# Building Trust in Synthetic Health Data: Methodological and Acceptance Challenges across Seven European Projects

Jens Declerck, Dipak Kalra, Antti Airola, Ahmed Youssef Ali Amer, Christos Chatzichristos, Maria del Mar Mañu, Bruno M. de Brito Robalo, Francesco Ghini, Alberto Gutierrez-Torre, Sem Hoogteijling, Susanne Hultsch, Jan Ramon, Sara Reidel, Francesco Regazzoni, Luís Silva, Inês Silveira, Christophe Maes

Submitted to: Journal of Medical Internet Research
on: September 01, 2025

# *Table of Contents*

# Building Trust in Synthetic Health Data: Methodological and Acceptance Challenges across Seven European Projects

Jens Declerck[1, 2] MSc; Dipak Kalra[1] Prof Dr; Antti Airola[3]; Ahmed Youssef Ali Amer[4]; Christos Chatzichristos[5]; Maria del Mar Mañu[6]; Bruno M. de Brito Robalo[7]; Francesco Ghini[8]; Alberto Gutierrez-Torre[9]; Sem Hoogteijling[10]; Susanne Hultsch[11]; Jan Ramon[12]; Sara Reidel[13, 14]; Francesco Regazzoni[15]; Luís Silva[16]; Inês Silveira[16]; Christophe Maes[1, 2]

[1] The European Institute for Innovation through Health Data Oosterzele BE
[2]Unit of Medical Informatics and Statistics Department of Public Health and Primary Care The University of Ghent Ghent BE
[3] Department of Computing University of Turku Turku FI
[4] Innovative Medicine Johnson & Johnson Beerse BE
[5] Department of Electrical Engineering KU Leuven Leuven BE
[6] University Hospital Vall d'Hebron Barcelona ES
[7]Department of Obstetrics and Gynaecology Erasmus MC University Medical Center Rotterdam NL
[8] Laboratory of Translational Research Reggio Emilia IT
[9] Supercomputing Center Barcelona ES
[10] Department of Neurology and Neurosurgery University Medical Center Utrecht Brain Center Utrecht NL
[11] VEIL.AI Helsinki FI
[12] INRIA Lille FR
[13] Vall d'Hebron Research Institute University Hospital Vall d'Hebron Barcelona ES
[14] BarcelonaTech (UPC) Universitat Politècnica de Catalunya Barcelona ES
[15] University of Amsterdam Amsterdam NL
[16] LIBPhys-UNL NOVA School of Science and Technology Caparica PT

**Corresponding Author:**
Jens Declerck MSc

The European Institute for Innovation through Health Data
Merebaaistraat 10
Oosterzele
BE



## *Abstract*

**Background:** Synthetic data generation (SDG) is increasingly used in health research to address privacy concerns and data sharing barriers, particularly in cross-border and multi-institutional contexts. Despite growing interest, questions remain around the methodological soundness, legal status, and real-world trustworthiness of synthetic health data.

**Objective:** This paper examines how SDG is operationalized across seven European projects in the HealthData4EU cluster. It identifies common methodological, compliance, and acceptance challenges, and highlights emerging strategies to ensure the quality and utility of synthetic health data.

**Methods:** A qualitative multiple case study approach was used, combining workshop presentations with structured input from each project. Projects were analyzed across four aspects: methodological challenges, data quality assurance, trust and transparency strategies, and contributions to the wider synthetic data ecosystem.

**Results:** Results reveal recurring challenges across projects: tension between privacy and utility, lack of shared quality validation standards, institutional readiness gaps, and legal ambiguity under GDPR. Projects employ varied strategies, including federated machine learning and validation pipelines, differential privacy, co-design processes, and metadata transparency tools, but implementation remains uneven. The absence of harmonised legal and evaluation frameworks continues to hinder broader uptake.

**Conclusions:** To scale SDG from isolated pilots to trusted European infrastructure, four priorities are critical: shared evaluation metrics, legal clarity, sustainable platforms, and embedded stakeholder engagement. While rooted in specific projects, the

insights offer broader relevance for global SDG efforts, as the core tensions identified are likely to recur across health systems.

(JMIR Preprints 01/09/2025:83369)
DOI: https://doi.org/10.2196/preprints.83369

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

# Original Manuscript

# Building Trust in Synthetic Health Data: Methodological and Acceptance Challenges across Seven European Projects

Abstract

**Background:** Synthetic data generation (SDG) is increasingly used in health research to address privacy concerns and data sharing barriers, particularly in cross-border and multi-institutional contexts. Despite growing interest, questions remain around the methodological soundness, legal status, and real-world trustworthiness of synthetic health data.

**Objective:** This paper examines how SDG is operationalized across seven European projects in the HealthData4EU cluster. It identifies common methodological, compliance, and acceptance challenges, and highlights emerging strategies to ensure the quality and utility of synthetic health data.

**Methods:** A qualitative multiple case study approach was used, combining workshop presentations with structured input from each project. Projects were analyzed across four aspects: methodological challenges, data quality assurance, trust and transparency strategies, and contributions to the wider synthetic data ecosystem.

**Results:** Results reveal recurring challenges across projects: tension between privacy and utility, lack of shared quality validation standards, institutional readiness gaps, and legal ambiguity under GDPR. Projects employ varied strategies, including federated machine learning and validation pipelines, differential privacy, co-design processes, and metadata transparency tools, but implementation remains uneven. The absence of harmonised legal and evaluation frameworks continues to hinder broader uptake.

**Conclusions:** To scale SDG from isolated pilots to trusted European infrastructure, four priorities are critical: shared evaluation metrics, legal clarity, sustainable platforms, and embedded stakeholder engagement. While rooted in specific projects, the insights offer broader relevance for global SDG efforts, as the core tensions identified are likely to recur across health systems.

Introduction

As healthcare systems are largely digitized and become increasingly data-driven, access to large-scale, high-quality data is essential for clinical research [1-3], digital health innovation [4], and evidence-based policymaking [5]. However, accessing patient-level health data remains a persistent challenge [6]. Privacy regulations [7], data quality issues [8, 9], and data fragmentation [10] continue to limit data access and sharing, especially in cross-organizational studies [9, 11]. These barriers not only slow innovation but also risk underrepresenting certain diseases and populations, such as rare conditions, in secondary data use.

Synthetic data generation (SDG) has emerged as a promising response to the barriers limiting access and reuse of patient-level health data [12]. By producing synthetic datasets that resemble the statistical and structural properties of real health data, without exposing personal identities, SDG can drive innovation while maintaining privacy. It enables safer collaboration across different types of health data (e.g. electronic health records, genomics, and medical imaging) and offers value in areas such as AI development [7, 13], rare disease research [14], and model testing [13]. The main aim of SDG is to retain the essential characteristics of original datasets, while retaining privacy (in most use cases). These can be grouped into quantitative [15] (e.g., statistical metrics, analytical patterns, signal-based features), qualitative [15] (e.g., expert-driven assessments), and domain-specific attributes [16] that depend on the data type and specific use case. Once identified, these characteristics guide the validation methods and evaluation metrics applied to the synthetic data.

SDG proves useful when real-world data is limited or imbalanced, such as in rare diseases or underrepresented subpopulations, making it valuable for machine learning training [17]. Moreover, it offers a path to privacy-preserving data reuse by removing direct and indirect identifiers while maintaining health data utility [7, 17]. But still, challenges remain, specific in ensuring fidelity [18] (e.g.,

statistical similarity), utility [18] (e.g., task performance), privacy [18] (e.g., how well it protects against the disclosure of sensitive or identifiable information) and diversity [19] (e.g., variation and representativeness) is complex. Generative models must be tailored to specific needs, and robust evaluation frameworks are essential to determine when synthetic data is truly fit for purpose. Beyond the technical aspects, trust in synthetic data also requires transparency, legal clarity (e.g., GDPR, AI Act), and engagement with clinicians, patients, and other stakeholders. Without this broader trust, adoption will remain limited among the healthcare stakeholder community.

Currently, several EU-funded projects are now testing SDG in real-world contexts, each facing a common question: “how can we ensure synthetic data is trustworthy, useful, and safe?” Despite growing interest, there has been no cross-project analysis of how these initiatives are tackling shared challenges. This paper addresses that gap by analysing seven EU-funded synthetic data projects. Through cross-case comparison, we examine how each initiative approaches methodological rigor, data quality, and stakeholder trust, and what this means for the future of synthetic data in European health research.

## Methodology

### *Research design*

This study employed a qualitative multiple case study design to investigate the methodological and acceptance challenges associated with SDG in the health domain. A case study approach is well suited to capture the contextual complexity and domain-specific nuances of SDG practices, which are highly sensitive to clinical requirements, technical infrastructure, and regulatory environments. This study design supports both in-depth understanding of individual initiatives and cross-care synthesis of emerging patterns, thereby offering insights into both project-level and system-level dynamics.

The choice of a multiple case study approach is particularly relevant given the novelty of the SDG field. At present, there is no consolidated body of good practices in SDG methodologies, governance, quality assessment, or utilisation. This emerging context positions case studies as a valuable means to examine the diversity of current strategies and practices, as well as to highlight early patterns of convergence. The analysis presented here does not attempt to define best practices but instead draws attention to potential consensus methods that may, over time, form the basis of future standards or guidance.

### *Case selection*

This study is based on seven EU-funded projects that constitute the HealthData4EU cluster, a Horizon Europe initiative focused on advancing synthetic data generation and secure data use in health research across Europe. Together, these projects cover complementary areas of expertise (e.g., SDG generation, federated learning, secure data sharing, and AI-driven diagnostics), forming an ecosystem that supports the development of a trustworthy, inclusive, and innovation-friendly European Health Data Space (EHDS).

These seven projects represent the most prominent synthetic data initiatives currently funded by the European Commission under Horizon Europe and the Innovative Health Initiative. All addressing shared challenges around data quality, privacy, interoperability, and real-world application of synthetic health data. While coordinated under a common EU framework, they span diverse clinical domains, data modalities (e.g., tabular, imaging, longitudinal), and technical approaches. This

diversity within a coordinated structure provides a unique opportunity to compare how SDG is being adapted across healthcare settings. And to identify emerging patterns that could inform future adoption within the EHDS.

At the time of analysis, all projects were still in progress. By including the entire HealthData4EU cluster, the study minimizes selection bias and provides a balanced overview of the European synthetic health data landscape. Figure 1 illustrates the variation in project timelines and marks the two key data collection points that structured the analysis.

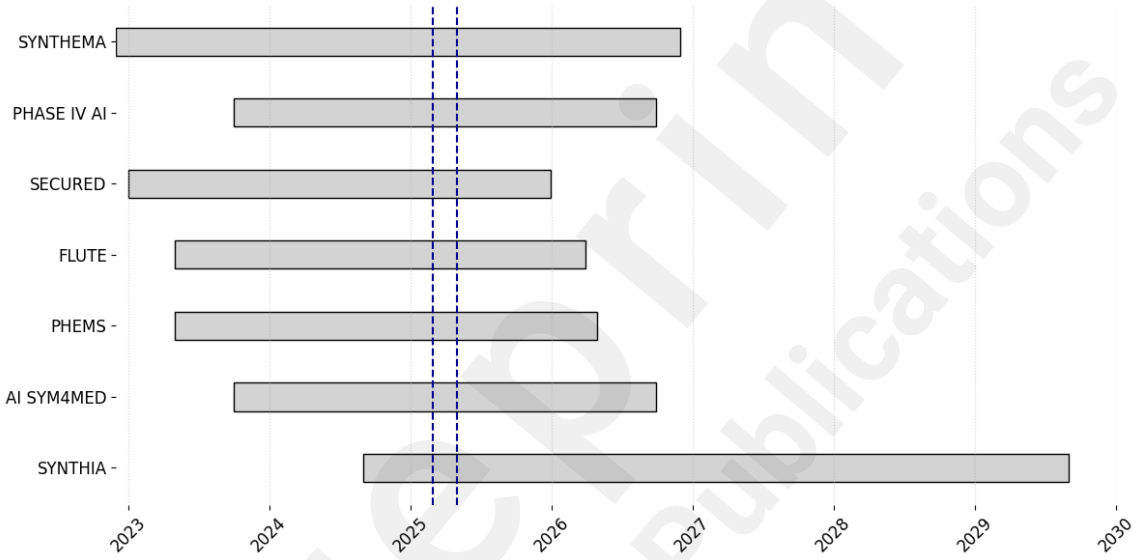


Figure 1: This study employed a qualitative multiple case study design to investigate the methodological and acceptance challenges associated with synthetic data generation (SDG) in the health domain

The projects analysed were: AISym4MED (*Synthetic and Scalable Data Platform for Medical Empowered AI*), FLUTE (*Federated Learning and Multi-party Computation Techniques for Prostate Cancer),* PHEMS (Paediatric *Hospitals as European Drivers for Multi-party Computation and Synthetic Data Generation Capabilities Across Clinical Specialities and Data Types*), PHASE IV AI (*Privacy Compliant Health Data as a Service for AI Development*), SECURED (*Scaling Up Secure Processing, Anonymization, and Generation of Health Data for EU Cross-Border Collaborative Research and Innovation*), SYNTHIA (*Synthetic Data Generation Framework for Integrated Validation of Use Cases and AI Healthcare Applications*), SYNTHEMA (*Synthetic Generation of Haematological Data Over Federated Computing Frameworks*).

These projects represent a broad spectrum of clinical domains, data types (e.g., tabular, imaging, signal, multimodal), and methodological approaches to synthetic data generation. Table 1

summarizes the main clinical domains and use cases addressed by each project.

| Project | Clinical domain(s) | Example use cases |
|---|---|---|
| AISym4MED | Neurological and chronic diseases (i.e., diffuse large B-cell lymphoma, lung cancer, multiple myeloma, Alzheimer's disease, type II diabetes, breast cancer). | Synthetic data for multimodal applications. Support for AI model testing. |
| FLUTE | Oncology (prostate cancer) | Risk stratification. Federated AI validation for prostate cancer diagnosis. |
| PHASE IV AI | Oncology and neurology (i.e., lung cancer, prostate cancer, ischaemic stroke). | Early lung cancer detection. Synthetic MRI for stroke and model testing. |
| PHEMS | Paediatric diseases (i.e., congenital cardiac conditions, sepsis, haemophilia). | Sepsis prediction in paediatric ICUs. Haemophilia care optimization. |
| SECURED | Multiple domains (i.e., mammography, histopathology, chest X-ray, cardiotocography). | Privacy-preserving SDG. Federated AI tool development. |
| SYNTHEMA | Rare haematological diseases (i.e., sickle-cell disease, acute myeloid leukaemia). | Synthetic datasets for rare disease research. Federated validation. |
| SYNTHIA | Multimodal and personalized medicine. | Benchmarking and validation of SDG methods across diverse clinical scenarios. |

Table 1 Overview of projects, clinical domains, and example use cases involving synthetic data

*Analytical framework*

To structure the analysis, a comparative framework was developed that reflects the central themes of the research question and enables systematic exploration of project-level differences and similarities. The framework comprises four core thematic dimensions: methodological challenges, data quality, stakeholder trust, and ecosystem contribution. First, it examines the specific technical, clinical, or infrastructural problems that each project seeks to address through SDG. Second, it investigates how data quality is defined, operationalized, and evaluated, with attention to dimensions such as fidelity, utility, completeness, and bias mitigation. Third, it explores the strategies for building stakeholder trust and acceptance, including stakeholder engagement, transparency measures, and ethical or regulatory compliance. Finally, it highlights broader contributions to the synthetic data ecosystem,

such as the development of tools, standards, validation methods, or collaborative infrastructures. Each project was analysed using a comparative matrix comprising 22 variables aligned with these four dimensions, enabling both granular project-level analysis and cross-case thematic synthesis.

*Data sources and collection*

Data collection followed a two-stage process. First, initial information was gathered during a dedicated HealthData4EU workshop held at the Health Data Summit in March 2025 in Brussels, where all seven projects formally presented their objectives, methods, and early outcomes. This event provided a structured baseline for cross-project comparison. In the second phase, individual consultations were conducted in May 2025 with representatives from each project. These consultations were used to validate and supplement the initial information, clarify methodological details, and complete a standardized comparative template comprising 22 variables. This process ensured the completeness and consistency of the dataset, while also enabling a nuanced, project-level analysis.

## Results

The findings are presented as a cross-case thematic synthesis that draws on structured data collected from the seven projects. Rather than describing each project individually, the results are organised according to the comparative framework introduced in the methods section. A more detailed breakdown of project responses, which forms the basis for this synthesis, is provided in Appendix I.

*Methodological challenges addressed*

Across the seven HealthData4EU projects, methodological challenges in SDG are closely tied to the specific goals each project pursues. Projects with similar aims tend to encounter overlapping technical, infrastructural, or regulatory barriers. This section groups the projects by shared objectives and highlights the methodological challenges that emerge from these common ambitions.

***Objective: Enabling privacy-preserving access to distributed health data***

Projects such as FLUTE, PHASE IV AI, PHEMS, SECURED, and SYNTHEMA, aim to support synthetic data generation in decentralized or federated infrastructures, where sensitive health data

must remain locally stored. These projects work with diverse data types, including MRI imaging, EHRs, time-series signals, and OMICs data, and apply generative techniques such as generative adversarial networks (GANs), variational auto encoders (VAEs), diffusion models, marginal-based methods for tabular data, and instance-based methods depending on the modality.

The primary methodological challenge in this group involves building interoperable, privacy-preserving pipelines that can function across heterogeneous hospital systems and national jurisdictions. This includes harmonizing data structures, managing computational constraints in federated training, and validating synthetic data quality without centralizing real data. To address this, projects are experimenting with alternative strategies such as proxy benchmarking and federated validation frameworks. Table 2 summarises the data modalities addressed by each project and the corresponding methodological challenges.

| Project | Data modalities used | Main methodological challenges |
|---|---|---|
| AISym4MED | Imaging, genomics, multimodal data. | Rare disease data sparsity.<br>Validation of multimodal outputs. |
| FLUTE | EHR, imaging. | Federated training across countries.<br>Interoperability of hospital IT systems. |
| PHASE IV AI | EHR, imaging. | Synthetic imaging (CT, MRI).<br>Ensuring privacy in decentralized infrastructures. |
| PHEMS | EHR, imaging, time-series data. | Cross-border federated pipelines.<br>Variation in data models across sites. |
| SECURED | Imaging, time-series data, tabular data. | Secure multiparty computation.<br>Integration into heterogeneous infrastructures. |
| SYNTHEMA | OMICs, imaging, tabular data. | Small cohort size.<br>Data scarcity and fragmentation and validation framework. |
| SYNTHIA | EHR, imaging, m-health. | Benchmarking and validation across diverse modalities. |

Table 2 Data modalities and methodological challenges across projects

***Objective: Generating complex and multi-modal synthetic data in prevalent or rare disease synthetic data***

Projects like AISym4MED, PHASE IV AI, SECURED, SYNTHIA, and SYNTHEMA focus on generating synthetic data from complex, high-dimensional sources including histopathological images, genomics, metabolomics, time-series, and clinical text. These efforts often place particular emphasis on underrepresented or clinically challenging contexts. For example, AI SYM4MED addresses rare conditions, while PHASE IV AI explores synthetic imaging modalities including CT for lung cancer and MRI for ischemic stroke. PHEMS, with its focus on paediatric diseases, also contributes to this space and includes a proof-of-concept for medical image anonymization.

These initiatives require modality-specific architectures, such as segmentation-based models for cytological imaging, VAEs for genomic data, and feature-level generators for radiomics. Common challenges include data sparsity, particularly in rare disease contexts, the technical complexity of aligning multiple data modalities into coherent synthetic records, and the lack of shared standards for evaluating the realism and consistency of multi-modal synthetic outputs. In many cases, expert input is often required to validate cross-modal fidelity and ensure clinical plausibility.

***Objective: Building trust, usability, and transparency in synthetic data ecosystems***

Projects including AISym4MED, PHASE IV AI, PHEMS, SYNTHEMA and SYNTHIA emphasize the development of synthetic data infrastructures that are ethically aligned, transparent, and user centred. Their methods extend beyond generation, incorporating bias detection, privacy-preserving mechanisms, and explainability tools to build stakeholder trust. Some projects only cover part of these, e.g., FLUTE covers generation, privacy preservation, regulatory compliance and transparency.

Challenges in this area revolve around the absence of standardised quality metrics, the need to communicate the reliability and limitations of synthetic datasets to clinicians, researchers, and regulators, and persistent regulatory ambiguity. Ensuring transparency across the entire lifecycle of data synthesis (e.g., metadata, model documentation, and interpretability frameworks) is seen as essential for real-world acceptance. One cross-cutting issue across all projects is the legal status of synthetic data. While synthetic datasets are often assumed to fall outside the scope of GDPR, project teams report varied interpretations by data controllers, ethics boards, and national authorities. This lack of harmonized guidance creates uncertainty around the use, sharing, and reuse of synthetic health data. Clarifying the legal boundaries, especially when and how synthetic data can be considered fully anonymized, would reduce institutional friction and accelerate adoption across clinical and research contexts.

***Objective: Developing scalable, reusable synthetic data platforms and standards***

Projects such as FLUTE, PHASE IV AI, SECURED, SYNTHEMA and SYNTHIA aim to develop general-purpose synthetic data infrastructures that can be reused across diseases, modalities, and institutional contexts. These platforms are designed to support scalability, modularity, and regulatory readiness, aligning with initiatives like the EHDS.

Methodologically, these efforts face challenges in designing configurable, distributed and federated generation pipelines, creating benchmarking frameworks that apply across clinical domains, and ensuring interoperability with existing health data standards (e.g., OMOP CDM, SNOMED-CT). Long-term sustainability and regulatory uptake, especially in clinical or commercial settings, are additional concerns.

*Approaches to data quality*

Ensuring the quality of synthetic data is a fundamental challenge across all projects, directly tied to both technical robustness and stakeholder trust. While there is no single definition of data quality in this domain, several recurring dimensions emerge across projects, most consistently statistical fidelity, clinical or analytical utility, and privacy protection. Fairness and completeness are also noted, though less consistently emphasised.

Fidelity is understood as the degree to which synthetic data replicates the statistical structure of the real-world datasets [20]. Most projects apply multi-level resemblance metrics, including univariate resemblance analysis (URA), multivariate resemblance analysis (MRA), and dimensional resemblance analysis (DRA). Other include data labelling tests (DLA) to determine whether synthetic and real data can be reliably distinguished. These methods evaluate distributional alignment, variable dependencies, and latent structure consistency.

Utility is typically operationalised through performance of downstream machine learning task [20]. The most frequently used metric is Train on Synthetic, Test on Real (TSTR), which evaluates whether models trained on synthetic data generalise well to real-world scenarios. Several projects also employ replication studies, in which the same analytic workflow is run on both real and synthetic data to assess comparability of results. These approaches are especially important in validating the clinical relevance of synthetic outputs.

Privacy is treated both as a constraint and as a quality attribute [20]. Techniques such as differential privacy [21] are used to provide formal guarantees during data generation are most used across the projects. Other projects, such as SECURED, assess disclosure risk using methods such as Similarity Evaluation Analysis (SEA), Membership Inference Attacks (MIA), and Attribute Inference Attacks (AIA). Additional metrics like k-anonymity and estimated re-identification risk are sometimes used to approximate identifiably

in tabular outputs, especially when aligning with existing anonymization frameworks. These approaches are not formal privacy guarantees but rather exploratory tools to gauge memorization risk and record-level similarity, especially when differential privacy is not used.

In contrast, regulatory frameworks like EMA Policy 0070 [22], developed for anonymizing clinical trial documents, allow both quantitative and qualitative justifications for anonymization assessment, without prescribing specific metrics. While the guidance mentions a re-identification risk threshold of 0.09, it clarifies that this is not a strict requirement and that context-specific justifications are acceptable. However, this policy does not apply to synthetic data, which currently lacks equivalent regulatory guidance. As a result, projects are left to rely on custom interpretations and often adopt conservative assumptions when evaluating privacy risks.

Bias and fairness receive less consistent attention but appear in projects focused on paediatric and rare disease domains. Here, data completeness, representativeness, and alignment with ethical and clinical norms are emphasised as part of quality control. Validation strategies range from standard quantitative metrics to expert-led evaluation. Several projects, such as PHASE IV AI and SYNTHEMA, combine statistical validation with domain expert review while others implement federated validation frameworks like SAFE, allowing synthetic data to be evaluated locally without pooling real datasets. These approaches are particularly valuable in privacy-sensitive or cross-border contexts.

Custom tools for quality assessment also vary in maturity. Some projects have developed dedicated pipelines for federated validation, privacy risk scoring, or clustering-based evaluation, while others rely on open-source libraries like SDMetrics [23], PyMDMA [23], SynthEval [23]. Table 3 summarises how projects combine standard tools with custom-built pipelines and highlights the quality dimensions they prioritise, illustrating the current fragmentation and lack of widely accepted evaluation standards.

| Project | Standard tools used | Custom tools/frameworks | Dimension emphasised |
|---|---|---|---|
| AISym4MED | SDMetrics, PyMDMA | Modular benchmarking pipeline | Fidelity, utility, privacy. |
| FLUTE | Custom built tool | Federated validation framework | Fidelity, privacy. |
| PHASE IV | SDMetrics | Replication studies | Fidelity, utility, privacy |

| | | | |
|---|---|---|---|
| AI | | Task-specific validation | |
| PHEMS | Custom built tool | SAFE validation pipeline<br>Expert review | Fidelity, utility, privacy, fairness |
| SECURED | SynthEval, SDMetrics | Privacy risk scoring tools | Fidelity, utility, privacy |
| SYNTHEMA | PyMDMA | SAFE framework integration | Fidelity, utility, privacy |
| SYNTHIA | SDMetrics, SynthEval | SYNTHIA platform benchmarking workflows | Fidelity, utility, multimodal consistency |

*Table 3 Quality evaluation approaches across projects*

*Strategies for trust and acceptance*

Establishing trust in synthetic data is critical for its adoption in health research, particularly where legal, clinical, and ethical considerations converge. Across the examined projects, trust-building strategies focus on three key areas: stakeholder engagement, transparency, and regulatory alignment.

Stakeholder engagement is a central mechanism for building legitimacy and ensuring synthetic data meets real-world needs. Most projects involve end users, including clinicians, data scientists, data managers, patients, and institutional partners early in the design and validation phase. Approaches include the creation of Communities of Practice, direct input from hospital IT teams and legal experts, and wide-ranging participation through hackathons, webinars, and user consultations. In some cases, clinical and patient representatives also help define research questions and validate outcomes, or a patient survey investigates the interest, role and consent of patients. This collaborative orientation ensures that synthetic data is not only technically sound, but clinically grounded and socially accepted.

Transparency is addressed through a variety of methods that aim to make synthetic data generation and evaluation auditable, interpretable, and safe to reuse. Several projects (e.g., AISem4MED, PHASE IV AI, SYNTHIA) provide model cards and dataset cards, or embed metadata detailing data origins, generation methods, and quality metrics across the fidelity, utility, and privacy dimensions. Others outline their validation logic clearly or use standardised reporting frameworks. Notably, projects such as SYNTHIA, plan to make their platforms accessible to external users, reinforcing openness and transparency at the ecosystem level.

Ethical and regulatory alignment is widely emphasised, with all projects signalling compliance with GDPR, and several also referencing the EU AI Act and national regulatory frameworks. Common mechanisms include local pseudonymisation/anonymization, consent-based inclusion, and privacy-

by-design principles. Despite this, there is still variation in how regulatory thresholds are interpreted across settings, highlighting the need for clearer, shared guidance.

Collectively, these strategies reflect a shift from synthetic data as a purely technical product to a socially embedded infrastructure. While projects show strong commitment to trust and acceptance, variability remains in how transparency and engagement are operationalised. As synthetic data moves toward broader implementation, the development of shared standards for co-design, documentation, and compliance will be essential to ensure lasting legitimacy and uptake. This also raises important domain-specific questions for health: “should physicians be informed if an AI model or diagnostic tool was trained on synthetic data?”. What level of transparency is owed to patients when synthetic data contributes to clinical decision-making? Addressing these uncertainties will be crucial to building trust in real-world applications.

*Innovation and ecosystem contributions*

In addition to addressing their individual use cases, many HealthData4EU projects contribute to a broader synthetic data ecosystem through technical innovation, interoperability, and resource sharing.

Technical contributions include cross-border ETL pipelines, federated validation architectures, privacy-preserving tools, and benchmarking platforms. Projects like SYNTHIA and SYNTHEMA develop full SDG pipelines and evaluation frameworks, while others provide libraries for assessing data quality and enabling federated generation workflows.

Interoperability is a shared focus. Most projects adopt the OMOP CDM and standard vocabularies (e.g., SNOMED-CT, LOINC), supporting integration with other health data infrastructures. The use of Python-based ML and federated learning libraries further enhances reusability and alignment with broader AI ecosystems.

Reusable assets range from harmonised datasets and eCRFs to anonymisation toolkits and domain-specific validation tools. Several initiatives plan to make platforms and resources openly accessible, supporting uptake beyond their original scope.

Nonetheless, challenges remain. These include scalability to broader disease areas, privacy risk certification, and the complexity of maintaining semantic coherence across multimodal data.

Regulatory acceptance and long-term governance also remain open questions.



## Discussion

The HealthData4EU projects provide a timely snapshot of how SDG is being operationalised in European health research. Despite differences in clinical focus, project maturity, and technical design, several recurring patterns emerge. Projects face ongoing dilemmas between privacy and data utility, a lack of shared validation frameworks, institutional barriers (especially in federated settings), and often operate within a fragmented innovation landscape. This discussion elaborates on these patterns and synthesises how projects are beginning to confront them, considering the research question concerning the methodological and acceptance challenges of trustworthy SDG.

### *Methodological challenges: Quality and Infrastructure*

Synthetic data quality is most assessed using tools such as SynthEval, SDMetrics, and PyMDMA, which focus on comparing synthetic data to real datasets in terms of statistical fidelity, diversity, utility, and privacy risk. However, many projects also rely on custom-built evaluation pipelines and ad hoc metrics tailored to their specific modalities and use cases, highlighting the absence of accepted standards. And these tools often assume that real data is itself of high quality and representative, without accounting for potential biases, gaps, or inconsistencies in the source data. As

a result, synthetic datasets may reproduce, and in some cases amplify, underlying flaws in the original data. Moreover, current assessments focus on output-level comparisons and cover only a small part of the data lifecycle. In federated settings in particular, limited visibility into early-stage data generation and preprocessing may have an impact on quality issues that influence downstream synthetic data outputs.

One key gap is the limited focus on representativeness. Despite SDG's promise in supporting data augmentation for underrepresented populations, many projects do not explicitly evaluate how well synthetic data reflects diverse patient cohorts. Expanding federated networks across more countries and care settings could help address this issue over time. If representativeness is not the primary objective, for instance in predictive model training, alternative performance measures such as task-specific model accuracy may be more appropriate than fidelity alone.

Federated SDG architecture introduces additional methodological complexity. While several projects explore decentralised, privacy-preserving pipelines, their implementation is often hindered by institutional factors documented across initiatives. For example, SECURED and FLUTE report difficulties integrating federated learning into heterogeneous hospital IT systems. And PHEMS highlights variation in data models and interoperability issues between clinical sites. Previous studies have also noted that many hospitals lack the technical infrastructure and data quality processes needed to support such pipelines [9, 24]. Achieving data maturity at the level of data sources is therefore crucial for implementing federated SDG. Projects that engage institutional partners (e.g., clinicians, legal teams, and IT leads) early in the design process are more resilient, but technical and regulatory barriers often surface later in development. This highlights a broader need to embed SDG innovation within more mature, integrated data ecosystems.

*Strategies for trust and acceptance*

Building stakeholder trust is a crucial yet inconsistently addressed dimension across the projects. Most initiatives signal GDPR compliance and ethical awareness, but fewer actively involve end users, such as clinicians or patients, throughout the design and validation lifecycle. While some projects implement stakeholder co-design, transparency mechanisms, or clinician-led validation, others limit engagement to compliance checks or internal review. Transparency efforts also vary, only a handful of projects publish detailed metadata, datasets cards, or model documentation, and few openly disclose quality metrics or generation logic.

Regulatory ambiguity compounds these challenges. In many contexts, it remains unclear whether synthetic data qualifies as anonymized under GDPR, creating uncertainty for data controllers and slowing uptake. While synthetic datasets are often assumed to fall outside the scope of GDPR, in practice, regulators and ethics boards may require strong evidence that re-identification is not possible, particularly when working with rare conditions or sensitive cohorts. Recent examples, such as Finnish regulatory guidelines [25] on publishing machine learning models, explicitly highlight differential privacy and synthetic data as mechanisms to mitigate disclosure risks, underscoring how approval processes can differ across jurisdictions. Within the HealthData4EU cluster the PHEMS project illustrates this approach. Despite generating synthetic data through a federated architecture that aligns with GDPR's privacy-by-design principles, it treats synthetic outputs as potentially personal data unless proven otherwise. This highlights the wider uncertainty across the ecosystem, where the absence of harmonised guidance leads institutions to face challenges on the side of legal conservatism. An EU-level clarification or certification pathway for synthetic data would provide needed legal certainty and reduce institutional friction.

Beyond compliance, the ethical use of synthetic data in clinical practice raises new domain-specific questions. For example: Should physicians be informed if an AI model or diagnostic tool was trained on synthetic data? What level of transparency is owed to patients when synthetic data contributes to clinical decision-making? These questions touch on autonomy, trust, and professional accountability. The absence of consensus on such disclosure practices risks undermining public legitimacy and trust. Addressing these gaps will require clear, context-sensitive policies that define when and how synthetic data usage should be communicated to end users.

*Tackling challenges: Emerging responses*

Despite these challenges, several initiatives offer emerging responses. Federated validation pipelines such as SAFE allow quality assessments without centralizing data, and libraries like PyMDMA offer reproducible benchmarking metrics. Some projects, such as SYNTHIA, are building publicly accessible platforms that include generation tools, benchmarking workflows, and guidelines for reuse. Others are contributing to interoperability by aligning data models with OMOP CDM or clinical vocabularies like SNOMED-CT and LOINC.

Still, these innovations vary in maturity and lack consistent integration. No shared consensus has yet emerged regarding what constitutes trustworthy SDG across clinical domains, modalities, or governance models. More coordinated mechanisms are needed to evaluate which tools, standards,

and practices are most effective and under which conditions.

*Synthesis and adoption*

To move from experimental pilots to sustainable ecosystem adoption, four priorities stand out:

- Establish shared quality evaluation frameworks, covering fidelity, utility, privacy, and representation.
- Clarify the legal status of synthetic data under evolving frameworks like the GDPR and AI Act.
- Support infrastructure through sustainable governance, including versioning, interoperability, and long-term maintenance.
- Embed stakeholder engagement throughout the SDG lifecycle, including patients, clinicians, data stewards, and regulators.

Progress to date shows that SDG holds promise but is not yet system ready. The challenge now is to coordinate technical, institutional, and regulatory efforts to move beyond isolated innovation toward a trusted, scalable, and inclusive synthetic data ecosystem for European health research.

Looking ahead, we envision a future in which synthetic data can be evaluated and trusted through a common set of validated, community-endorsed metrics. These would span statistical fidelity, task-specific utility, and measurable privacy risk. If a synthetic dataset can meet well-defined thresholds (such as resilience to membership inference or re-identification) it could be recognized as GDPR-safe with legal certainty. Such alignment would not only streamline compliance but also reduce institutional hesitancy and accelerate the real-world uptake of synthetic data tools.

Limitations

This study is based on a cross-sectional analysis of projects still in development, meaning that many findings reflect interim states rather than final outcomes. Furthermore, while the HealthData4EU cluster offers a representative overview of European SDG efforts, its insights may not capture practices emerging in non-EU settings or in private-sector initiatives. The reliance on project self-reporting and structured consultations may also introduce bias or omit evolving project components. Finally, this analysis focuses primarily on design and implementation challenges, rather than long-term outcomes or impacts. Future research should examine how these projects perform in practice and assess their contributions to real-world adoption, reuse, and public trust in synthetic health data.

Conclusions

This paper examined the methodological and acceptance challenges faced by seven major European projects developing synthetic health data. Our analysis reveals a range of barriers, from data quality validation gaps and institutional readiness issues to regulatory ambiguity and inconsistent trust strategies. While each project addresses these challenges in context-specific ways, the broader synthesis points to the need for greater standardization, alignment, and stakeholder inclusion.

To scale SDG from project-level pilots to system-wide practice, four priorities must be addressed:

- Shared evaluation metrics
- Legal clarity on GDPR-safe synthetic data
- Sustainable infrastructures and standards
- Meaningful stakeholder engagement throughout the SDG lifecycle

While not all projects adopt the same standardization approach, many integrate existing frameworks such as OMOP CDM, HL7 FHIR, and established clinical vocabularies (e.g., SNOMED-CT, LOINC, HPO) to support interoperability within modular platforms.

As SDG continues to mature, establishing shared evaluation frameworks, clear legal guidance, and sustainable infrastructures will be essential. Collectively, these steps can support the development of a trustworthy, scalable, and inclusive synthetic data ecosystem across Europe. While this paper draws on seven projects, the conclusion offers guidance. The tensions identified (e.g. privacy versus utility, centralization versus federation, trust versus complexity) are not unique to this cluster. They are likely to recur in any future synthetic data initiative. Recognizing and addressing them now will be

critical to shaping the responsible use of synthetic data in health research.

Funding

This study draws on seven research projects that are part of the HealthData4EU cluster. Funding sources for each project are as follows:

- SYNTHEMA, PHASE IV AI, SECURED, FLUTE, and AI SYM4MED are funded by the Horizon Europe programme.
- PHEMS is funded by Horizon Europe and UK Research and Innovation (UKRI).
- SYNTHIA is funded through the Innovative Health Initiative (IHI).

Acknowledgements

The authors would like to thank all seven HealthData4EU projects for their valuable input, collaboration, and contributions to this paper. Their openness and insights were essential to shaping the cross-case analysis. The authors are also grateful to Iraia Nuñez and Anna Lorenzini for their dedicated coordination and communications support. While not involved in the technical content of this study, their efforts in organizing exchanges and facilitating collaboration across the projects played a vital role in enabling this synthesis.



1. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Australian Health Review. 2018;42(5):568. doi: 10.1071/ah17255.
2. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A Survey of the Literature on Unintended Consequences Associated with Health Information Technology: 2014-2015. Yearb Med Inform. 2016 Nov 10(1):13-29. PMID: 27830227. doi: 10.15265/iy-2016-036.
3. Declerck J, Lee J, Sen A, Palmeri A, Oostenbrink R, Giannuzzi V, et al. The Potential to Leverage Real-World Data for Pediatric Clinical Trials: A Proof-of-Concept Study. J Med Internet Res. 2025;27:e72573. doi: 10.2196/72573.
4. Wang Z, Penning M, Zozus M. Analysis of Anesthesia Screens for Rule-Based Data Quality Assessment Opportunities. Stud Health Technol Inform.

2019;257:473-8. PMID: 30741242.
5. Wiebe N, Xu Y, Shaheen AA, Eastwood C, Boussat B, Quan H. Indicators of missing Electronic Medical Record (EMR) discharge summaries: A retrospective study on Canadian data. Int J Popul Data Sci. 2020 Dec 11;5(1):1352. PMID: 34007880. doi: 10.23889/ijpds.v5i3.1352.
6. Declerck J KD, Vander Stichele R, Coorevits P. Frameworks, dimensions, definitions of aspects and assessment methods for the appraisal of quality of health data for secondary use: a review of reviews. JMIR Medical Informatics. 2024.
7. Shabani M, Marelli L. Re‐identifiability of genomic data and the GDPR. EMBO reports. 2019;20(6):e48316. doi: https://doi.org/10.15252/embr.201948316.
8. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244. PMID: 27713905. doi: 10.13063/2327-9214.1244.
9. Declerck J, Vandenberk B, Deschepper M, Colpaert K, Cool L, Goemaere J, et al. Building a Foundation for High-Quality Health Data: Multihospital Case Study in Belgium. JMIR Med Inform. 2024 Dec 20;12:e60244. PMID: 39727158. doi: 10.2196/60244.
10. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):219-24. PMID: 22249968. doi: 10.1136/amiajnl-2011-000597.
11. Oja M, Tamm S, Mooses K, Pajusalu M, Talvik HA, Ott A, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. JAMIA Open. 2023 Dec;6(4):ooad100. PMID: 38058679. doi: 10.1093/jamiaopen/ooad100.
12. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. PLOS Digital Health. 2023;2(1):e0000082. doi: 10.1371/journal.pdig.0000082.
13. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digital Medicine. 2023 2023/10/09;6(1):186. doi: 10.1038/s41746-023-00927-3.
14. Al-Dhamari I, Abu Attieh H, Prasser F. Synthetic datasets for open software development in rare disease research. Orphanet Journal of Rare Diseases. 2024 2024/07/15;19(1):265. doi: 10.1186/s13023-024-03254-2.
15. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. arXiv preprint arXiv:180603384. 2018.
16. Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional gans. arXiv preprint arXiv:170602633. 2017.
17. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Medical Research Methodology. 2020 2020/05/07;20(1):108. doi: 10.1186/s12874-020-00977-1.
18. Stenger M, Leppich R, Foster I, Kounev S, Bauer A. Evaluation is key: a survey on evaluation measures for synthetic time series. Journal of Big Data.

2024 2024/05/07;11(1):66. doi: 10.1186/s40537-024-00924-7.
19. Alaa A, Van Breugel B, Saveliev ES, Van Der Schaar M, editors. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. International conference on machine learning; 2022: PMLR.
20. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic Data--what, why and how? arXiv preprint arXiv:220503257. 2022.
21. Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. Found Trends Theor Comput Sci. 2014;9(3–4):211–407. doi: 10.1561/0400000042.
22. 2022; Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf.
23. ; Available from: https://github.com/UrbanInstitute/syntheval.
24. Tzortzis IN, Gutierrez-Torre A, Sykiotis S, Agulló F, Bakalos N, Doulamis A, et al. Towards generalizable Federated Learning in medical imaging: A real-world case study on mammography data. Computational and Structural Biotechnology Journal. 2025 2025/01/01/;28:106-17. doi: https://doi.org/10.1016/j.csbj.2025.03.031.
25. Producing anonymous results. Findata (Finnish Social and Health Data Permit Authority); 2025 [2025 Aug 7]; Available from: https://findata.fi/en/services-and-instructions/producing-anonymous-results/#other-result-types.

# Supplementary Files

# Multimedia Appendixes

Overview of all HealthData4EU projects, their objectives, data modalities, use cases, and methodological challenges.
URL: http://asset.jmir.pub/assets/8437d456752fc17aeb87051d57ee2181.docx