European Network of Expertise (NoE) on Omics

(Joint Action JANE-2 – GA 101183265 – WP9)

Contribution to

"Data Minimisation, Pseudonymisation, Anonymisation and Synthetic Data" (*Public consultation, November 2025*)

Public consultations - Tehdas

1. PURPOSE

The European Network of Expertise on Omics (NoE Omics), established under Joint Action JANE-2 and coordinated by Sciensano, Belgium, welcomes the opportunity to provide feedback on the TEHDAS2 draft guideline *Data Minimisation, Pseudonymisation, Anonymisation and Synthetic Data*.

Omics technologies (genomics, transcriptomics, epigenomics, proteomics, metabolomics, and their integration) are reshaping cancer care, rare disease management, and personalised medicine. They produce datasets of exceptional granularity and scientific value, but also of heightened sensitivity under the GDPR.

This contribution evaluates the TEHDAS2 draft in light of these omics realities. It highlights the strengths of the proposed framework, identifies critical gaps, and offers recommendations that address the needs of the wider health data community while reflecting the specific challenges of omics.

The draft guideline (M7.2) establishes an operational framework for the safe and secure secondary use of health data within the European Health Data Space (EHDS) Regulation (EU 2025/327). It anchors data minimisation in GDPR Article 5(1)(c), recognises pseudonymisation as a safeguard under Articles 25 and 32, and correctly frames anonymisation as non-binary and non-permanent under Recital 26. It also acknowledges synthetic data as a potential instrument for research and innovation.

2. Strengths of the draft guideline

The TEHDAS2 guideline provides a coherent operational framework for secondary use of health data under the EHDS Regulation (EU 2025/327). It anchors data minimisation in GDPR Article 5(1)(c), recognises pseudonymisation as a safeguard under Articles 25 and 32, and correctly frames anonymisation as non-binary and non-permanent under Recital 26.

The guideline structures the EHDS user journey, data discovery, preparation, access, use, and finalization, and assigns responsibilities to Health Data Access Bodies (HDABs). It requires minimisation and re-identification risk assessments before permits are issued, and mandates pseudonymisation before data enters secure processing environments (SPEs). This phased approach strengthens trust and interoperability.

The draft also acknowledges synthetic data as a potential tool, referencing EDPB Opinion 28/2024 on AI models. This recognition is important for future AI-driven research.

3. Challenges in the Omics Context

Omics data pose challenges that the draft does not fully address. Whole genome sequences and multi-omics profiles remain inherently identifying even when pseudonymised. The CJEU ruling in Case C-413/23 P (EDPS v SRB) confirmed that pseudonymised data may still qualify as personal data depending on context. For omics, identifiability is persistent.

While the likelihood of successful re-identification from omics data alone, particularly from methylation patterns or isolated mutational profiles, may be low and would typically require disproportionate effort, recent literature nonetheless highlights that certain high-resolution datasets produced in precision medicine can share characteristics with rare-disease cohorts. In such settings, the combination of very small sample sizes with highly specific molecular or clinical descriptors can increase the possibility that some data points function as quasi-identifiers. This does not imply that methylation or variant data independently enable re-identification; rather, risks generally arise when such granular biological features are combined with demographic or clinical details that narrow the cohort sufficiently. Consequently, a number of studies caution that standard anonymisation measures such as k-anonymity or simple aggregation may not always be adequate for very small or highly specific cohorts, and may require careful, context-specific reinforcement within a broader safeguards framework (as also reflected in EFPIA's existing risk-based approach to secondary use of clinical trial data).

In the future, as health data becomes predominantly electronic and digitally linked across systems, the potential for unintended re-identification, profiling, or misuse will grow. Linkage across datasets, whether clinical, genomic, behavioral, or administrative, can amplify privacy risks, especially when quasi-identifiers (e.g. rare disease codes, variant combinations, or location-time stamps) function as direct identifiers. These risks must not remain hidden in technical silos or buried in consent forms. Instead, they should be made transparent to all stakeholders, and mitigated through robust technological safeguards, governance frameworks, and ethical oversight.

These challenges should not remain hidden in technical details or confined to lengthy consent forms. Instead, they deserve to be openly acknowledged and addressed in ways that build trust among patients, citizens, and researchers. Authorities and supervisory bodies can play a crucial role by offering clear guidance, robust recommendations, and practical frameworks that evolve alongside technological progress. Such support should ensure that protective measures are not only legally sound but also technically feasible, funded, and adaptable to the realities of modern research.

We need to combine transparency with investment in technical solutions so we safeguard privacy while enabling innovative research that advances preventive care and improves outcomes for patients and citizens.

The draft underspecifies *how* minimisation should apply to omics. Practical measures, limiting granularity, restricting time windows, masking rare variants, require explicit guidance. Without this, HDABs will apply inconsistent standards.

Synthetic data offers promise for AI model training, but omics synthetic datasets must be validated for both privacy and utility. The draft mentions documentation of model resistance to re-identification but does not provide omics-specific criteria.

Further gaps include:

- **Risk categorisation:** The draft lacks a framework distinguishing identity disclosure, attribute disclosure, and group disclosure risks. Omics datasets require this differentiation.
- **Omics and rare disease safeguards in precision medicine:** Guidance for managing identifiability risks in ultra-rare or highly specific sub-cohorts in precision-medicine research remains limited. Although re-identification generally depends heavily on demographic or contextual leakage rather than on biological features alone, certain ultra-rare disease codes or very specific variant constellations may, when combined with other contextual information, increase the risk of singling out individuals. Existing frameworks, including EFPIA's safeguards for secondary use of clinical trial data, already rely on layered, risk-proportionate controls; however, additional clarity would support consistent application when dealing with exceptionally small or unique cohorts.
- **Proportional governance:** The draft does not integrate proportionality tools to calibrate safeguards according to data sensitivity, cohort rarity, and AI usage. Networks such as JANE2 can contribute evidence and assessments to design and refine such tools, with the proportionality grid offering a concrete model for implementation.
- **Cross-border pseudonymisation:** No harmonised schema exists across Member States, creating duplication and linkage failures.
- **Metadata and provenance:** Omics requires standardised metadata and provenance logs for auditability. The draft does not address this.
- **Clinically significant findings:** Workflows for Health Data Access Bodies (HDABs) could be designed to ensure that clinically significant omics findings are routed to clinicians or patients in line with opt-in/opt-out preferences. Because omics datasets often generate inferred data, such as pathogenic variants or risk profiles, governance should build on existing frameworks (GDPR, EHDS, CTR, IVDR, bioethics laws) and clarify what protections apply, who holds responsibility for communication, and how patient preferences are respected. Embedding these safeguards would strengthen trust and support the promise of personalised medicine.
- **Federated analysis patterns:** SPEs need reference designs for privacy-preserving federated analytics. These are absent.

4. Recommendations

4.1 GENERAL RECOMMENDATIONS

- Introduce a framework distinguishing identity, attribute, and group disclosure risks.
- Require re-identification risk reviews for longitudinal and multi-omics datasets.
- Provide HDABs with operational templates: checklists, DPIA exemplars, pseudonymisation standards.
- Develop reference patterns for SPEs enabling federated queries, differential privacy, and synthetic data use.
- Standardise metadata catalogues and provenance logs to support FAIR principles and auditability.
- Strengthen transparency and opt-out rights, and clarify communication of clinically significant findings.

4.2 OMICS-SPECIFIC RECOMMENDATIONS

- Integrate proportionality model to calibrate safeguards according to omics data sensitivity and cohort rarity.
- Harmonise pseudonymisation across Member States, with trusted third-party orchestration, to enable lawful linkage.
- Require omics-specific assessments and validation paths of synthetic datasets for both privacy and utility.
- Define HDAB criteria and pathways for routing validated pathogenic variants to clinicians, with patient consent options.

4.3 LEGAL AND CONTRACTUAL CONSIDERATIONS

- Add model contractual language covering ownership, custodianship, publication rights, intellectual property, return or destruction of data, and breach notification.
- Define minimum contractual and technical assurances for Trusted Third Parties and Trusted Data Holders, including service level agreements, audit rights, custody rules, jurisdictional governance, and liability allocation.
- Publish illustrative parameter ranges (e.g., k-anonymity values, epsilon ranges for differential privacy, uniqueness thresholds).
- Provide accreditation checklists for tools, including validation, provenance management, and metadata signing.
- Introduce a machine-readable metadata schema aligned with HealthDCAT-AP and EHDS Article 78 fields.
- Clarify export governance with stepwise approval flows for different classes of outputs (tables, images, trained models), reviewer roles, timelines, appeal loops, and record retention periods.
- Address operational costs and timeliness through clear resource estimation, defined service-level expectations, and pragmatic procedures for large or complex extractions. These parameters must be standardised within the guideline and not left solely to the discretion of individual stakeholders, as

inconsistent practices would undermine predictability, efficiency, and trust across the European Health Data Space.

RARE DISEASES AND OMICS: CONTRIBUTION TO TEHDAS2 GUIDELINES

Personalisation and Uniqueness

Rare diseases and omics technologies are inseparable. For most rare conditions, clinical signs alone cannot provide certainty. Genomics, transcriptomics, epigenomics, proteomics, metabolomics, and multi-omics integration are the only tools that can reveal pathogenic variants, stratify patients, and guide targeted therapies. Omics enables personalised medicine, where treatment decisions are tailored to the unique molecular profile of each patient. For rare disease patients, this personalised approach is not optional, it is the only path to diagnosis and care.

The uniqueness of omics data is both its strength and its vulnerability. Rare disease datasets are small, granular, and often singular. A single variant, methylation profile, or proteomic signature can identify an individual. Under GDPR, such data are “special categories” requiring the highest protection. The Court of Justice of the EU (Case C-413/23 P) confirmed that pseudonymised data may still be personal data if re-identification is reasonably possible. In rare diseases, re-identification is inherent because the data itself is unique. Disease codes or variant combinations can act as direct identifiers in ultra-rare cohorts.

The TEHDAS2 draft guideline on data minimisation, pseudonymisation, anonymisation, and synthetic data provides a strong foundation for secondary use of health data under the EHDS. It anchors minimisation in GDPR Article 5(1)(c), requires pseudonymisation before data enters secure processing environments, and recognises anonymisation as non-binary and non-permanent. It also acknowledges synthetic data as a potential tool. These principles are essential, but without omics-specific safeguards they risk undermining the personalised care rare disease patients depend on.

Minimisation must be adapted: reducing granularity or masking rare variants can erase the diagnostic signal and compromise personalised medicine. Pseudonymisation cannot eliminate identifiability when the dataset itself is unique. Anonymisation collapses immediately in small cohorts. Synthetic data offers promise for AI model training, but omics-specific validation is needed to ensure both privacy and preservation of biological signal.

Proper implementation of TEHDAS2 is therefore critical. Rare disease research depends on pooling omics data across Member States. National divergences in consent, biobank governance, and localisation rules already hinder collaboration. If TEHDAS2 is implemented inconsistently, omics and rare disease consortia will face fragmentation, delays, and loss of trust. Conversely, a harmonised, proportionate, and

omics-aware implementation will enable secure cross-border data sharing, accelerate personalised diagnostics, and deliver equitable access to innovation.

Omics consortia such as JANE2 bring forward practical approaches that can strengthen TEHDAS2 in addressing the realities of rare disease and personalised medicine. They provide proportionality models to calibrate safeguards according to cohort rarity, comparative assessments to highlight national divergences, and contractual frameworks to enable federated biobanking and cross-border collaboration. Integrating these outputs into TEHDAS2 will ensure that omics data, unique by nature and essential for personalised care, are governed responsibly, balancing protection with utility. Rare disease omics exemplifies the tension between privacy and diagnostic value: without tailored safeguards, minimisation and anonymisation risk erasing clinically meaningful signals. Embedding omics-specific approaches into TEHDAS2 will allow Europe to handle these datasets responsibly, balancing protection with innovation, and securing trust across borders.

7. Conclusion

The TEHDAS2 draft guideline provides a strong foundation for the secondary use of health data under the EHDS. It rightly emphasises minimisation, pseudonymisation, and anonymisation, and introduces synthetic data as an emerging tool.

However, omics data and precission medicine contexts require additional safeguards and proportional governance. Contributions from omics consortia such as JANE2 can help the guideline better reflect these realities by offering regulatory insights, comparative assessments of national practices, proportionality models, and contractual frameworks for federated biobanking and data sharing. Integrating such outputs will ensure that omics innovation progresses responsibly, balancing privacy with utility, and fostering citizen trust.

The European Network of Expertise (JANE2) on Omics therefore recommends that TEHDAS2 include an omics-specific annex, adopt risk-based and proportional approaches, harmonise pseudonymisation across Member States, and strengthen metadata, logging, and workflows for clinically significant findings. These measures will benefit the wider health data community while addressing the unique challenges of omics.

On behalf of the NoE on Omics in Cancer,

Contact details:

NoE on Omics Technologies: JANEWP9 JANEWP9@sciensano.be

- Hélène Antoine-Poirel Helene.Antoine-Poirel@sciensano.be
- Inge Smeers Inge.Smeers@sciensano.be

Task 9.3 - Legal and Ethical Aspects:

- Wannes Van Hoof Wannes.VanHoof@sciensano.be
- Roxana Albu Roxana.Albu@sciensano.be