


SHARE: towards usable, trustworthy and interoperable synthetic health data for rare diseases

Richard Noll ,¹ Philipp Koch,² Benedikt Langenberger,³ Philipp C Stoffers,^{4,5} Ruth Biller,⁶ Andreas Goldschmidt,⁷ Sadegh Mohammadi,⁸ Michele Zoch ,⁹ Gabriela Gan,¹⁰ Benjamin Szilagyi,¹¹ Nicolai Dinh Khang Truong,¹² Richard Röttger,¹² Gennadi Rabinovitch,¹³ Andreas Ekelhart,^{14,15} Daniela Martinez-Duarte,¹⁴ Rudolf Mayer,¹⁴ Holger Storf,¹ Jannik Schaaf¹

To cite: Noll R, Koch P, Langenberger B, et al. SHARE: towards usable, trustworthy and interoperable synthetic health data for rare diseases. *BMJ Health Care Inform* 2026;**33**:e101757. doi:10.1136/bmjhci-2025-101757

Received 08 August 2025
Accepted 11 January 2026

MOTIVATIONS AND GOALS

The growing demand for accessible, high-quality and privacy-preserving health data has led to increased interest in *synthetic health data* as a promising solution to overcome data scarcity and legal barriers.¹ Synthetic data refer to information that has been created artificially to mimic real-world observations. This is particularly relevant in the context of *rare diseases*, where real-world data are often fragmented, siloed or insufficient for robust *artificial intelligence (AI)* development and clinical research.² This paper summarises the outcomes of a multidisciplinary Sandpit workshop involving experts with lived experiences in rare diseases, as well as experts in clinical medicine, data science, cybersecurity and medical informatics. The goal was to define a shared vision and roadmap for a synthetic health data repository (*SHARE*).

STAKEHOLDER BENEFITS

SHARE provides realistic yet privacy-preserving datasets that enable *clinicians, researchers* and *AI developers* to develop, validate and benchmark digital health solutions without accessing real patient records. *Policymakers* and *regulators* gain transparency and reproducibility for algorithm evaluation, supporting safe experimentation within regulatory sandboxes. *Educators, students* and ultimately *patients* benefit from improved training resources and more robust, trustworthy and inclusive digital health innovations.

DESIGN AND IMPLEMENTATION OF THE SYNTHETIC DATA GENERATION APPROACH

As an initial demonstrator, SHARE will focus on *arrhythmic right ventricular cardiomyopathy*

(*ARVC*), a hereditary cardiac disorder characterised by progressive fibrofatty replacement of the right ventricle and a high risk of malignant arrhythmias.³ Because ARVC is rare, heterogeneous and frequently misdiagnosed or underdiagnosed, it represents an ideal use case for synthetic data generation. The objective is to *create synthetic patient cohorts that reflect characteristic disease trajectories*, including ECG abnormalities, imaging and biopsy findings (eg, echo and MRI parameters, endomyocardial biopsy), genetic variants and longitudinal clinical outcomes. Representing these trajectories synthetically may help *clarify diagnostic boundaries* with overlapping conditions such as myocarditis or cardiac sarcoidosis, thereby supporting more *precise diagnostic reasoning*.

To ensure clinical realism and representativeness, SHARE will leverage *multimodal datasets* from expert centres within the *ERN GUARD-Heart network* (<https://guardheart.ern-net.eu/>), which aggregates high-quality genetic data on inherited cardiac diseases across Europe. These real-world ARVC cohorts provide the reference distributions, correlations and temporal dynamics needed to train and validate the synthetic data generation models. The synthetic data pipeline will be built entirely on the *HL7 FHIR (Fast Healthcare Interoperability Resources)* standard.⁴ Structured ARVC data will be represented using modular FHIR resources (<https://build.fhir.org/resourcelist.html>) such as Observation, Condition, Procedure, ImagingStudy and accompanying *terminology bindings* (eg, Human Phenotype Ontology, Systematized Nomenclature of Medicine, Logical Observation Identifiers Names and Codes, Anatomical Therapeutic Chemical/Defined Daily Dose Classification) to ensure



© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to

Dr Richard Noll;
noll@med.uni-frankfurt.de

semantic consistency and *machine-actionable interoperability*. *Generative modelling approaches*, including time-series models, generative adversarial networks, diffusion models and variational autoencoders, will be explored to synthesise the multimodal ARVC data (numerical, categorical and textual elements).⁵ The choice between centralised versus federated generation will be made within a clear governance framework that balances data protection, institutional autonomy and technical feasibility.

Evaluation will focus on fidelity, utility, bias detection and privacy preservation using metrics tailored to multimodal EHR data. We will compare real and synthetic distributions for numerical, categorical, temporal and textual variables, and assess *subgroup-level biases* (eg, sex, age, genotype) through stratified analyses. Utility will be quantified in a clinically meaningful downstream task by training risk-prediction or phenotype-classification models on synthetic data and testing them on real patients (*train-on-synthetic, test-on-real*), assessing whether synthetic trajectories preserve clinically relevant diagnostic boundaries. Privacy risks will be evaluated through *nearest-neighbour analyses* and *membership-inference checks*.

All components will be shared through a two-step publication strategy. A public *GitHub repository* (<https://github.com/>) will host the *codebase* for the generator, FHIR profiles and mappings, documentation and example bundles, enabling collaborative development. In parallel, *versioned releases* of the synthetic ARVC datasets will be deposited on *Zenodo* (<https://zenodo.org/>), providing a DOI-assigned, stable and citable archive. This combination ensures both agile iteration and long-term *FAIR-compliant* (Findable, Accessible, Interoperable and Reusable) accessibility.⁶ SHARE will adhere to the guidance of the European Data Protection Board, according to which synthetic data may fall outside the scope of the *General Data Protection Regulation* if re-identification can be ruled out.^{2 7}

Author affiliations

¹Goethe University Frankfurt, University Medicine Frankfurt, Institute of Medical Informatics, Frankfurt, Germany

²German Research Center for Artificial Intelligence, Lübeck, Germany

³Hasso Plattner Institute, Chair of Digital Health, Economics and Policy, Potsdam, Germany

⁴Chair of Digital Health, Hasso Plattner Institute, Chair of Digital Health, Economics and Policy, Potsdam, Germany

⁵Department of Hepatology and Gastroenterology, Charité -Universitätsmedizin Berlin, Berlin, Germany

⁶ARVC-Selbsthilfe e.V, Unterschleissheim, Germany

⁷Goethe University Frankfurt, University Medicine Frankfurt, Institute for Occupational, Social and Environmental Medicine, Frankfurt am Main, Germany

⁸Medical Imaging & AI, Bayer AG, Leverkusen, Germany

⁹Institute for Medical Informatics and Biometry, TUD Dresden University of Technology, Dresden, Germany

¹⁰Open Innovation in Science Center, Ludwig Boltzmann Gesellschaft, Vienna, Austria

¹¹Data4Trust AG, Maisprach, Switzerland

¹²Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

¹³Institute of Medical Informatics, D4L Data4Life gGmbH, Potsdam, Germany

¹⁴SBA Research, Vienna, Austria

¹⁵Research Group Security and Privacy, ³ University of Vienna, Vienna, Germany

Contributors Richard Noll drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version to be published.

Funding The sandpit was funded by the Wuebben Science Foundation and was conducted as a 3-day event in June 2025. As this was a workshop-format exploratory funding, no grant number was assigned.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer-reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <https://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Richard Noll <https://orcid.org/0000-0002-5240-8484>

Michele Zoch <https://orcid.org/0000-0002-5577-7760>

REFERENCES

- 1 Rajotte JF, Bergen R, Buckeridge DL, *et al*. Synthetic data as an enabler for machine learning applications in medicine. *iScience* 2022;25:105331.
- 2 Mendes JM, Barbar A, Refaie M. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Front Digit Health* 2025;7:1563991.
- 3 Krahn AD, Wilde AAM, Calkins H, *et al*. Arrhythmogenic Right Ventricular Cardiomyopathy. *JACC Clin Electrophysiol* 2022;8:533–53.
- 4 Vorisek CN, Lehne M, Klopfenstein SAI, *et al*. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. *JMIR Med Inform* 2022;10:e35724.
- 5 Bond-Taylor S, Leach A, Long Y, *et al*. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Trans Pattern Anal Mach Intell* 2022;44:7327–47.
- 6 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018:160018.
- 7 European Data Protection Board. Opinion 05/2014 on anonymisation techniques. 2014. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf