BETTER: Better rEal-world healTh-DaTa distributEd analytics Research platform: Project Progress and Emerging Research Challenges

Adrián García[1,*], Diana Martínez-Minguet[1,*], Ana León[1] and Oscar Pastor[1]

[1]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV), 46022, Valencia, Spain

Abstract

Over the last few years, data-driven medicine has gained increasing importance in terms of diagnosis, treatment, and research due to the exponential growth of healthcare data. The linkage of cross-border health data from various sources, including genomics, and analysis via innovative Artificial Intelligence (AI) approaches will allow a better understanding of risk factors, causes, and the development of optimal treatment in different disease areas. However, the reuse of patient data is often limited to data sets available in a single medical center. The main reasons why health data are not shared across institutional boundaries rely on ethical, legal, and privacy aspects and rules. Therefore, in order to (1) enable the sharing of health data across national borders, (2) fully comply with the current GDPR privacy guidelines/regulations, and (3) innovate by pushing research beyond state of the art, the BETTER project proposes a robust decentralized privacy preservation infrastructure which will empower researchers, innovators, and healthcare professionals to exploit the full potential of larger sets of multisource health data through tailored AI tools useful to compare, integrate, and analyze in a secure, cost-effective fashion; with the end goal of supporting the improvement of citizen health outcomes. In detail, this interdisciplinary project proposes the co-creation of three clinical use cases involving seven medical centers located in the EU and beyond, where sensitive patient data, including genomics, are made available and analyzed in a GDPR-compliant mechanism via a Distributed Analytics (DA) paradigm called the Personal Health Train (PHT). The main principle of the PHT is that the analytical task is brought to the data provider (medical center), and the data instances remain in their original location. In this project, two mature implementations of the PHT (PADME and Vantage6), already validated in real-world scenarios, are integrated to build the BETTER platform. At the current stage, BETTER is in an advanced integration and testing phase. Processing pipelines across the clinical use cases are nearing completion, supported by substantial progress in data standardization. PADME is operational and being deployed at clinical centers for final technical tests, while research focuses on AI algorithm selection and multimodal data fusion in local and distributed settings. The next phase will finalize integration, enable real-world data onboarding for federated analytics, and support an initial release of privacy-preserving synthetic datasets reflecting real-world data trends.

Keywords

Health Data, Personal Health Train, Artificial Intelligence, Distributed Analytics, PADME, Vantage6

1. Introduction

1.1. Context and Motivation

Integrating vast arrays of health data from genomics and electronic health records through advanced artificial intelligence (AI) technologies has revolutionized our understanding of diseases, risk factors, and therapeutic strategies [1]. However, the utility of data in medical research is profoundly dependent on its volume and diversity. This is especially true when studying rare diseases and could be solved by sharing data among clinical centers. Nevertheless, sharing patient data across institutions is conditioned by ethical, legal, and privacy concerns [2].

RCIS 2026: Companion Proceedings of the 20th Conference on Research Challenges in Information Science: RCIS Research Projects and Workshops, May 26-29, 2026, Toulouse, France

*Corresponding author.

adgaran1@vrain.upv.es (A. García); dmarmin@vrain.upv.es (D. Martínez-Minguet); aleon@vrain.upv.es (A. León); opastor@dsic.upv.es (O. Pastor)

0009-0004-5103-9770 (A. García); 0009-0002-3191-1969 (D. Martínez-Minguet); 0000-0003-3516-8893 (A. León); 0000-0002-1320-8471 (O. Pastor)

Current data protection regulations, such as the General Data Protection Regulation (GDPR), prohibit data centralization for analysis because of privacy risks, such as the accidental disclosure of personal data to third parties. Overcoming these challenges requires moving from a centralized to a decentralized paradigm that enables the secure and efficient exchange of health information across borders while ensuring compliance with privacy regulations.

This shift entails complex technical challenges, including orchestrating decentralized computation, ensuring secure audibility, and harmonizing heterogeneous clinical and genomic datasets. Addressing these challenges requires a coordinated, long-term effort involving the development, deployment, and validation of federated learning infrastructures that can operate in real-world clinical settings.

This paper extends our previous work presenting the BETTER project [3], where the main focus was on the overall vision of the project, its distributed infrastructure, and the initial deployment strategy. The present version reflects a more advanced project stage, with substantial progress in system integration, data standardization, processing pipeline development, and early experimentation on multimodal data fusion. In addition to reporting this progress, the updated contribution also brings into focus methodological and operational challenges that become more visible as the project moves from conceptual design to real operational deployment. To make this shift explicit, Table 1 summarizes how the emphasis has evolved from 2025 to 2026 across project stage, technical priorities, and validation context.

Dimension	2025 emphasis	2026 emphasis
Project stage	Initial deployment and infrastructure setup	Advanced integration and early analytical validation
Data management	Ethical approvals, FAIR database, and ETL setup	Cross-site standardization progress and pipeline integration
Technical focus	Infrastructure connectivity and platform testing	Operational deployment, multimodal fusion, and genomic feature engineering
Analytical work	Definition of analytical tasks	Experimentation on AI algorithms and local/distributed fusion
Validation context	Synthetic data for controlled testing	Preparation for real-world onboarding and preliminary internal results

Table 1
Evolution of the BETTER project focus from 2025 to 2026.

1.2. Project Approach

The BETTER project[1] proposes a decentralized infrastructure that uses Distributed Analytics (DA) through a mechanism known as the Personal Health Train (PHT) [4]. The PHT model ensures that analytical processes are executed at the data provider's site, allowing data to remain securely within its original location. This model can be illustrated using a railway system analogy, where the key infrastructure components include Trains, Stations, and a Central Service (Figure 1).

Trains encapsulate code to execute analytical tasks at distributed data nodes, known as Stations. As they travel from one Station to another, they process data locally, leveraging the available information at each stop to incrementally build the final analysis result. A Station is a node (institution, hospital, or department) within the distributed architecture that securely stores confidential data and executes Train operations. Each Station functions as an independent and autonomous unit, managing permission requests to control access to its confidential data. The Central Service includes procedures for Train orchestration, operational logic, business logic, data management, and discovery. The Central Service offers: (1) a metadata repository for efficient data discovery; (2) management tools for Train creation, secure transmission to Stations, orchestration, monitoring, and debugging; and (3) a repository of

[1] https://www.better-health-project.eu/

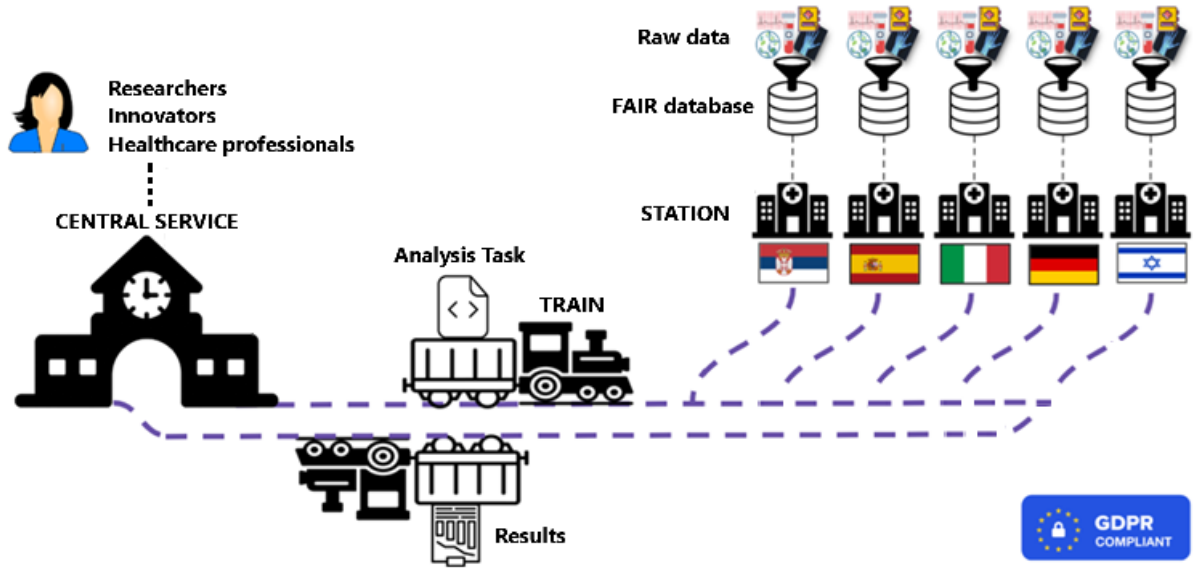

Figure 1: Main infrastructure of the PHT model adapted to the BETTER project.

pre-trained Trains that healthcare professionals can directly apply to their data, enabling them to obtain results from an AI-based method trained iteratively on data from multiple institutions.

At the current stage, the work is structured around three main lines: (1) final coordination for onboarding real-world data across partners; (2) deployment and standardization of processing pipelines across use cases; and (3) validation of multimodal analytical workflows in both local and distributed settings.

1.3. The BETTER Project Technology

The BETTER project integrates two established implementations of the PHT: PADME (Platform for Analytics and Distributed Machine Learning for Enterprises)[2] and Vantage6 (priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange) [5]. Figure 2 shows how these two platforms are connected within the BETTER infrastructure.

Both implementations have already demonstrated their effectiveness in various clinical settings, including oncology [6], diabetes [7], and cardiovascular diseases [8]. In BETTER, they are therefore treated as a mature baseline for execution rather than as an object of conceptual discussion. At the current project stage, the focus is on deployment of the release version at clinical partner centers for end-to-end technical tests, as well as on methodological analyses of multimodal fusion across modalities and distributed stations.

[2]https://padme-analytics.de/

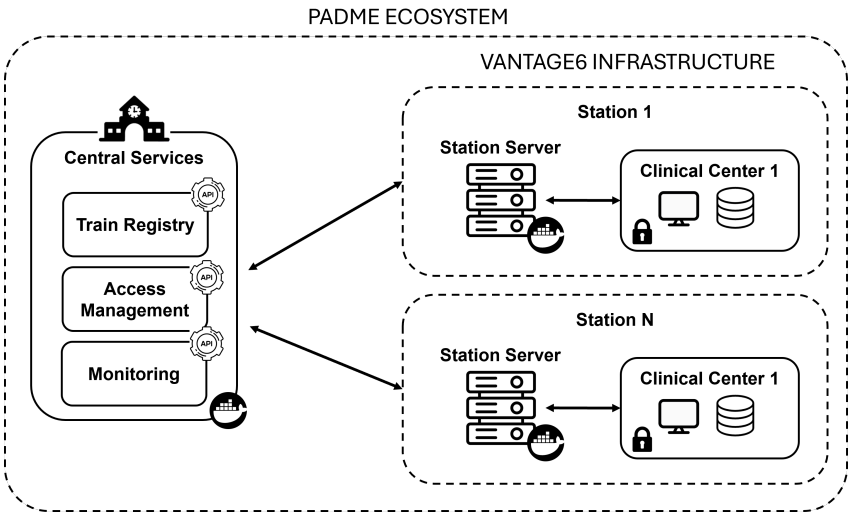

Figure 2: BETTER infrastructure.

1.4. Participants

BETTER is a 42-month Horizon Europe project running from December 2023 to May 2027, with the participation of 14 organizations from eight countries, including seven clinical institutions and seven technological centers:

- *Clinical Institutions*: Klinikum Der Universitaet Zu Koeln (UKK - Germany), Fundació de Recerca Sant Joan de Déu (FDSJD - Spain), Azienda Socio-Sanitaria Territoriale Fatebenefratelli Sacco (BUZZI - Italy), Fundació Docència i Recerca Mutua de Terrassa (Spain), Instituto de Investigación Sanitaria - Hospital Universitario y Politécnico La Fe (Spain), Institut Za Molekularnu Genetiku I Geneticko Inzenjerstvo (IMGGE - Serbia), and Hadassah Medical Organization (HMO - Israel).
- *Technological Centers*: Datrix Spa (Italy), Universiteit Maastricht (UM - Netherlands), Politecnico di Milano (POLIMI - Italy), Universitat Politècnica de València (UPV - Spain), Universitetet i Tromsø - Norges Arktiske Universitet (UiT - Norway), Rheasoft ApS (Denmark), and Noosware Bv (Netherlands).

1.5. Clinical Use Cases

The project aims to apply these innovative DA methodologies to three clinical use cases: Pediatric Intellectual Disability, Inherited Retinal Dystrophies, and Autism Spectrum Disorders. The overarching goal is to harness the full potential of multisource health data, enabling researchers, healthcare professionals, and innovators to conduct comprehensive analyses, integrate disparate data types, and derive meaningful insights securely and cost-effectively.

Across the three use cases, seven medical centers will integrate, validate, and use the digital tools developed within the BETTER platform, enabling the fusion and analysis of heterogeneous data sources from different centers to improve clinical outcomes.

1.5.1. Use Case 1: Integration of Genomic and Phenotypic Data from Pediatric Rare Diseases to Decipher Pathways of Intellectual Disability

Intellectual disability (ID) is a common disorder characterized by significant limitations of cognitive functions and adaptive behavior, with onset before age 18. This use case aims to (1) evaluate and correlate the phenotypic, genomic, multi-omic, and clinical parameters between early-diagnosed and later-diagnosed patients; (2) Improve diagnosis by identifying new genetic biomarkers that can be used

in newborn screening protocols; (3) Develop new tools based on Digital Twins Model to define new diagnostic biomarkers, pathways, and therapeutic molecular targets.

To this end, clinical data, brain images, genomic data (whole-exome and whole-genome sequences), and biological data (cellular and molecular pathways) will be integrated. The participants in this use case are the Hospital Sant Joan de Deu (medical leader), IMGGE, the Children's Hospital Vittore Buzzi, and the Politecnico di Milano (technological leader).

1.5.2. Use Case 2: Accelerate Inherited Retinal Dystrophies Diagnosis using AI

Inherited Retinal Diseases (IRDs) are a group of disorders characterized by the generally progressive death or dysfunction of photoreceptors and retinal pigment epithelium (RPE) cells, leading to loss of visual function, sometimes leading to legal blindness. An early molecular diagnosis is necessary to confirm the clinical diagnosis and offer adequate care to patients. In addition, developing new genetic analysis tools that allow the precise identification of the molecular cause of disease is essential to improve the understanding of the pathophysiological mechanisms at the base of the symptoms and open the doors to future therapies. This study aims to (1) identify pathogenic genes and variants responsible for the IRDs, and (2) define existing genotype-phenotype correlations to better understand the prognosis of patients and improve their clinical management.

To this aim, genomic data (gene panels, clinical exome, whole exome, whole genome), clinical reports, and images will be integrated. The participants in this Use Case are the Hospital Universitario La Fe de Valencia (medical leader), the Hadassah Medical Center, and the Polytechnic University of Valencia (technological leader).

1.5.3. Use Case 3: Predicting the Risk of Self-Harm and Suicidal Behaviors in Patients with Autism Spectrum Disorders

Autism Spectrum Disorders (ASD) are neurodevelopmental disabilities characterized by social, communication, and behavioral challenges. Children and adolescents with ASD are at a substantially higher risk of self-injurious and suicidal behavior compared to the general population (up to 9 times). However, the causes of this increased risk remain largely unknown, and there is little knowledge about the potential role of phenotypic, metabolic, genomic, and environmental factors. This use case aims to (1) identify predictive phenotypic, genomic, and environmental risk factors of suicidality and self-injury in ASD individuals; (2) Personalise prevention intervention plans to reduce self-injury and suicidality in each ASD individual; and (3) Develop monitoring strategies to recognize signs of vulnerability in ASD individuals that will lead to prevent strategies at an earlier stage and thereby further reduce risk of self-harm and suicidal intentions.

To this aim, clinical, metabolic, environmental and demographic data, patient interviews, and genomic data (epigenome and whole genome sequencing) will be integrated. The participants in this Use Case are Hospital Universitario Mutua Terrassa (medical leader), Children's Hospital Vittore Buzzi, and the Klinikum Der Universitaet Zu Koeln (UKK). UKK also participates as the technological leader.

2. Project Objectives

The BETTER project pursues five main objectives: (1) overcome cross-border barriers to health data integration, access, FAIRification, and preprocessing; (2) ensure health data fusion and integration; (3) deploy a distributed analytics framework for cross-border data processing and analysis; (4) develop distributed tools that leverage artificial intelligence capabilities; and (5) incorporate ethical, legal, and societal aspects (ELSA) throughout the AI lifecycle. Together, these objectives align with core RCIS research topics, particularly data and information management, distributed information infrastructures, and AI-enabled analytics for healthcare information systems.

2.1. Objective 1: Overcome Cross-Border Barriers to Health Data Integration, Access, FAIRification, and Preprocessing

The main aim of this first objective is to guide medical centers in collecting patients' data following a common schema to promote interoperability and the reuse of datasets in scope. This includes collecting legal, ethical, and data protection authorizations and using well-established and widely understood ontologies. Data pseudonymization will be performed as a default preprocessing step to mitigate the risk of personal data leaks. Finally, a BETTER station will be installed at each medical center, ensuring access to the relevant local datasets.

2.2. Objective 2: Ensure Health Data Fusion and Integration

To gain the maximum from data, one of the important steps is integrating multiple data sources to produce more consistent, accurate, and useful information than any single data source. The ambition is to fuse several dimensions, including laboratory analysis, medical reports, drug therapy, imaging, genomics, socio-demographic, geographical, and medical questionnaires.

BETTER uses standardized ontologies (e.g., NCIT, LOINC, ICD-11) and a shared metadata schema to ensure semantic alignment of data across sites. This enables data fusion and integration using the proposed distributed framework in two directions: within a single medical center (local data fusion) and across centers (distributed data fusion) by leveraging each other's historical datasets. Local data fusion involves integrating data from multiple sources within a single institution. This type of data fusion is useful when the data sources are heterogeneous, such as genomic, clinical, and phenotypic data of the same patient. Distributed data fusion integrates data from multiple institutions, a fairly novel discipline that removes potential biases due to different collection protocols or techniques.

2.3. Objective 3: Deploy a Distributed Analytics Framework for Cross-Border Data Processing and Analysis

The ambition of this objective regards the deployment of BETTER, a privacy-by-design infrastructure, to all medical centers connecting FAIR data sources and allowing federated data analysis and machine learning. To effectively exploit multiple datasets via AI, a common schema and ontology should be applied.

2.4. Objective 4: Development of Distributed Tools Leveraging Artificial Intelligence Capabilities

To properly answer clinical needs and push data analysis boundaries beyond state-of-the-art, tailored tools must be developed to exploit DA and AI within each use case. The tools are developed using a co-creation methodology where medical end-users closely collaborate with researchers and technology providers, enabling the development of new concepts. In this context, current efforts focus on selecting AI algorithms and genomic feature sets, and on advancing multimodal data fusion strategies that combine imaging and tabular data in both local and distributed settings.

2.5. Objective 5: Include Ethical, Legal and Societal Aspects (ELSA) in the AI Lifecycle

Most data science projects do not co-create or co-develop using a methodology that includes the ethical, legal, and societal aspects (ELSA) involved in the data science lifecycle. In this objective, the BETTER project will develop ELSA-awareness tools and methods for co-creating and co-developing AI models and apply them to the proposed use cases. This will ensure the appropriateness and clinical effectiveness of the developed AI tools while considering the safety, value, and sustainability of the AI.

3. Impact and Expected Outcomes

Overcoming the current barriers to data sharing and utilization, the BETTER project opens the way to more accurate diagnoses, tailored treatments, and a deeper understanding of complex diseases. The project's target groups are healthcare professionals, researchers, innovators, health policymakers, and citizens.

BETTER promotes a hands-on and experience-building approach towards the implementation of cross-border data-sharing partnerships in the area of real-world health data. This contribution paves the way for a European medical center data sharing and analysis network. Preliminary internal experiments suggest that local multimodal data fusion can provide better predictive performance than single-modality approaches, reinforcing the value of integrated analytics in the project use cases. The expected outcomes for the BETTER project are:

- A public release of the platform implementation, which will reinforce two open-source projects, namely PADME and Vantage6.
- Publication of the FAIRification pipelines, data catalogs, and ontologies to unleash the potential of data exchange and reuse.
- Release of synthetic datasets to the community. This is particularly valuable for developing and benchmarking models in data-restricted scenarios.
- Finally, cross-border health data secure exchange and reuse require a solid and compliant legal, ethical, and data protection framework. By enhancing existing templates, BETTER will consolidate and publish a documentation folder useful and applicable for future initiatives.

4. Current Project Status and Future Work

4.1. Milestones

The BETTER project is currently in an advanced integration and testing phase. The following milestones have been achieved or are close to completion:

- Final discussions are ongoing to complete agreements for onboarding real-world data into the platform.
- Processing pipelines and standardization for the three clinical use cases are reaching integration closure.
- The exploitation of data, particularly genomic data, is under active methodological study across use cases.
- Data fusion research is underway both locally and across nodes, with current experiments combining imaging and clinical tabular data.
- The release version of PADME is finalized and currently being deployed at clinical partners for final technical tests.
- The web interface for platform navigation and visualization of onboarded datasets is close to completion and under internal validation.
- Preparation for the final synthetic data release is ongoing, with the aim of generating privacy-preserving datasets that reflect real project data trends.

Current research activities focus on AI algorithm selection, genomic feature engineering, and multi-modal fusion under local and distributed settings, including IID and non-IID assumptions. In parallel, the effects of class imbalance are being explored. Preliminary internal results indicate that local fusion settings can outperform separate single-modality pipelines.

4.2. Emerging Research Challenges

As BETTER moves from infrastructure deployment towards integration and early analytical validation, several research challenges are becoming more clearly visible.

One important challenge concerns distributed data fusion in scenarios where participating centers expose heterogeneous feature spaces and only partially overlapping class distributions. In such settings, combining multimodal information across sites is not only a technical interoperability issue, but also a methodological challenge for robust and clinically meaningful model design.

Emerging research challenges also concern the effective exploitation of data within the analytical workflows of the project. Beyond data processing, an important open question is how to define meaningful feature engineering strategies that allow information sources, such as genomic data, to contribute to the clinical objectives of each use case. This is particularly challenging because the added value of genomic data is not always directly observable and may instead emerge only when informative features are properly identified or when it is combined with other sources of information, such as clinical, phenotypic, or imaging data. At the same time, these methodological questions must be addressed under the operational constraints of privacy-preserving distributed infrastructures, where data access, computational demands, particularly in genomic data exploitation, and delays in data-sharing agreements directly affect the timing and scope of real-world validation. Altogether, these aspects show that, beyond platform development itself, deploying federated analytics in cross-border healthcare settings requires addressing tightly connected challenges in data representation, multimodal integration, methodological robustness, and operational feasibility.

Alongside these research challenges, the project also faces operational risks related to coordination and deployment across participating sites. These include potential delays in data-sharing agreements, infrastructure-related constraints, and the need to ensure stable and robust execution of data fusion workflows in real-world settings. The consortium is currently evaluating mitigation strategies, while technical solutions for these issues remain under active investigation.

In the next project phase, BETTER will complete pipeline integration, onboard and process real-world data through the federated infrastructure, and operationalize AI trains across clinical nodes using the PADME release setup. In parallel, the consortium will finalize the dataset navigation interface and prepare an initial synthetic data release that captures clinically relevant trends while preserving privacy.

5. Conclusion

The BETTER project aims to provide a decentralized paradigm to enable the secure and efficient exchange of health information across borders while ensuring compliance with privacy regulations. Using the Personal Health Train model implemented by two open implementations (PADME and Vantage6), the project will explore the feasibility of federated learning in three clinical use cases (Pediatric Intellectual Disability, Inherited Retinal Dystrophies, and Autism Spectrum Disorders). BETTER is a 42-month project financed by the European Union's Horizon Europe research and innovation program, with 14 participants from eight countries, including seven clinical institutions and seven technological centers. The overarching goal is to harness the full potential of multisource health data, enabling researchers, healthcare professionals, and innovators to conduct comprehensive analyses, integrate disparate data types, and derive meaningful insights securely and cost-effectively. At its current stage, the project not only demonstrates substantial progress in integration and deployment, but also exposes methodological and operational challenges that are central to the future development of federated analytics in cross-border healthcare settings.

Acknowledgments

This work is part of the Horizon Europe project BETTER. The BETTER project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement

No. 101136262. https://www.better-health-project.eu/.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] S. Welten, L. Neumann, Y. U. Yediel, L. O. B. da Silva Santos, S. Decker, O. Beyan, Dams: A distributed analytics metadata schema, Data Intelligence 3 (2021) 528–547.

[2] T. M. Deist, F. J. Dankers, P. Ojha, M. S. Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, et al., Distributed learning on 20 000+ lung cancer patients–the personal health train, Radiotherapy and Oncology 144 (2020) 189–200.

[3] A. Palacio, J. Reyes Román, O. Pastor, Better: Better real-world health-data distributed analytics research platform, Ceur Workshop Proceedings 3987 (2025).

[4] O. Beyan, A. Choudhury, J. van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, M. R. Karim, M. Dumontier, S. Decker, L. O. B. da Silva Santos, A. Dekker, Distributed Analytics on Sensitive Medical Data: The Personal Health Train, Data Intelligence 2 (2020) 96–107. doi:10.1162/dint_a_00032.

[5] A. Moncada-Torres, F. Martin, M. Sieswerda, J. Van Soest, G. Geleijnse, Vantage6: an open source privacy preserving federated learning infrastructure for secure insight exchange, in: AMIA annual symposium proceedings, volume 2020, 2021, p. 870.

[6] S. Theophanous, P.-I. Lønne, A. Choudhury, M. Berbee, A. Dekker, K. Dennis, A. Dewdney, M. A. Gambacorta, A. Gilbert, M. G. Guren, et al., Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomcat2 study, Diagnostic and prognostic research 6 (2022) 14.

[7] C. Sun, J. van Soest, A. Koster, S. J. Eussen, M. T. Schram, C. D. Stehouwer, P. C. Dagnelie, M. Dumontier, Studying the association of diabetes and healthcare cost on distributed data from the maastricht study and statistics netherlands using a privacy-preserving federated learning infrastructure, Journal of Biomedical Informatics 134 (2022) 104194.

[8] B. Scheenstra, A. Bruninx, F. van Daalen, N. Stahl, E. Latuapon, M. Imkamp, L. Ippel, S. Duijsings-Mahangi, D. Smits, D. Townend, et al., Digital health solutions to reduce the burden of atherosclerotic cardiovascular disease proposed by the carrier consortium, JMIR cardio 6 (2022) e37437.