# Synthetic data and European General Data Protection Regulation: Ethics, quality and legality of data sharing

Received: 8th May, 2024

**Shalini Dwivedi**
VP, Head of Medical Writing and Clinical Trial Transparency, Krystelis, India

Shalini is VP, Head of Medical Writing and Clinical Trial Transparency at Krystelis. She is an M. Pharm, with a specialisation in phytochemistry. Shalini has over 17 years of academic and clinical research experience. Her current focus includes overseeing the delivery of clinical content development, regulatory writing and clinical trial transparency, including redaction and anonymisation projects for dataset requests, EMA Policy 0070, Health Canada PRCI, and EU-CTR submissions. Shalini has published papers on pharmacy, pharmacognosy, phytochemistry, medical writing and clinical research.

F-42, DLF Ultima, Sector-81, Gurugram, India 122004
Tel: +91 9873407199; E-mail: shalini.dwivedi@krystelis.com

**Abstract** Synthetic data is increasingly being used across the financial services, clinical research, manufacturing and transport industries. In clinical research, use cases for synthetic data include secondary analysis to identify novel treatment pathways, to develop healthcare policies, to evaluate research methods and, importantly, to evaluate research hypotheses without exposing real patients to potentially harmful experimental treatments. Methods for creating synthetic data in a manner that can reconcile the privacy of clinical trial participants while preserving the utility of data for analysis are rapidly evolving. However, challenges remain that include obtaining appropriate consent for the use of real patient data in the creation of synthetic datasets, eliminating bias in synthetic data and ensuring that data privacy concerns can be addressed.

KEYWORDS: synthetic data, GDPR, anonymisation, personal data, data sharing, clinical trial transparency, data privacy, data protection, data transfer, reidentification risk

DOI: 10.69554/LDIY2897

## INTRODUCTION

Several industry and regulatory initiatives have driven the growth in the use of synthetic data in clinical research. Over the last two decades, initiatives have been implemented to promote clinical research transparency. In particular, the European Medicines Agency (EMA) has played a leading role in driving clinical trial data sharing.[1] The EMA launched EMA Policy 0043 in 2010 to support requests to clinical trial sponsors for access to documents related to medicinal products for human and veterinary use. In 2014, the EMA extended the scope to provide public access to clinical documents when it launched Policy 0070 on the publication of clinical data for medicinal products for human use.[2] The main objectives of this policy are to enable public scrutiny and the application of new knowledge to future research. This policy was intended to be implemented in two phases — Phase 1: clinical trial document sharing and Phase 2: sharing of individual patient data. The EMA has defined rules and standards to protect the privacy of clinical

trial participants and others involved in the clinical trial, such as trial support staff. Therefore, all data must be anonymised before publication under the policy. This requires significant effort including conducting a reidentification risk assessment. To support pharmaceutical companies, the EMA also released a detailed guidance document setting out its expectations for the implementation of the policy.[3] Phase 1 of the policy was launched in 2016; however, Phase 2, related to individual patient data sharing is yet to be implemented.

As discussed below, there are numerous benefits of providing wider access to clinical documents and data. Hence, stakeholders including academic journals, research funders and regulatory bodies are driving and supporting transparency initiatives. Industry groups such as the European Federation of Pharmaceutical Industries Association (EFPIA) and the US-focused Pharmaceutical Research and Manufacturer's Association (PhRMA) are also committed to supporting responsible clinical trial data sharing. Despite this, researchers often face difficulties in accessing high-quality clinical data. One study has cited that the success rate for obtaining individual-level data for research projects from authors varies significantly, ranging between 0 per cent and 58 per cent.[4] One barrier to data sharing is increasingly strict data protection regulations, such as the General Data Protection Regulation (GDPR), in the EU.

Creating synthetic data, which cannot be linked to any individual but retains high utility in a specific context, could be an effective solution, although, as described later in this paper, this is not without its data privacy challenges.

## ADVANTAGES OF SYNTHETIC DATA VERSUS ANONYMISED DATA

To meet data privacy requirements when the data needs to be published, pharmaceutical companies apply anonymisation methodologies to eliminate or alter personal information. This process employs techniques such as data aggregation, generalisation, noise-addition, pseudonymisation, date-offsetting and suppression. The primary objective is to modify the data so that it is not possible to associate it directly with an individual. However, literature suggests that a residual risk of reidentification can remain even after robust anonymisation techniques have been applied.[5–7] In addition, the expectation from anonymisation is, that despite modifications, the data should retain a high degree of analytical utility.

The complexity of achieving absolute anonymisation must be acknowledged, given that advancements in data reidentification methods can potentially jeopardise privacy, even with anonymised datasets. As synthetic data originates from artificial sources, privacy concerns normally linked with the use of real-world data could be mitigated.

Synthetic data also minimises data storage costs. As it is easy to create and is prepared 'fit-for-purpose', redundant data need not be collected or stored.[8]

## USE CASES OF SYNTHETIC DATA IN CLINICAL RESEARCH

The application of synthetic data to clinical trials is a significant innovation. It supports regulatory authorities' initiatives and the pharmaceutical industry's commitment to advance clinical research, addresses privacy concerns, encourages data sharing and paves the way for more streamlined, efficient and cost-effective research. Synthetic data has multiple use cases in the clinical research and healthcare industry, some of which are illustrated in Figure 1, and some more detailed possibilities are set out below.

*Conducting secondary research*: Conducting secondary analysis using data from completed clinical studies can be used to validate results and offer new insights, including identifying new drug safety concerns and evaluation of research bias.

*Healthcare policy development*: A real-world evidence study conducted using synthetic data

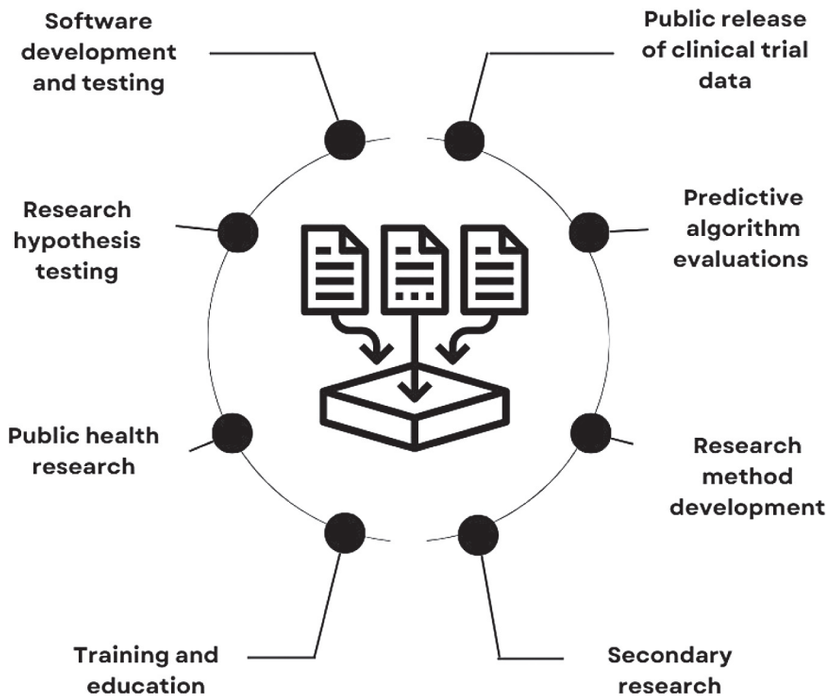


**Figure 1:** Uses of synthetic data

applied a novel micro-simulation approach to successfully create a synthetic dataset from several data sources and presented quantifiable scenario options for healthcare policy development and implementation.[9]

*Identifying and improving patient treatment pathways*: A study utilised a natural language processing (NLP) model trained using synthetic datasets to assess a patient's disease and anticipated treatment pathways.[10]

Synthetic data has also been utilised to build patient-specific models that could increase the effectiveness of treatment programmes and achieve better patient outcomes.[11]

*Research model validation*: Synthetic datasets can help validate research methods, without compromising sensitive patient information, through developing and testing codes, algorithms and statistical methods before deploying them on real datasets.[12]

*Identifying health trends*: Using synthetic data, researchers can combine information from multiple domains (eg genetic data, social determinants of health and electronic health records) to observe new health trends.[13]

*Control groups in clinical trials*: Synthetic data is being used to create control groups for clinical trials when there is a lack of available data, particularly for uncommon or novel diseases. The EMA and the US Food and Drug Administration (FDA) have acknowledged the challenges of studies in rare diseases and small populations and have taken steps to allow innovative methods using synthetic external control data.[14,15]

Researchers can assess the comparative efficacy of treatments using these artificial control groups instead of depending exclusively on patient data.[16]

*To fulfil data sharing requests*: Clinical trial sponsors receive requests from researchers to share data obtained from their clinical trials for secondary research. Such data is likely to contain sensitive information about the trial participants. Additionally, participants may not have consented to share their data

for analysis beyond the scope of the trial in which they have participated. Therefore, despite a commitment to data sharing, the companies may be unable to fulfil the request for data to be used for a different purpose. Synthetic data can be a viable alternative in these cases.

*Creation of digital patient profiles*: Synthetic data can support the creation of digital patient profiles (digital twins) to simulate the characteristics of the target patient population for a clinical trial and help stakeholders forecast patient recruitment challenges, population diversity and optimise inclusion and exclusion criteria. The validity of synthetic data has been evaluated through a synthetic data generator (Synthea) using clinical quality measures. For this, the researchers created a synthetic patient population with the objective of statistically mirroring various parameters of the real population, such as demographics, disease burdens, vaccinations, medical visits and social determinants. The synthetic data generated by Synthea was found to be reliable in modelling demographics and probabilities of services provided in an average healthcare setting.[17]

*Clinical trial design and planning*: Synthetic data has supported decision making during design of clinical trials, selection of investigator sites, feasibility studies for participant enrolment and accurate forecasting of trial enrolment cycle time.[18] By using predictive analytics on the synthetic data, researchers can simulate various scenarios to optimise trial designs. This includes determining the most efficient dosing regimens, identifying potential risks and estimating trial duration. One clinical study utilised real-world data, coupled with deep and innovative analysis using a Trial Accelerator™ platform (Phesi), to improve decision making and ensure patient safety in a hematologic malignancies clinical trial exploring a CAR-T treatment.[19] A key challenge in this trial was to monitor its safety, specifically the incidence and grade of cytokine release syndrome (CRS), a known serious side effect of CAR-T treatment. This approach supported decision making by the pharmaceutical company and was used to guide discussions on reducing patient and investigator site burden, cost and trial duration. This case study is presented in Figure 2.

## PROCESS FOR SYNTHETIC DATA CREATION

Several steps are required to generate synthetic datasets to ensure that they mimic the statistical properties and characteristics of real-world data.[20] These steps vary depending on the techniques and methods employed. Typical steps are summarised in Figure 3.

Initially, real-world data is collected from various sources. This is the foundation for understanding the statistical patterns, structures and relationships that the synthetic data should emulate.

The real-world data is analysed to identify statistical properties, distributions, correlations and patterns and to develop an understanding of the characteristics that need to be replicated in the synthetic data. This is essential for capturing the complexity and diversity of the original dataset. Based on this analysis, a suitable generative model or method is selected to create the equivalent synthetic data. Common approaches for this include statistical models, machine learning models (eg generative adversarial networks or autoencoders) or rule-based methods. The choice of the model depends on the specific requirements and characteristics of the original data.

Machine learning approaches are trained using the real data. During training, the model learns the underlying patterns, structures and relationships present in the original dataset. Once the model is trained, it can be used to generate synthetic data that shares statistical properties with the original dataset. Synthetic

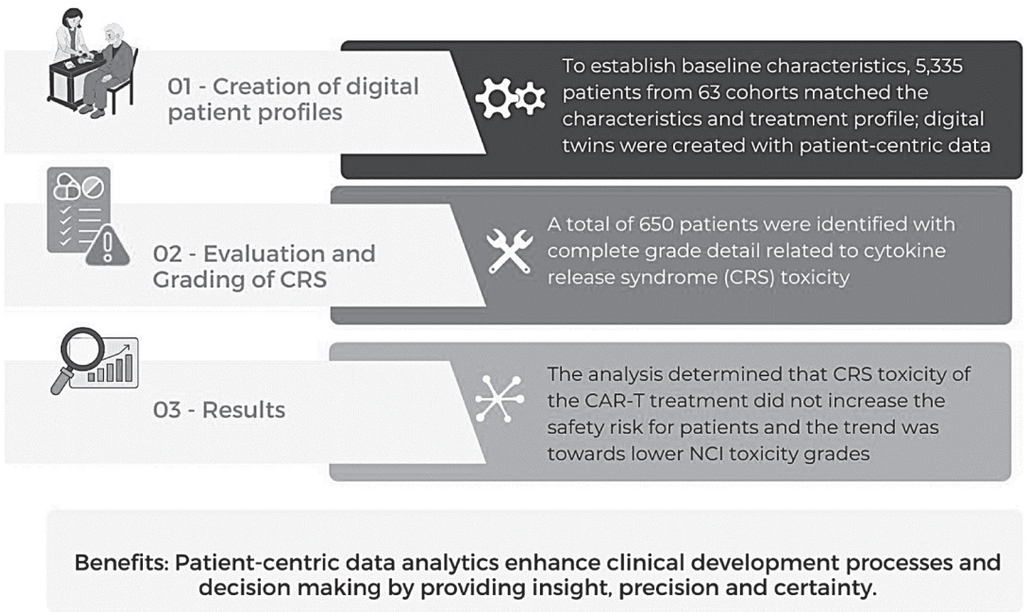


**Figure 2:** Case study: Synthetic data to support healthcare decision making in clinical trials (adapted from Li, G., 'Protocol Planning')

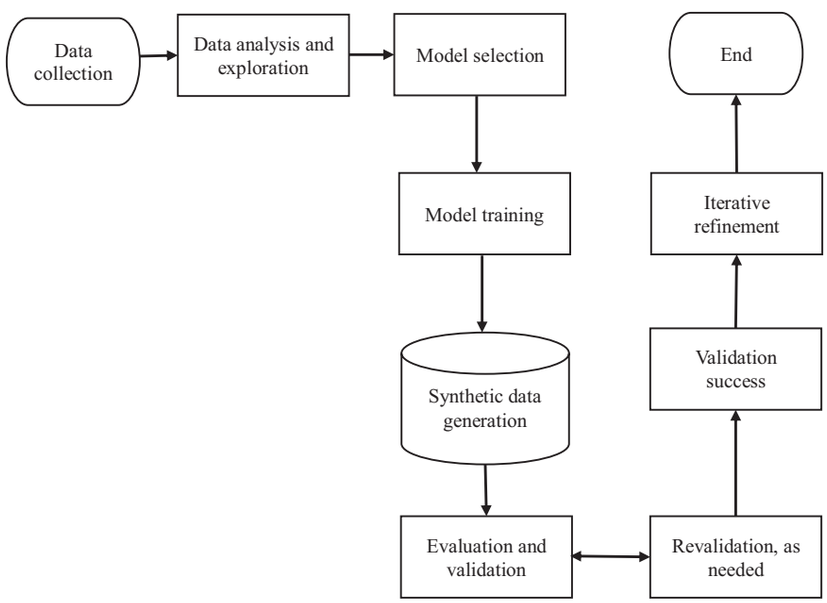


**Figure 3:** Synthetic data generation and validation process

data created in this way should be diverse and representative, capturing the variability observed in the real data.

The generated synthetic data undergoes an evaluation and validation process to ensure its quality and reliability. This involves comparing the statistical properties of the synthetic data with those of the original data, including parameters such as mean, variance and correlation. Although the purpose of

generating synthetic data is to address privacy concerns, additional steps are taken to pseudonymise or anonymise the synthetic data without diminishing the utility of the data.

Three specific methodologies have been used to generate synthetic data using this process — transformation of the data collected, using a data simulator and, generative adversarial networks.[21] Iterative adjustments to the model parameters or the choice of generative model may be made to improve the quality and realism of the synthetic data.

### Factors affecting synthetic data generation

Data complexity, coupled with the diversity of features, relationships and patterns, can increase the difficulty of creating a realistic synthetic dataset. Diverse types of data (eg structured, unstructured) and their format (eg tabular, textual, image) may require the application of specific synthetic data generation techniques tailored to the data characteristics.

If biases were present in the original data, these will be reflected in the synthetic data. In addition, synthetic data cannot accurately reflect all healthcare scenarios and patient demographics if the source data does not comprehensively cover a range of real-world patient populations. Therefore, multiple iterations may be required to identify and remove biases in a synthetic dataset.[22]

The method used to create synthetic data can significantly affect the quality and realism of the generated dataset. Models, such as statistical models (transformation model), machine learning models (eg generative adversarial networks or autoencoders) or rule-based methods, all have their strengths and limitations. In addition, the availability of computational resources, including processing power and memory, can impact the choice of synthetic data generation methods, especially for computationally intensive models.

## LEGALITY OF SYNTHETIC DATA CONSIDERING THE GDPR IN THE EU

In May 2018, the EU implemented GDPR to introduce rules for the protection of natural persons regarding the processing of their data and the free movement of personal data. Article 4 of the GDPR also defines 'personal data' and 'data processing'.[23] The definitions of these terms denote that, before sharing/transferring any personal data of clinical trial participants, it should be rendered anonymous. Data controllers may follow different anonymisation methodologies, but they all must consider the residual risk of reidentification. For clinical trial data sharing, EMA has defined an acceptable risk threshold limit for reidentification of 9 per cent.[24]

To create synthetic data, especially in healthcare and clinical research, original data usually needs to be shared with a third party. Although the GDPR should not apply to the resultant synthetic data, it does apply to the original real data and its transfer. Therefore, the concept of risk inherent in the EU data protection framework and, more specifically, in the concept of personal data — a legal construct for which the risk of reidentification is central to establishing its legal nature — can impact the degree of accessibility to real datasets. As the risk may evolve over time, based on the context and nature of the data release, it is important to understand the legal nature of the synthetic data in the context of the GDPR.

### Personal data

The first part of the 'personal data' definition in the GDPR says '*any information*'. Of course, synthetic data use cases are intended to provide information, regardless of the type and nature of the data. The GDPR also defines 'personal data' but not the 'information'. Contextually, it is important to understand that the word 'data' generally indicates 'structured information' which is more easily understood and communicated.

The second part of the GDPR definition says, 'relating to *an identified or identifiable natural person ('data subject')*', which indicates that if the synthetic data has been generated based on 'any information' from an 'identified or identifiable natural person', it will be considered personal data, unless appropriately anonymised.

The GDPR also includes the concept of pseudonymisation, which is defined as data that cannot be linked to a particular individual without additional information. Nonetheless, if synthetic data is produced using an exact one-to-one transformation of the original dataset in such a way that each data point in the synthetic data parallels its real data point, and the original source characteristics would be substantially maintained, it would be defined as pseudonymous data, and the GDPR would apply.

For the publication or sharing of clinical research data and documents, EMA refers to the Article 29 Working Party recommendations, which suggest three criteria to be fulfilled to ensure successful anonymisation. These are, no possibility of singling out an individual, linking records relating to an individual, and/or making inference concerning an individual.[25]

In the External Guidance for EMA Policy 0070, EMA says:

> . . . data that have been altered using techniques to mitigate risks of re-identification of the individuals concerned but have not attained the threshold required by Article 2(a) and recital 26 of Directive 95/46/EC are not considered anonymised data. Therefore, such approach is only appropriate for limited disclosure for re-use by screened parties but not for public disclosure and re-use under open licence. Recital 26 signifies that to anonymise any data, the data must be stripped of sufficient elements such that the data subject can no longer be identified. More precisely, the data must be processed in such a way that it can no longer be used to identify a natural person by using 'all the means likely reasonably to be used' by either the controller or a third party . . . .[26]

The phrase, 'reasonably likely' weighs strongly here, as it denotes that a risk-assessment should be conducted to ensure that reidentification risk is low. It is therefore important that, while generating synthetic data, care is taken to ensure that the generated data is not attributable to an 'identified or identifiable natural person' and is rendered anonymous.

The third part of the definition says, '. . . reference to an identifier [. . .] identity of that natural person'. Within a clinical trial disclosure context, two main types of identifiers have been defined — *Direct* (eg name, address, phone number, e-mail address) and *Indirect* (eg age, gender, medical history, country of residence). In clinical documents, additional sensitive information about trial participants and family members may be present (including newsworthy information such as a motor vehicle accident, drug abuse history). This must be assessed and protected.

Identity disclosure, attribution disclosure or membership disclosure are the most common privacy risks that could apply to synthetic data.[27] Identity disclosure may occur if the data is partially synthesised through modification of only a few variables, and therefore, one-to-one matching of the synthetic data to the real data is possible.[28] Fully synthetic data is at more risk of inference about an individual through attribution disclosure.[29] Membership disclosure may occur when an adversary has confidence that an individual was in the real dataset from which the synthetic data has been generated.[30,31] Therefore, a reidentification risk assessment of synthetic datasets should be carried out. If no link exists between records in a synthetic dataset and records in the original dataset, data will not be considered as personal data as defined by the GDPR.

### Data processing

The definition of data processing covers aspects of: '. . . collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction'.[32] For these purposes, a 'controller' who determines the purposes and the means of processing of personal data may share the data with a 'processor', who processes the personal data on behalf of the controller, and so, the aspect of data sharing should be considered.

Articles 24 and 28 of the GDPR define the responsibilities of the controller and the processor, respectively. While the 'controller' has the responsibility 'to implement appropriate technical and organisational measures', they are also responsible for selecting a 'processor' who 'can provide sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of GDPR and ensure the protection of the rights of the data subject'.[33]

Based on the above, the contractual obligations agreed between the 'controller' and the 'processor' should consider the latter's responsibility to maintain data confidentiality, follow the data controller's instructions, account for the nature of processing and demonstrate compliance with the GDPR.

Additional GDPR principles to be considered are data minimisation and data quality.[34] As synthetic data is usually generated upon request with a well-defined specific use, data holders should not collect excess data but limit collection to only what is required.

### Data quality

Data quality is a key consideration when clinical data has been requested for independent secondary analysis. Most data sharing requests received by pharmaceutical companies are to support meta-analysis of clinical trial data or for independent reanalysis of clinical trial results.[35] Maintaining an optimal balance that preserves privacy and allows the data to be evaluated for socially and scientifically valuable causes is important. In the clinical research context, the requirement of data sharing is such that it should provide maximum information while protecting privacy, because the main objective of data sharing is to gain more insights into the diseases and new treatments under research. Most anonymisation techniques achieve privacy protection by redaction, noise addition, perturbation or generalisation, which break the association between attributes. However, the process can disrupt meaningful data points and reduce the overall data utility.

Highest utility can be achieved when the synthesised data is indistinguishable from and retains all the statistical properties of the original data. However, this confers a risk of one-to-one matching between the real and the synthetic data. This is especially relevant to a dataset containing information on rare disease trial participants or in situations of multiple patients with different outliers. In a recent paper, Stadler *et al.*, have shown that the data characteristics preserved by generative models can be easily used by an adversary to reconstruct information about an individual.[36] Therefore, similar to the previously described evaluation of reidentification risk for anonymised datasets, risk should also be routinely evaluated for the synthetic datasets.

## CONCLUSION

Synthetic data can be a valuable tool to address privacy concerns, promote data sharing, drive healthcare decision making and support clinical research and development activities. While synthetic

data offers many benefits, it is important to consider the goals of its application, potential risks and the necessity for transparency and documentation in its generation to support its utility in public health and clinical decisions. Clinicians may be reluctant to use synthetic data due to concerns that synthetic datasets do not represent a diverse population. Therefore, data scientists must take care to incorporate diverse data covering wider demographics, geographies and socio-economic backgrounds into synthetic datasets.

Concerns that insufficient ethical and legal restrictions are applied to the creation of synthetic datasets must also be considered, especially in relation to how the input data is processed and interpreted by the data production algorithms. These algorithms may unintentionally reproduce and even amplify biases that exist in the real-world data sources. This issue is particularly sensitive in situations with potential racial or gender biases.[37] Therefore, it is important to conduct a thorough bias analysis of the original personal data to identify and then mitigate any biases. Data augmentation methods, such as synthetic minority oversampling techniques (known as SMOTE) should be utilised to normalise the original data and reduce bias by adjusting for over or under-represented elements.[38] This can help achieve a more representative distribution.

Another challenge is that a one-to-one match between synthetic and original data may be possible and thus compromise data privacy. This is especially true when synthetic data is used for rare disease or small population studies. To control this, anomaly detection methods can be applied to identify outliers and protect data privacy.[39] Differential privacy techniques can prevent the reconstruction of the individual records while still providing accurate aggregate information. Synthetic data augmentation approaches can also help by blending real and synthetic data to minimise privacy risks while maintaining the utility of the dataset. Furthermore, instead of directly synthesising individual records, data scientists can create aggregated features or statistical summaries related to rare diseases. Through a combination of these methods and considering the unique characteristics of rare disease datasets, a synthetic dataset that balances privacy with coherent data utility can be created.

El Emam has discussed seven utility assessment methods for synthetic data: study replication, subjective expert assessment, general utility metrics, evaluation of bias and stability, structural similarity, comparison with privacy enhancing technologies and comparison with public aggregate data.[40] These methods can support the utility assessments and improvements required during synthetic data generation.

An ethical question may arise if artificial intelligence (AI) techniques are used to produce synthetic datasets that lead to the spread of inaccurate information that could have a negative impact on society.[41] This may also occur when synthetic data deviates significantly from reality or inaccuracies lead to erroneous conclusions or decisions. Furthermore, issues related to data ownership and control may arise when synthetic data is generated from proprietary or sensitive datasets. Individuals may have legitimate concerns about sharing their data or how synthetic data generated using their data will be used. Therefore, data scientists and organisations involved in sharing clinical trial participants' data and generating synthetic data must be aware of the GDPR and operate under its legal framework. Regulators and policy makers should evaluate and address the questions of ethical standards and data ownership, control and governance to ensure fair and responsible use of synthetic data by creating appropriate policies and regulatory guidelines.

When synthetic data is used to support decision making, review by an ethics committee could be considered prudent.

Establishing practice guidelines for creating and utilising synthetic data in AI training are important to guarantee its dependability and quality. Furthermore, strong security standards should be formed to safeguard the integrity of AI training procedures and to prevent synthetic data from being traced back to the original data.

As the synthetic data domain continues to evolve, researchers, regulators and institutional review boards will be important stakeholders shaping the guidelines and best practices for its use in healthcare research. This will be supported by initiatives to improve data standardisation, interoperability and data-sharing infrastructure. Researchers will find it easier to work together, share ideas and build on previous research if healthcare data is more freely accessible and compatible. For this, it is important to regulate synthetic data especially data generation models and methods, parameters and correlations between the original data and the synthetic data generated from it.

When any personal data is shared, especially in clinical research, the consent process should be transparent and strengthened to avoid the misuse of synthetic data.[42] Clinical trial participants should be informed about the use of their data for generating synthetic data, and consent should be appropriately obtained. Transparency about the methods used to generate synthetic data and transparency about its intended purpose are also important for building trust.

The factors mentioned in this paper need to be carefully considered to ensure that synthetic data is fit for its intended purpose and aligns with ethical and legal standards. Understanding the specific context in which synthetic data is applied is essential for ensuring its appropriate use and addressing ethical and legal considerations. This will promote responsible innovation, guarantee accountability and increase confidence in research.

## AUTHOR'S NOTE

The author would like to thank Stuart Donald (Krystelis) for his valuable input, critical review and editing of the paper.

## References

1. European Medicines Agency (4th October, 2018) 'Policy 0043: European Medicines Agency Policy on Access to Documents', available at https://www.ema.europa.eu/en/documents/other/policy-43-european-medicines-agency-policy-access-documents_en.pdf (accessed 25th February, 2024).
2. European Medicines Agency (21st March, 2019) 'Policy 0070: Policy on Publication of Clinical Data for Medicinal Products for Human Use', 2019 revision, available at https://www.ema.europa.eu/en/documents/other/policy-70-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf (accessed 25th February, 2024).
3. European Medicines Agency (15th October, 2019) 'External Guidance on the Implementation of the European Medicines Agency Policy on the Publication Of Clinical Data for Medicinal Products for Human Use', Revision 4, available at https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use-version-14_en.pdf (accessed 3rd June, 2024).
4. Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. and El Emam, K. (2021) 'Can Synthetic Data Be a Proxy for Real Clinical Trial Data? A Validation Study', *BMJ Open*, Vol. 11, No. 4, Article e043497.
5. Benitez, K. and Malin, B. (2010) 'Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule', *Journal of the American Medical Informatics Association*, Vol. 17, No. 2, pp. 169–77.
6. Janmey, V. and Elkin, P. L. (2018) 'Re-identification Risk in HIPAA De-identified Datasets: The MVA Attack', in AMIA Annual Symposium Proceedings 2018, Vol. 2018, American Medical Informatics Association, Bethesda, MD, p. 1329.
7. Arbuckle, L. and Ritchie, F. (2019) 'The Five Safes of Risk-based Anonymization', *IEEE Security & Privacy*, Vol. 17, No. 5, pp. 84–9.
8. Gal, M. and Lynskey, O. (2023) 'Synthetic Data: Legal Implications of the Data-Generation Revolution', *109 Iowa Law Review*, LSE Legal Studies Working Paper No. 6/2023, available at http://dx.doi.org/10.2139/ssrn.4414385.
9. Davis, P., Lay-Yee, R. and Pearson, J. (2010) 'Using Micro-simulation to Create A Synthesised Data Set and Test Policy Options: The Case of Health Service Effects under Demographic Ageing', *Health Policy*, Vol. 97, pp. 267–74.
10. Ive, J., Viani, N., Kam, J., Yin, L., Verma, S., Puntis, S., Cardinal, R. N., Roberts, A., Stewart, R. and Velupillai, S. (2020) 'Generation and Evaluation of

Artificial Mental Health Records for Natural Language Processing', *NPJ Digital Medicine*, Vol. 3, Article 69.
11. Giuffrè, M. and Shung, D. L. (2023) 'Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy', *NPJ Digital Medicine*, Vol. 6, Article 186, available at https://doi.org/10.1038/s41746-023-00927-3.
12. Kokosi, T. and Harron, K. (2022) 'Synthetic Data in Medical Research', *BMJ Medicine*, Vol. 1, No. 1, Article e000167. doi:10.1136/bmjmed-2022-000167.
13. *Ibid*.
14. US Food and Drug Administration (September 2022) Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products' available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drug-and-biological-products (accessed 15th March, 2024).
15. European Medicines Agency (2006) 'Clinical Trials in Small Populations — Scientific Guideline', available at https://www.ema.europa.eu/en/clinical-trials-small-populations-scientific-guideline#current-effective-version-9414 (accessed 15th March, 2024).
16. Anju (1st September, 2020) 'Insights on Using Synthetic Data for Clinical Trial Leaders', available at https://www.anjusoftware.com/insights/eclinical/decentralized-trials/using-synthetic-data/ (accessed 6th March, 2024).
17. Chen, J., Chun, D., Patel, M., Chiang, E. and James, J. (2019) 'The Validity of Synthetic Clinical Data: A Validation Study of a Leading Synthetic Data Generator (Synthea) Using Clinical Quality Measures', *BMC Medical Informatics and Decision Making*, Vol. 19, pp. 1–9.
18. Li, G. (n.d.) 'Protocol Planning: Inotuzumab Ozogamicin in Acute Lymphoblastic Leukemia (ALL)', Phesi, available at https://www.phesi.com/case/protocol-planning-inotuzumab-ozogamicin-in-acute-lymphoblastic-leukemia-all/ (accessed 16th March, 2024).
19. Li, G. (n.d.) 'Digital Patient & Digital Twin: CAR-T Cytokine Release Syndrome (CRS)', Phesi, available at https://www.phesi.com/case/protocol-planning-inotuzumab-ozogamicin-in-acute-lymphoblastic-leukemia-all/ (accessed 16th March, 2024).
20. El Emam, K., Mosquera, L. and Hoptroff, R. (2020) 'Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data', O'Reilly Media, Inc, Sebastopol, CA.
21. Wehmeyer, C. (11th February, 2021) 'How Do You Generate Synthetic Data?', Statice, available at https://www.statice.ai/post/how-generate-synthetic-data (accessed 6th March, 2024).
22. QuestionPro (n.d.) 'Synthetic Data in Healthcare: Role in Research & Innovation', available at https://www.questionpro.com/blog/synthetic-data-in-healthcare/ (accessed 15th March, 2024).
23. Parliament and Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), available at https://gdpr-info.eu/ (accessed 26th February, 2024).
24. European Medicines Agency, ref 3 above.
25. Article 29 Data Protection Working Party (2014) 'Opinion 05/2014 on Anonymisation Techniques', available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed 26th February, 2024).
26. European Medicines Agency, ref 3 above.
27. El Emam *et al.*, ref 20 above.
28. El Emam, K., Mosquera, L. and Bass, J. (2020) 'Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation', *Journal of Medical Internet Research*, Vol. 22. No. 11, Article e23139.
29. Elliot, M. (2014) 'Final Report on the Disclosure Risk Associated with the Synthetic Data produced by the SYLLS Team', Manchester University, available at https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf (accessed 3rd June, 2024).
30. El Emam, K., Mosquera, L. and Fang, X. (2022) 'Validating a Membership Disclosure Metric for Synthetic Health Data', *JAMIA Open*, Vol. 5, Article ooac083.
31. El Kababji, S., Mitsakakis, N., Fang, X., Beltran-Bless, A. A., Pond, G., Vandermeer, L., Radhakrishnan, D., Mosquera, L., Paterson, A., Shepherd, L. and Chen, B. (2023) 'Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets', *JCO Clinical Cancer Informatics*, Vol. 7, Article e2300116.
32. GDPR, ref 23 above.
33. *Ibid*.
34. *Ibid*.
35. Unpublished data. Presentation at PHUSE Data Transparency Winter Event by Shalini Dwivedi, Krystelis. February 2024.
36. Stadler, T., Oprisanu, B. and Troncoso, C. (December 2020) 'Synthetic Data — A privacy Mirage', arXiv, available at https://arxiv.org/abs/2011.07018v2 (accessed 3rd June, 2024).
37. Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z. and Tang, H. (2024) 'Synthetic Data in AI: Challenges, Applications, and Ethical Implications', arXiv, available at https://arxiv.org/abs/2401.01629 (accessed 3rd June, 2024).
38. Smith, T. (29th June, 2023) 'Got Data? How SMOTE and GANs Create Synthetic Data', DZone, available at https://dzone.com/articles/got-data-how-smote-and-gans-create-synthetic-data#:~:text=SMOTE%20is%20a%20data%20augmentation%20technique%20that%20is,will%20then%20identify%20k%20of%20its%20nearest%20neighbors. (accessed 6th May, 2024).
39. Mayer, R., Hittmeir, M. and Ekelhart, A. (2020) 'Privacy-preserving Anomaly Detection Using

Synthetic Data', in 'Data and Applications Security and Privacy XXXIV: 34th Annual IFIP WG 11.3 Conference, DBSec 2020, Regensburg, Germany, June 25–26, 2020, Proceedings', Springer International Publishing, Cham, pp. 195–207.
40. El Emam, K. (2020) 'Seven Ways to Evaluate the Utility of Synthetic Data', *IEEE Security and Privacy*, Vol. 18, No. 4, pp. 56–9.
41. Hao *et al.*, ref 37 above.
42. Unpublished data, ref 35 above.