Comments on EDPB Guidelines 02/2026 on Anonymisation: synthetic data, generative models and the evidence required to demonstrate anonymity

Submitted by Linas Vastakas (Faculty of Law, University of Bergen) and Dr Qinyi Liu (School of Education, City University of Macau). Submitted in a personal academic capacity; the views expressed do not necessarily reflect those of our institutions.

We welcome the draft guidelines, and particularly paragraph 82, which clarifies that both synthetic data and AI models can allow drawing inferences about individuals, and so are subject to the No Inference criterion. We would like to provide some further comments from a legal and data-science perspective regarding a question that the draft guidelines have currently arguably left open: what evidence a controller must present to show that a dataset satisfies the three criteria and against what it must be checked. We would like to make the following points:

1. The assessment should account for access to the trained generative model, as well as the released synthetic data.

We believe that paragraph 82 correctly observes that inferences can be made by prompting models or by querying synthetic data. Where recipients can also access the trained generative model, whether through a query interface or access to its parameters, assessing a single released synthetic dataset may not capture the full privacy risk. Model access may enable repeated or adaptive queries, or attacks that are not possible using the released dataset alone. Conversely, where only a fixed synthetic dataset is released, the assessment should distinguish the recipient's access from the information and capabilities retained by the generating controller. Opinion 28/2024 already suggests that AI models trained on personal data "cannot, in all cases, be considered anonymous" (para 34), that they are by default "likely to require a thorough evaluation of the likelihood of identification" (para 43), and that "a mere assertion of anonymity of the model is not enough" (para 134).

We would suggest an addition to the guidelines: namely, the guidelines should explicitly distinguish the anonymity assessment of released synthetic data from that of the trained generative model. Where the model is shared or made queryable, the assessment should consider the information obtainable from the released data and the model together. Where the model is retained internally, its relevance should be assessed in light of the access and capabilities of the party in relation to which anonymity is being evaluated. Paragraph 37 of the draft Guidelines already provides that where information is anonymised for the recipient "but not for the anonymising controller, that controller will still have to treat the given data as personal data". Our proposal is that the guidelines make explicit the following:

1. Where the trained generative model is published, shared or made queryable, the assessment should examine whether information obtainable through that access, alone or in combination with the released synthetic data and relevant auxiliary information, enables record isolation, linkage or inference. The tests should reflect the access provided, including query access or access to model parameters.

2. The controller should document whether the source data and trained generative model are retained, who can access them, and under what technical and organisational restrictions. The assessment should explain how these arrangements affect the anonymity conclusion for the controller and for the intended recipients.
3. Deleting the trained model or effectively restricting access to it may reduce the additional risks arising from model access and should be considered where relevant (much as paragraph 39 treats deletion of the original data as a positive factor in the legitimate interest balancing test). Such measures do not, however, establish the anonymity of synthetic data already released, and their effectiveness depends on whether copies of the model remain accessible.

We believe this addition is consistent with recent technical work. Ganev and De Cristofaro show that an attacker with black-box access to a fitted generator and its privacy metrics can reconstruct most outliers in the training data (IEEE S&P 2025; see point 2), and argue that anonymity assessments should account for the generative model wherever it can be queried or interacted with ("Rethinking Anonymity Claims in Synthetic Data Generation: A Model-Centric Privacy Attack Perspective", WPES 2026). Research on large language models also demonstrates that specific training records can sometimes be extracted from trained models (Carlini et al., "Extracting Training Data from Large Language Models", 2021). These findings support assessing the risks associated with the model access actually provided.

2. Similarity and matching metrics are not sufficient evidence that the three criteria are met.

The metrics that are commonly offered as proof of anonymity for synthetic data are similarity-based and matching-based: distance measures such as distance to closest record (DCR) and nearest-neighbour distance ratio (NNDR), as well as measures of uniqueness, identifiability and presence disclosure, such as k-anonymity and the identifiability score. Several of these are built into widely used open-source evaluation toolkits such as Synthcity (Qian, Cebere and van der Schaar, 2023).

However, in our view, recent technical literature shows that, while these measures may provide useful diagnostic information, they capture only particular aspects of disclosure risk. A favourable score does not, by itself, establish that the No Record Isolation, No Linkage and No Inference criteria are satisfied. For example, datasets that fulfil these metrics can still be vulnerable to membership inference (Yao et al., 2025), and these metrics themselves are vulnerable to exploitation, allowing the recovery of 78–100% of outliers in one study under an attack setting involving black-box access to a fitted generator and its privacy metrics (Ganev and De Cristofaro, "The Inadequacy of Similarity-based Privacy Metrics: Privacy Attacks against 'Truly Anonymous' Synthetic Datasets", 2025).

Moreover, in some cases attribute inference attacks can be successful against ordinary records too, not only against outliers (Annamalai, Gadotti and Rocher, 2024, "A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data", USENIX Security).

Furthermore, some of these metrics can produce opposite conclusions about the same generator (Liu et al., BJET 2025).

Therefore we suggest that the guidelines clarify that the reliance on such metrics by a controller does not, in itself, conclusively show that the No Record Isolation, No Linkage and No Inference criteria are met. Instead, such a demonstration should require attack-based tests under a stated threat model, following the approach Opinion 28/2024 adopts for AI models, where supervisory authorities consider the controller's testing against attacks such as attribute and membership inference and reconstruction (para 55). As that same paragraph also notes, successful testing "can only be evidence for the resistance to those attacks", so the failure of the tested attacks should not be treated as conclusive evidence of anonymity. Negative results must be interpreted in light of the attacks' capabilities, the auxiliary information assumed, and the scope and limitations of the evaluation.

3. The guidelines should specify the required contents of the documented synthetic data assessment.

The draft guidelines currently seem to expect "adequate documentation of the anonymisation processing" (para 41) and to recommend periodic reassessment (para 35). However, the guidelines do not clearly specify what the documentation should contain. Opinion 28/2024 already does this for AI models (which, in our view, covers the generator itself, paras 25–26), listing what supervisory authorities should look for in the controller's documentation, including the threat model, "reports on how the model has been tested (by whom, when, how and to which extent)" and "the results of the tests" (para 58). We suggest a comparable list for synthetic data.

Drawing on the technical evaluation literature, we propose a baseline set of documentation requirements. Crucially, these elements must be evaluated holistically. For instance, a standalone privacy parameter like ε is insufficiently informative, as the same value can yield different privacy outcomes depending on the dataset, generator, and task (Trudslev et al., 2025). The proposed minimum content should include:

1. The generator type, training configuration and (in case differential privacy is used) the privacy parameters (ε, δ) and whether the privacy guarantee is central or local and the stages of the generation and release process covered by the guarantee.
2. The assumed threat model, including the recipients and any auxiliary information.
3. The attack-based evaluations conducted, as appropriate to the threat model, including membership inference, attribute inference, reconstruction and linkage; the tools and versions used; the evaluation setup, relevant baselines, results and uncertainty; and the limitations of the tests.
4. The release architecture, including whether access is to a fixed dataset, a query interface or model parameters; the volume and frequency of releases or queries; and any applicable access restrictions.
5. Any anomalies observed, such as unexpected memorisation.

6. The review trigger, i.e., the events that will require the assessment to be repeated; the guidelines already acknowledge that anonymity may lapse “over time” (para 35), and that a security incident may lead to a reassessment (para 42). We believe it would be consistent with this if the documentation specified foreseeable triggers for reassessment in advance - such as the publication of a new attack, the availability of new auxiliary data, changes to recipients or release architecture, and the retraining, sharing, or opening of the generator, as well as the aforementioned security incidents.
7. The recipient for which the conclusion holds, since data may be anonymous for a recipient while remaining personal data for the controller that generated them (para 37 of the Guidelines; C-413/23 P EDPS v SRB).

We remain at the EDPB’s disposal for any further information or clarification.

References

EDPB documents

European Data Protection Board (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024.

European Data Protection Board (2026). Guidelines 02/2026 on Anonymisation, Version 1.0, adopted on 7 July 2026 (version for public consultation).

Case law

Case C-413/23 P, European Data Protection Supervisor v Single Resolution Board, Judgment of the Court of Justice of 4 September 2025.

Literature

Annamalai, M.S.M.S., Gadotti, A. and Rocher, L. (2024). “A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data”. In Proceedings of the 33rd USENIX Security Symposium (USENIX Security 2024). USENIX Association.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A. and Raffel, C. (2021). “Extracting Training Data from Large Language Models”. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 2021). USENIX Association.

Ganev, G. and De Cristofaro, E. (2025). “The Inadequacy of Similarity-based Privacy Metrics: Privacy Attacks against ‘Truly Anonymous’ Synthetic Datasets”. In Proceedings of the 46th IEEE Symposium on Security and Privacy (IEEE S&P 2025). arXiv:2312.05114.

Ganev, G. and De Cristofaro, E. (2026). “Rethinking Anonymity Claims in Synthetic Data Generation: A Model-Centric Privacy Attack Perspective”. In Proceedings of the 25th Workshop on Privacy in the Electronic Society (WPES 2026), ACM CCS 2026. arXiv:2601.22434.

Liu, Q., Shakya, R., Jovanovic, J., Khalil, M. and de la Hoz-Ruiz, J. (2025). “Ensuring privacy through synthetic data generation in education”. British Journal of Educational Technology, 56(3), 1053–1073. https://doi.org/10.1111/bjet.13576

Qian, Z., Cebere, B.-C. and van der Schaar, M. (2023). “Synthcity: facilitating innovative use cases of synthetic data in different data modalities”. arXiv:2301.07573.

Trudslev, F.M., Lissandrini, M., Rodriguez, J.M., Bøgsted, M. and Dell’Aglio, D. (2025). “PrivEval: a tool for interactive evaluation of privacy metrics in synthetic data generation”. Proceedings of the VLDB Endowment, 18(12), 5271–5274. https://doi.org/10.14778/3750601.3750649

Yao, Z., Krčo, N., Ganev, G. and de Montjoye, Y.-A. (2025). “The DCR Delusion: Measuring the Privacy Risk of Synthetic Data”. arXiv:2505.01524.