to clarify both the role of the IDAGC in dispute resolution and the flexibility provided by the federated governance model when divergent national interpretations occur.

Competing Interests: No competing interests were disclosed.