synthia

Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications

Tackling Public Trust for Synthetic Data in Health

SYNTHIA is a multi-stakeholder Public-Private Partnership focused on developing practical guidance for the responsible generation and use of synthetic data in health and personalised medicine. The consortium aims to bridge legal, technical, and governance perspectives to support privacy-preserving data use that is demonstrably fit for purpose and trustworthy.

This paper examines public trust as a cross-cutting challenge for the adoption of synthetic data in healthcare, proposing implementable commitments and policy recommendations to support trustworthy adoption.

Policy 02 / Trust Paper

Synthetic data are increasingly recognized as a practical enabler of health data reuse in contexts where privacy constraints, fragmented governance, limited cross-border mechanisms, and limited availability restrict access to real-world data on a scale.

When properly generated and governed, synthetic data can expand safe access to realistic datasets for research, innovation, education, and selected policy applications while reducing disclosure risks associated with sharing raw patient-level information.

Within the six disease domains addressed by SYNTHIA, synthetic data offer opportunities to enable cross-border collaboration and support research and innovation. At the same time, their promise must be considered alongside limitations related to privacy, utility, bias, transparency, and public trust.

Trust in synthetic data is therefore not only a technical property but an outcome of governance, assurance, and social legitimacy.

Leverage synthetic data to propel personalized medicine to new heights

- **SYNTHIA is a multidisciplinary collaboration of 39 consortium partners bringing together synthetic data generation developers, FAIR data experts, clinical researchers, developers of therapies and data-based tools, legal experts, socio-economic analysts, regulatory, policy advocacy, and communication experts to provide a 360° vision on how to advance healthcare applications through the use of synthetic data.** As the first Innovative Health Initiative (IHI) project dedicated to synthetic data in healthcare, SYNTHIA aims to revolutionize personalized medicine by developing and validating robust synthetic data generation methods that enable privacy-preserving access to high-quality health data. By addressing privacy concerns, ensuring data quality, and building trust among stakeholders, the project seeks to accelerate research and innovation while safeguarding patient information.
- **Through six high-impact clinical use cases (lung cancer, breast cancer, multiple myeloma, diffuse large B-cell lymphoma, Alzheimer's disease, and type 2 diabetes) SYNTHIA demonstrates how synthetic data can support research, AI development, and clinical decision-making across diverse healthcare settings**. The project develops validated methods for generating synthetic data across multiple data modalities, including clinical, imaging, genomic, laboratory, and longitudinal data, while establishing robust evaluation frameworks for privacy, quality, utility, and bias mitigation. Together with its federated platform, transparent governance framework, and strong multidisciplinary collaboration, SYNTHIA aims to unlock the full potential of synthetic data to advance biomedical research, foster innovation, and support responsible data use across Europe.

  www.ihi-synthia.eu

Table of Content

1. Executive Summary 6

1.1 The promise, paradox and trust challenge of synthetic health data 6

1.2 Cross-cutting challenges shaping public trust 6

2. Background and Rationale 10

2.1 Why synthetic data requires a trust-first approach 10

2.3 Opportunities and trust challenges across disease contexts 11

2.4 Early trust-enhancing use cases 11

3. The Public Trust Challenge for Synthetic Health Data 13

4. Organisational and Governance Lens 14

4.1 Key organisational and governance challenges 14

4.2 Organisational recommendations 14

5.1 Technical trust challenges 16

5.2 Technical recommendations 16

6. Regulatory and Ethical Lens 18

6.1 Regulatory and ethical problematique 18

6.2 Regulatory and ethical recommendations 18

7. Resources, Workforce and Investment Lens 20

7.1 Capacity and resource gaps 20

7.2 Resource and workforce recommendations 20

7.3 Investment opportunities and economic rationale 20

8. Citizen, Patient and Societal Lens 22

8.1 Public attitudes and expectations 22

9. Implementation Roadmap 22

10. Conclusion 24

“Public trust cannot simply be assumed where data are labelled “synthetic.”

For some stakeholders, the term may evoke concerns about artificiality, opacity, or manipulation. In certain contexts, synthetic data may need to be sufficiently realistic to be fit for purpose. Depending on how fidelity and privacy are defined and implemented, higher fidelity may raise concerns about residual risks such as re-identification or the inference of information about individuals represented in the source data (see also Kulynych et al., 2026 for a discussion of the relationship between fidelity measures and privacy guarantees).

In other contexts, datasets that are not fully realistic may still be appropriate for specific downstream uses, such as training analytical models. This results in a context-dependent balance between utility and privacy, rather than a single threshold of adequacy (Adams et al., 2025). Additional risks include the reproduction or amplification of structural bias inherited from the real data sources, inappropriate secondary use, or misinterpretation in high-stakes contexts. In settings where synthetic data may inform consequential scientific, regulatory, or policy decisions, these risks require particular scrutiny. If risks are insufficiently assessed or governed, synthetic data may be used to inform decisions without adequate validation, transparency, or accountability.

Trust in synthetic data is therefore not only a technical property but an outcome of governance, assurance, and social legitimacy. Legal compliance alone may be insufficient to sustain public confidence, particularly where data reuse practices are perceived as opaque or misaligned with societal expectations.

- This policy paper is developed in the context of the SYNTHIA project, which aims to accelerate the ethically responsible and scalable use of synthetic data in healthcare by advancing generation and validation methods, operational assurance approaches, and governance models that can operate across institutional and cross-border settings, in line with the broader objectives described in the project’s official documentation.
- While this paper does not assess the scientific adequacy of synthetic data for specific evidentiary or regulatory claims, it frames public trust as a cross-cutting challenge spanning organizational, technical, regulatory, resource, and societal dimensions. It proposes implementable commitments and policy recommendations to support trustworthy adoption. Its central message is that synthetic data can strengthen trust in health data sharing only if it is governed with explicit safeguards, independent scrutiny for high-stake uses, and meaningful transparency and redress.
- This work is complemented by the SYNTHIA Policy White Paper which translates these trust principles into concrete governance and implementation pathways for the use of synthetic data in public health policymaking.

Visit the SYNTHIA website to find the policy materials: www.ihi-synthia.eu/stakeholders/policy-makers



1. Executive Summary

1.1 The promise, paradox and trust challenge of synthetic health data

Synthetic data offers significant opportunities for health systems and research ecosystems. It can expand access to realistic datasets for research and AI development, including in areas where real data are scarce, such as rare diseases or small sub-populations, by amplifying small or fragmented datasets, enabling controlled reuse where direct sharing is constrained, and supporting cross-site analysis without exposing raw records (Rajotte et al., 2022).

Unlike traditional anonymisation techniques, which modify or suppress elements of real datasets, synthetic data involves the generation of new datasets that replicate statistical properties of source data without containing direct records of real individuals, thereby decoupling analytic access from identifiable patient-level information. Synthetic data can also support testing, validation, training, and exploratory analyses that are difficult or impractical to conduct using real data alone (Giuffrè & Shung, 2023).

At the same time, synthetic data introduces a paradox for public trust. While it may reduce certain privacy risks, it can also create false assurance if treated as inherently safe or automatically anonymous. Trust challenges arise not only from data protection considerations, but also from the validity, reliability, and appropriate interpretation of insights derived from synthetic data (Emam, 2020; Koul et al., 2025). Poorly generated or insufficiently evaluated synthetic data may still leak sensitive information, reproduce historical biases, distort rare patterns or support decisions without adequate evidentiary grounding. Over-simplified narratives risk framing synthetic data as a “silver bullet”, increasing the likelihood that failures will undermine trust not only in synthetic data, but in health data sharing more broadly (Kaabachi et al., 2025).

Trust in synthetic health data must therefore be actively earned. It depends on clear governance, demonstrable assurance, appropriate use-case restrictions and meaningful transparency (Vallevik et al., 2024). Without these conditions, synthetic data risks being perceived as a workaround rather than a responsible innovation, particularly in contexts where decisions informed by data have significant individual or societal consequences.

1.2 Cross-cutting challenges shaping public trust

Across organisational, technical, regulatory and societal context, a set of recurring challenges undermines confidence in synthetic data deployment:

- **Unclear governance and accountability,** particularly in cross-border and multi-partner ecosystems, where roles and responsibilities across the synthetic data lifecycle *(i.e generation, validation, sharing, and continuous governance)* are often fragmented.
- **Uncertain privacy guarantees,** reflecting concerns that synthetic data may still enable re-identification or inference about individuals whose real-world data were used to generate the models, particularly where privacy safeguards are weak or where generation methods do not provide formal privacy guarantees.
- **Uncertain fitness for purpose**, where affected populations, decision-makers, or downstream users lack confidence that insights derived from synthetic data are applicable, accurate, or appropriate for the intended use, especially in high-stakes or population-level decision-making contexts.
- **Bias, fairness and equity concerns,** including the replication of historical inequities, the erasure of rare but clinically relevant patterns, *and the risk of intentional manipulation (e.g. through data poisoning) that can distort subgroup performance and lead to unfair downstream impacts.* While it may be impossible to eliminate all bias, a lack of transparency about known limitations and population coverage undermines trust in both the data and the decisions informed by it.
- **Regulatory ambiguity**, as existing legal frameworks remain largely technology-neutral and provide limited operational guidance for synthetic data in high-stakes health contexts and lack a widely accepted legal definition of "synthetic data". For comparative purposes, synthetic data-specific regulatory guidance from jurisdictions including the United Kingdom, Singapore, South Korea and Canada has been analysed in recent literature (Pilgram, Ko, et al., 2025; Reich et al., 2025: Pilgram, Fineberg, et al., 2025).
- **Low literacy and limited engagement**, with technical and project-centric communication failing to support broader public understanding and legitimacy.
- **Policy and implementation lag**, whereby synthetic data is recognised as a privacy preserving technique but rarely governed in sufficient detail to provide reassurance and accountability in practice. Some exceptions occur where clarification efforts have been attempted, for example by the United Kingdom (Office for National Statistics, 2021), Singapore (Personal Data Protection Commission Singapore, 2025), South Korea (Korea, 2024), and the United Nations (De Wilde et al., 2024).
- **Over or unwarranted confidence** in synthetic data and downstream AI outputs, from an illusion of comprehensive and diverse datasets, lacking traceability and public accountability (Shanley et al., 2024).

Together, these challenges show that public trust in synthetic data cannot be addressed through technical measures alone. It requires coordinated action across governance, regulation, resources, and societal engagement, as well as attention to additional risks such as epistemic validity (e.g. correlations mistaken for causality or unwarranted confidence in outputs), security and adversarial threats (including leakage or manipulation of generative models), intellectual property and ownership uncertainty, and liability considerations when synthetic data informs high-stakes healthcare decisions.

1.3 Strategic objective

The objective of this policy paper is to support policymakers, regulators, funders, health systems, and innovators in developing credible and implementable frameworks that:

- Protect individuals and communities from privacy, bias and group-level harms;
- Enable trustworthy innovation and uptake of synthetic-data-enabled tools aligned with their intended use;
- Use synthetic data to strengthen, rather than erode, public trust in health data sharing and AI

1.4 High level priority recommendations

To strengthen public trust in the use of synthetic data in health, policy action should focus on a limited number of high-impact priorities across governance, technical assurance, regulation, resources and societal engagement, inspired by related work published by the European Medicines Agency (European Medicines Agency, 2026) **and World Economic Forum** (World Economic Forum, 2025).

Table 1. summarises these priority areas and associated policy actions, providing a concise overview of the core interventions required to support trustworthy and accountable deployment. Together, these priorities can be operationalised through a coherent "Trust Stack" and a set of minimum Trust Commitments, which provide a policy-ready summary of the conditions required for trustworthy deployment of synthetic data in health.

Priority area	Key actions
Organisational and governance	Establish multi-stakeholder Synthetic Data Governance Boards at EU level including patient and citizen representation: These boards may be embedded within existing data governance bodies, such as national health data access bodies under the European Health Data Space, institutional data access committees, or competent authorities overseeing high-impact digital health deployments. Clarify roles, responsibilities and liability across the synthetic data lifecycle (from seed data to downstream decision use); adopt transparent incident response and trust-preservation playbooks aligned with EU data governance standards.
Technical and infrastructural	Mandate proportionate validation and assurance aligned with risk and intended use; require standard documentation and controlled-access ("safe settings") for policy-relevant datasets; invest in shared benchmarks and supervised sandboxes/testbeds.
Regulatory and ethical	Publish operational guidance on the legal status of synthetic data including clarification at EU level (e.g. through the European Data Protection Board (EDPB)); define value-added use cases for synthetic data relative to real-world data, and require explicit justification in regulatory and Health Technology Assessment (HTA) contexts - including coordination at EU level through bodies such as the HTA Coordination Group (HTACG); clarify ownership, licensing and reuse rights.
Resources and workforce	Invest in capacity-building across regulators, assessors, clinicians, ethics committees and patient representatives; protect funding for assurance, review, engagement and monitoring; support shared-service models such as Data Generation as a Service.
Citizen and patient trust	Embed a social-licence approach combining participation, transparency and redress; publish citizen-facing transparency summaries; monitor trust and equity impacts over time, including grievance handling and corrective actions.

——— Table 1. Priority policy actions for trustworthy use of synthetic health data.

The Trust Stack (what must be in place, together)

1. **Technical assurance** (validation, documentation, benchmarks, monitoring)
2. **Governance** (boards, accountability matrix, incident response playbook, auditability)
3. **Legal clarity** (status, permitted uses, licensing, transparency duties)
4. **Social licence** (participation, citizen-facing transparency, redress)
5. **Operational sustainability** (funding, shared services, workforce capacity)

Importantly, the suggested Trust Stack is not intended to recreate the access constraints associated with real-world health data. Its elements should be applied in a proportionate, risk-based manner, with lighter requirements for low-risk uses and enhanced safeguards reserved for high-stakes, decision-relevant deployments.

Trust Commitments (minimum expectations for any synthetic data deployment in health)

1. No "risk-free" claims; only **risk-reduced** claims supported by evidence.
2. A **baseline assurance package** for all decision-support uses of synthetic data (i.e. uses beyond internal method development or exploratory testing).
3. **Independent review** for high stakes uses, proportionate to risk and impact.
4. A clear **accountability and liability map** across the lifecycle, specifying responsible actors for generation, validation, release, monitoring, and decision use.
5. A standing **grievance and redress** mechanism.
6. Public-facing **transparency summaries** (what, why, safeguards, limitations).
7. Equity-by-design, including proportionate representativeness checks and subgroup performance assessment aligned with intended use and comparable standards applied to real-world data.
8. Continuous monitoring and **re-validation** of datasets or generative methods where synthetic data is used in ongoing, high-impact, or evolving context

After setting out the trust challenge and strategic objectives, *Section 2* presents the rationale for a trust-first approach to synthetic health data and situates this paper within the broader SYNTHIA trust pathway. *Sections 3 to 8* examine the public trust challenge through complementary organisational, technical, regulatory, resource and societal lenses, translating identified risks into concrete, implementable policy recommendations. Each section addresses different actors across the health data ecosystem, reflecting the distributed nature of trust in synthetic data. *Section 9* outlines a phased implementation roadmap, distinguishing near-, medium- and longer-term actions. The paper concludes by highlighting the conditions under which synthetic data can strengthen, rather than erode, trust in health data sharing and AI.

2. Background and Rationale

2.1 Why synthetic data requires a trust-first approach

Health systems face a structural tension: society expects innovation (better prediction, earlier diagnosis, more equitable care) while also expecting strong privacy protection and accountability for data use.

Synthetic data is increasingly positioned as a mechanism to reduce friction in this tension by enabling wider analytic access without broad exposure of raw patient-level data. However, reducing access barriers alone is insufficient to secure trust, particularly where synthetic data informs decisions with real-world consequences. Evidence from health data-sharing initiatives (Carter et al., 2015) and the literature on social licence (Mueller et al., 2021) highlights several recurring conditions for building and maintaining public trust, including:

- **Clarity:** people understand what is happening and why.
- **Control:** there are credible safeguards, oversight, and constraints on misuse.
- **Reciprocity:** benefits are visible, and communities can influence priorities.

Synthetic data must therefore be integrated into this trust logic. It cannot rely on technical framing alone, nor on assumptions that privacy protection automatically translates into legitimacy or confidence in downstream use.

2.2 SYNTHIA's trust pathway

SYNTHIA operationalises a complete trust pathway for synthetic health data, spanning data generation methods, evaluation and assurance, legal and ethical governance, and implementation and stakeholder engagement. This integrated structure enables trust to be addressed as a system-level challenge rather than a single technical or legal issue.

Within this pathway **SYNTHIA Policy Engagement Materials Series** (www.ihi-synthia.eu/stakeholders/policy-makers) focuses on translating project outputs into policy-ready practices. This includes defining minimum assurance expectations, developing governance and accountability templates, and identifying scalable implementation mechanisms that can be adopted across institutions and borders. In this way, SYNTHIA functions not only as a technical initiative, but as a practical testbed for the trustworthy and accountable deployment of synthetic data in healthcare.

This contribution is grounded in empirical learning rather than abstract assumptions. An internal requirements analysis across six diverse health use cases has already shown that data availability, data quality, feasibility of synthetic data generation and associated risks vary substantially by disease area, data modality and institutional context. The analysis also highlights persistent challenges related to multimodal data integration, residual bias, operational dependencies and governance complexity, even at early design stages. These findings reinforce the need for proportionate, use-case-specific assurance and governance approaches, and underline why trust in synthetic health data must be actively designed and resourced rather than assumed.

2.3 Opportunities and trust challenges across disease contexts

Trust challenges associated with synthetic data recur across disease areas, despite differences in data modalities and clinical pathways. Table 2 illustrates how the value of synthetic data and the associated trust challenges vary across selected disease contexts, highlighting recurring patterns.

Use case	Value of synthetic data	Key trust challenges
Rare diseases	Generates larger, statistically realistic cohorts for exploratory analyses and feasibility work; enables cross-border collaboration while reducing exposure of small, easily identifiable groups.	Ensuring rare patterns are not "smoothed away"; avoiding misrepresentation; enabling communities to contest and correct harms.
Oncology	Augments limited evidence for rare tumour subtypes or biomarker-defined populations; supports AI development in imaging and pathology; contributes to feasibility modelling for trials.	Ensuring representativeness across biomarker-defined and molecularly stratified subgroups; validating synthetic imaging and pathology data (e.g. tumour segmentation, radiomics features) to prevent artefacts or loss of clinically relevant patterns; clarity on acceptability for evidence evaluation; avoiding institutionalisation of biased treatment pathways.
Chronic conditions and wearables	Generates realistic time-series data for algorithm development and simulation without exposing raw device streams or behavioural data, which may be highly granular, continuous, and susceptible to re-identification due to distinctive temporal or mobility patterns.	Avoiding overfitting to digitally engaged populations; transparency on data collection; protection against downstream profiling.
Neurodegenerative diseases	Supports multimodal cohort construction and "what-if" modelling for screening and referral pathways.	Managing expectations; preventing stigma or insurance-related harms; ensuring cohorts reflect diverse ageing trajectories rather than reinforcing inequities.

Table 2. Priority health use cases for synthetic data and associated trust challenges

Across contexts, synthetic data strengthens trust only when validation, governance and communication are aligned with the expectations of affected communities and with the intended scope and limitations of use.

2.4 Early trust-enhancing use cases

Early deployments suggest that synthetic data can strengthen institutional and regulatory trust when it is in clearly defined, lower-risk contexts and accompanied by explicit assurance. Examples include use of synthetic data to:

- Provide safe, realistic datasets for training, method development and controlled testing of analytics pipelines, without replacing validation on appropriate real-world data;
- Enable controlled collaboration where direct sharing of raw microdata is not feasible;
- Demonstrate that privacy risk and fitness for intended use are actively evaluated and documented rather than assumed.
- Support supervised sandboxes in which innovators, health authorities and regulators learn together under agreed evaluation criteria.

These are directional examples: they show pathways to trust, not guaranteed outcomes.



3. The Public Trust Challenge for Synthetic Health Data

This section frames the public trust challenge as a system-level issue. While public trust is a societal outcome, responsibility for shaping it lies with identifiable actors across the health data ecosystem. The challenges outlined here provide the foundation for the actor-specific responsibilities defined in Sections 4 to 8. They draw on emerging academic literature, comparative regulatory guidance, and empirical insights generated within the SYNTHIA project, including cross-use-case requirements analysis and stakeholder discussions. Table 3 consolidates the main public trust challenges associated with synthetic health data and their implications for governance and policy.

Trust challenge	Description and implications
Ambiguous privacy promises	Trust is undermined when synthetic data is described as automatically anonymous. Privacy risk depends on generation method, seed data, release modality, adversary model and linkability with external data. Risk management must be explicit, evidence-based and proportionate.
Uncertain fitness for clinical and policy use	Statistically similar synthetic data may still be clinically implausible, omit key correlations or distort rare events. When used for model calibration, system stress-testing or policy modelling, poor fitness can mislead decisions. Fitness must be defined relative to intended use, with documented limitations.
Bias, fairness and group-level harms	Synthetic data can reproduce historical inequities, amplify underrepresentation by smoothing rare patterns, and distort subgroup performance, leading to unfair downstream impacts. Trust requires equity-by-design, including subgroup checks, representativeness assessments and community engagement.
Institutional fragmentation and unclear accountability	Multi-party pipelines risk “everyone involved, no one responsible”. Trust depends on explicit assignment of duties for validation, monitoring, incident response and accountability when synthetic-data-driven outputs influence decisions.
Communication gaps and low literacy	Lack of understanding among citizens and professionals about what synthetic data is, why it is used and what safeguards exist increases suspicion. Communication must be plain-language, consistent, and supported by transparency and redress mechanisms, not only technical explanations.

Table 3. Public trust challenges for synthetic health data

4. Organisational and Governance Lens

Primary actor: Institutional governance owners

This chapter is addressed to institutional leaders, data controllers, and governance bodies responsible for overseeing synthetic data initiatives across their lifecycle. These actors are accountable for establishing oversight structures, assigning responsibilities, managing risk, and ensuring that synthetic data deployments operate within defined governance and accountability frameworks.

4.1 Key organisational and governance challenges

- Multi-actor synthetic data pipelines across borders or sectors, with unclear allocation of decision rights and lifecycle accountability, particularly where synthetic data is derived from seed data governed under different legal regimes.
- Inconsistent governance maturity across institutions and regions.
- Lack of independent scrutiny for high impact uses.
- Limited operational readiness for incidents (privacy leakage, bias failures, decision-support failures).

4.2 Organisational recommendations

Table 4. translates these challenges into concrete organisational recommendations and operational requirements. Implementation should occur at the level where decision-making authority over the synthetic data deployment resides (institutional, national, or cross-border) proportionate to scale and impact.

Objective	Purpose	Key requirements and actions
O1. Establish multi-stakeholder Synthetic Data Governance Boards	Provide formal oversight for high-impact synthetic data use cases at national, regional or large-institution level (or embedded within existing data councils).	Minimum membership: patient/citizen representatives, clinicians, technical experts, ethics and legal experts, and where relevant regulatory/HTA and payer perspectives. Minimum mandates: approve high-impact use cases and conditions of use; oversee assurance requirements and independent reviews; require public-facing transparency summaries; publish annual lessons learned, including failures.
O2. Clarify roles and responsibilities across the synthetic data lifecycle	Ensure accountability and risk management from seed data to downstream use.	Require consortium agreements, data-sharing agreements and institutional policies to identify and document the respective roles of controllers, processors and successive independent controllers, as determined by their actual functions and by who determines the purposes and essential means of each processing operation; allocate responsibilities for risk assessment, fairness checks, documentation, monitoring, incident response, and handling data subject requests, including access, rectification and erasure and the implications of erasure or withdrawal of consent concerning seed data for generative models and synthetic datasets already generated; and clarify liability allocation for misuse and harm.
O3. Embed citizens and patients in governance as a standing function	Strengthen legitimacy, acceptability and trust through continuous public involvement.	Establish a Citizen and Patient Advisory Panel with authority to review public-facing materials; advise on acceptability by disease context; and provide input on grievance handling and incident response expectations.
O4. Adopt an Accountability and Liability Matrix (operational tool)	Operationalise accountability for each synthetic data deployment.	Require a one-page matrix for each significant deployment covering: seed data holder; synthetic data generator; validator/auditor; platform operator; end user/decision-maker. For each role, specify obligations for privacy risk assessment and safeguards; fairness and representativeness checks; documentation and disclosure; monitoring and re-validation; incident response and communications; and withdrawal or containment authority.
O5. Implement an Incident Response and Trust-Preservation Playbook	Enable rapid, transparent and accountable response to failures affecting trust.	Define three incident classes and actions: 1. Privacy leakage / re-identification: immediate containment, independent review, public communication principles, decision on withdrawal, revision or restricted re-release. 2. Bias or fairness failure: trigger thresholds, mitigation plan, community engagement, monitoring and follow-up reporting. 3. Decision-support failure: traceability review, corrective actions, governance review of ongoing acceptability.

——— Table 4. Governance objectives and operational requirements for synthetic data in health



5. Technical and Infrastructural Lens

Primary actor: Synthetic data developers and technical implementers

This chapter is addressed to synthetic data generators, data scientists, model engineers, and platform operators responsible for designing, validating, documenting, releasing, and monitoring synthetic datasets. The recommendations in this section define the technical and assurance obligations that must be fulfilled before synthetic data is relied upon by governance bodies, regulators, HTA assessors, or decision-makers.

5.1 Technical trust challenges

- Variable methods and inconsistent evaluation practices.
- Limited shared benchmarks for comparability.
- Increasing complexity (multimodal data, imaging, longitudinal trajectories) with higher infrastructure and compute demands.
- Inadequate documentation and weak reproducibility.
- Lack of controlled release modalities appropriate for high-stakes contexts.

5.2 Technical recommendations

Table 5. translates these challenges into concrete organisational recommendations and operational requirements.

The fulfilment of these technical obligations enables governance bodies (Section 4) and regulators and HTA assessors (Section 6) to make informed decisions regarding acceptance and deployment.

Technical recommendation	Purpose	Key requirements and actions
T1. Operationalise a validation and assurance approach for synthetic data	Ensure synthetic data is safe, reliable and fit for purpose.	Adopt a structured assurance approach covering privacy risk (including residual disclosure risk under realistic adversary assumptions); utility and fitness for intended tasks; statistical fidelity and stability; semantic and clinical plausibility; and subgroup performance, representativeness and rare-event handling.
T2. Require a Baseline Assurance Package (mandatory for all non-internal uses)	Establish a minimum, standardised assurance baseline supported by standardised transparency templates and tools for synthetic datasets intended for consistent external or decision-support use.	Require a synthetic dataset datasheet that provides a life cycle overview (provenance, population scope, limitations, permitted and prohibited uses); generator model card (method, assumptions, failure modes, key controls); validation report (privacy risk, fidelity, utility and fairness/subgroup checks); governance log (approvals, residual risks, mitigations, monitoring plan); and access terms (safe-setting modality, logging, audit rights, enforcement measures).
T3. High-stakes upgrade (mandatory)	Strengthen safeguards for high-impact and decision-relevant uses.	For high-impact or regulatory-relevant uses, consider proportionate review and periodic re-validation where appropriate alongside decision traceability documentation (dataset → model → output → decision context) consistent with comparable real-world data
T4. Invest in secure, interoperable infrastructures and controlled release ("safe settings")	Reduce misuse risk and enable scrutiny while supporting reuse.	Ensure repositories and platforms provide strong access controls and logging; controlled analysis environments for sensitive use cases; and interoperability and metadata completeness to support reuse and independent scrutiny.
T5. Promote shared benchmarks and supervised sandboxes/testbeds	Enable comparability, regulatory learning and continuous improvement.	Establish shared benchmark libraries for privacy risk, fidelity, utility and fairness; supervised sandboxes for testing methods and use cases under governance and, where appropriate, regulator participation; and publication of lessons learned, including negative findings.
T6. Plan for scale prerequisites (including compute and multimodal ambitions)	Prepare systems for advanced and large-scale synthetic data use.	For advanced synthesis use cases (imaging, multimodal cohorts, high-frequency time series), plan for infrastructure and compute capacity; federated and privacy-enhancing approaches where seed data cannot move; and workforce capability for evaluation, monitoring and audit.

Table 5. Technical recommendations for trustworthy synthetic data deployment in health



6. Regulatory and Ethical Lens

Primary actor: Rule-setters and evidence evaluators

This chapter is addressed to legislators, regulators, data protection authorities, ethics committees, and health technology assessment (HTA) bodies responsible for defining the legal status, evidentiary acceptance criteria, and oversight conditions for synthetic data use in healthcare. These actors determine when and under what conditions synthetic-data-augmented evidence may inform regulatory, reimbursement, or policy decisions. This section defines how competent authorities evaluate and set conditions for the outputs generated under the technical and governance responsibilities described in Sections 4 and 5.

6.1 Regulatory and ethical problematique

Regulatory neutrality can create practical ambiguity: stakeholders lack clear expectations on how to treat synthetic data, when it may still be personal data, and what is required for acceptance in high-stakes processes. Ethical concerns also expand beyond privacy into fairness, group harms, benefit sharing, and perceived circumvention of protections.

6.2 Regulatory and ethical recommendations

Table 6. translates these challenges into concrete organisational recommendations and operational requirements.

Regulatory / ethical recommendation	Purpose	Key requirements and actions
R1. Clarify legal status and duties for synthetic data	Provide legal certainty and consistent application of obligations by competent public authorities.	Issue operational guidance defining criteria for when synthetic data should be treated as personal or non-personal data; clarify lawful basis, transparency requirements, DPIA expectations and rights implications; and provide health-specific examples and risk tiers.
R2. Require disclosure of synthetic data use in model development and evidence	Ensure transparency and informed evaluation of AI and analytics.	For high-impact healthcare or policy uses, require disclosure of whether synthetic data was used for training, testing, validation or augmentation; what assurance evidence supports use; and known limitations and constraints.
R3. Embed synthetic data into HTA-adjacent evidence evaluation	Enable consistent and credible assessment of synthetic-data-augmented evidence.	Require HTA bodies and regulatory authorities to adopt structured evaluation criteria for synthetic-data-augmented evidence is acceptable; how robustness is evaluated; and how uncertainty and limitations must be communicated.
R4. Strengthen consent and ethics committee practices	Improve ethical oversight and participant understanding.	Update consent templates and review processes to explain potential synthetic data generation and downstream uses; define meaningful opt-out options where appropriate, including for secondary uses not based on consent where such options can be implemented without undermining the lawful purpose or integrity of the processing; and require SD-specific review questions on privacy risk, group harms and governance.
R5. Address group-level harms and justice	Prevent inequities and collective harms.	Require proportionate subgroup representativeness and performance reporting, aligned with standards applied to comparable real-world data uses; mitigation of disparities; and engagement with affected communities in interpreting and governing sensitive uses.
R6. Resolve ownership, licensing and reuse rights	Remove governance ambiguity and barriers to scale through public-interest-oriented legal frameworks while protecting the public interest.	Standardise licensing options for synthetic assets derived from public investment; rules for derivative synthetic datasets and downstream model training; and public-interest clauses and transparency expectations to avoid perceived appropriation without benefit sharing.
R7. Accountability for synthetic-data-driven decision support	Ensure responsibility remains with decision-makers and systems.	Require explicit accountability assignment to decision-making bodies; decision traceability and monitoring obligations; and clear thresholds for suspension, correction and public notification when failures occur.

Table 6. Regulatory and ethical recommendations for synthetic data in health

7. Resources, Workforce and Investment Lens

Primary actor: Funders and system capacity planners

This chapter is addressed to public and private funders, health ministries, HTA agencies, and system-level planners responsible for allocating resources, building workforce capacity, and sustaining the infrastructure required for trustworthy synthetic data use. These actors ensure that assurance, review, engagement, and monitoring functions are adequately resourced.

7.1 Capacity and resource gaps

- Shortage of skilled evaluators who can interpret assurance results.
- Under-resourced institutions and regions lack infrastructure and safe settings.
- Engagement, transparency, and monitoring are often unfunded or deprioritised.
- Limited independent review capacity for high stakes uses.
-

7.2 Resource and workforce recommendations

Table 7. translates these challenges into concrete organisational recommendations and operational requirements.

Recommendation	Focus	Core elements
W1. Fund and mandate structured training programmes	Capacity-building	Tailored modules for clinicians, researchers, regulators, HTA assessors, DPOs, ethics committees, and patient/citizen leaders.
W2. Fund "synthetic data translators"	Bridging roles	Dedicated roles connecting technical, clinical, legal and public dimensions of synthetic data use through public–private partnerships and public funding mechanisms.
W3. Protected budgets for assurance and engagement	Sustainable delivery	Treat assurance, independent review, engagement and monitoring as core costs; earmark funding for assurance implementation, transparency reporting, and re-validation.
W4. Communities of practice	Shared learning	Support networks for tools, templates, benchmarks, case studies and collective learning.
W5. Data Generation as a Service	Equity and scale	Trusted shared services providing generation, validation, documentation and controlled release; consistent assurance quality; affordable access for under-resourced regions.

——— Table 7. Resource and workforce recommendations for trustworthy synthetic data use

7.3 Investment opportunities and economic rationale

Strategic investment in synthetic data can:

- accelerate responsible innovation in areas of high unmet need;
- reduce duplication through shared standards and services;
- strengthen Europe's ability to evaluate and govern evidence;
- prevent costly downstream failures by funding trust infrastructure upfront.



8. Citizen, Patient and Societal Lens

Primary actor: Public-facing institutions accountable for legitimacy

This chapter is addressed to health authorities, public data custodians, and institutions responsible for transparency, public engagement, and redress mechanisms. These actors are accountable for maintaining social licence and ensuring that synthetic data deployments align with societal expectations and community interests.

8.1 Public attitudes and expectations

Public trust depends on perceived legitimacy: whether people believe synthetic data is being used for meaningful public benefit, under effective safeguards, and with genuine accountability. Trust is fragile when governance is opaque, benefits are unclear, or communities feel excluded. The following table translates these challenges into concrete organisational recommendations and operational requirements.

Recommendation	Focus	Key elements
S1. Co-design policies and use cases	Participation and legitimacy	Use citizen juries, deliberative workshops and advisory boards to define contextual acceptability and benefit-sharing expectations.
S2. Citizen-facing "transparency minimum dataset"	Meaningful transparency	Publish plain-language disclosures for each deployment covering purpose and use of synthetic data; expected benefits and beneficiaries; safeguards and oversight bodies; limitations and uncertainties; prohibited uses; and routes for concerns and redress.
S3. Feedback and grievance mechanisms	Accountability and redress	Provide accessible channels for questions and concerns; define service-level expectations including response timelines, escalation pathways and public reporting of outcomes.
S4. Ongoing trust metrics	Continuous monitoring	Track awareness and understanding; grievance volume, resolution time and outcomes; audit pass/fail rates and remediation actions; public summaries of use cases approved or rejected; and subgroup performance flags with mitigation actions.

Table 8. Societal recommendations for building and sustaining public trust in synthetic data

9. Implementation Roadmap

Table 9 presents a phased implementation roadmap, distinguishing near-, medium- and longer-term actions.

Time horizon	Key actions	Primary responsible actors
Near term (0–2 years)	• Operationalise validation and assurance templates; • Establish governance boards and citizen/patient panels; • Publish Trust Commitments and mandate the Baseline Assurance Package; • Implement accountability/liability matrices and incident response playbooks; • Launch initial supervised sandboxes/testbeds.	❖ Synthetic data developers and platform operators (validation templates) ❖ Institutional leaders and data controllers (governance boards, accountability structures) ❖ Regulators and Health Technology Assessment (HTA) bodies (adoption of Trust Commitments in high-impact contexts) ❖ Public funders and system planners (resourcing sandboxes and review capacity)
Medium term (2–5 years)	• Align EU and national guidance • Scale sandboxes, benchmarks and independent review capacity; • Embed synthetic data governance into national and cross-border mechanisms; • Expand workforce programmes, communities of practice and shared-service “Data Generation as a Service” models.	❖ European Data Protection Board (EDPB), European Commission services, and national competent authorities (guidance alignment) ❖ EU Health Technology Assessment Coordination Group (HTACG) and national HTA agencies (evaluation criteria harmonisation) ❖ Public funders and ministries of health (capacity-building and shared services) ❖ Institutional governance bodies (embedding oversight structures)
Longer term (5+ years)	• Promote international convergence on assurance and governance; • Expand to broader health system applications with built-in trust and equity mechanisms; • Regularly update guidance based on real-world evidence; • Integrate trust metrics into routine performance reporting.	❖ EU institutions and international regulatory networks ❖ National health authorities and HTA bodies ❖ Institutional governance boards ❖ Public reporting and oversight bodies

Table 9 Phased implementation roadmap



10. Conclusion

Synthetic data has the potential to transform how health systems learn from data while easing tensions between privacy protection and data use. But public trust will not be achieved by the mere fact that data are "synthetic."

Trust must be earned through:

- robust assurance and transparent documentation;
- clear governance with explicit accountability and credible redress;
- independent scrutiny for high-stakes uses;
- sustained investment in skills, infrastructure, and shared services;
- genuine engagement that treats social legitimacy as a delivery requirement.

If synthetic data is governed everywhere in principle but nowhere in detail, adoption will remain uneven and trust will remain fragile. If, instead, synthetic data is integrated into a coherent trust stack- technical, organisational, legal, societal, and operational - *through coordinated action by policymakers, regulators, health systems, and other responsible actors,* Europe can use synthetic data as a lever to strengthen trust in health data sharing and AI, while enabling innovation that delivers tangible benefit for patients and health systems.

Annex 1 - Glossary

Accountability matrix: A structured allocation of roles and responsibilities across the synthetic data lifecycle, specifying obligations for validation, monitoring, documentation, incident response, and liability.

Adversary model: An explicit description of the assumed knowledge, capabilities, and intent of a potential attacker attempting re-identification or inference from synthetic data.

Baseline Assurance Package: A minimum set of documentation and validation materials accompanying a synthetic dataset, including provenance, methodology, privacy risk assessment, utility evaluation, limitations, governance approvals, and permitted uses.

Community engagement: Structured involvement of affected communities, patients, or public representatives in the design, governance, interpretation, and oversight of synthetic data use.

Decision traceability: The ability to document and reconstruct how a synthetic dataset contributed to a model and how that model informed a specific decision context.

Equity-by-design: An approach that integrates fairness, subgroup performance evaluation, and bias mitigation into the development, validation, and governance of synthetic data.

Fidelity: The degree to which synthetic data reproduces the statistical structure, distributions, or relationships present in the source data.

Fitness for purpose: The suitability of synthetic data for a clearly defined intended use, considering accuracy, limitations, uncertainty, and associated risks.

Grievance mechanism: A structured and accessible process through which individuals or communities can raise concerns and seek clarification, correction, or remediation regarding synthetic data use.

Group-level harms: Adverse impacts affecting identifiable communities or demographic groups, even where no individual is directly identifiable.

High-stakes use: Applications in which synthetic data informs regulatory, reimbursement, clinical, or public programme decisions with material consequences for individuals or populations.

HTA-adjacent evidence: Evidence considered in health technology assessment (HTA), reimbursement, or coverage decisions, including economic modelling, comparative effectiveness analysis, and performance evaluation.

Incident response playbook: A predefined set of procedures and communication principles activated in the event of privacy breaches, bias failures, or decision-support errors involving synthetic data.

Independent review: Assessment conducted by an entity not directly involved in the development of the synthetic dataset or model, particularly for uses with elevated impact or risk.

Microdata: Individual-level records, as opposed to aggregated statistics or summary-level information.

Proportionate, risk-based approach: Application of safeguards, validation requirements, and oversight measures calibrated to the level of risk and impact associated with the intended use of synthetic data.

Redress: A formal mechanism through which individuals or communities can seek correction, explanation, remediation, or accountability where synthetic data use has caused harm, error, or justified concern.

Release modality: The way synthetic data is made available to users, such as open publication, licensed sharing, controlled access, or use within a secure analysis environment.

Re-identification risk: The possibility that synthetic data could be linked, directly or indirectly, to real individuals.

Representativeness assessment: Evaluation of whether a synthetic dataset adequately reflects the diversity, distribution, and characteristics of the population relevant to its intended use.

Residual disclosure risk: The remaining risk that individuals or sensitive attributes could be inferred or reconstructed from synthetic data after safeguards have been applied.

Safe settings: Controlled technical environments that restrict, monitor, and log access, use, and outputs of datasets to reduce misuse and disclosure risk.

Seed data: Real-world data (e.g. clinical, research, trial, or administrative data) used to train, calibrate, or generate synthetic datasets.

Social licence: Ongoing public acceptance and perceived legitimacy of data use, extending beyond formal legal compliance.

Subgroup checks: Analyses evaluating whether synthetic data or downstream models perform consistently across relevant demographic, clinical, or population subgroups.

Synthetic data: Artificially generated data that reproduces selected statistical properties or patterns of real-world data without directly containing identifiable records of real individuals.

Synthetic-data-augmented evidence: Evidence in which synthetic data has been used for training, testing, validation, modelling, or augmentation of analytical outputs.

Transparency minimum dataset: A concise, plain-language disclosure summarising the purpose, safeguards, limitations, governance arrangements, and redress mechanisms associated with a synthetic data deployment.

Utility: The extent to which synthetic data is suitable for a specific analytical, modelling, or decision-support task.

Annex 2 - References

Adams, T., Birkenbihl, C., Otte, K., Ng, H. G., Rieling, J. A., Näher, A. F., Sax, U., Prasser, F., & Fröhlich, H. (2025). On the fidelity versus privacy and utility trade-off of synthetic patient data. *iScience*, *28*(5), 112382. https://doi.org/10.1016/j.isci.2025.112382

De Wilde, P., Arora, P., Buarque de Lima Neto, F., Chin, Y., Thinyane, M., Stinckwich, S., Fournier-Tombs, E., & Marwala, T. (2024). Policy Guideline: Recommendations on the use of synthetic data to train AI models. https://unu.edu/publication/recommendations-use-synthetic-data-train-ai-models

Emam, K. E. (2020). Seven Ways to Evaluate the Utility of Synthetic Data. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9138552&utm_source=sciencedirect_contenthosting&getft_integrator=sciencedirect_contenthosting

European Medicines Agency. (2026). *Guiding principles of good AI practice in drug development* https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, *6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3

IHI SYNTHIA. (2025). *Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications.* https://www.ihi-synthia.eu/

Korea, T. P. I. P. C. S. (2024). *The Synthetic Data Creation and Utilisation Guide*. Retrieved from https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=10871

Koul, A., Duran, D., & Hernandez-Boussard, T. (2025). Synthetic data, synthetic trust: navigating data challenges in the digital revolution. *The Lancet Digital Health*, *7*(11). https://doi.org/10.1016/j.landig.2025.100924

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, *8*(1), 60. https://doi.org/10.1038/s41746-024-01359-3

Kulynych, B., Stadler, T., Raisaro, J. L., Troncoso, C. (2026). Should I use Synthetic Data for That? An Analysis of the Suitability of Synthetic Data for Data Sharing and Augmentation. *ArXiv, https://arxiv.org/html/2602.03791v1*

Office for National Statistics. (2021). *Synthetic data policy*. Retrieved from https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/syntheticdatapolicy

Personal Data Protection Commission Singapore. (2025). *Privacy enhancing Technology (PET): Proposed Guide On Synthetic Data Generation*. Retrieved from https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/other-guides/proposed-guide-on-synthetic-data-generation.pdf

Pilgram, L., Fineberg, A., Jonker, E., & El Emam, K. (2025). An assessment of synthetic data generation, use and disclosure under Canadian privacy regulations. *AI and Ethics*, *5*(6), 6225-6240. https://doi.org/10.1007/s43681-025-00819-0

Pilgram, L., Ko, H., Tung, A., & Emam, K. E. (2025). Protecting patient privacy in tabular synthetic health data: a regulatory perspective. *npj Digital Medicine*(8). https://www.nature.com/articles/s41746-025-02112-0

Rajotte, J.-F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., & Strome, E. (2022). Synthetic data as an enabler for machine learning applications in medicine. *iScience*, *25*(11). https://doi.org/10.1016/j.isci.2022.105331

Reich, V., Mitchell, C., Hill, E. R., Myles, P., Branson, R., Pearson, R., & Members of the Expert Working Group. (2025). *Synthetic data for development of AI as a medical device (AIaMDs): Regulatory Considerations* (MHRA & PHG Foundation, Issue. https://doi.org/10.61599/vhac1

Shanley, D., Hogenboom, J., Lysen, F., Wee, L., Lobo Gomes, A., Dekker, A., & Meacham, D. (2024). Getting real about synthetic data ethics. *EMBO Reports*, *25*(5), 2152-2155. https://doi.org/10.1038/s44319-024-00101-0

Vallevik, V. B., Babic, A., Marshall, S. E., Elvatun, S., Brøgger, H. M. B., Alagaratnam, S., Edwin, B., Veeraragavan, N. R., Befring, A. K., & Nygård, J. F. (2024). Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare. *International Journal of Medical Informatics*, *185*, 105413. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2024.105413

World Economic Forum. (2025). *Synthetic Data: The New Data Frontier*. https://reports.weforum.org/docs/WEF_Synthetic_Data_2025.pdf

This publication was prepared by:

Authors

- Denis Horgan, EAPM
- Sofia Pazzagli, EAPM
- Serena Elizabeth Marshall, DNV

Contributors

- Davide Cirillo, BSC
- Olga Startseva, UNIVIE
- Boris Kandov, UNIVIE
- Eleonora Iascone, Humanitas
- Bogdan Kulynych, CHUV
- Stephen Duffield, NICE
- Christophe Maes, i-HD
- Matthias Müllenborn, Novo Nordisk
- Ellen de Waal, MATICAL
- Rea Tambic, Science Affairs

Direct link to the SYNTHIA Policy Engagement Materials:
www.ihi-synthia.eu/stakeholders/policy-makers

September 2026

www.ihi-synthia.eu

IHI SYNTHIA

contact@ihi-synthia.eu

IHI-SYNTHIA

@IHI SYNTHIA

ihi_synthia

@ihi-synthia.bsky.social

SYNTHIA, Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications, is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101172872. The JU receives support from the European Union's Horizon Europe research and innovation programme and COCIR, EFPIA, Europa Bio, MedTech Europe, and Vaccines Europe and DNV. The UK consortium partner, The National Institute for Health and Care Excellence (NICE) is supported by UKRI Grant 10132181.