Sender contact
RISE Research Institutes of Sweden
Rickard Brännvall
Digital Systems
+46-730 753 713
rickard.brannvall@ri.se

To the European Data Protection Board

Feedback on the EDPB Guidelines 02/2026 on Anonymisation

RISE Research Institutes of Sweden is Sweden's state-owned research institute. This reply was prepared within the project PPAI×REG (Privacy-Preserving AI under Regulatory Pressure), whose participants work on measuring privacy leakage from data and machine learning models, on privacy-enhancing technologies and on secure processing environments for sensitive data. We welcome Guidelines 02/2026 on anonymisation (EDPB, 2026) and the opportunity to comment on them. The contextual and simplified approaches, and the inclusion of AI models in scope, address the questions we meet in our work. Our comments are technical and concern how the guidelines can be applied in practice. Where they touch on legal questions, we do not take a position.

Our comments concern mainly the No Inference criterion (para 67 ff.), which is the most relevant criterion for aggregate outputs and trained models. Reliance on an individual's data can be shown by comparing the inferences drawn from the given data with and without that individual's data, A(D) and A(D′) (para 71b, footnotes 43–44). In our reading, this comparison underlies both differential privacy (Dwork and Roth, 2014), which bounds how much any inference can differ between the two, and membership inference (para 79), which tests whether an attacker can tell them apart. We distinguish three types of evidence. *Formal evidence*, such as differential privacy, bounds leakage from above for any recipient. *Empirical evidence* from attack-based tests shows what a concrete attack achieves and so bounds leakage from below. We call all other evidence *indirect evidence*: documentation of how the data was processed or a model trained, the key factors in paras 85–86 and simple risk indicators, which inform a judgement about the risk without measuring or bounding it.

We raise seven points, each with a short summary, an example, a question to the EDPB and our rationale.

1. **Validation of re-identification tests:** how should the detection limit, bias and uncertainty of a test be documented when the test is used as evidence of anonymity (para 41)?
2. **Evidence for individual records:** what can be concluded about a single individual when test results are aggregated over many records?
3. **Secure processing environments with an airlock:** could an example show the contextual approach applied inside a secure processing environment, with output review at the airlock?
4. **Release types for AI models:** how does the comparison in para 71b apply to the release of model weights compared with access through an API?
5. **Releases that cannot be withdrawn:** how should reasonably foreseeable developments in re-identification techniques (para 29(e)) be taken into account at the time of release?
6. **Formal guarantees and the simplified approach:** can differential privacy support the conclusion that the No Inference criterion is met under the simplified approach?
7. **Indirect evidence:** when can documentation and simple risk indicators be sufficient without attack-based testing or a formal guarantee?

Points 1 and 2 concern empirical evidence, points 3 to 5 the setting and form of a release, point 6 formal evidence and point 7 indirect evidence.

1. Validation of re-identification tests

Summary: A failed re-identification attempt is a measurement result, not a proof of anonymity. Its weight as evidence depends on what the test can detect. We recommend that the documentation of testing (para 41) states the detection limit, bias and uncertainty of the test, established by validating the method on cases where the outcome is known.

Example: A membership inference attack on a trained model is first validated in a reference setting where the leakage is known, for example with inserted canaries (artificial records whose membership is known). Suppose it wrongly flags one in a thousand non-members. Among 1,000 training records, it then needs at least four hits before the result differs from chance at the 5% level. Even if every leaking record is flagged, leakage must affect about 0.7% of the records to be detected with 95% probability. A result at chance level therefore only shows that leakage above this limit is unlikely.

Question: Could the EDPB clarify how the performance of a re-identification test, in particular its detection limit, bias and uncertainty, should be documented when the test is used as evidence of anonymity (para 41)?

Rationale: The guidelines judge a re-identification technique by whether its results are close enough, and held with sufficient confidence, to distinguish individuals (para 87). By the same reasoning, a test used to show that such techniques fail should state how confident its conclusion is. Existing standards cover the validation of test methods (ISO/IEC 17025:2017), the expression of measurement uncertainty (JCGM 100:2008) and the role of uncertainty when a result is compared with a limit (JCGM 106:2012). Other regulated fields require validated methods before the results can support decisions, for example the analytical performance of in vitro diagnostic devices, including their limit of detection (Regulation (EU) 2017/746, Annex I, para 9.1a), and the independent validation of internal models in banking (Federal Reserve and OCC, 2011). The open question is how these principles apply to re-identification tests. The depth of testing can be proportionate to the risk of the release.

2. Evidence for individual records

Summary: The likelihood of re-identification must be insignificant for every individual (para 36). Attack-based tests typically gain their statistical power by aggregating over many records and say much less about any single individual. It is not clear what evidence is expected at the level of the individual.

Example: A company tests its model with membership inference before release and finds no significant leakage across the training records. After release, a data subject withdraws consent and requests erasure. The pre-release test has little statistical power for this one individual, and once the original training data is deleted, the comparison in para 71b can no longer be repeated.

Question: Could the EDPB add a worked example of what evidence is expected for individual records when test results are aggregated over many records, and of what can be concluded when a test cannot resolve the risk for a single individual?

Rationale: A technique can be successful even if it succeeds against only a single individual (para 87, Example 20). Tests can report results for the most vulnerable records, for example the true positive rate at a low false positive rate (Carlini et al., 2022), and repeated experiments can add evidence for individual records, but their interpretation depends on how well the experiments represent the release. When few or no successes are observed, the result gives an upper confidence bound on the success of the tested attack, not a claim of zero risk. Testing before release, while the original data is still available, gives the best opportunity to gather such evidence. This may also bear on erasure obligations (cf. para 10).

3. Secure processing environments with an airlock

Summary: A secure processing environment (SPE) separates the entities working inside, whose access and means are known and controlled, from the recipients of the outputs it releases. Inside, the contextual approach applies naturally; at the airlock, outputs are reviewed before they leave.

Example: A company developing an AI model obtains de-identified data from a telecom operator and trains the model inside a secure processing environment. The company works in defined roles with logged access, cannot import external data and has no access to the telecom operator's additional information. Under the contextual approach, the data may therefore be anonymous from the company's perspective, although it remains personal data for the telecom operator. Before the trained model leaves through the airlock, it is tested with attacks on the data still inside.

Question: Could the EDPB add an example of the contextual approach applied to a secure processing environment, distinguishing the entities inside from the recipients of its outputs, with an output review at the airlock deciding what may be released?

Rationale: Inside the SPE, security controls limit access to authorised users (para 89), and entities without access need not be assessed (para 9). Such controls are not sufficient by themselves (para 90), so the outputs still need review. An output that goes to a known recipient under comparable controls can be assessed from that recipient's perspective, while a published output may reach any entity. Established procedures exist for checking aggregate outputs (Griffiths et al., 2019), and recent work extends them to trained models (Jefferson et al., 2022; Smith et al., 2025). Our question concerns how the contextual approach applies inside the environment and how outputs are assessed at the airlock.

4. Release types for AI models

Summary: The guidelines bring AI models into scope (paras 77, 82) and describe the A(D)/A(D′) comparison (para 71b), but do not say whether the given data of a model is its weights or the answers it gives. Released weights can be copied and attacked with full access, while a model behind an API stays under the provider's control.

Example: A company has trained a model on its customers' data and wants to offer it to business customers. It can either ship the model weights for use in the customers' products, or offer access through an API with authentication, rate limits and answers restricted to labels.

Question: Could the EDPB add an example applying the comparison in para 71b to a trained model, contrasting the release of model weights with access through an API, and showing how controls on the service enter the assessment?

Rationale: Opinion 28/2024 notes that the conclusion of an assessment may differ between a publicly available model and an internal one, and that different levels of testing may be considered (EDPB, 2024, para 46). Recipients of the weights face no limits on who they are or what they do with the model, including inference attacks, whereas for a service the provider decides who may send queries, how many and what the answers contain, which restricts access to authorised users (para 89) and limits what they can obtain from the model; a test can then mirror the service, with its output format and query budget. Restricting answers to labels reduces but does not remove the risk, since attacks that see only labels can still succeed (Choquette-Choo et al., 2021). A service open to the public restricts every user's means but not who the users are, which may call for the simplified approach. The provider can also retrain or withdraw the model, unlike published weights (point 5), and a retrained model is a new release that needs its own assessment (cf. para 35).

5. Releases that cannot be withdrawn

Summary: Re-identification techniques improve over time (para 35), and the assessment should consider reasonably foreseeable technological developments (para 29(e)). Periodic reassessment cannot correct a release that cannot be withdrawn, such as openly published record-level data or model weights.

Example: A public agency publishes a record-level dataset under an open licence, and a company publishes model weights in a public repository. Both were tested against the attacks known at the time. A few years later, a stronger attack was published. Neither release can be recalled.

Question: For releases that cannot be withdrawn, could the EDPB clarify how reasonably foreseeable developments in re-identification techniques (para 29(e)) should be taken into account at the time of release, under the contextual and the simplified approach?

Rationale: If the likelihood of re-identification later becomes significant, the released data should again be considered personal data (para 35). Attacks can improve quickly: language models, for example, now make linkage of free-text records feasible at scale (Lermen et al., 2026). Under the contextual approach, the EDPB recommends adequate safety margins so that the assessment remains robust as the entities' capabilities evolve (para 88). Because a published release cannot be withdrawn or changed, every mitigation has to be applied before the release. A model shared with a known partner under contract cannot be recalled either, and the partner may later obtain new additional information, for example through a data leak. The simplified approach disregards differences in means between entities (para 48), but it is not clear whether it also covers attacks that are not yet known. A margin can be added by testing with stronger attacker assumptions than those of the expected attacker. Another option is evidence whose bound does not depend on the strength of the attack, which is the subject of point 6.

6. Formal guarantees and the simplified approach

Summary: Differential privacy bounds how much any recipient's posterior belief about an individual can change depending on whether that individual's data was included, whatever additional information the recipient holds. Because the bound holds for any recipient, it appears to fit the simplified approach.

Example: A statistics office publishes counts and averages by region from a sensitive register, produced with a differentially private mechanism, and argues under the simplified approach that the release protects against any recipient.

Question: Could the EDPB clarify whether, and under what conditions, a formal guarantee such as differential privacy can support the conclusion that the No Inference criterion is met under the simplified approach, and how the choice of its privacy parameters should be justified?

Rationale: Differential privacy still allows inference about the population at large, which the guidelines accept (para 70). Cohen and Nissim (2020) formalise the GDPR's notion of singling out and show that differential privacy protects against it. Its meaning depends on choices such as the privacy parameter, the unit of privacy and the variant of the definition, and no data protection authority has set values (cf. NIST, 2025). Which values are small enough for a posterior belief not to be “significantly different” (footnote 44) may itself depend on the context of the release, even under the simplified approach. The guarantee holds only for a correct implementation, which can be checked by privacy auditing (Steinke et al., 2023). Such audits give a lower bound on the actual privacy loss and are themselves tests that need validation as in point 1. Several releases from the same data add up (para 77) and need to be accounted for together. Differential privacy is not mentioned in the guidelines, although several replies to the 2025 consultation on pseudonymisation raised it (EDPB, 2025).

7. Indirect evidence

Summary: Not every release needs an attack-based test or a formal guarantee. Indirect evidence, such as the properties of the data and documentation of how a model was built, informs a judgement about the risk without measuring or bounding it. The guidelines already reason this way in the key factors of paras 85 and 86.

Example: A hospital releases a logistic regression model with a handful of coefficients trained on several thousand patients with strong regularisation. Its performance on training and test data is almost the same. Documentation of the training and these indicators may be enough to show that the risk is low, without running attack-based tests.

Question: Could the EDPB clarify whether, and for which releases, indirect evidence such as the key factors in paras 85 and 86, documentation of how a model was trained and simple risk indicators can be sufficient to conclude anonymity without attack-based testing or a formal guarantee?

Rationale: Indirect evidence is widely used in practice. Before a model may leave its secure processing environment, Findata, the Finnish authority for the secondary use of health and social data, asks applicants to complete a questionnaire on how the model was built and is to be used, in which attack testing is one question among many (Findata, 2025). The SACRO-ML tools check model types and settings for disclosure risk before training and run attacks after training (Smith et al., 2025). The conditions proposed by Jebreel et al. (2026) for when an attack poses a realistic threat can serve as a similar check. In a staged assessment, indirect evidence can screen out low-risk releases and give context for empirical tests where the risk is higher. Guidance on when it suffices would help make the depth of assessment proportionate to the risk.

We would be happy to clarify any of these comments.

Yours sincerely,

RISE Research Institutes of Sweden AB
Digital Systems

Stella Riad — Department manager

Rickard Brännvall — Senior Researcher

Other contributors (RISE Research Institutes of Sweden)

Aslak Felin, Senior Project Manager;

Henrik Forsgren, Research and Development Engineer;

Michael Popoff, Senior Researcher;

Susanne Stenberg, Senior Researcher and legal expert.

References

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A. and Tramèr, F. (2022). Membership inference attacks from first principles. IEEE Symposium on Security and Privacy.

Choquette-Choo, C. A., Tramèr, F., Carlini, N. and Papernot, N. (2021). Label-only membership inference attacks. Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139.

Cohen, A. and Nissim, K. (2020). Towards formalizing the GDPR's notion of singling out. Proceedings of the National Academy of Sciences, 117(15), 8344–8352.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

EDPB (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted 17 December 2024.

EDPB (2025). Public consultation on Guidelines 01/2025 on pseudonymisation: published responses.

EDPB (2026). Guidelines 02/2026 on anonymisation, version 1.0. Adopted 7 July 2026 for public consultation.

Federal Reserve and OCC (2011). Supervisory guidance on model risk management (SR Letter 11-7). Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency.

Findata (2025). Anonymity form for AI and machine learning models. Finnish Social and Health Data Permit Authority.

Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A. and Woods, C. (2019). Handbook on statistical disclosure control for outputs. Safe Data Access Professionals Working Group.

ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories.

JCGM 100:2008. Evaluation of measurement data – Guide to the expression of uncertainty in measurement.

JCGM 106:2012. Evaluation of measurement data – The role of measurement uncertainty in conformity assessment.

Jebreel, N., Sánchez, D. and Domingo-Ferrer, J. (2026). A critical review on the effectiveness and privacy threats of membership inference attacks. arXiv:2603.22987.

Jefferson, E. et al. (2022). GRAIMATTER Green Paper: Recommendations for disclosure control of trained machine learning (ML) models from trusted research environments (TREs). Zenodo, doi:10.5281/zenodo.7089491.

Lermen, S., Paleka, D., Swanson, J., Aerni, M., Carlini, N. and Tramèr, F. (2026). Large-scale online deanonymization with LLMs. arXiv:2602.16800.

NIST (2025). Guidelines for evaluating differential privacy guarantees. NIST Special Publication 800-226.

Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices.

Smith, J. et al. (2025). Safe machine learning model release from trusted research environments: the SACRO-ML package. arXiv:2212.01233v4.

Steinke, T., Nasr, M. and Jagielski, M. (2023). Privacy auditing with one (1) training run. Advances in Neural Information Processing Systems (NeurIPS).