Subject: Clarification of re-identification risk, practical safeguards, and a proportionate interpretation of scientific research requirements under the GDPR

1. General comment

The draft Guidelines provide welcome clarification on the notion of scientific research, the legal bases available under the GDPR, the relevance of Article 89 safeguards, and the role of transparency, accountability and governance in research activities. In particular, the Guidelines rightly confirm that scientific research under the GDPR must be interpreted broadly and may include privately funded and commercial research, as long as the relevant activities are genuinely scientific and meet the key indicative factors identified in the draft.

At the same time, the draft Guidelines would benefit from further clarification on one recurring difficulty in practice: the assessment of re-identification risks, including through data correlation, dataset combination, or longitudinal and multi-source analysis. The current text already recognises the relevance of such risks, including when datasets are combined, when data are retained over time, or when pseudonymised data are further used in research environments.

However, in practice, diverging interpretations at national level may lead to an overly restrictive understanding of such risks. This may create legal uncertainty for controllers and processors carrying out legitimate scientific research, including in highly regulated environments. For this reason, the Guidelines could usefully provide stronger clarification that the GDPR requires **risk management and proportional safeguards**, rather than the elimination of any purely theoretical possibility of re-identification. This would be consistent with the general architecture of the GDPR, and with the central role of Article 89(1), which expressly requires appropriate safeguards adapted to the concrete research context.

2. Re-identification through correlation should not be treated as a disqualifying factor in itself

The Guidelines already acknowledge that controllers must consider risks of re-identification, including where personal data are combined with other datasets or retained over time, and that safeguards must be assessed in light of the nature, scope, context, purposes and risks of the processing.

We respectfully suggest that the final Guidelines should go one step further and clarify that:

The mere theoretical possibility of re-identification through data correlation, data combination or linkage should not, in itself, prevent lawful processing for scientific research purposes, nor automatically exclude reliance on protective techniques such as anonymisation or pseudonymisation where those techniques are effective in context.

Such a clarification would be especially useful because many forms of modern scientific research necessarily involve:

- longitudinal analysis,
- pattern detection across multiple variables,
- cross-validation of datasets,
- and controlled combination of information originating from different sources.

These features are not anomalies; they are often intrinsic to robust scientific methodology. The Guidelines themselves recognise that genuine scientific research may include technological development, applied research, privately funded research, and the use of research data infrastructures, ancillary processing, and future research use within a defined research area.

Accordingly, the existence of a correlation risk should not be treated as a quasi-automatic obstacle. It should instead trigger a structured assessment of necessity, proportionality and safeguards.

3. The Guidelines should more explicitly reject any implicit "zero-risk" standard

The draft takes an important step by clarifying that anonymised data should be used where possible, and that pseudonymisation is required where the research purposes cannot be fulfilled through anonymised information. It also emphasises that directly identifying data should only be used where this is strictly necessary and proportionate.

This is a sound and balanced approach. However, the final version could usefully make even clearer that:

The GDPR does not require research controllers to eliminate every conceivable residual risk, but to reduce such risk, through appropriate technical and organisational safeguards, to a level that is acceptable, proportionate, and not reasonably likely to materialise in practice.

This point is particularly important for scientific research, because the combination of:

- retained datasets,
- reproducibility requirements,
- quality control,
- external scrutiny,
- and future or secondary scientific use,

means that some residual risk may remain even in well-governed research environments. The Guidelines already recognise that data may legitimately be stored for longer periods for scientific research purposes, including where future research projects are reasonably foreseeable within a given research area, provided that Article 89 safeguards are implemented and necessity is reviewed over time.

A clarification on acceptable residual risk would therefore improve legal certainty and would better reflect the operational reality of scientific research without weakening data subject protection.

4. Correlation is not merely a source of risk; in many contexts it is a methodological necessity

Without adopting a sector-specific framework, the Guidelines could usefully acknowledge that in many fields of scientific research, including but not limited to medical, public health, technological and data-intensive research, the controlled correlation of variables is an intrinsic component of the scientific method.

The draft already recognises that scientific research includes technological development, demonstration, applied research, privately funded research, and research infrastructures designed to support future studies. It also recognises that scientific value lies in contributing to knowledge, applying knowledge in novel ways, and generating verifiable results through systematic methods.

In that light, the final Guidelines could include language such as:

The use of correlated, longitudinal, or multi-source datasets may be scientifically necessary in order to achieve valid and reproducible research results. The existence of such methodological necessity should be duly taken into account in the assessment of necessity and proportionality under the GDPR.

This would not diminish the need for safeguards. Rather, it would help ensure that data protection analysis remains connected to the real conditions under which serious, methodologically sound research is actually carried out.

5. Practical guidance on safeguards would materially improve the usefulness of the Guidelines

One of the strongest elements of the draft is its insistence on safeguards under Article 89(1), including anonymisation or pseudonymisation, ethical oversight, secure processing environments, privacy-enhancing technologies, confidentiality arrangements, conditions for further use, role allocation, and governance mechanisms adapted to the context.

However, in practice, controllers often need more concrete guidance in order to assess whether their safeguards are sufficient where some residual re-identification or correlation risk remains. The final Guidelines would therefore be significantly strengthened by including **non-exhaustive practical examples** of safeguards that the Board considers relevant and effective.

In particular, the Guidelines could recommend, where appropriate:

a) Functional separation of roles

- separation between the entity or team holding identifiers and the team carrying out the scientific analysis;
- separation between data custody, data preparation and data analysis functions;
- preventing analysts from accessing identity keys or correspondence tables, except where strictly necessary and duly controlled.

This would be fully consistent with the emphasis placed in the draft on role allocation, pseudonymisation, confidentiality, and access restriction as part of appropriate safeguards.

b) Limiting effective re-identification capacity of analysis teams

The final Guidelines could usefully recognise that re-identification risk is materially reduced where the teams analysing pseudonymised datasets:

- do not have access to identity information;
- do not have practical means to re-link records;
- and, where relevant, do not possess external knowledge or contextual information that would significantly increase re-identification capacity.

This could be formulated carefully, for example by referring to **limitations in access, contextual knowledge and practical re-linking capacity**, rather than by creating a rigid sector rule. Such clarification would be particularly valuable in contexts where health or otherwise sensitive data are analysed by technical, statistical, or data science teams that are organisationally separated from clinical or operational functions.

c) Clear data lifecycle management

The draft already underlines the need to determine retention periods or criteria, to regularly review continued necessity, and to assess in which format data should be retained over time.

The final Guidelines could build on this by recommending that controllers:

- define retention periods linked to the scientific objective;
- erase or irreversibly anonymise data as soon as continued identification is no longer necessary;
- conduct periodic necessity reviews;
- and ensure that the risk of future correlation does not persist longer than scientifically justified.

This would be especially helpful to prevent the mistaken view that research retention is by nature indefinite.

d) Secure hosting by the controller where safeguards are ensured

The Guidelines rightly mention secure processing environments, role-based access controls, confidentiality arrangements, and federated or otherwise controlled access models.

To avoid unnecessarily restrictive interpretations, the final text could explicitly state that:

Personal data may, where justified and adequately protected, be securely hosted by the controller itself, including in the controller's own controlled environment, provided that appropriate technical and organisational measures effectively ensure confidentiality, integrity, access limitation and compliance with Article 89(1).

This would be very helpful in practice. It would make clear that the GDPR does not require a specific hosting architecture as such; rather, it requires effective safeguards adapted to the risks.

e) Restrictions on downstream use and re-linking

The Guidelines could also recommend:

- contractual bans on unauthorised re-identification attempts,
- strict purpose limitations for recipients,
- logging and audit trails,
- restrictions on export or local download,
- and sanctions or technical barriers against unauthorised linkage.

These examples are already broadly reflected in the spirit of the draft, especially in the sections on secure environments, data access conditions, confidentiality arrangements, and contractual or governance safeguards.

6. Strong pseudonymisation should be recognised as a robust protective standard in many research contexts

The draft appropriately states that anonymisation should be used where possible, but that pseudonymisation is the relevant measure where anonymisation would prevent the scientific purpose from being achieved. It also underlines that pseudonymisation is a technical and organisational measure that helps controllers and processors meet their obligations under Articles 25, 32 and 89 GDPR.

In our view, the final Guidelines could usefully articulate more clearly that:

In many types of scientific research, robust pseudonymisation, combined with strict governance, access controls, secure environments and role separation, may provide a sufficiently high level of protection even where theoretical re-identification scenarios remain conceivable.

This would not equate pseudonymised data with anonymous data. But it would appropriately reflect the role that pseudonymisation is intended to play under the GDPR, particularly in legitimate research environments where continued linkability is scientifically necessary for validation, quality control, longitudinal follow-up, or reproducibility.

7. Importance of consistency across Member States

The draft repeatedly emphasises accountability, demonstrability, lawful basis assessment, transparency, safeguards, and consistent role determination among controllers, processors and joint controllers.

To complement this, it would be very valuable if the final Guidelines also encouraged a genuinely harmonised supervisory approach. We would therefore suggest adding language such as:

Supervisory authorities should ensure a consistent application of these principles and avoid interpretations under which purely theoretical or highly remote re-identification scenarios are treated as automatically preventing otherwise legitimate scientific research, despite the effective implementation of appropriate safeguards.

This would support:

- legal certainty,
- cross-border research projects,
- the European Research Area,
- and the ability of researchers and research-supporting organisations to design compliant governance structures with confidence.

8. The interpretation of Article 89 should remain innovation-enabling, provided safeguards are real and demonstrable

The draft already recognises the major societal value of scientific research, including where conducted by private entities and on a commercial basis, and acknowledges that such research may rely on legal bases such as legitimate interest or public interest where the conditions are met.

In that spirit, the final Guidelines could more expressly confirm that:

The interpretation of the GDPR in the context of scientific research should protect data subjects effectively while remaining compatible with scientific progress, methodological robustness and innovation, provided that appropriate safeguards are implemented and documented.

Such wording would remain faithful to the structure of the GDPR and to the draft's own internal logic:

- research is valuable,
- not all future purposes can always be precisely known from the outset,
- processing may continue over time,
- and safeguards are the mechanism through which rights and freedoms are protected in context.

9. Suggested drafting addition

If the Board considers it useful, the following paragraph could be inserted, for example in the section on safeguards or risk analysis:

The Board recalls that the risk of re-identification through data correlation, dataset combination or longitudinal use is inherent in many scientific research activities. Such risk must be assessed in light of the concrete context of processing and of the means reasonably likely to be used in practice, taking due account of the safeguards implemented pursuant to Article 89(1) GDPR. The mere theoretical possibility of re-identification does not, in itself, render processing unlawful nor exclude the use of anonymisation or pseudonymisation techniques where those techniques are effective in the relevant context. Appropriate safeguards may include, inter alia, functional separation of roles, strict access controls, secure hosting and processing environments, contractual and technical restrictions on re-linking, and data lifecycle limitations, including erasure or anonymisation once identification is no longer necessary. Where such safeguards are effectively implemented, a residual risk that is not reasonably likely to materialise should be regarded as compatible with the GDPR framework for scientific research.

This addition would, in our view, significantly enhance the practical utility and harmonising value of the final Guidelines.