Synthetic data: Promise and hidden peril in academic research

NEIL TANNEN AND VICTOR LOBO

In the rapidly evolving landscape of artificial intelligence and data science, synthetic data has emerged as a groundbreaking tool in academic research. But what exactly is synthetic data? Simply put, it is artificially generated data designed to mimic real-world data, often used when actual data is difficult to obtain or privacy is a concern. For example, a dataset simulating medical records can help doctors train AI models without compromising patient confidentiality. While synthetic data appears to offer a solution to data accessibility and privacy challenges, its hidden risks threaten the accuracy, ethics, and credibility of academic research.

Synthetic data enables researchers to create datasets resembling real-world data using techniques like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). It is increasingly employed across fields such as healthcare, political science, and economics, allowing researchers to work without breaching privacy. However, beneath this seemingly risk-free innovation lies a host of problems that, if not carefully managed, could undermine the very foundations of academic research.

One of the main advantages of synthetic data is its ability to protect privacy, especially in sensitive fields like healthcare. Researchers can use synthetic medical datasets to study diseases such as cancer or heart disease without revealing personal patient information. Institutions like the Mayo Clinic have even tested synthetic health records to train AI for diagnosing various conditions. Yet, as a 2023 report in *The Lancet Digital Health* reveals, synthetic data often fails to capture the full complexity of human biology. Without critical biological and environmental variations, AI models trained on synthetic data may overlook rare diseases or misclassify conditions, potentially leading to harmful misdiagnoses. While privacy is vital, prioritising it over data accuracy in medical research can endanger lives.

In political science, synthetic data has been used to forecast voter behaviour, but it carries its own set of pitfalls. A 2022 study in the *Journal of Political Science Methods* examined synthetic voter data used to predict US election outcomes. The models, based on past voting trends, failed to account for new political dynamics or shifts like the Covid-19 pandemic. As a result, predictions were off by 16%, skewing forecasts in favour of traditional political parties and ignoring the rising influence of independent candidates. This example highlights the dangers of relying on synthetic data to draw conclusions from historically biased or incomplete datasets, as it risks reinforcing outdated ideas and producing inaccurate insights.

Beyond issues of bias and accuracy, synthetic data poses a serious threat to the reproducibility of academic research, a cornerstone of scientific integrity. A 2021 study in *Nature Machine Intelligence* found that nearly half of the 50 papers reviewed, which used synthetic data, couldn't be replicated because the data was either unavailable or generated using proprietary methods. This lack of transparency makes it impossible for other scholars to verify or replicate findings, eroding the credibility of such research. Clearer guidelines are needed to ensure that synthetic datasets and the methods used to create them are openly disclosed, enabling peer review and preserving scientific rigour.

Synthetic data is also increasingly used in economic research and policymaking, where it is used to model trends such as inflation or labour market dynamics. However, errors in these predictions can lead to misguided policies with real-world consequences. For instance, the Bank of England used synthetic labour market data to forecast a quick post-pandemic recovery, only to find that unemployment remained stubbornly high, especially in vulnerable sectors. This discrepancy arose because synthetic models failed to account for important variables, such as inflation and worker migration. As a result, policies based on these flawed models had serious implications, underscoring the risks of over-reliance on synthetic data in public policy decisions.

Ethical concerns further complicate the use of synthetic data. Since the algorithms generating synthetic data are often proprietary and not transparent, questions about accountability and responsibility arise. Who verifies the accuracy of this data? Should journals require full disclosure of synthetic datasets used in research? And if research based on synthetic data leads to false conclusions, who is held accountable? These ethical dilemmas came to a head in 2023 when a researcher at a major European university fabricated synthetic climate change data to support a predetermined hypothesis. This scandal serves as a stark reminder of how synthetic data can be manipulated to advance misleading narratives, making academic fraud harder to detect.

Despite these challenges, synthetic data is not without merit. It holds great promise in areas where privacy and data scarcity are significant concerns. However, to prevent misuse and protect the integrity of academic research, tighter regulations are essential. Institutions and funding bodies must create clear guidelines for using synthetic data, including transparency about how it is generated, requirements for reproducibility, and ethical oversight. Only then can synthetic data be used responsibly, ensuring it advances research rather than distorting it.

While synthetic data offers exciting potential for academic research, it must be handled with caution. From bias and ethical concerns to challenges in replicability and transparency, the risks are substantial. To unlock its full potential without compromising scientific integrity, researchers must use synthetic data responsibly, with full disclosure and rigorous validation. If done right, synthetic data can be a powerful tool for advancing knowledge; if misused, it risks distorting truth and misleading research.

(Neil is an assistant professor at the Department of Political Science, and Victor is Vice-Chancellor, St Joseph's University, Bengaluru)