Synthetic data and the innovation, assessment, and regulation of AI medical devices

Puja Myles, MPH, PhD
Johan Ordish, MA
Richard Branson, MSc, MA

Synthetic data are artificial data that mimic the properties of and relationships in real data. It shows promise for facilitating data access, validation, and benchmarking, addressing missing data and under-sampling, sample boosting, and the creation of control arms in clinical trials. The UK Medicines and Healthcare products Regulatory Agency (MHRA) is using its current research into the development of high-fidelity synthetic data, to develop its regulatory position on AI medical devices trained on synthetic data, and on synthetic data as a tool for the validation and benchmarking of AI medical devices.

Introduction

The predicted rise of artificial intelligence (AI) for health and social care implies that AI as a medical device (AIaMD) will become an ever more prominent subcategory of medical device.[1] It is therefore increasingly important that medical device regulations are fit for purpose for AI and that manufacturers understand as well as comply with their obligations, chief of which is to demonstrate a favourable benefit-risk ratio for their AIaMD.[2] Robust datasets are core to demonstrating the performance of AIaMD and are often the chief blocker to the development of such devices.[3] It is incumbent on medical device regulators to ensure manufacturers have at their disposal the tools necessary to comply with these obligations and provide wider support to encourage the development of such innovative devices. The development of synthetic datasets may well constitute such an assistive tool. This paper outlines efforts by the MHRA to research and develop synthetic data and consider its use in the context of wider reforms to ensure medical device regulation is fit for purpose for AI.

The synthetic data landscape

Recent years have witnessed a growing interest in synthetic data due to a number of factors, including potential ease of access in a world with stricter data governance regulations, protection of patient privacy, benchmarking and validation capabilities in the context of machine-learning algorithms, and the ability to address limitations of real data, such as missing data, under-sampling, and small samples.[4] More importantly, although the potential applications of synthetic data have been discussed for many years, it is only recently that methodologic advancements in synthetic data generation have been able to yield high-quality synthetic data.[5]

Defining synthetic data

Conceptually, synthetic data are artificial data that mimics the properties of and relationships in real data. The quality of synthetic data depends on the approach taken to synthetic data generation. The quality of synthetic data is often described in terms of its "utility" or "fidelity." A synthetic dataset that is able to capture complex inter-relationships between various data fields and the statistical properties of real data can be referred to as a "high-utility" or "high-fidelity" synthetic dataset. In the context of patient health care data, a high-fidelity synthetic dataset would be able to capture complex clinical relationships and be clinically indistinguishable from real patient data. The generation of high-utility synthetic data tends to be highly resource intensive and depending on the application for which synthetic data are required, it may be acceptable to use low or medium utility synthetic data.

It is important to note that there is a trade-off between utility and privacy when generating synthetic datasets. While the general assumption is that a high-utility synthetic data are associated with a

higher privacy risk because it is closer to the real data, this may not always be the case and is dependent on the synthetic data generation approach used.

Synthetic data generation approaches

Synthetic data generation methods can be broadly categorized into three groups: generating synthetic data based on statistical properties of real data; adding noise to real data; and using machine-learning techniques to generate synthetic data.

Generating synthetic data based on statistical properties of real data

This approach relies on statistical properties of real data such as population distributions – for example, mean values, standard deviation, and value ranges for a data field such as blood pressure or blood glucose measurements or known prevalence of a disease in various subgroups. This approach is useful when the real data are difficult to access, or the distribution of events is highly imbalanced in the available real data sample. A key limitation of this approach is that, while each synthetic data field will have the statistical properties of real data, the complex relationship between data fields will be difficult to capture. Synthetic data fields generated using this approach will be completely artificial and thus, should not pose a privacy risk.

Adding noise to real data

This approach involves perturbation of some of the data fields in real data in different ways including substitution of real values with other realistic values, random shuffling of data values within a particular data field or application of a random numeric variance (e.g., plus/minus 10% to all data values in a field such that the data distribution is preserved). Substitution of real values can also be approached by swapping data within a data field with another sample from the same distribution.[6] These techniques can provide medium utility data while posing a slightly higher privacy risk as compared with other synthetic generation methods.

Machine-learning techniques to generate synthetic data

These techniques can be used to generate both semisynthetic and fully synthetic data. Machine-learning techniques include Hidden Markov models, Bayesian networks (BNs), and deep-learning approaches such as generative adversarial networks (GANs) to learn patterns in real data. The learned patterns are then used as an input for the synthetic data generator to yield synthetic data. The utility of fully synthetic data generated using machine-learning techniques tends to be high because they are able to capture complex relationships between various data fields, with low privacy risks.

The actual choice of machine-learning algorithm is dependent on the specific requirements for synthetic data. For instance, when transparency is a key requirement, BN approaches are preferable to GANs. Unpublished findings from the MHRA's synthetic data research team suggest that GAN-based approaches may perform better than BN approaches for numerical data fields and vice versa for categorical/nominal data fields. Hidden Markov models on the other hand, have been particularly useful for taking into account missing values in the real data.[4,7]

The cost of synthetic data generation

Based on the MHRA's own experience of generating synthetic datasets, the cost of generating synthetic data is inversely proportional to both the fidelity and level of privacy assurances required. In general, machine learning approaches to generating synthetic data are more expensive because they require more computational power. The agency has not noted a meaningful difference in processing requirements for GAN versus BN approaches in head-to-head comparisons based on the same real data extract (ground truth data). However, the use of structural expectation maximization to model and recreate missing data in the ground truth data is computationally very expensive. The

nature of data being generated also influences the cost, for example, learning the structure and relationships in a large ground truth dataset that has millions of observations will require more computational power. A complex ground truth dataset that needs some customization of the synthetic data generation methods would additionally incur a greater human resource cost.

Another way of considering costs is by comparing the costs of synthetic data generation with the costs of collecting and curating comparable real data. This depends on the real data source and our reflections are based on the main data source we have used, a routinely collected UK primary care electronic healthcare record database that is made available for secondary research in an anonymized format within an established data governance framework. In general, generating synthetic data based on publicly available statistical properties of real data is the least expensive approach to synthetic data generation with the caveat that this will not yield high-fidelity synthetic data. This approach would be cheaper than accessing real data. Perturbation approaches will invariably be slightly more expensive than the real data to which perturbation has been applied as there is a base data cost. This is because perturbation approaches could be viewed as value-added services applied to real data, generally, for privacy preservation.

On the one hand, machine learning-based approaches can be significantly more expensive than real data given the current technology and our resource modelling suggests that synthetic data generated using these approaches can be between two to four times the cost of comparable real data. On the other hand, if these methods are used to generate synthetic data that can boost sample sizes in a clinical trial, there may be significant cost-efficiencies. However, as will be discussed in the next section, this particular synthetic data application needs further development and testing. Thus, synthetic data is not necessarily a cheaper option to real data; rather, high-fidelity synthetic data should be viewed as a valuable tool for specific applications where real data is not available, challenging to procure, or unsuitable.

Potential applications of synthetic data

This section provides a more detailed overview of the most promising synthetic data applications.

Facilitating data access

Access to individual patient level data is subject to strict data governance requirements and can be challenging for the health technology sector. In some instances, it may be possible to justify access to these data based on a clear public benefit from the use of these data, for example, to support the development of risk prediction algorithms that improve patient care and outcomes. However, even in such cases, once a promising algorithm is developed, there may be restrictions preventing the use of real data for backend software developments to create a risk prediction application on this basis. The decision on whether low-, medium-, or high-fidelity synthetic data should be used is dependent on the intended application. Where the synthetic data are being used as a proxy for real data for exploratory analyses or to train a machine-learning algorithm, it would need to be high fidelity. Where the key requirement is to understand the structure of the data, low- or medium-fidelity synthetic data may be sufficient. Open-source medium-fidelity synthetic data could also be used to design analyses and create analysis programmes that could then be run on the real data by researchers working within, or in partnership with, the data controller organisation.[8]

Validation and benchmarking

High-fidelity synthetic data could be especially valuable for validation of machine-learning algorithms that are trained using real data when alternative real data are not available for external validation purposes. In such cases, for meaningful external validation, the synthetic data should not merely replicate the real data used for training the machine-learning algorithm. Synthetic data used for validation purposes should either be based on a different real data source or be generated to be

intentionally distinct from the training data to reveal potential biases in algorithm performance. This can be accomplished by using a conditional generation approach to synthetic data generation so that it corrects for known or suspected biases in the real data. Synthetic data can be particularly helpful in correcting biases due to under-sampling of certain population subgroups as is discussed in the next subsection.

Addressing missing data and under-sampling

Missing data is a common issue with health care datasets and can present in two ways. The first is as missing values within a specific data field, for example, body mass index values may be missing for a proportion of subjects in the dataset. The second type of missing data issue could be viewed as an under-sampling problem whereby certain population subgroups are completely absent from the data – for example, very elderly patients, or patients from a particular ethnic group. Many methods have been developed to deal with missing values in real data, and one such technique is called multiple imputation (MI). MI essentially infers the missing value based on other known characteristics of the patient, based on comparable patient data. Thus, some values for some patients in the real data would be "synthetic."

Further work has been undertaken by Draghi et al. (2022)[9] to explore whether synthetic data could be used to address the under-sampling problem, more specifically, by correcting for biases due to under-sampling by boosting underrepresented groups using synthetic data. The authors validated their work by using a biased subset of data and comparing the bias-corrected data to the 'full' dataset. Additional testing of their approach using other datasets would be ideal but this initial work itself is very promising and could help improve the generalisability of any machine-learning algorithms to population subgroups that are missing in real data.

Sample boosting

Synthetic data may be used to boost the sample size in scenarios when the real data are based on a small sample. Wang and colleagues[4] demonstrated experimentally how synthetic data could be used to increase the sample size ten-fold, from 583 to 5,830 subjects. They were able to show that the correlational direction between the data variables was well preserved even after the scaling up of size. This application could be extremely beneficial when dealing with small patient cohorts relating to rare diseases and outcomes. However, further experimental evidence is required to establish whether the increase in sample size is informative and adds statistical value over and above the real data.

In silico clinical trial methods

Finally, synthetic data generation methodologies could be applied to generate virtual patient cohorts for clinical trials. This builds on the concept of in silico trials, which use computer simulation methods to demonstrate efficacy and safety of medical products compared with traditional experimental (in vitro or in vivo) approaches to evidence generation. In silico trials may use historical or contemporary data from other clinical trials or real-world data sources to create virtual patient cohorts. While in silico trial methods have been considered in the context of medical devices as a whole rather than specifically in the context of AIaMD, there may be opportunities to apply these methods to this sub-domain in the future.

Regulatory applications

From a regulatory perspective, there are two primary questions to consider how synthetic data might fit with AIaMD: Under what circumstances (if any) would medical device regulators and approved bodies accept AIaMD trained on synthetic data versus ground truth data; and, as outlined above, are there opportunities to use synthetic data for regulatory purposes, for instance, to validate or benchmark AIaMD? We consider each in turn.

With respect to the acceptability of synthetic data as training data for AIaMD, the MHRA is in the early stages of considering under what circumstances that would be acceptable, what requirements might be necessary for the synthetic dataset itself (such as fidelity), and what supportive guidance might assist manufacturers exploring the use of synthetic data for this purpose and underpin future research. First, real data are currently the default option and will likely remain the default option for the foreseeable future. Second, MHRA is open-minded about the use of synthetic data for validation purposes while acknowledging that more experimental exemplars are needed before this is accepted as the norm. Third, it is important that synthetic data not compound already existing issues, becoming an "old lady that swallowed a fly" problem, where demonstrating the fidelity of a synthetic dataset becomes more problematic than issues of validating in a comparable real dataset. In short, while it has become plain that generating high-fidelity synthetic datasets for a variety of data types is possible, what is less clear is the position of these datasets within a safety-critical space such as AIaMD. It is this clarity that MHRA hopes to bring in the not-so-distant future.

Innovation should not be limited to the market which is being regulated but also present in the methods and thinking of the regulator itself. Given this, the MHRA is exploring whether synthetic data might be used to "validate" or "benchmark" AIaMD. There are a number of elements necessary to consider the position of synthetic data as a regulatory tool. First, are there "upstream" issues that would block the effective use of synthetic data as a regulatory tool. For instance, there are multiple shades of meaning of "validation" of AI systems, in which case, the precise meaning of validation or further specificity of what benchmarking is required of AIaMD needs to be sought to unlock the potential of any innovative tools.[10] Second, for what precise purpose would synthetic data be useful? For example, one of the primary challenges for AI is the generalisability (or lack thereof) of models. Generalizability translates in the context of medical device regulation as the requirement that clinical evidence demonstrates that the device is fit for purpose in all populations in which the device is intended to be used.[11] In this respect, the malleability of synthetic data might transcend the limitations of ground truth data to test whether a model remains robust under a variety of scenarios and in different population subgroups. Given the above, the MHRA is cautiously optimistic that synthetic data might not only be of use for manufacturers but also as a regulatory tool to assist with validation and benchmarking of AIaMD.

Conclusion

High-fidelity synthetic data has the potential to assist with some of the major challenges of AIaMD. Namely, facilitating access to training data often without such a steep cost to patient privacy, providing better access to validation or benchmarking datasets, filling gaps in data that would otherwise exist, and boosting sample sizes. In addition, there are two primary regulatory questions that emerge from such techniques. That is, under what circumstances, if any, would it be acceptable for AIaMD to be trained or tested upon synthetic data versus real data, and are there opportunities to use synthetic data to better validate or test AIaMD models? The MHRA is committed to exploring such opportunities and questions to ensure synthetic data supports the burgeoning UK AIaMD market.

Abbreviations

AI, artificial intelligence; **AIaMD,** artificial intelligence as a medical device; **BN,** Bayesian networks; **GAN,** generative adversarial networks; **MHRA,** Medicines and Healthcare products Regulatory Agency; **MI,** multiple imputation

About the authors

Puja Myles, MPH, PhD, is Director of the MHRA's specialist research data services centre, Clinical Practice Research Datalink (CPRD). She initially joined the MHRA as Head of Observational Research, CPRD in 2017 and prior to this, trained as a public health specialist and was a public health academic at the University of Nottingham, UK. She is a fellow of the Faculty of Public Health (UK), a senior fellow of the Higher Education Academy (UK), and has a doctorate in epidemiology. She can be contacted at puja.myles@mhra.gov.uk

Johan Ordish, MA, is Head of Software and AI, Innovative Devices, MHRA. Software Group is responsible for most aspects of regulating software as a medical device in the UK. He is also an Honorary Associate Professor in the College of Medical and Dental Sciences, University of Birmingham and an Associate of Hughes Hall, University of Cambridge. He has a law degree and a postgraduate qualification in political sciences. He can be contacted at johan.ordish@mhra.gov.uk

Richard Branson, MSc, MA, is Analysis, Planning and Reporting Specialist, Innovative Devices, MHRA and has project managed the MHRA's synthetic data research work. He has two masters' degrees, in business administration and psychoanalytic approaches to consultation and the organisation. He can be contacted at richard.branson@mhra.gov.uk

References

1. Topol E. High-performance medicine: The convergence of human and artificial intelligence. https://www.nature.com/articles/s41591-018-0300-7 Published online 7 January 2019. Accessed 28 April 2021.
2. European Commission. Council Directive 93/42/EEC, Annex I. https://www.legislation.gov.uk/eudr/1993/42/annex/I Last updated 31 December 2020. Accessed 6 May 2020.
3. Vollmer et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. https://www.bmj.com/content/368/bmj.l6927 Published 20 March 2020. Accessed 28 April 2021.
4. Wang Z, et al. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. https://onlinelibrary.wiley.com/doi/10.1111/coin.12427 Published online 3 January 2120. Accessed 28 April 2021.
5. El Emam K. Accelerating AI with synthetic data. O'Reilly; 2020.
6. Patel N, Patel S. A study on data perturbation techniques in privacy preserving data mining. https://www.irjet.net/archives/V2/i9/IRJET-V2I9242.pdf Published December 2015. Accessed 28 April 2021.
7. Tucker A, et al. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. https://www.nature.com/articles/s41746-020-00353-9 Published 9 November 2020. Accessed 28 April 2021.
8. Health Data Insight. The Simulacrum. https://healthdatainsight.org.uk/project/the-simulacrum/ First released, 28 November 2018. Accessed 28 April 2021.
9. Draghi, B., Wang, Z., Myles, P., Tucker, A. (2022). BayesBoost: Identifying and Handling Bias Using Synthetic Data Generators. SSRN. Bayesboost: Identifying and Handling Bias Using Synthetic Data Generators by Barbara Draghi, Zhenchen Wang, Puja Myles, Allan Tucker :: SSRN. Published 8 March 2022. Accessed 14 November 2022.
10. Hand D, Khan S. Validating and verifying AI systems. https://www.sciencedirect.com/science/article/pii/S2666389920300428 Published 12 June 2020. Accessed 28 April 2021.
11. European Commission. Guidelines on medical devices [MEDDEV 2.7/1, revision 4; June 2016]. https://www.medical-device-regulation.eu/wp-content/uploads/2019/05/2_7_1_rev4_en.pdf Dated June 2016. Accessed 28 April 2021.