European Network of Expertise (NoE) on Omics in Cancer

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

Contribution to the EDPB Guideline entitled

" Guidelines 2/2026 on Anonymisation, version 1.0, adopted 7 July 2026"

Operational clarification of the EDPB anonymisation criteria for omics and multi-omics research

30 October 2026

TABLE OF CONTENTS

JANE2 OMICS TECHNOLOGIES IN CANCER RESEARCH AND CARE 2
EXECUTIVE SUMMARY 3
1. WHY HEALTH AND OMICS NEED OPERATIONAL GUIDANCE 6
2. KEY CLARIFICATIONS REQUESTED FROM THE EDPB 7
3. WHY OMICS NEEDS ITS OWN PRACTICAL TOOLBOX 11
4. PRESERVE SCIENTIFIC UTILITY WHILE REDUCING EXPOSURE 15
5. PROPOSED LIFECYCLE, STAKEHOLDER AND RESPONSIBILITY MATRIX 19
6. A PRACTICAL REASSESSMENT MODEL 21
7. CONSOLIDATED REQUESTS TO THE EDPB 22
8. CONCLUSION 24
APPENDICES: 25
PARTICIPANTS IN THE JANE-2 WP9.3 ELSI TASK FORCE THEMATIC MEETING 25
DATA-TYPE-SPECIFIC IDENTIFIABILITY CHECKS: MODALITY AND REPRESENTATION 25

THE CENTRAL REQUEST

The JANE-2 Network of Expertise on Omics Technologies asks the EDPB to translate its principles into a practical, **proportionate methodology for health and research using the various types of omics high-throughput data**, with clear assessment criteria, worked examples, minimum documentation and **an omics-specific toolbox**. The EDPB should also promote harmonised, field-specific **Good Data Protection Practices** that embed effective safeguards into research workflows, clarify responsibilities and preserve scientific utility. The goal should be to protect individuals while enabling responsible data reuse, avoiding **unnecessary restrictions** and ensuring that **data protection and scientific progress work hand in hand** for the benefit of patients and society. This also requires recognising the importance of balancing data protection with data traceability, quality assurance, clinical utility, and the rights of patients and citizens in health research and care. For those working on, applying or advising on legal, ethical and information security aspects of health data use, clarity in authoritative guidance is essential. Principles alone are not sufficient: guidance must be practical and supported by concrete examples that reflect the specific needs, risks and operational realities of the relevant field.

On behalf of:
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:
- Roxana Albu Roxana.Albu@sciensano.be
- Wannes Van Hoof Wannes.VanHoof@sciensano.be

JANE2 Omics Technologies in Cancer Research and Care

What is the JANE-2 Network of Expertise on Omics Technologies? JANE-2 (Joint Action Networks of Expertise in Cancer, https://jane-2.eu/) is an ambitious initiative, stemming from Europe's Beating Cancer Plan, to implement seven new European Networks of Expertise in different cancer conditions, addressing: 1) complex and poor prognosis cancers; 2) palliative care; 3) survivorship; 4) personalized primary and secondary prevention; 5) omics technologies, 6) hi-tech medical resources; and 7) adolescents and young adults with cancer. It represents 121 partners across 29 European countries. Omics technologies are revolutionizing the management of patients with cancers, haematological malignancies, as well as rare diseases, by allowing for the comprehensive study of various biological molecules (DNA – genomics - , RNA – transcriptomics - , proteins – proteomics - , etc.) within cells or fluids. By analysing different layers of biological information, multi-omics enables a deeper understanding of how these molecules interact and influence health and disease. In cancer, omics can reveal new biomarkers for risk assessment and early diagnosis, identify therapeutic targets, and personalize treatments and monitoring based on a patient's unique molecular profile. Genomics and transcriptomics are currently the most widely used in clinical practice. However, emerging fields such as methylomics, proteomics, and metabolomics, as well as multi-omics integration, are rapidly advancing and poised to play a significant role in the future. The ambition of the Network of Expertise on Omics (https://jane-2.eu/omics-technologies/) is to support the integration of innovative omics technologies into the standard of care at the different steps of cancer management in a sound and sustainable manner, achieving equitable access to these services for all EU citizens.

Why this matters for the anonymisation Guidelines. Omics molecules are very heterogeneous and do not have the same risk of reidentification, from constitutive DNA to proteins expressed in cancer cells. Omics technologies workflows span discovery, validation, clinical implementation and research reuse. They may involve in-house analytical pipelines, laboratory-developed methods, evolving reference resources and quality systems, including ISO-based accreditation and institutional data governance. The EDPB guidance should complement these arrangements rather than duplicate them, while making clear that quality accreditation or ethics oversight does not by itself establish that data are anonymous. A proportionate, context-sensitive approach is especially important for rare diseases, small molecularly-defined entities, cross-border collaboration and longitudinal studies, where transformations that reduce identifiability may also undermine scientific validity.

Addressing practical challenges in anonymisation and pseudonymisation of omics data under the GDPR. Omics research involves diverse and evolving data types and processing methods, each presenting different identification risks depending on the omics modality (e.g. genomics, transcriptomics, proteomics or metabolomics), the format and granularity of the data, data volume, number of omics layers, availability of auxiliary information, linkage possibilities and processing context. Applying anonymisation requirements without considering these differences may lead to inconsistent assessments, unnecessary restrictions and loss of scientific validity, particularly in rare diseases and cancer research. Existing quality standards, ethical oversight and institutional safeguards provide important context but do not establish anonymity under the GDPR. The EDPB should clarify how to assess the three criteria (No Record Isolation, No Linkage and No Inference) in realistic omics scenarios, recognising pseudonymisation as risk reduction rather than anonymisation. Practical guidance, supported by worked examples and adaptable templates, should enable consistent, proportionate assessments across Europe, preserve scientific utility and clarify when reassessment is needed following material changes in data, access conditions or linkage possibilities. The objective is to protect individuals' rights and strengthen their ability to exercise control over their health data, while enabling responsible cross-border research and secondary use of health data.

Executive summary

The JANE-2 Network of Expertise on Omics Technologies welcomes the EDPB's efforts to clarify when information can be considered anonymous under the GDPR. The draft's contextual approach and its three assessment criteria (No Record Isolation, No Linkage and No Inference) provide a valuable foundation.

However, the multidisciplinary expert Task Force of the above-mentioned network of experts identifies a key challenge: ***translating these principles into practical, consistent, and reproducible assessments for health data and omics research in a risk-based approach.***

The secondary use of Omics technologies for research purposes highlights why a single, generic checklist may not be sufficient. The assessment needs to account for both the type of omics data, the number of omics layers, and how those data are represented and processed. For example, a whole-genome sequence or raw sequencing file, a small fragment of DNA sequence, an individual-level gene-expression or protein-abundance matrix, an aggregated dataset, and the integration of multiple layers of omics (ie genomics, transcriptomics and methylomics) may present different identification risks. These risks also depend on the characteristics of the cohort, such as its size and rarity of the condition under study (=denominator), the availability of additional information, and the realistic possibilities for linking or inferring information about individuals (sets of arguments). As highlighted in the Task Force discussions, the assessment should therefore distinguish between omics data types, modalities and representations, rather than assuming that all data within a given omics category carry the same risk. The extent to which re-identification risks have been studied also varies across omics modalities. These risks have been extensively investigated in genomics, to a lesser extent in transcriptomics and methylomics, while other omics modalities remain comparatively underexplored.

Data Protection Guidance should explain how to apply the three criteria to these different scenarios, taking into account the relevant actor's access to information and technical capabilities, the intended purpose and sharing conditions, and the potential for identification through linkage or inference. This would support more consistent and proportionate assessments while preserving the scientific value of omics data.

Balancing data protection, traceability, scientific quality and patient rights in health research and care

In health research ability to reconnect data to an individual is sometimes essential (in addition to link different datasets at the individual level for research purposes. Maintaining a link to the source or additional information can also enable data linkage for specific research questions, where justified and subject to appropriate safeguards. This can help researchers and healthcare professionals verify data, perform quality checks, correct errors, and strengthen the accuracy and reliability of research findings.. This is not only important for scientific validity but also for complying with established healthcare codes of practice and respecting other patient rights, such as access to information, correction of inaccurate data and, where appropriate, receiving clinically relevant results (according to their informed consent for incidental findings). These processes can have direct consequences for the health and treatment of the individual concerned (even if it is not the primary goal of research studies) and may also inform decisions affecting other patients with similar conditions and findings. For the sake of clarity, in the health sector, incidental findings refer to results that are not the primary focus of the research but whose clinical significance or utility is already known at the time of the study. Discovery findings, in contrast, arise from the research itself and may reveal previously unknown associations, variants or other clinically relevant information. While such discoveries may not yet have established clinical utility for the individual concerned, they may generate knowledge that could inform future research, clinical practice or decisions affecting other patients with similar characteristics or conditions.

The EDPB should therefore recognise that privacy and security safeguards on a risk-based approach must be built into the way data are made usable for their intended research, rather than assuming that removing the possibility of reconnecting data to an individual is always the appropriate objective. Health research already operates within established regulatory and quality frameworks: for example, clinical trial sponsors and service providers must comply with Good Clinical Practice and other applicable requirements, while health care professionals, data stewards, bio-IT experts and personnel accessing health data are subject to training, confidentiality obligations, institutional policies and relevant ethical and legal standards. Controlled access, secure systems, distinct data workflows for care and research uses, governance procedures and professional accountability are therefore important parts of the protection framework and should be considered in the overall risk assessment. They do not automatically make data anonymous, but they are relevant safeguards when determining what level of protection is appropriate and proportionate.

The work undertaken within the JANE-2 Network of Expertise on Omics Technologies provides a practical proof of concept in this regard. Focusing on cancer research and the integration of genomic and other omics data into cancer care, it illustrates how risk-based safeguards, controlled access and appropriate governance arrangements can support the responsible use of sensitive health data while preserving their scientific and clinical value. Although developed in the context of cancer, this approach could be adapted to other disease areas, taking into account differences between rare and common diseases and between germline and somatic data. These distinctions matter because the potential for identification, linkage and inference may vary depending on the disease context, the type of omics data and the availability of related information. The experience therefore offers a basis for developing practical guidance that can be applied beyond oncology, while remaining sensitive to disease-specific risks.

This also addresses the comments by making clear that:

- The example concerns clinical research and the integration of omics into cancer care, rather than clinical care alone.
- The JANE2 work is presented as a proof of concept, not as a universal model.
- Transferability to other disease areas requires consideration of the differences between rare and common diseases.

WHAT WE ASK THE EDPB TO DO

The guidance should clarify the key legal and methodological questions, define the minimum elements of a reproducible assessment, and provide practical, usable examples for the use of the various omics data in health and research. It should establish overarching principles while addressing the specific identifiability risks of individual omics modalities and the additional risks arising from their combination in multi-omics datasets. Given the highly regulated nature of this field (health), the guidance should be developed with input from all relevant stakeholders, including data protection officers, IT security specialists, data managers and researchers, to ensure that requirements are both legally sound and workable in practice. It should clarify responsibilities across the data lifecycle, including when reassessment is warranted, and help translate fundamental rights and data protection requirements into proportionate measures that safeguard individuals without imposing unjustified restrictions that offer no meaningful additional protection. The objective should be to enable data to be made as accessible as possible, within the applicable legal and ethical safeguards, to support valuable health research and ultimately improve health outcomes.

Five priority requests

Priority	Requested clarification
1. Operationalise the framework	Provide a step-by-step assessment method, practical questions, model documentation, worked examples and an indicative risk matrix.
2. Explain the context and actor perspective	Show how to assess realistic means for each relevant actor, including private sector, independent recipients, processors, trusted third parties and connected organisations.
3. Make the three criteria usable	Explain how record isolation, linkage and inference interact; distinguish a unique record, a narrowed cohort or an inferred attribute from a realistic ability to identify a natural person. Include a concrete linkage assessment question and worked examples.
4. Address omics-specific risk pathways	Include examples regarding omics technologies (large-scale germline genomics, vs small-scale, somatic genomics, molecular tumour expression profiles, single-omics vs multi-omics), disease conditions (cancer, rare vs common diseases), propositus vs family relatives and clinical/registry linkage.
5. Make reassessment and responsibilities explicit	Define material change triggers and offer a lifecycle responsibility matrix covering assessment, access, safeguards, approval, monitoring and review.

Scope and status of this contribution. *This document is based on the JANE-2 WP9.3 ELSI Task Force thematic meeting of 22 October 2026 on the EDPB Guidelines 02/2026 on Anonymisation (Version 1.0, adopted for public consultation on 7 July 2026). The participant list at the end records who attended the discussion.*

1. Why health and omics need operational guidance

The JANE-2 WP9.3 discussion brought together multidisciplinary expertise spanning data protection and law, data engineering, molecular biology, clinical genetics, oncology, biobanking, hospitals and research. These perspectives highlighted the diversity of data types, actors, and access arrangements involved in health and omics research.

The practical challenge extends beyond determining whether names or codes have been removed. Stakeholders need to assess what information is being considered, from whose perspective, for what purpose, in which environment, and with what realistic possibilities for linkage or inference. They must also distinguish the legal assessment of identifiability from the subsequent decisions on appropriate safeguards and access arrangements. Clear, practical guidance is needed to support consistent assessments across these different contexts and throughout the data lifecycle.

THE PRINCIPLE

Anonymity should be assessed in relation to the type of Omics data, the relevant actor, the means reasonably likely to be used, the processing context, the data usability and the safeguards. The conclusion should state its assumptions and the conditions under which it remains valid.

What is already valuable in the EDPB draft?

- It recognises that the status of information may differ depending on the relevant entity and the circumstances.
- It distinguishes pseudonymisation from anonymisation and treats anonymisation as a process rather than simply deleting identifiers.
- It provides three useful analytical lenses: No Record Isolation, No Linkage and No Inference.
- It recognises that technical, organisational and contextual factors influence whether identification is reasonably likely.



What needs to be clarified for implementation?

The draft would be more useful if it explained how to translate those principles into a reproducible assessment, how to evaluate plausible scenarios without requiring organisations to speculate about every imaginable attack, how to balance identifiability with scientific utility, and how to decide whether the appropriate response is anonymisation, pseudonymisation, controlled access or analysis in a secure environment.

2. Key clarifications requested from the EDPB

Define a practical assessment sequence

The Guidelines should include a common minimum sequence that organisations can apply and document. The method should support comparable assessments without imposing one universal numerical threshold or assuming that one technical technique works for every dataset.

Stage	*Minimum question / output*
01 Define	*Purpose, dataset, data granularity, intended output and scientific utility to preserve.*
02 Map	*Actors, roles, access rights, keys, data flows, external information and foreseeable linkage routes.*
03 Test	*Plausible isolation, linkage and inference scenarios for the specific data and context.*
04 Evaluate	*Means reasonably likely to be used: availability, access, legality, effort, time, expertise, cost and safeguards.*
05 Decide	*Document the conclusion for each relevant actor and select a proportionate data/access model.*
06 Revisit	*Review after material changes and at a proportionate, documented interval where appropriate.*

Clarify the contextual approach and relevant actors

The Guidelines should distinguish Omics-informed actors who are independent recipients with no realistic access path to identifying information from actors who remain functionally connected through a controller, ongoing linkage service, shared infrastructure or other practical access arrangements. It should also explain how the assessment changes where a processor acts on behalf of a controller, where organisations jointly determine purposes or means, or where a trusted third party holds linkage keys.

Actor / stakeholder	Questions the method should require
Data holder / originating hospital or laboratory	What identifying information and source records remain available? What purpose and legal context govern the data?
Controller / joint controllers	Who determines the purpose, data configuration, recipients, safeguards, and review conditions? Who documents the overall assessment?
Processor/platform / infrastructure	What can this actor actually access or derive? Can it combine datasets, export records, or access keys? What instructions and controls apply?
Trusted third party/linkage service	What linkage information is held? Is separation of duties real and technically enforced? Can the service or another actor reconnect records?
Research recipient / secondary user	What data and auxiliary information can the recipient reasonably obtain? What tools, expertise and access rights are available?
DPO, information security, ethics and governance functions	Who advises, challenges, approves access or safeguards, manages incidents and ensures the rationale is reviewable?

Clarify the three criteria without turning them into automatic labels

The three criteria (No Record Isolation, No Linkage and No Inference) provide a useful structure for assessing identifiability, but their practical application to omics data requires further clarification. The EDPB should explain how the criteria relate to one another and how their findings contribute to the overall assessment, without treating them as automatic labels or assuming that a single indicator, such as the uniqueness of a molecular profile, determines whether data are anonymous. Practical examples should help distinguish between singling out a record, linking it to a person and inferring information about an identifiable individual. This would support more consistent, transparent and proportionate assessments across different omics modalities and research contexts, in line with the risk-based approach under Recital 26 GDPR.

The proposed approach is to start with germline whole-genome sequencing (WGS), as one of the most sensitive and widely recognised categories of omics data, and use it as an initial reference case for risk assessment. The assessment should then be extended to other omics data types, examining their respective characteristics and risks. Depending on the data type, context, and potential for identification or linkage, some may warrant a lower level of protection and could be considered closer to basic biological data. Finally, particular attention should be given to integrated or multi-omics datasets, since combining different data types may create additional risks that cannot be adequately assessed by considering each omics layer in isolation.

EDPB criterion	*Clarification requested*	*Illustrative omics question*
No Record Isolation	*Explain the difference between distinguishing a record and identifying the natural person to whom it relates. Record uniqueness alone should not automatically settle the legal assessment.*	*Can a rare variant or unusual molecular profile distinguish one record? What realistic means connect that record to a person?*
No Linkage	*Explain how to identify realistic linkage routes and assess availability, accessibility and likely use of other information. Re-test the combined data after linkage and identify which additional tool is needed to reidentify an individual.*	*Could a molecular tumour profile be linked to clinical records, a registry, a biobank or published information?*
No Inference	*Distinguish inferring a characteristic from identifying the person to whom it relates; clarify the relevance of reliability and practical consequences and identify which additional tool is needed to reidentify an individual.*	*Can a profile reveal a disease characteristic or treatment response, and can that information reasonably be attributed to an identifiable person?*

For instance, in the case of the molecular tumour profile: should the distinction be made between genomic profiling and gene-expression profiling? These represent different biological information and have distinct risk profiles. Genomic variations are generally stable over time, whereas gene-expression profiles are more dynamic and depend on the tissue or cell type, physiological state, and specific conditions at the time of sampling. This distinction should be reflected in the risk assessment rather than treating all omics data as a single category.

The Guidelines should explain whether the three criteria are independent tests, cumulative conditions or complementary analytical perspectives. Where a criterion is not met, the guidance should explain what further analysis is needed and how the result informs the overall assessment under Recital 26 GDPR.

Define reasonably likely means and foreseeable scenarios

Organisations need a predictable way to distinguish plausible identification pathways from purely hypothetical possibilities. The Guidelines should ask assessors to document practical accessibility of additional information, realistic opportunities for access, legal and organisational barriers, expected effort, time, cost, expertise, computational capacity and the likely benefit of identification. They should clarify how security threats and unauthorised access fit into the anonymisation assessment without treating every imaginable future breach as conclusive proof that data can never be anonymous.

Reflection point: How should anonymisation risk assessments reasonably anticipate technological developments that may significantly change the possibilities of accessing, reading, linking or interpreting personal data? For example, technologies that may currently seem speculative could become feasible within the next three to five years. The Guidelines should consider how to account for foreseeable technological evolution without relying on purely hypothetical or unrealistic scenarios.

Clarify when reassessment is expected

The Guidelines should distinguish a legal requirement to assess identifiability in the relevant processing context from recommended risk-based governance practices for reviewing a prior conclusion. The guidance should explain that reassessment is particularly important when material changes affect the data, actors, access, linkage possibilities, or relevant capabilities; it should not imply a duty to continuously monitor every hypothetical technological development.

Material change / trigger	***When should an assessment be performed, and what criteria should guide it to ensure a proportionate approach without excessive or unjustified requirements?***
New omics type or layer or added variables	*Does the combined profile become more distinctive or reveal new attributes?*
Linkage to EHR, registry, biobank, imaging or another omics dataset	*Reassess the combined dataset, not just each source in isolation.*
New recipient, role or access model	*Reassess means available to that actor, including onward access and practical connections to other parties.*
New purpose or secondary use	*Check whether the data configuration, users, analytical needs or access arrangements change.*
New analytical technique, reference data or material capability change	*Review whether a previously implausible linkage or inference pathway has become realistic.*
Security incident or newly discovered exposure	*Review promptly in light of the incident and the information or access it creates.*
No material change identified	*Use a proportionate review schedule under the organisation's documented governance approach; do not imply that a fixed interval is a universal legal rule.*

What should change in the flowchart that summarises the 3 criteria ?

The EDPB should consider adding a short omics-specific assessment layer alongside the common decision tree. This would not require a separate legal test for every modality. Instead, it would explain how to answer the existing questions in light of the data representation, the scientific context and the relevant actor's realistic means and the resulting risk assessment.

In particular, the flowchart or its accompanying examples provided in the current EDPB Guideline should clarify:

- **What exactly is being assessed?** The assessment should distinguish between raw data, processed individual-level data, aggregated outputs, and linked multi-omics datasets. It should also account for the type of omics data, the volume and granularity of the data, and the number of omics layers combined, as these factors may lead to different risk profiles.
- **What does a "yes" actually establish?** A finding of uniqueness, possible linkage or meaningful inference should trigger the appropriate further analysis; it should not automatically determine the legal status of the data. Any resulting authorisation could also be time-limited and subject to clearly defined criteria and periodic review.
- **What is the relevant unit of assessment?** The guidance should address individual records, samples, cells, donors and combined datasets, including how these relate to the natural person, if applicable.
- **When should the assessment be repeated?** Linkage to new datasets, combining modalities, changing recipients or releasing additional outputs may alter the result.
- **How should uncertainty be handled?** Worked examples should identify the evidence needed to support a conclusion, rather than leaving stakeholders to interpret the yes/no branches without a clear decision rule.

The objective should be a **reproducible risk-assessment method** that remains sensitive and adaptive to scientific and technical differences and evolutions. This would help prevent both underestimation of genuine identification risks and disproportionate restrictions based on assumptions about an entire category of omics data. It would also support the wider objective of making data as accessible as possible for responsible research, within the safeguards required to protect individuals.

3. Why omics needs its own practical toolbox

The JANE-2 WP9.3 discussion highlighted that omics is not one uniform data category. The assessment should distinguish the biological modality from the data representation and processing level. Whole-genome sequence data, raw RNA-seq reads, processed expression matrices, protein-abundance matrices and cohort-level summaries do not carry identical identification pathways. At the same time, no representation should be declared anonymous solely because it has been processed, aggregated, or stripped of direct identifiers. The relevant question is what a realistic actor could isolate, link, or infer in the actual context.

COMMON CORE + MODALITY-SPECIFIC MODULES

When should an assessment be performed, what criteria should guide it, and how can a proportionate approach be ensured without unnecessary restrictions? The guidance should adopt a stepwise, risk-based approach. As a starting point, germline whole-genome sequencing (WGS), widely recognised as one of the most sensitive forms of omics data, could serve as an initial reference case for developing and testing the assessment approach. This should then be extended to other omics modalities to evaluate their respective risk profiles, recognising that some may warrant a lower level of protection, potentially closer to that applicable to basic biological data. Finally, specific attention should be given to integrated and multi-omics datasets, for which the combination of different data types may create additional risks requiring a dedicated assessment. It should define who is responsible for conducting, contributing to, and validating these assessments, and how their practical implementation should be tested.

To achieve this, the EDPB could facilitate a multidisciplinary dialogue with relevant regulatory bodies, including the EMA, data protection authorities, research organisations, pharmaceutical companies, various omics and bio-IT experts, data managers, IT security specialists, technology and innovation developers, and citizen representatives. Such engagement would help establish shared principles, clarify responsibilities, and co-develop and validate practical tools, templates and examples that reflect both regulatory requirements and scientific and technical realities. The objective should be accurate, usable, and proportionate guidance that protects fundamental rights while enabling responsible data use and innovation.

The EDPB should establish one shared legal framework and a common assessment methodology, with technical criteria adapted to the specific omics modality and the risks arising from data integration. Neither single-omics nor multi-omics data should be presumed anonymous or identifiable solely because of their category.

The assessment should be proportionate to the actual processing context and supported by evidence. It should distinguish the legal determination of whether data remain personal data from the subsequent choice of safeguards and access arrangements. This would help avoid both underestimating identification risks and imposing restrictions that unnecessarily compromise scientifically valuable research.

Dimension	Single-omics assessment	Multi-omics assessment
Scientific scope	One omics modality is assessed at a time.	Two or more omics modalities are combined, potentially with clinical, phenotypic or longitudinal data.
Scientific examples	Germline Whole-genome sequencing (WGS), Cancer WGS; RNA-expression analysis in specific cell types; DNA methylation in germline vs somatic conditions; proteomics; metabolomics.	Germline genomics combined with transcriptomics; genomic and proteomic profiles linked to clinical outcomes; longitudinal multi-omics cancer datasets.
Data representations	Raw sequencing files, variant files, individual-level expression matrices, protein-abundance tables or aggregated results.	Linked individual-level datasets, integrated molecular profiles, longitudinal records or combined analytical outputs.

Potential identification pathways	Rare variants, distinctive molecular profiles, cohort characteristics or linkage with external information.	Additional identification pathways may arise from the combination of modalities, overlapping records, shared characteristics or inference across datasets.
Key assessment question	What could reasonably be inferred or identified from this particular dataset, considering the relevant actor and context?	Does combining datasets create additional opportunities for linkage, inference or identification that would not be apparent when assessing each dataset separately?
Scientific utility	Would anonymisation or other transformations preserve the features needed for the intended analysis?	Would transformations or separation of datasets compromise cross-modal associations, longitudinal relationships or the validity of integrated analyses?
Data protection measures	Depending on the risk, measures may include data minimisation, aggregation, removal of direct identifiers, pseudonymisation, access controls and secure processing.	Measures may additionally include controlled linkage, separation of identifiers and analytical datasets, restricted access to integrated data, and secure analysis environments.
Assessment criteria	Apply No Record Isolation, No Linkage and No Inference to the specific data representation and processing context.	Apply the same criteria to the combined dataset, including risks arising from interactions between modalities and the availability of auxiliary information.
Reassessment triggers	Changes in data granularity, cohort composition, available external datasets, technology, recipients or intended use.	Addition of a new modality, linkage with another dataset, expansion of the cohort, new analytical capabilities or changes in access and sharing conditions.
Relevant stakeholders	Researchers, molecular and clinical experts, data managers, data protection officers (DPOs), IT security specialists and data custodians.	All relevant single-omics stakeholders, plus experts responsible for data integration, cross-modal analysis, linkage, infrastructure and access governance.
Responsibility and validation	Document who conducts the assessment, who supplies technical evidence, who reviews the safeguards, and who approves access or sharing.	Clarify responsibilities across contributing data holders and integration teams, including who assesses the combined dataset and validates the impact of new linkages.
Expected guidance from the EDPB	Modality-specific examples explaining how the common legal principles apply to different data types and representations.	Examples explaining how to assess combined datasets, identify additional risks, allocate responsibilities and determine when a new assessment is necessary.



Illustrative use cases for the Guidelines

Use case raised / reflected in the discussion	*What the EDPB should illustrate*	*Key assessment questions*
Somatic exoression ? or genomic ? tumour profile	*A specific tumour alteration or molecular profile used for classification or treatment selection.*	*Is the alteration common or rare in the relevant cohort? Can it be connected to a patient through a clinical record, treatment timeline, centre or other available information?*
Germline genomics/ rare disease[1]	*Sequencing data or a rare inherited variant in a small population.*	*Could the combination of variant, phenotype, family information and rare-disease context single out or link to a person? What family or kinship implications arise?*

[1] *Tumor DNA and cfDNA sequencing although primarily performed to identify somatic mutations for cancer diagnosis or treatment selection, may also reveal variants of potential germline origin. The risk assessment should account for this possibility, including the implications for patients and their relatives.*

Use case raised / reflected in the discussion	*What the EDPB should illustrate*	*Key assessment questions*
Population-level genomic data	*Large-scale sequencing intended for public health or population research.*	*Does scale reduce or increase risk in the specific release? What detail is retained, and could subgroups or unusual profiles be isolated or linked?*
Single-omics versus multi-omics	*A single molecular layer compared with combined genomic, transcriptomic, proteomic, metabolomic, epigenomic and clinical data.*	*What additional identification or inference routes emerge from combining layers? Does the combined dataset require a fresh assessment?*
Clinical and longitudinal linkage	*Omics linked to diagnoses, treatment, outcomes, imaging, registry or hospital data over time.*	*Do dates, rare trajectories, combinations of variables or cross-source consistency enable linkage?*
Data held outside the hospital	*The hospital may retain only a clinical report while underlying sequencing data are stored by an external provider or archive.*	*Which organisation holds the actual files, who controls access, and what can each actor realistically obtain? Is the proposed access pathway clear?*
Secondary use / AI analysis	*Data reused for a new research question, model development or validation.*	*Are the variables and data access necessary for the stated scientific question? Could model outputs, inference or new linkages alter the risk?*



Questions that should appear in an omics assessment template

- What is the scientific question, and which variables are necessary to answer it? Identify the minimum scientifically sufficient dataset, not merely the smallest dataset imaginable.
- Could the analysis inadvertently reveal inherited (germline) genomic information, even when the test is intended to identify only somatic variants, for example in tumour-only sequencing or circulating tumour DNA (ctDNA) testing?
- What is the risk for the different types of biological molecules that can be analysed in a high-throughput manner, other than genomics? Which general criteria can be used to identify the specific risks?
- What modality and data representation are involved? Record both the omics modality (genomics, transcriptomics, proteomics, metabolomics, epigenomics, or multi-omics) and the processing level: raw reads (FASTQ), aligned reads (BAM/CRAM), variant calls (VCF), individual-level processed matrices, aggregated outputs, or model parameters.
- What population and cohort structure matters: rare disease, small cohort, rare variant, molecular tumour subgroup, family relationship, or longitudinal follow-up?
- What direct identifiers, quasi-identifiers, metadata, timestamps, centre details and distinctive combinations remain?
- What other information can each relevant actor realistically access, including EHRs, registries, biobanks, clinical trial records, reference databases or published material?
- What are the realistic isolation, linkage, and inference scenarios, and what evidence supports or weakens each scenario?
- What access model is proposed: open release, controlled access, secure processing environment, trusted third-party linkage, federated analysis or release of checked outputs only?
- Which safeguards are effective in practice, not just promised in a contract, and what residual risk remains?
- What scientific utility would be lost through aggregation, suppression, generalisation or other transformations? Is a less exposing access model preferable to distorting the data?
- What change would trigger reassessment, who owns that review, and what evidence will be retained?



Avoid a one-size-fits-all technical threshold

The request is not for the EDPB to classify all (gen)omic data as identifiable or to establish a single universal threshold for anonymity. Uniqueness is a relevant signal, but it is not, on its own, proof that an individual can be identified. Using the example of a fingerprint, being distinctive is not the same as being identifiable: the practical question is whether that distinctive feature can realistically be linked to a person, taking into account the information and means available to the relevant actor. The assessment must therefore distinguish between the uniqueness of a data point, the possibility of singling out an individual, and the realistic possibility of identifying them.

The Guidelines should consequently establish a tiered, data-representation-specific and context-dependent assessment that considers the omics modality, processing level, cohort size and rarity, available auxiliary information, actor capabilities, intended purpose, access arrangements and safeguards. Raw genomic sequences, aligned sequencing files, variant-call files, individual-level gene-expression or protein-abundance matrices, and aggregated outputs should not automatically be treated as equivalent. Nor should processed or non-genomic omics data be presumed anonymous simply because they are not whole-genome sequences. Multi-omics combinations also require specific consideration because integrating otherwise distinct datasets may create additional opportunities for linkage or inference.

The EDPB should translate these distinctions into practical criteria and worked examples that help stakeholders determine when distinctiveness becomes a realistic identification risk, what evidence supports that conclusion, and which safeguards are appropriate. This would avoid both underestimating genuine risks and imposing blanket restrictions based on uniqueness alone, while preserving the GDPR's fundamental protections and the scientific value of omics research.

4. Preserve scientific utility while reducing exposure

The discussion raised a practical concern: making data maximally difficult to identify can also make them less useful for valid research, clinical classification, or understanding treatment response. The right question is not simply whether data can be made anonymous, but whether the scientific objective can be achieved with anonymous data, and, if not, what minimum data and access architecture is necessary while maintaining proportionate safeguards. We also questioned the practical value of presenting three broad scenarios (higher risk/open access, higher scientific utility/controlled access, and very high-risk/secure analysis or synthetic data) as if they were readily distinguishable categories. Without operational definitions, these scenarios are difficult to apply consistently and may be inapplicable to real research workflows. The final Guidelines should not rely on these labels as a decision rule. They should instead explain how to move from the evidence-based identifiability assessment and the minimum scientifically necessary data to a justified choice of release or access model.



Replace broad access scenarios with a practical decision procedure

We therefore ask the EDPB to clarify the relationship between anonymisation, pseudonymisation and access controls without implying that controlled access makes personal data anonymous, or that open access is the default objective. The access model is a consequence of the assessment and the research need; it is not a substitute for assessing whether the data are personal for the relevant actor.

A practical procedure should ask, in this order:

- What is the precise research question, intended output, and minimum scientifically sufficient dataset?
- For each relevant actor, what identifying information and auxiliary data are realistically accessible, including through other consortium members, clinical systems, registries, biobanks, public resources or commercial/reference databases?
- What specific, plausible pathway could connect the data or a distinctive profile to an identifiable natural person? Identify the variables, combinations, access route, and evidence; do not treat a merely theoretical possibility as sufficient.
- What do the three EDPB criteria show in this context: No Record Isolation, No Linkage and No Inference, and how does each finding contribute to the overall assessment?
- Which transformation or safeguard would interrupt the identified pathway, and what would it cost in scientific utility?
- On that evidence, should the proposed activity involve release of data assessed as anonymous, or should it instead use personal data under an appropriate GDPR framework with pseudonymisation, time-limited controlled access, a secure processing environment, federated analysis, output checking or another safeguard?
- Who is accountable for the decision, what assumptions and residual risks are recorded, and what material change would trigger reassessment?



Clarify what a linkage risk actually means

The draft should distinguish the possibility of combining datasets from a demonstrated or reasonably foreseeable route to identifying a person. A dataset may contain variables that can be associated with one another without revealing a person's name or otherwise making the person identifiable. Likewise, narrowing a population or inferring an attribute does not, by itself, establish that a natural person has been identified. The assessment should specify whether the combination enables a realistic connection to an identifiable person, what additional

information makes that connection possible, and which actor can access it. This clarification is particularly important for rare variants, family relationships, distinctive phenotypes, clinical timelines and multi-omics profiles.

Suggested practical question for the final Guidelines

Could this omics dataset, in the actual processing context and considering auxiliary information realistically accessible to the relevant actor, be combined with another dataset in a way that identifies a natural person or makes identification reasonably likely? If yes, which variables or combinations provide the linkage route, what additional information is needed, who can access it, and what evidence supports the scenario? If the combination only narrows a group or permits an inference without connecting the information to an identifiable person, explain why that does, or does not, affect the legal assessment.



Questions for which the Guidelines should provide worked answers

1. What does it actually mean for omics data to be "anonymous" under the GDPR, and how can this be demonstrated in practice?
2. How should we assess in practice for different omics and multi-omics whether individuals can still be singled out, linked or identified through inference when genomic, clinical, phenotypic and longitudinal data are considered together?
3. When does reducing data granularity meaningfully reduce identifiability, and when does it simply make the data scientifically unusable?
4. How should pseudonymisation be understood as a safeguard that reduces identification risk without changing the fact that the data remain personal data?
5. How should the risk of re-identification through linkage be assessed when pseudonymised omics data are combined with other datasets?
6. When can synthetic data genuinely reduce the need to access real patient-level data, and how should risks such as memorisation, record replication and membership inference be assessed?
7. How should we distinguish data that are genuinely necessary for scientific validity from data that are simply useful or desirable, particularly in rare disease research?
8. What concrete, reproducible procedure is needed to make these assessments consistently across different omics use cases and European research environments?

A practical assessment framework could be developed progressively, starting with germline whole-genome sequencing (WGS), as one of the most sensitive and widely recognised forms of omics data, to establish an initial reference case for assessing identifiability and residual risks. The Guidelines or authoritative guidance in the field should then explain how this approach can be tested and adapted to other omics modalities, taking into account their specific characteristics and risk profiles. Some data types may present substantially lower identification risks and, depending on the context, may justify a less restrictive assessment or level of protection, potentially closer to that applicable to basic biological data. In addition, the Guidelines or authoritative guidance in the field should address integrated and multi-omics datasets, recognising that combining different omics layers, particularly with clinical, phenotypic or longitudinal information, may generate additional identification risks that require assessment at the level of the combined dataset, rather than of each modality in isolation. This progressive approach would provide a practical basis for developing consistent, proportionate and scientifically meaningful assessment criteria across different omics applications.

Option	When it may be considered	What could/must still be assessed
Anonymised data/outputs	*Where the evidence supports that people are not identifiable using means reasonably likely to be used in the relevant context.*	*Residual isolation, linkage and inference risk; utility impact; actor-specific release context. A conclusion must be supported by the assessment, not by the label applied to the data.*
Pseudonymised personal data	*Where longitudinal linkage or individual-level analysis is necessary, and controlled re-identification may be required for legitimate purposes.*	*GDPR requirements, key separation, access controls, roles, legal basis and safeguards.*
Controlled access / secure processing environment	*Where detailed or personal-level data are scientifically necessary, and broader release is not justified. This is an access model, not proof that the data are anonymous.*	*Who can access and do what; export/output checks; logging; governance; reassessment after changes.*
Trusted third-party or federated analysis	*Where linkage or analysis can be performed without giving every party direct access to underlying identifiers or source data.*	*Separation of duties, key management, linkage governance, model/output risks and responsibility allocation.*
Synthetic or aggregated data	*Where a substitute can support development, testing, demonstration or some analytical purposes.*	*Whether synthetic outputs reproduce, memorise or disclose information about real individuals; whether utility is sufficient.*

What decision rule and criteria should the EDPB recommend to ensure a proportionate, scientifically meaningful assessment?

Could the assessment begin with the research purpose and the data necessary to achieve it, then evaluate identifiability in the specific processing context and determine the least exposing data representation and access model that remains scientifically workable? How should the EDPB ensure that this approach protects fundamental rights and complies with the GDPR, while avoiding unnecessary restrictions that do not deliver a demonstrable protection benefit? The objective should be to apply data protection in a way that enables responsible research and serves patients' interests, rather than allowing compliance processes to displace the scientific purpose they are intended to safeguard.

The argument should also recognise that health research does not operate in an unregulated environment. Health data are already subject to a complex landscape of EU and national legislation, professional confidentiality obligations, ethical oversight, institutional codes of conduct, and technical and organisational safeguards. Depending on the country, institution, and research context, these may include restricted access to clinical records, role-based permissions, secure research environments, pseudonymisation, data-access committees, research ethics review, and data-sharing agreements. These existing safeguards do not automatically make data anonymous under the GDPR, nor do they replace the need to assess identifiability.

However, they are relevant to understanding the actual conditions under which data can be accessed, linked and used, and should be considered when evaluating realistic identification risks and the proportionality of additional requirements. Relevant EU-level examples include the GDPR's research safeguards, national rules on medical confidentiality, professional codes of conduct, institutional governance and controlled-access procedures, and the EHDS framework for permitted secondary use, Health Data Access Body authorisation and secure processing environments. The European Commission's review of Member State rules confirms that national requirements remain relevant and can differ across jurisdictions.

Suggested addition to the paragraph on uniqueness and identifiability:

These existing safeguards should form part of the practical context for the assessment, without being treated as automatic proof of anonymisation. The EDPB should clarify how to account for them when assessing realistic identification risks, and how to distinguish the legal status of data from the safeguards required to access and use them. This would help ensure that the Guidelines reflect the realities of a highly regulated health research environment, protect fundamental rights, and avoid duplicating existing controls or imposing additional restrictions without a clear, evidence-based justification.

5. Proposed lifecycle, stakeholder and responsibility matrix

A recurring concern in the JANE-2 Task Force discussions on the ELSI aspects of Omics Technologies was how to draw a proportionate line between protecting individuals and enabling scientifically meaningful research. Health research already operates within a dense regulatory and governance framework, and further requirements should provide a clear, demonstrable benefit rather than add layers of compliance without meaningful additional protection. Data protection, data usability and scientific value should go hand in hand: the aim is neither to minimise safeguards nor to impose restrictions beyond what is necessary, but to identify the point at which protection is effective and proportionate while allowing responsible research to progress.

Because the relevant context can change as data move between organisations, are linked or are reused, the EDPB should encourage a lifecycle matrix that identifies who provides information, who assesses risks, who decides on access, who implements safeguards and who reassesses the situation when circumstances change. Such a matrix could also help clarify where additional safeguards are genuinely needed and where existing measures are sufficient. The exact allocation will depend on the legal and operational arrangements; the matrix below is an illustrative model for adaptation, not a universal assignment of GDPR roles.

Lifecycle stage	*Key stakeholders to involve*	*Core responsibility / decision*	*Evidence / output*
1. Define purpose and data needs	*Research lead, clinician/oncologist, molecular/clinical genetics expert, statistician/bio-IT, data scientist, ethics/governance*	*Specify research question, scientifically necessary variables, population and intended outputs.*	*Protocol; data specification; utility rationale.*
2. Map source data and custodians	*Hospital/lab, biobank, registry, external archive/provider, data holder*	*Identify where raw files, reports, keys, metadata and linked datasets are held; clarify access rights.*	*Data-flow and custody map.*
3. Assess identifiability	*Controller(s), DPO, information security, omics/data experts, relevant recipient*	*Evaluate actor-specific means and the isolation, linkage and inference scenarios.*	*Assessment template; assumptions; scenario evidence.*
4. Select access and safeguards	*Controller/data holder, recipient, infrastructure operator, trusted third party, security team*	*Choose anonymisation, pseudonymisation, controlled access, secure environment, federated analysis or another proportionate option.*	*Decision rationale; safeguard plan; agreement/instructions.*
5. Approve access / release	*Relevant governance body, access committee/HDAB where applicable, ethics or security review where required*	*Confirm the proposed route, conditions and responsibilities under applicable rules and governance.*	*Permit/approval; conditions; role allocation.*
6. Link, integrate and analyse	*Linkage service, data engineers, researchers, platform operator*	*Reassess the combined data and new derived information; enforce separation of duties and access restrictions.*	*Post-linkage assessment; logs; output checks.*
7. Share results or enable secondary use	*Research team, recipient, data holder, governance and security functions*	*Assess the actual release or output; check whether purpose, recipient, data or access model has changed.*	*Release decision; output review; reuse conditions.*
8. Monitor and reassess	*Controller(s), DPO, security, data holder, infrastructure operator, recipient*	*Trigger review after material change, incident or new linkage/capability; document whether the prior conclusion remains valid.*	*Review log; updated assessment; corrective actions.*

Questions for the EDPB on responsibility allocation and who decides on the context

To make the guidance practical and proportionate, the EDPB should clarify how responsibilities are allocated across the data lifecycle, including who determines and documents the relevant processing context and whose perspective should guide the identifiability assessment. Where several organisations and actors are involved, it should be clear who makes this determination, what information they must consider, and how differences in access, technical capabilities, intended purposes and onward sharing are addressed. The aim should be to establish clear accountability without creating unnecessary duplication or unrealistic obligations. The following questions highlight where further clarification could help stakeholders apply the assessment consistently, ensure effective safeguards and avoid requirements that do not provide meaningful additional protection.

- Who is responsible for defining and documenting the relevant actor perspective where a controller, processor, joint controllers, data holder, infrastructure and recipient are all involved?
- What should the transmitting organisation reasonably know about the recipient's capabilities and onward sharing, and what information should the recipient be required to provide?
- How should responsibility be allocated where a trusted third party performs linkage or where the same institution performs different roles in different parts of the workflow?
- Which decisions require a new assessment: addition of data, change of purpose, new recipient, change in access rights, new analytical technique, or release of derived outputs?
- How can contractual obligations support governance while being complemented by technical and organisational measures that actually constrain access and re-identification?

6. A practical reassessment model

The Guidelines should make clear that a documented conclusion is tied to the facts and assumptions on which it was based. Reassessment is most useful when there is a material change capable of affecting those assumptions. A risk-based review schedule may complement event-triggered review, but should not be described as a universal fixed legal deadline unless the law provides one.

BASELINE	TRIGGER	REASSESS	DECIDE	RECORD
Initial assessment before release, sharing, linkage or new use.	New data/layer • linkage • recipient/role • purpose • access model • technology/reference data • incident.	Revisit actor perspective, reasonably likely means, three criteria, safeguards and residual risk.	Maintain conclusion / add safeguards / restrict access / change output / treat as personal data.	Document trigger, evidence, decision-maker, date, actions and next review condition.



Illustrative reassessment scenarios

The identifiability assessment should not be treated as a one-off exercise. In omics research, changes in data composition, linkage opportunities, access conditions or intended use may alter the realistic possibility of identifying individuals. The EDPB should therefore clarify when reassessment is warranted, who is responsible for initiating it, and what changes are sufficiently significant to justify revisiting the original conclusion. The following scenarios illustrate how the assessment may need to evolve over the data lifecycle, without requiring a complete reassessment for every minor operational change. The aim is to ensure that protection remains effective as circumstances change, while avoiding unnecessary duplication and administrative burden.

Scenario	*Why the old assessment may no longer be sufficient*
Single-omics dataset becomes multi-omics including WGS, vs not	*Combining layers can create more distinctive profiles or enable new inferences.*
Rare variant linked to a clinical condition	*A record that was difficult to connect in isolation may become linkable through the additional clinical context.*
Research dataset moves to a new recipient or platform	*The recipient's access, technical capability, additional information or onward sharing may differ.*
A new purpose uses a more detailed subset	*The variables retained and the utility/risk balance may change.*
A previously external data source becomes accessible	*The availability of auxiliary information may alter the realistic linkage assessment.*
A security incident exposes keys or additional information	*The practical means available to identify people may materially change.*

7. Consolidated requests to the EDPB

The requests below consolidate the main recommendations of the JANE-2 Task Force on the ELSI aspects of Omics Technologies. They aim to help the EDPB translate the draft Guidelines into practical, consistent and proportionate guidance for health and research contexts, moving beyond general principles towards an operational methodology supported by worked examples, clear responsibilities and decision-making criteria across the data lifecycle.

The EDPB should work closely with research communities, scientific societies and relevant legal, regulatory and technical experts to ensure that good data protection practices are embedded in research workflows from the outset, rather than addressed in a separate compliance process. The priority is to turn legal principles into workable solutions that help stakeholders determine what protection is necessary, which safeguards are effective, what level of data detail is justified and required, and where additional restrictions would provide no meaningful benefit. This requires a proportionate approach that goes beyond definitions and focuses on the critical questions of identifiability, scientific utility, effective protection and responsible data reuse. The final Guidelines should establish common principles across Europe while allowing assessments to reflect the characteristics of the data, the processing context and the safeguards already in place.

Data protection, data usability and scientific value should go hand in hand. The aim is to embed protection into good research practice, avoid unnecessary duplication and barriers, and enable responsible research that safeguards individuals while delivering meaningful benefits for patients and society.

Request	*What the final Guidelines should provide*
1. Publish an operational methodology	*A reproducible sequence, practical questions, minimum evidence, and documentation template.*
2. Provide a health and research annex	*Worked examples for hospitals, clinical trials, registries, biobanks, research consortia and secondary use.*
3. Add an omics-specific toolbox	*Separate question sets and examples for genomics (germline vs somatic), transcriptomics, proteomics, metabolomics, epigenomics and multi-omics, with attention to rare disease, somatic/germline data, family implications, linkage and scientific utility. For instance, include examples of variants of potential germline origin arising from tumor and cfDNA sequencing, and clarify how such findings should be considered when assessing identifiability and selecting appropriate safeguards.*
4. Clarify the three criteria	*Define record isolation, linkage and inference in practical terms and explain how the criteria inform the overall legal assessment.*
5. Clarify contextual assessment	*Address controllers, processors, joint controllers, independent recipients, trusted third parties and infrastructure operators.*
6. Set out a reasonableness/foreseeability method	*Explain how accessibility, legal and organisational barriers, effort, time, cost, expertise, technology and safeguards inform realistic scenarios.*
7. Clarify reassessment	*Distinguish legal requirements from recommended governance review; specify material change triggers and avoid implying continuous monitoring of all hypothetical developments.*
8. Address the data lifecycle	*Include assessment before release, after linkage or integration, before new secondary use and when access conditions or recipients change.*
9. Explain technical and organisational safeguards	*Describe strengths and limitations of aggregation, suppression, generalisation, synthetic data, secure processing, controlled access, federated approaches, output checks and key separation.*
10. Include a responsibility matrix	*Identify who assesses, supplies information, decides, implements safeguards, documents, monitors and reassesses at each lifecycle stage.*

Request	*What the final Guidelines should provide*
11. Address scientific utility and proportionality	*Explain how to identify the minimum scientifically sufficient dataset and compare the impact of transformations against realistic identification pathways. Avoid broad risk/access scenarios unless their criteria and practical application are defined.*
12. Support consistency across Europe	*Provide common minimum principles and documentation while permitting proportionate adaptation to data type, purpose, access model and national governance context.*

8. Conclusion

JANE2 Network of Expertise in Omics Technologies supports the EDPB's objective of improving legal certainty on anonymisation and the distinction between personal and anonymous data. The draft provides an important foundation, but its **value for the use of omics technologies in research will depend on whether organisations can translate it into clear, defensible and proportionate decisions**. Our reflection and brainstorm reinforced that a useful methodology must start from the research purpose and actual data environment; identify relevant actors and realistic means; test No Record Isolation, No Linkage and No Inference; and document safeguards, residual risk and reassessment triggers. The discussion brought together perspectives from data protection, law, data engineering, molecular biology, clinical genetics, oncology, research and project coordination.

OVERARCHING RECOMMENDATION

The EDPB should complement its principles-based framework with an operational, lifecycle-oriented and risk-based methodology for **use of omics technologies in** research, and actively support the development and harmonisation of field-specific **Good Data Protection Practices applied to Health Research**. While common principles must apply across sectors, their practical implementation needs to reflect the characteristics of the data, the scientific purpose, the research environment and the realistic risks of identification. The EDPB should work with research communities, scientific societies, multidisciplinary expertise, data protection authorities and relevant legal, technical and regulatory experts to translate these principles into practical procedures, templates, worked examples and a stakeholder/responsibility matrix.

Such practices should address the critical decisions throughout the data lifecycle, from data collection and processing to sharing, linkage, secondary use and release of derived outputs. They should clarify who assesses identifiability, who determines the relevant processing context, what evidence supports a conclusion, when reassessment is needed and which technical and organisational safeguards are proportionate. This would help embed data protection in research workflows rather than leave it as a separate compliance exercise, reduce inconsistent interpretations and unnecessary duplication, and support more transparent and reproducible decisions across Europe.

The objective should be not to create additional layers of requirements, but to operationalise existing data protection principles in scientifically meaningful ways, legally sound and workable in practice. We encourage and promote harmonised, field-specific good practices, and here the EDPB can help ensure that the protection of individuals and the responsible use of data reinforce one another, enabling research that serves patients and society without imposing restrictions that offer no demonstrable additional protection.

Appendices:

1. Participants in the JANE-2 WP9.3 ELSI Task Force thematic meeting

***Meeting**: 22 October 2026, thematic discussion on pseudonymisation and the special status of genetic data.*

This document presents requests for clarification and practical guidance informed by the multidisciplinary discussion on pseudonymisation, genetic data and identifiability in health and omics research. It does not imply that every participant endorsed every recommendation.

Participant	Institution/affiliation
Almudena Sánchez	*Hospital Spain*
Saptarshi Chkarabarti	*KU Leuven*
Glykeria Gkoliou	*National Hellenic Research Foundation (NHRF)*
Annalisa Musola	*Institution not clearly established in the supplied transcript*
Alexandra Voutsina	*National Hellenic Research Foundation (NHRF)*
Alex Pintzas	*National Hellenic Research Foundation (NHRF)*
Els Hermans	*VIB*
J. Matt McCrary	*University Hospital Würzburg (UKW-DE)*
Tanja Jutzi	*German Cancer Research Center (DKFZ); National Center for Tumor Diseases*
Michail Artemis	*National Hellenic Research Foundation (NHRF)*
Wenkai Li	*Vrije Universiteit Brussel (VUB)*
Henriett Butz	*National Institute of Oncology, Hungary*
Patrizio Giacomini	*Alliance Against Cancer (ACC)*
Hélène Antoine-Poirel	*Belgian Cancer Centre / Sciensano*
Inge Smeers	*Belgian Cancer Centre / Sciensano*
Laurent Prévotat	*Luxembourg Institute of Health (LIH)*
Nancy Frederickx	*Belgian Cancer Centre / Sciensano*
Roxana Albu	*Sciensano (Data Governance, ELSI unit)*
Wannes Van Hoof	*Sciensano (Data Governance, ELSI unit)*

2. Data-type-specific identifiability checks: modality and representation

A key clarification requested from the EDPB is that the assessment decision tree should distinguish both (1) the omics modality and (2) the data representation/processing level. The word "transcriptomics" or "proteomics" alone is not enough to determine the relevant identification risk.

Data representation	***Illustrative examples***	***Identifiability questions for the assessment***
Sequence- or genotype-bearing data	*WGS/WES FASTQ or BAM/CRAM; VCF; raw RNA-seq reads that may retain genotype information* *Blood vs tumor samples*	*Can sequence variation or distinctive profiles be linked to reference or external datasets? Are germline, somatic, family or rare-variant signals present? Who can realistically perform the linkage?*
Individual-level derived omics matrices	*RNA expression count/TPM matrix; methylation matrix; protein-abundance*	*Could high-dimensional patterns support linkage or inference, including through eQTL relationships, distinctive*

	or metabolomics matrix; single-cell count matrix	*profiles, variant peptides or external reference data? What cohort and metadata are available?*
Aggregated outputs	*Cohort summaries; aggregated differential-expression results; stratified statistics*	*Could small cells, rare subgroups, repeated queries, differencing or combinations with public information reveal individuals or sensitive attributes? Are output checks required?*
Linked multi-omics and clinical data	*Genomic + expression + proteomic data linked with diagnosis, age, centre, family or longitudinal records*	*Does the combination create a materially stronger identification pathway than each component alone? Which variables and actors enable the link, and can the pathway be interrupted without invalidating the research?*
Model-derived data and synthetic data	*Model parameters, embeddings, generated patient-level records, synthetic omics data*	*Could memorisation, record replication, membership inference or extraction expose information about training individuals? What testing and documentation support the claim of reduced risk?*

Illustrative contrast. *A genomic example may better illustrate the distinction. For instance, germline whole-genome sequencing (WGS) data are relatively stable over an individual's lifetime and may carry persistent identifying characteristics, whereas RNA-expression profiles are highly dynamic and dependent on the tissue or cell type, physiological state, and conditions at the time of sampling. Consequently, RNA-expression data may present different and potentially lower re-identification risks when considered in isolation, although these risks may increase substantially when combined with genomic data or other identifying information. The Guidelines or authoritative advice should demonstrate how such differences are reflected in the practical assessment, rather than assigning a risk or status solely on the basis of the omics modality.*

Proposed assessment sequence by data representation. *One avenue could be to record the research purpose, intended recipient(s), intended use (research/secondary use or direct clinical care), and minimum scientifically sufficient data.*

1. *Describe the modality and exact representation/processing level, including whether sequence reads, genotype-bearing variants, individual-level matrices, metadata, linked records or model outputs are present.*
2. *Identify the relevant actor(s) and what auxiliary information each can realistically access, including public reference databases, clinical records, registries, biobanks, trial data, and other consortium datasets.*
3. *Assess No Record Isolation, No Linkage, and No Inference as complementary lenses. For each plausible pathway, name the variables or patterns involved, the actor, the access route, the evidence and the practical barriers.*
4. *Consider whether the data can be transformed or released in a different form, or whether access should instead be controlled. Evaluate the scientific consequences of aggregation, suppression, generalisation or removal of variables.*
5. *Document the conclusion, safeguards, residual risk, rationale, uncertainty and reassessment triggers. Do not treat pseudonymisation or controlled access alone as proof of anonymisation.*

Scope clarification: research reuse versus direct clinical care.

A proposed decision tree for omics and multi-omics should primarily support research sharing, access, and secondary use of omics data. The EDPB should state this scope explicitly. Processing needed for direct clinical care may follow a different legal and governance pathway, including workflows where identifying information is necessary for safe diagnosis, treatment, follow-up or return of results. The anonymisation assessment should not be interpreted as requiring the removal of clinically necessary links merely to label data anonymous. Where data remain identifiable or pseudonymised, the controller should apply the relevant GDPR framework and safeguards for that processing.

The guidance should also clarify how to assess a dataset's status when the same data move between research and clinical contexts, or when a research dataset is later used for a clinically relevant purpose. The purpose, actors, access conditions and practical identification pathways should be reassessed; the label attached at an earlier stage should not be assumed to settle the later context.

Specific clarification requests for the EDPB

1. *Please distinguish genomics and other omics modalities without assuming that all data within a modality carry the same risk. In particular, clarify how sequence-bearing data, derived individual-level matrices, aggregated outputs, and linked multi-omics datasets should be assessed.*
2. *Please add "data representation/processing level" as an explicit assessment field, covering raw reads, aligned reads, variant calls, individual-level processed matrices, aggregated results, and model parameters.*
3. *Please provide worked examples for contrasting datasets, including a large normalized bulk RNA-expression matrix without demographic variables versus raw RNA-seq BAM files from a small rare-tumour cohort with age and diagnosis.*
4. *Please explain how to assess realistic linkage and inference pathways for derived expression, methylation, protein-abundance, metabolomics and single-cell matrices, including where genotype or identity information may be inferred indirectly.*
5. *Please clarify how No Record Isolation, No Linkage and No Inference interact, and why uniqueness or the possibility of linking records does not automatically establish that a natural person is identifiable.*
6. *Please make clear whether the proposed assessment pathway is aimed at research/secondary-use sharing and reuse, and explain how it relates to processing for direct clinical care where identity links may be necessary.*
7. *Please provide a reproducible, proportionate procedure that records the actor, data representation, accessible auxiliary information, evidence, safeguards, scientific utility trade-offs, and residual risk, without imposing one universal technical threshold.*