

Synthetic patient data in health care: a widening legal loophole

Artificial intelligence in health care has received widespread attention, focused on deductive systems that analyse datasets to learn patterns that would be infeasible to programme. However, more recently, another form of artificial intelligence has emerged: generative adversarial networks (GANs). GANs is a form of artificial intelligence with the purpose of creating high-fidelity fake data. The artificial intelligence system is provided with a dataset of real data and learns to produce new data that retains the overall properties of the original dataset but is artificial. Synthetic data has received considerable attention as a method of protecting patient privacy and augmenting clinical research. Synthetic data carries the ability to create fake patient records and fake medical imaging that is truly non-identifiable because the data does not relate to any real individual. In a sense, the synthetic data is a derivative of the original real data but no synthetic datapoint can be attributed to a single real datapoint.¹

The creation of synthetic data carries great promise to protect patient privacy, diversify datasets, and enhance clinical research. However, the rise of synthetic data has also heralded an industry of companies seeking to monetise fake data and enable cross-border data sharing beyond the confines of data protection legislation.² Concerningly, there are no robust and objective methods of determining whether a synthetic dataset is sufficiently different from the original real dataset to be classed as truly anonymous. The absence of legislation covering synthetic data presents potential risks to consumers. For example, it could give insurance companies free rein to

buy and sell synthetic consumer data that is technically non-identifiable but retains all the properties of the original dataset required to adjust premiums for specific consumer groups. Furthermore, although technological companies are bound by data protection legislation when handling customer data for targeted advertising, there are no obvious restrictions to disseminating synthetic representations of such sensitive data.

At present, there is no clear legislation surrounding the use of synthetic data. Although there has been emerging interest in the positive use cases of synthetic data, it is also important that consumers and policy makers are aware of the potential drawbacks. It has been predicted that synthetic data for building algorithms will overtake real data by 2030 and financial investment into synthetic data is increasing rapidly.^{2,3} The potential benefits and risks of synthetic data might be unclear, but that does not mean they are unlikely to occur.

AnmA declares honorary research roles with NHS England and NHS Improvement and Moorfield's Eye Hospital as well as panel membership for the National Institute for Health Research and Health Data Research UK. AnaA declares no competing interests.

***Anmol Arora, Ananya Arora**
aa957@cam.ac.uk

School of Clinical Medicine, University of Cambridge, Cambridge, CB2 0SP, UK

- 1 Zerdick T. Is the future of privacy synthetic? European Data Protection Supervisor. July 14, 2021. https://edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en (accessed Jan 28, 2022).
- 2 Andrews G. What is synthetic data? June 8, 2021. <https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/> (accessed Jan 28, 2022).
- 3 Schoettler J. Council post: what investors need to know about "fake" data companies. Forbes. May 10, 2021. <https://www.forbes.com/sites/forbesbusinesscouncil/2021/05/10/what-investors-need-to-know-about-fake-data-companies/> (accessed Jan 28, 2022).



Published Online
March 28, 2022
[https://doi.org/10.1016/S0140-6736\(22\)00232-X](https://doi.org/10.1016/S0140-6736(22)00232-X)

Submissions should be made via our electronic submission system at <http://ees.elsevier.com/thelancet/>