Ileana Montoya Perez, Parisa Movahedi, Valtteri Nieminen, Antti Airola, Tapio Pahikkala

Response to letter by Dehaene et al. on synthetic discovery is not only a problem of differentially private synthetic data

2025

https://doi.org/10.1055/a-2540-8346

Final draft

Montoya Perez, Ileana, et al. “Response to Letter by Dehaene et al. on Synthetic Discovery Is Not Only a Problem of Differentially Private Synthetic Data.” Methods of Information in Medicine, vol. 63, no. 05/06, 2024, pp. 205–06. https://doi.org/10.1055/a-2540-8346

Response to letter by Dehaene et al. on synthetic discovery is not only a problem of differentially private synthetic data

Ileana Montoya Perez[1], Parisa Movahedi[1], Valtteri Nieminen[1],

Antti Airola[1], Tapio Pahikkala[1]

[1]Department of Computing, University of Turku, Finland

Dear Editor,

We thank for the opportunity to respond to the commentary letter by Dehaene et al on our recent article, “Does Differentially Private Synthetic Data Lead to Synthetic Discoveries?” [1] published in Methods of Information in Medicine. We appreciate the commentators’ interest in our work and their contribution to an important and ongoing discussion on the utility of synthetic data and its implications for statistical inference.

The letter from Dehaene et al raises a concern about two possible interpretations of the results in our article, namely that the risk of unacceptably high false-positive findings from synthetic data can be simply countered by increasing the amount of original data enough, or by stepping away from differentially private (DP) synthetization methods. Referring to simulation results in Decruyenaere et al,[2] they note that even for non-DP methods and large original sample sizes, this risk can remain high, especially when using deep learning based generation methods. We find that Dehaene et al raise an important point and their observations are compatible also with our results. While reducing amount of DP noise and increasing original sample size are positively correlated with the utility of generated synthetic data, these alone are not enough if the generator is a misspecified parametric model or suffers from what Decruyenaere et al[2] refers to as the regularization bias.

As the authors note, citing Chen et al: “synthetic data are artificial data that (attempt to) mimic the original data in terms of statistical properties, without revealing individual records”.[3] Obviously, if privacy would not be of concern and reliable prior information on true distribution of data absent, this would be achieved simply by using the original data. Indeed, some DP data release methods reconstruct the original data in the limit of epsilon approaching infinity. In our experiments, the DP perturbed and DP smoothed histograms have such property. Accordingly, these methods demonstrate a clear trade-off between similarity to original data, privacy level, and the amount of original data, with inferential utility of the synthetic data typically increasing both with respect to original sample size and inversely with respect to privacy level. On the other hand, the synthetic

data generated by Multiplicative Weights Exponential Mechanism (MWEM) and Private-PGM (Private-Probabilistic Graphical Model) may diverge from the distribution of original data in the limit due to approximating higher dimensional data with low-dimensional marginals. Hence, the trade-off may be less clear, if the statistical property of interest changes not only due to privacy level but also due to approximation. In some of our results, this is reflected by the utility increasing as a function of decreasing privacy level only up to a certain limit but not achieving the utility of the original data. A similar effect can take place if the synthetization methods make incorrect parametric assumptions. At the other extreme of this continuum of methods, there are synthesizers having regularization bias aimed for purposes other than reproducing the original data. For example in our experiments the DP GAN method had very different behavior compared to the other methods, the risk of false discoveries even increasing as a function of decreasing privacy level.

Accordingly, we agree with the main message of Dehane et al that the inferential utility level of the original data is not necessarily achieved simply by decreasing the privacy level or with larger amounts of original data, but is very method dependent. Hence caution is certainly always warranted when performing statistical inference on synthetic data, with different methods having different trade-offs and some demonstrating systematic biases that are not easy to counter.

Sincerely,

The authors

Acknowledgments

This research has received funding from European Union's Horizon Europe research and innovation programme (grant number 101095384). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Conflict of Interest

Not applicable

Ethical considerations

Not applicable

References

1 Montoya Perez I, Movahedi P, Nieminen V, Airola A, Pahikkala T. Does Differentially Private Synthetic Data Lead to Synthetic Discoveries? Methods Inf Med 2024;63(01/02):035–051

2 Decruyenaere A, Dehaene H, Rabaey P, et al. The Real Deal Behind the Artificial Appeal: Inferential Utility of Tabular Synthetic Data. The 40th Conference on Uncertainty in Artificial Intelligence. Barcelona; 2024

3 Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering 2021 5:6 2021;5(6):493–497