Case study

Governance considerations for the use of synthetic data in health and care research

In the rapidly evolving landscape of health and care research, the integration of synthetic data introduces novel and promising avenues for inquiry, while necessitating careful data and research related governance considerations. Synthetic data, generated through sophisticated algorithms that mimic the statistical properties of real-world data sets without containing any actual patient information, holds potential for advancing research in health and care domains. However, legal, regulatory, and ethical considerations around its creation, disclosure and use requires careful attention.

As researchers increasingly turn to synthetic data to navigate challenges related to privacy, security, and timely data accessibility, we must clarify how governance frameworks apply to its creation and use. Understanding these frameworks ensures that the benefits of using synthetic data encourage innovation and knowledge generation can be realised by researchers, whilst safeguarding the rights and privacy of patients who are the core of health and care research, along with upholding the duty of confidentiality that health and care providers have to them.

Understanding synthetic data

The Information Commissioner's Office (ICO) defines synthetic data as 'artificial' data that replicates patterns and the statistical properties of real data (which may be personal information). Synthetic data is generated from real data using a model trained to reproduce its characteristics and structure, or through redaction (replacing sensitive information) and substitution (replacing sensitive information with irrelevant but plausible data) for simpler data structures. This means that an analysis of the synthetic data should produce very similar results to analysis carried out on the original real data.

Methods for generating synthetic data

There are several approaches that can be used to generate synthetic data. The choice of approach depends on the specific characteristics and requirements of the data, including its dimensionality, structure, and the level of the required privacy protection (dependent in turn on why and how the data will be used), as described in the select examples (methods described are not exhaustive) below:

Generative Adversarial Networks (GANs)

GANs consist of two neural networks, a generator, and a discriminator, which are trained simultaneously. The generator creates synthetic data, and the discriminator evaluates how well the synthetic data resembles the real data (Xie et al., 2018). This adversarial training process results in the generation of realistic synthetic data.

Rule-based approaches

This method involves the generation of synthetic data on predefined rules and constraints and can be effective when specific patterns or relationships in original data need to be retained in the synthetic data (Kaur et al., 2021).

Data masking

This method involves the addition of random noise to numerical values or categorical variables (Goncalves et al., 2020). The approach aims to preserve the general patterns and characteristics of the source data while introducing sufficient variability to protect individual privacy.

With any method for generating synthetic data, there are several evaluation approaches in the literature to make sure that the statistical properties and structures of the original data are sufficiently preserved whilst also protecting privacy. For example, utility metrics derived from the distributions (Maximum Mean Discrepancy (MMD), Hellinger distance (HD), Classifier Two Sample (C2S metric), or measures based on classification performance on real and generated data (Borji, 2019; El emam et al., 2022).

Synthetic data governance considerations

Some synthetic data is neither personal data nor confidential patient information. Unlike personal data under data protection law, confidential patient information, subject to the common law duty of confidentiality, includes any data about National Health Service (NHS) patients (living or dead) from which they are identifiable. Learn more about these distinctions on the Health Research Authority (HRA) website.

This is the case if the synthetic data does not contain any of the underlying real data on which it was trained. This type of synthetic data is not subject to data protection legislation or the common law duty of confidentiality, as an individual cannot be identified solely from analysing such data.

Where data is created artificially from confidential patient information or personal data, the act of creating it through a process of information synthesis is subject to data protection legislation and the common law duty of confidentiality, in the same way that the process of anonymisation is covered by these legal frameworks. Read more about personal definitions on the ICO website What is personal data? | ICO and types of health and care information and the legal frameworks protecting them on the HRA website.

Where synthetic data is generated to be statistically consistent with a real data set that it replaces, an assessment should be carried out regarding the likelihood of individuals being reidentified from the synthesised data by the research team. This assessment should consider the risk of model inversion or membership inference attacks, and attribute disclosure risk. If necessary, additional safeguards may be needed to make sure that any reidentification risks (or other privacy risks) are mitigated until they are sufficiently remote (PHG Foundation, 2023).

Synthetic data governance risk mitigation measures

The methods used to generate synthetic data can raise concerns about residual identification risks and how likely it is that patients could be reidentified by those analysing synthetic data for research purposes. While the methods are powerful tools for creating realistic data, they can potentially retain identifiable characteristics from the original data set.

Therefore, it is crucial that those responsible for synthetic data creation and usage for research purposes do a thorough risk assessment to identify and quantify the risk of reidentification. Practical methods such as linkage attacks and attribute inference attacks can be employed. Linkage attacks attempt to link synthetic records to records in the original data set, while attribute inference attacks attempt to infer sensitive attributes from the synthetic data. In turn, carrying out this risk assessment exercise will help determine whether the level of residual identifiability risk would be sufficiently low to researchers accessing the synthetic data (such that they would not be able to reidentify patients using all means reasonably available to them), rendering the data effectively anonymised.

Case study: leveraging synthetic data for clinically valid lung cancer risk prediction

This case study relates to a research project, limited to the use of data, which generated synthetic data to evaluate the potential utility in developing risk prediction models for lung cancer. The synthetic data, generated from patient data using a Generative Adversarial Network (GAN), replicates the statistical distributions of real data. Researchers can then employ this synthetic data to construct models potentially capable of predicting the risk of lung cancer mortality close to the discriminative accuracy of risk

prediction models developed on real data. In the UK, lung cancer screening trials employ sophisticated algorithms or simple smoking-age criteria for eligibility determination. However, both methods necessitate access to NHS health data, the process of gaining access to which is often a time-consuming and resource-intensive endeavour.

The need for large-scale data

Lung cancer, a complex disease afflicting millions worldwide, demands large-scale, high-quality data sets for accurate analysis and prediction. Enhancing early cancer detection necessitates timely access to real-world data and the exchange of research cohorts. By generating synthetic data that replicates the statistical distributions of real data, researchers can access large, high-quality data sets covering diverse populations against which analytic code can be developed.

Synthetic data: a privacy-preserving solution

This study has demonstrated the feasibility of analysing health data without compromising sensitive information. This approach helps real time health data access, safeguards confidentiality through the application of controls tightly limiting access to raw patient data for research analysis purposes and enables the dissemination of aggregated results. Synthetic data plays a pivotal role in this approach.

Anonymity and statistical similarity evaluation

The study meticulously evaluates both the anonymity and statistical similarity of the generated synthetic data. K-anonymity metrics and the diversity metric proposed by (Alaa et al., 2022) are employed to assess the strength of anonymity techniques applied. These metrics effectively quantify the likelihood of reidentifying individuals from the synthetic data. Wasserstein distance, Jensen-Shannon distances, and alpha-precision and beta-recall metrics are used to evaluate statistical similarity. These metrics can also be used to gauge the synthetic data's retention of the original data set's statistical properties, ensuring accurate representation of the underlying population.

Data privacy considerations

Data security is paramount in the study analysis. The study refrains from permitting access to either the original or synthetic data without prior authorisation and strict adherence to data protection and information governance guidelines. Additionally, data is managed securely in accordance with data privacy principles underpinned by a relevant Data Access Agreement (DAA). Data access is managed through an ISO-certified secure environment within the host firewalls.

Recommendations for synthetic data release

The following measures are suggested by the study in the event of synthetic data release:

- Restrict access to bona fide researchers, i.e., only verified researchers with a legitimate research purpose and a demonstrated understanding of data privacy concerns should be granted access to the synthetic data.
- Implementation of a DAA. A concise DAA should outline contractual obligations, including explicit prohibitions on sharing, reidentification, or linking of the data sets. This agreement serves as a legally binding document to safeguard the privacy of the anonymised data and prevent potential misuse.

Conclusion

Advances in machine learning have enabled the generation of high-quality synthetic data sets, paving the way for clinical risk prediction models nearly as accurate as those developed using original data. This opens the possibility of generating readily available data sets for model development and the approach allows for model benchmarking while preserving real-world data sets for validation and calibration. Generating anonymous synthetic data to be shared, has the potential to expedites researchers' access to data, enabling near real-time health insights. This system aligns better with patients’ expectations and data minimisation requirements. By eliminating the need to transfer original health data between organisations, the process enhances security and maintains patient data confidentiality.

This study highlights the potential of synthetic data as a valuable resource in lung cancer research. By replicating the statistical characteristics of real patient data, synthetic data sets potentially offer a viable alternative for researchers to overcome privacy and data accessibility challenges. Existing synthetic data generators allow for the development of synthetic data sets that mimic characteristics but not copy real data. This capability signifies a move towards a simplification of data and research governance issues faced by researchers when using health data due to lower identifiability risks, so preserving individual privacy. Hence, it validates the utility of research incorporating synthetic data to generate meaningful insights. However, it is essential to acknowledge that while synthetic data can approximate real-world scenarios, such data may not fully capture the complexities and nuances of patient specific cases.

As with all anonymisation techniques, there is a trade-off between data utility and privacy protection that researchers must think about, which will ultimately be

determined by them depending on their overall research purposes. Furthermore, synthetic data generated from real data may inculcate biases present in the original data (Hao, S., 2024). For instance, data primarily based on individuals from specific racial or age groups may not be suitable for use in other groups. The same cautions and mitigations should therefore apply.

References

Alaa, A., Van Breugel, B., Saveliev, E.S. and van der Schaar, M., 2022. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative modelsv. Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR pp. 290-306.

Borji, A., 2019. Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding, 179, pp. 41-65.

Hao, S., 2024. Synthetic data in AI: challenges, applications, and ethical implications. Available at: https://arxiv.org/abs/2401.01629 (Accessed: 8 August 2024).

Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A. and Markuzon, N., 2021. Application of Bayesian networks to generate synthetic health data. Journal of the American Medical Informatics Association, 28(4), pp. 801-811.

El Emam, K., Mosquera, L., Fang, X. and El-Hussuna, A., 2022. Utility metrics for evaluating synthetic health data generation methods: validation study. JMIR Medical Informatics, 10(4), p. e35734.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L. and Sales, A.P., 2020. Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20(1), pp. 1-40.

PHG Foundation, Mitchell, C. and Hill, R.H., 2023. Are synthetic health data 'personal data'? Available at: https://www.phgfoundation.org/ (Accessed: 22 November 2023).

Xie, L., Lin, K., Wang, S., Wang, F. and Zhou, J., 2018. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739.

Disclaimer

This case study was developed by the Health Research Authority (HRA) in collaboration with the Information Commissioners Office (ICO) and features a data only research project which used synthetic data to develop a risk prediction model for lung cancer. This case study is not an endorsement for the use of synthetic data for product development and/or safety assessment purposes of medical device products.

This case study is intended to provide insights into individual experiences but does not reflect the views or recommendations of the AI and Digital Regulations Service partners (AIDRS). AIDRS emphasises that users should continue to seek and adhere to formal statutory guidance and legal requirements applicable to their specific circumstances. It is the responsibility of the legal manufacturer to comply with all applicable statutory regulations.