Associazione Farmaceutici Industria
Società Scientifica

Feedback on EDPB Guidelines 02/2026 on Anonymisation

This document provides constructive feedback and recommendations regarding **Guidelines 02/2026 on Anonymisation**. The feedback is organized into substantive conceptual issues, operational methodologies, and specific editorial/cross-referencing corrections.

1. Application of Guidelines to Synthetic Data and Emerging Technologies

The Guidelines formally adopt a technology-neutral approach, however most operational examples concern data retaining a direct or transformed relationship with source records. The Guidelines appear de facto tailored to such de-identification methods. While the Guidelines recognise that AI models and synthetic data may enable specific inferences, they do not provide sufficient guidance and examples on the application of the three criteria to generative synthetic data as a de-identification technology. In particular, in synthetic datasets, the one-to-one relationship between given records and real subjects is broken when using synthetic data, so that an isolated record does not (necessarily) correspond to a real data subject.

Recommendation: Include a dedicated subsection or specific examples in Section 3 explaining how the criteria apply to generative synthetic data. Clarify that synthetic records containing artificial attribute combinations do not constitute "record isolation" unless they allow re-identification of or inference about a real natural person in the source dataset.

2. Clarification of Singling Out, Record Isolation, and Continuous Variables

The relationship between unique rows, record isolation, and singling out requires further conceptual refinement.

Opaque Definition of Singling Out

Footnote 37 states that singling out "will be explained in detail in section 3.5.3"; however, the treatment in paragraphs 98–100 is brief and focuses primarily on matching isolated records with external attributes. The text should provide a clearer conceptual definition distinguishing **row isolation** (a structural property of the dataset) from **singling out** (the capability to distinguish and treat an individual differently), especially for de-identification techniques without a 1:1 correspondence between real and de-identified data (e.g. synthetic data).

Continuous Numerical Values vs. Record Isolation

In datasets containing high-precision continuous variables (e.g., exact timestamps, GPS coordinates, financial amounts), virtually every row is unique by mathematical necessity. The draft does not specify whether datasets with continuous unique values automatically fail the "No Record Isolation" criterion, or how controllers should handle continuous data under the simplified approach.

Recommendation: Explicitly clarify that continuous numerical uniqueness does not automatically constitute actionable record isolation if the values cannot be linked to natural persons.

3. Operational Feasibility, Progress Beyond Opinion 05/2014, and Risk Thresholds

While Guidelines 02/2026 update Article 29 Working Party Opinion 05/2014, the operational guidance remains overly qualitative for practical deployment. In order to facilitate the operational application of the guidelines, we suggest the following amendments.

Guideline Requirement	Operational Challenge	Recommendation
Par. 36: Risk must be "insignificant for all individuals".	"Insignificant" is undefined. Without objective thresholds or quantitative metrics, controllers cannot determine compliance.	Provide benchmarks (e.g., empirical re-identification probabilities). A universal numerical threshold may not be appropriate. Nevertheless, controllers need clearer guidance on suitable evidence, uncertainty and the treatment of high-risk individuals.
Par. 36, 54 - 72: Assessing the 3 criteria across entire datasets.	Exhaustively testing every record across massive datasets is computationally prohibitive (esp. Considering par. 71b requiring iterative application of the de-identification method).	Explicitly confirm that performance checks and technical audits may be performed on statistically representative random sub-samples or, where relevant, rare, exceptional or particularly exposed individuals and/or on records or groups presenting an elevated risk.

Guideline Requirement	Operational Challenge	Recommendation
		Relevant/rare/exceptional/exposed records may require distinct treatment from more homogeneous records (cf. Section 5 below)

The Guidelines should remain open to future technical work on validated metrics, testing protocols, uncertainty estimates and sector-specific benchmarks, particularly for generated data and trained models.

4. Include practical mechanisms to demonstrate compliance

While the guidelines provide three clear criteria for anonymity, it does not present practical ways to demonstrate compliance. Illustrating how compliance can be demonstrated, e.g. through concrete examples of metrics, would strengthen the Guidelines by making them easier to adopt in realistic scenarios. This is particularly relevant for the No Record Isolation criterion, as singling out can also result in the violation of the other criteria (No Linkage, No Inference). Metrics should be applicable to the various forms of de-identification that result in possibly anonymized datasets.

Recommendations:

1. Include practical manners in which to demonstrate compliance with the criteria through metrics;
2. Include examples of the proper use of such metrics, most importantly for the No Record Isolation criterion;

5. Proposed Overarching Testing Framework

Relating to the above, we propose that the Guidelines recognise that some techniques may represent a **unified overarching testing framework**.

The Guidelines recognize that a record's uniqueness may present a privacy hazard. This is corroborated by scientific literature that shows that membership inference attacks are (nearly) exclusively effective against highly unique and therefore vulnerable records [3,4,5] or that develops specific attacks against unique records (e.g. the singling out attack in [1]). As example 12 of the Guidelines shows, however, inference attacks do not require record-level uniqueness and may in fact exploit shared characteristics of small sub-populations. Homogeneous individuals are instead at risk if a set of records in the given data reveal information through

statistically significant similarities to original data and by exposing correlations in sub-groups of the population (cf. example 12).

We therefore propose a situational testing framework to protect both unique records and homogeneous records. In particular, we propose a case distinction between two situations:

- **Situation 1:** highly unique records are irrelevant or undesirable in the application of the given data. In this case, we propose outlier removal prior to de-identification. This removes records vulnerable due to their uniqueness. Similarity-based metrics (such as the Distance-to-the-Closest-Record, or DCR [6]) should then be applied to the resulting given data to ensure that statistical similarities between given and original data records are avoided;
- **Situation 2:** Outliers are important in the data's application or the data is sparse. This situation can for instance arise in the study of rare diseases. In this situation, outlier removal is undesirable, as it defeats the purpose of the given data. In this situation, we propose leveraging similarity-based metrics to protect homogeneous records, combined with an empirical worst-case membership inference attacks against vulnerable records.

Recommendations:

1. State that, under the simplified approach, a controller may demonstrate all three criteria at once by showing that the given data satisfies the conditions in the proposed testing framework;
2. Keep the text open to future validated criteria that satisfy the same counterfactual, consistent with the technology-neutral approach of the Guidelines.

6. Typographical, Cross-Reference, and Editorial Corrections

The following minor errors, missing text, and broken cross-references should be corrected prior to final publication:

1. **Annex 1 Flowchart Cross-References:**

    - The box "Could the given data be extended by information obtained from Linkage and/or Inference?" cites (see 73, 103 and 104). The document ends at paragraph 103; 104 does not exist.
    - The decision boxes citing (see 100-102) for No Record Isolation and Singling Out should refer to paragraphs 98–100.

2. Paragraph 97 Table: The table entry under No Record Isolation reads "See paras. 98 to 00 below", which contains a typo (00 instead of 100).

3. Paragraph 53: Missing preposition "in". Text currently reads: "...reintroduced sub-section 3.5." It should read: "...reintroduced [in] sub-section 3.5."

References

[1] Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2022). A unified framework for quantifying privacy risk in synthetic data. arXiv preprint arXiv:2211.10459.
[2] Houssiau, F., Jordon, J., Cohen, S. N., Daniel, O., Elliott, A., Geddes, J., ... & Szpruch, L. (2022). Tapas: a toolbox for adversarial privacy auditing of synthetic data. arXiv preprint arXiv:2211.06550.
[3] T. Stadler, B. Oprisanu, and C. Troncoso, “Synthetic data – anonymisation groundhog day,” 2022.
[4] M. Meeus, F. Guepin, A.-M. Cretu, and Y.-A. de Montjoye, “Achilles’ heels: Vulnerable record identification in synthetic data publishing,” 2023
[5] L. Pilgram, F. K. Dankar, J. Drechsler, M. Elliot, J. Domingo-Ferrer, P. Francis, M. Kantar cioglu, L. Kong, B. Malin, K. Muralidhar, P. Myles, F. Prasser, J. L. Raisaro, C. Yan, and K. E. Emam, “A consensus privacy metrics framework for synthetic data,” 2025.
[6] J. Hradec, M. Craglia, M. Di Leo, S. De Nigris, N. Ostlaender, and N. Nicholson, Multi purpose synthetic population for policy applications. JRC technical report, 2022.

Supported by Farminustria Italian national trade association representing the research- and development-based (R&D) pharmaceutical industry



Giorgio Cesare Maria Bruno
28/10/2026 11:40:59 UTC+0100

