

To democratize research with sensitive data, we should make synthetic data more accessible

Erik-Jan van Kesteren

March 2024

Abstract

For over 30 years, synthetic data has been heralded as a promising solution to make sensitive datasets accessible. However, despite much research effort and several high-profile use-cases, the widespread adoption of synthetic data as a tool for open, accessible, reproducible research with sensitive data is still a distant dream. In this opinion, Erik-Jan van Kesteren, head of the [ODISSEI Social Data Science team](#), argues that in order to progress towards widespread adoption of synthetic data as a privacy enhancing technology, the data science research community should shift focus away from developing better synthesis methods: instead, it should develop accessible tools, educate peers, and publish small-scale case studies.

Over 30 years ago, Rubin (1993) proposed synthetic data as a promising solution to make sensitive datasets available for reuse. The idea is as follows: instead of sharing a sensitive dataset, make available a synthetic version which has similar structure and statistical characteristics. This version can be used for research and analysis, but does not disclose private information of the participants. Indeed, increasing availability of sensitive data holds great promise: a wealth of information is hiding behind access barriers and privacy regulations. If all of this information could be made accessible through public synthetic datasets, this would be an enormous boon for a wide range of scientific disciplines – especially those involving human subjects (e.g., social, behavioral, medical sciences).

In my role as the head of the data science team of the Dutch social science data infrastructure ([ODISSEI](#)), I frequently encounter the promise of accessible sensitive data. I consult social scientists on a wide range of projects that would benefit greatly from reduced access barriers. For example, scientists need sensitive data access to (a) analyze historical microdata on residence, demographics, and occupation to research how inequality is distributed geographically, (b) combine multiple existing longitudinal studies to determine lasting psychological impacts of the COVID-19 pandemic, or (c) track school- and student-level metrics from educational registries to assess the impact of an educational policy. Impactful research often needs access to existing sensitive data – which is costly and time-consuming – so the promise of synthetic data to ease or even bypass access barriers captivates many social scientists.

Yet, this promise has not led to widespread uptake of synthetic data as a privacy-enhancing tool for improving the accessibility of sensitive data, despite much research effort in the statistics, machine learning, and computer science communities (for recent reviews, see Drechsler & Haensch, 2023; Fonseca & Bacao, 2023; Hu & Bowen, 2024; Jordon et al., 2022; Raghunathan, 2021; Reiter, 2023). High-profile cases do exist, such as the [Netherlands Cancer Registry](#), the [The CRPD Cardiovascular disease dataset](#), and [Eurostat's statistics on income and living public use files](#), but most of these have been large projects enacted by teams of data scientists and privacy experts – not by smaller research groups or individual researchers collecting sensitive data.

The main underlying cause for this is that it is complex, if not impossible, to create a privacy-safe synthetic version of a sensitive dataset that can be both openly publicized *and* productively

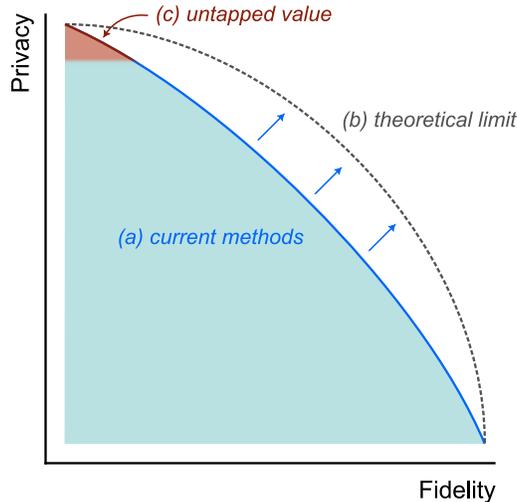


Figure 1: The inverse relation between privacy and fidelity in synthetic data. Several research lines are (a) pushing the currently available methods frontier closer to the theoretical boundary of this privacy-fidelity trade-off, and (b) defining how to measure privacy and fidelity and how to define this theoretical boundary, but they leave a large source of value untapped: (c) low-fidelity synthetic data, which can already be used as a privacy-enhancing technology.

used for research. The fundamental constraint is the *privacy-fidelity trade-off*: synthetic data is not inherently private (Jordon et al., 2022), and increasing the privacy level requires reducing the statistical closeness (“fidelity”) of the synthetic data to the real data.

A straightforward response is to suggest that we need to research better methods to create high-fidelity synthetic data that adheres to certain privacy standards. Several excellent tools exist which attempt this for a wide variety of underlying generative models and privacy definitions (e.g. Nowok, Raab, & Dibben, 2016; Patki, Wedge, & Veeramachaneni, 2016; Qian, Cebere, & van der Schaar, 2023) – their development and evaluation in terms of privacy disclosure risks are the subject of several active research lines. However, while this work is undeniably worthwhile, the best we can hope for with these tools is that they push the privacy-fidelity frontier towards its theoretical limit (Figure 1). As such, precluding a major revolution in our statistical definition and treatment of privacy, these advances are not bringing us closer to the widespread accessibility of sensitive data in a privacy-friendly way.

To put it more concretely, the current lines of data science research in synthetic data are focused too heavily on the technical aspects of synthetic data generation (Hu & Bowen, 2024). Putting synthetic data into real-world practice requires solving real-world issues, such as legal compliance, organizational reputation¹, education of data managers and owners, infrastructure development, training researchers to use this data, and more.

To start tackling these issues, I suggest a renewed focus from the data science community towards the safe side of the trade-off: create low-fidelity data with high privacy². I consciously use the word *fidelity* instead of the more commonly used term *utility*, because I think low-fidelity synthetic data is

¹Anecdotally, a national statistical agency publishing synthetic version of their microdata might be concerned with the question “*what if the public thinks this is real data and publishes analyses using this data?*”

of very high utility, for example in research code development, education, open science, reproducible research, and data accessibility. One of the biggest problems with research on sensitive data is that it takes time, effort, and money for researchers to be allowed access. Synthetic data can reduce these costs. With a public, low-fidelity mock version of a sensitive dataset, it is possible to inspect its columns, assess whether a certain research question can be answered, and even create analysis scripts that can run on the real data. In this way, the barriers to the first steps in the research process are greatly diminished, thereby democratizing research with existing sensitive data.

Potential ways in which low-fidelity synthetic data can democratize and improve research with sensitive data are as follows:

- Findability and accessibility. An organisation that produces sensitive data may release publicly a low-fidelity synthetic version of the data. This lowers the barrier for researchers to discover and use sensitive data, as it becomes easier to look up and test its contents and structure.
- Computational reproducibility. Upon publication of a project which uses sensitive data, researchers may publish a low-fidelity synthetic version of the data together with their public code repository (or point to where such data is available). This allows others to run, check, and learn from the code, even though the analysis results will be quite different from those in the original published paper.
- Security. If low-fidelity synthetic data is publicly available, researchers can assess exactly which parts they need before asking for access to the real data. With this information, a data owner can restrict subsequent access to only those columns and rows in specific tables. An even stronger security model would be “code-to-data”, which means running an analysis script on the sensitive data without researcher access. In this approach, low-fidelity synthetic data is essential to developing analysis scripts. These newly accessible security methods may convince data owners to enable researcher access where this was not possible before.
- Transparency and rigor. When analysis scripts are developed openly on a low-fidelity synthetic version of the data and then run only once on real data, it is harder for analysts to engage in common questionable research practices which may invalidate scientific findings. Common examples are manually removing data points or selecting variables to make results in line with hypotheses.

A perceived downside of low-fidelity synthetic data is that it still requires access to the real, sensitive data in order to be able to make valid conclusions. Indeed, the approach I am advocating renounces one of the most attractive promises of synthetic data, namely the ability to do analyses without needing real data access at all. At first sight, this appears to be a steep price to pay, but I argue that this promise cannot be fulfilled *in general*³: it would require the synthetic data to take into account any analysis, including nonlinear, multilevel, and outlier-dependent methods, all of which are hard (or impossible) to safely enable. For sensitive data providers, the goal of allowing access is often to enable valuable new research. In most cases, it is impossible know in advance which methods or statistics can answer valuable research questions, so access to the real data remains needed regardless of the fidelity of the synthetic data. In these cases, low-fidelity synthetic data is a great solution, when combined with other existing privacy enhancing infrastructure such as secure remote computing environments or systems that bring analysis code to the data.

²The UK’s Office for National Statistics would call such data *synthetically-augmented plausible* in their categorization (Bates, Spakulová, Dove, & Meador, 2019)

³For certain analysis types this promise can be fulfilled. For example, for many linear models it is sufficient for the synthetic data to accurately represent the means, covariance matrix, and sample size of the sensitive data, which are relatively easy to protect for public release. This can be done by (a) computing these statistics from the sensitive data, (b) performing an appropriate level of sufficient statistics perturbation (SSP) or disclosure control, and (c) generating multivariate normal synthetic data based on the perturbed statistics.



Figure 2: Steps to strategic change in science. Reprinted from <https://www.cos.io/blog/cos-celebrates-10-years> (accessed 2024-07-03).

To capitalize on the value of synthetic data as a privacy enhancing technology, we as a data science community should work on including it in real-world scientific practice. My proposal of focusing more on accessibility aligns with the theory of strategic change in science put forth by the center for open science (see Figure 2). The past 30 years in synthetic data research have focused on “making it possible” (step 1); in order to extract the value of synthetic data and democratize research with sensitive data it is now time to “make it easy” (step 2).

In conclusion, in addition to continuing existing research lines on the privacy-fidelity trade-off for synthetic data, the data science community should work directly on removing the barriers for data owners to create an open synthetic version of their data for public use. The community should build accessible tools (Jordon et al., 2022), develop educational materials (Hu & Bowen, 2024), and implement and publish pilot projects. Promising projects in this regard do already exist: auditable and accessible synthetic data tools such as *metasyn* (Schram, Spithorst, & van Kesteren, 2024), educational projects such as the Urban Institute’s trainings and associated reports (Pickens, Andre, & Morrison, 2023), and use-cases of synthetic data for computational reproducibility can be found on the Open Science Framework (e.g., Kirtley, Hussey, & Marzano, 2023). Let us expand these efforts in order to pave the way for more widespread use of synthetic data for open, accessible, reproducible research with sensitive datasets.

Acknowledgments

I am grateful for valuable feedback from Peter-Paul de Wolf, Thom Benjamin Volker, and Lara Rösler.

References

- Bates, A., Spakulová, I., Dove, I., & Meador, A. (2019). *Ons methodology working paper series number 16—synthetic data pilot*.
- Drechsler, J., & Haensch, A.-C. (2023). *30 years of synthetic data*.
- Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1), 115.
- Hu, J., & Bowen, C. M. (2024). Advancing microdata privacy protection: A review of synthetic data methods. *WIREs Computational Statistics*, 16(1), e1636. Retrieved from <https://>

- wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1636 doi: <https://doi.org/10.1002/wics.1636>
- Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., ... Weller, A. (2022). Synthetic data—what, why and how? *arXiv preprint arXiv:2205.03257*.
- Kirtley, O. J., Hussey, I., & Marzano, L. (2023, Mar). *Data processing, synthetic data, and analyses*. OSF. Retrieved from osf.io/asgck
- Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74(11), 1–26. Retrieved from <https://www.jstatsoft.org/index.php/jss/article/view/v074i11> doi: 10.18637/jss.v074.i11
- Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 399–410).
- Pickens, M., Andre, J., & Morrison, G. (2023). Generating a fully synthetic human services dataset. *Urban Institute*.
- Qian, Z., Ceber, B.-C., & van der Schaar, M. (2023). *Synthcity: facilitating innovative use cases of synthetic data in different data modalities*.
- Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, 8(1), 129–140. Retrieved from <https://doi.org/10.1146/annurev-statistics-040720-031848> doi: 10.1146/annurev-statistics-040720-031848
- Reiter, J. P. (2023). Synthetic data: A look back and a look forward. *Trans. Data Priv.*, 16(1), 15–24.
- Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, 9(2), 461–468.
- Schram, R., Spithorst, S., & van Kesteren, E.-J. (2024). *metasyn*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.7696031> doi: 10.5281/zenodo.7696031