Summary page:
TEHDAS2 public consultation on Draft guideline for Health Data Access Bodies on data minimisation, pseudonymisation, anonymisation and synthetic data

This consultation has 4 pages and 25 questions. The first and the second pages are common to all TEHDAS2 public consultations and cover demography of the responder and overall quality of the document. Pages 3 and 4 consist of questions specific to this document.

Demography

1. Country *

Other

2. Type of responder *

Academic or research organisation

3. Are you responding on behalf of several organisations? *

If yes: On behalf of how many organisations?

No

4. Sector *

Other
Academic research and policy center

5. Organisation size *

Small to medium enterprise (10–249 employees)

6. Professional role / function

Academic researcher and investigator responding on behalf of a research and policy center

Quality

7. Is the document easy to understand? *

1

8. How well does the document address the key issues related to its subject matter? *

1

9. How feasible do you find the guidelines or technical specifications to implement, as outlined in the document? *

1

10. Generic feedback

Do you have any suggestions for improving the document? Are there any additional topics or areas that should be covered? Max. 5000 characters.

While the aims and goals are clear, and the general guidance is welcome, several conceptual and technical challenges remain. Clarifying definitional boundaries, strengthening methodological guidance, ensuring alignment with GDPR, EDPB, the AI Act, and international, global considerations will be essential.

Ambiguity in Key Definitions and Regulatory Boundaries: Substantial uncertainty persists regarding the distinctions between pseudonymization, anonymization, and synthetic data. The guideline advises a risk-based approach but does not provide sufficiently detailed methodologies to support consistent decision-making across HDABs. Without these standards, data elements will lose important detail and will not be readily interoperable. Furthermore, the entity responsible for ensuring the sufficiency of anonymized data (and synthetic data) over time should be clear. (Who is responsible for future-proofing the datasets?)

Data Minimization versus Data Utility: The methodological consequences of minimization and aggressive de-identification, as well as their impact on analytic utility, are inadequately considered, particularly in research domains such as genomics, rare diseases, and minority populations (e.g., pediatric diseases). Data minimization risks diminishing scientific value and is insufficiently addressed.

Need for Explicit and Harmonized Technical Standards: While TEHDAS2 guidance is not prescriptive, more concrete guidance is needed to ensure consistent implementation (e.g., standardized pseudonymization techniques, metadata standards, etc.). Shared risk assessment templates/methods/metrics would facilitate harmonization. Re-identification risk assessments are necessary, but mechanisms for quantifying or evaluating such risk are not provided, and existing approaches (e.g., k-anonymity, differential privacy, etc.) are mentioned but not further explained or characterized. When would one use one versus another? What is acceptable as a standard? The absence of standards or processes risks significant heterogeneity across HDABs. More clarity is needed regarding timing, frequency, documentation of assessments, and monitoring.

Synthetic Data: Promise, Limitations, and Oversight Requirements: While the guideline acknowledges the potential for re-identification of synthetic data, it offers a limited discussion of evaluation utility, methods for detecting overfitting, and procedures for certifying synthetic datasets as sufficiently privacy-preserving. Validation frameworks and metrics for evaluating risk would be welcome.

Application to Clinical Trials and Research: Insufficient direction is provided for implementing EDHS (and Guideline 7) in multinational clinical trials and research. Various entities (e.g., sponsors, clinical trial sites, investigators, clinical and research labs, imaging facilities) each have either identifiable or pseudonymized data, but the dataset must be complete and remain intact to retain utility and ensure replicability. Who is responsible for what? When must the data (and/or dataset) be made available?

Governance, Accountability, and Resource Constraints: The guideline would benefit from more explicit articulation of governance expectations, including model documentation structures, audit trail requirements, and accountability mechanisms. The

logistics and operationalization of the various components of this guidance for anonymization and pseudonymization remain vague and, therefore, challenging. Details related to risks and mitigation of re-identification, as well as workflows specific to clinical research, would be welcome. Additionally, tools, training, and templates are needed to support the practical implementation of the guideline.
Additional Topics: The strong foundation provided in this document could be further improved by integrating additional clinical trial-specific considerations, as previously noted: use cases, reidentification risks, integration with existing ethical guidance and frameworks, data set linkage, training for users, cross-jurisdictional harmonization, audit frameworks, and issues related to imaging and genomic data. Additionally, further consideration of the special risks associated with small populations (e.g., pediatrics, rare diseases) is necessary, along with guidance on addressing these risks (e.g., statistical tools for small populations). The impact of AI on the EHDS generally, and on anonymization and pseudonymization specifically, require further attention.

The following questions are specific for TEHDAS2 draft guideline for Health Data Access Bodies on data minimisation, pseudonymisation, anonymisation and synthetic data

Part 1: General questions

11. What are you representing, according to the definition of the EHDS Regulation? *

Data User
Other
International academic research and policy center

12. From your perspective, how well does the guideline provide practical and actionable guidance for HDABs, data holders and data users regarding safe and secure processing of electronic health data within the EHDS?

2

Please elaborate on any areas where the guidance could be made more practical or actionable.
Max. 5000 characters.

Please see Q13, and:
Additional guidance: resources, tools, and checklists; training and real-time help (eg “help desk”); case studies to enhance usability for clinical research and trials; managing genomic and sensitive data, rare diseases, small cohorts; methods to permit data linkage. Guidance related to cross-border data transfers, dataset storage, and data analysis across more than one member state could be further developed to harmonize standards.

13. Are there any critical aspects or challenges regarding data minimisation, pseudonymisation, anonymisation, or synthetic data generation within the EHDS that you believe are not sufficiently addressed in the guideline?

Max. 5000 characters.

Clear, standardized technical requirements for secure processing environments and use: More concrete specifications are needed for EHDS-compliant secure processing environments, including baseline encryption standards, access control expectations, role segregation, logging and monitoring requirements, audit requirements, incident response procedures, and security testing and frequency. Practical expectations for secure-environment use (e.g., permitted operations, prohibited actions, training requirements, breach handling) would help ensure safe secondary use by data users.

Standardized and operational data minimization guidance: Tools and resources, including decision trees, checklists, and templates, would help HDABs and data holders implement minimization consistently. Use-case examples (e.g., EHR, claims, pharmacies, imaging facilities, registries) would clarify how different datasets should be reduced while retaining utility. The particular example of clinical trials and research should be considered as should other use cases such as pharmacovigilance.

Methods for re-identification risk assessment: The guideline should provide recommended methodologies, minimum documentation expectations, and examples illustrating how risk differs for rare diseases, small populations (e.g., pediatrics, insular communities), and multimodal datasets (e.g., EHR and genomics). Recommendations for the appropriate use of quasi-identifier analysis, k-anonymity, and/or differential privacy techniques (when, how often, under what circumstances) should be suggested. The responsibility for and frequency of reassessment should be clarified.

Standards for pseudonymization: Pseudonymization is often implemented inconsistently. More specific guidance is needed on hashing, tokenization methods, application of advanced encryption standards, secure storage, and pseudonymization at the source versus within HDABs. Guidance on cross-border pseudonymization matching would support EHDS interoperability. Minimum metadata standards, interoperability expectations for pseudonymization and minimization, and model agreements for cross-border access environments would facilitate secondary use.

Guidance on anonymization versus pseudonymization: Specific criteria and/or tools to assess when anonymization is technically feasible, and when health datasets must remain pseudonymized (e.g., genomics, rare disease data) would enhance the implementation of common approaches and their utility.

Synthetic data requirements: The guideline should outline technical standards for synthetic data generation (e.g., fidelity, quality), utility assessment, privacy testing, and provide templates for synthetic data documentation, as well as appropriate use cases.

Guidance on governance and legal compliance: The alignment of EHDS with GDPR, EDPB, Data Governance Act, and the AI Act should be clarified. Standardized expectations for privacy assessments, audit trails, and documentation for pseudonymization decisions, minimization justifications, and risk assessments would reduce variation in implementation.

Specific to clinical trials and research:

- Because clinical trials often cross jurisdictions, details related to cross-border data transfers could be further developed. Issues related to dataset storage, data analysis across more than 1 member state, harmonization standards could be further developed.
- In data sets that require long-term follow-up (genomic data, e.g.) issues related to safety data, follow-up, and pharmacovigilance may necessitate long-term access to participant-level data. Further clarifications related to pseudonymization (rather full anonymization) are warranted.
- Additional details related to regulatory inspections would be welcome, acknowledging that trial sponsors must maintain subject traceability for regulators, emphasizing the differing stakeholder interests.
- Risk of over-minimization in undermining the interpretability of clinical trial outcomes, particularly for subgroup analyses.
- Standardized, auditable methods for assessing re-identification risks that are feasible for sponsors could be specified.
- Include a framework for validating synthetic data
- Use of indirect identifiers in rare diseases and genomic data

14. To what extent do the guidelines offer clear and harmonised approaches for implementing the EHDS regulation’s requirements concerning data minimisation, pseudonymisation, anonymisation, and synthetic data across Member States?

2

What improvements would you suggest to enhance the overall clarity, comprehensiveness, and practical applicability of the guideline (i.e., specific sections, terms or concepts)?
Max. 5000 characters.

While the guidance is not prescriptive, in the absence of standards, variability in implementation across Member States will significantly impact the utility and range of EHDS data availability. Consistent, defined terminology is needed across Member States. Harmonization is necessary across the 27 member states to enhance clarity, reduce fragmentation, and facilitate the reuse of cross-border data. Sector-specific guidance for clinical trials, rare diseases, and genomic research is necessary.

Including illustrative workflows for anonymization and pseudonymization in real-world clinical trial contexts would be helpful, as would clarification of the HDAB’s role in determining what is acceptable anonymization for the reuse of trial data. Finally, but not exhaustively, expansion of synthetic data standards would be of value.

Specific to pediatric populations, further details on cross-border harmonization would be beneficial, acknowledging that many pediatric trials are conducted internationally to ensure adequate enrollment. Tools that allow for specific consideration of pediatric populations would be helpful; perhaps an overlay template/tool that raises additional considerations across member states would provide the additional assurances needed that the pediatric data sets have been properly managed, as would comparison maps between EHDS requirements and other regulatory regimes (GDPR, e.g.).

15. From your professional perspective, do you currently have the technical and organisational capacity to implement the recommendations (e.g., tools for data protection risk assessment, synthetic data generation)?

No answers

What capacity gaps or resource needs would require support?
Max. 5000 characters.

No answers

Part 2: Data minimisation

16. Does the guideline clarify when and by whom data minimisation should be performed throughout the data lifecycle (data collection, application assessment, data processing and result export)?

2

Do you have suggestions for improving clarity on roles and timings?
Max. 5000 characters.

The draft guideline conveys the principle that data minimization is expected “throughout the lifecycle,” but translating this into operationally clear responsibilities, stage-specific criteria, or role definitions may result in inconsistent implementation and compliance risks across Member States. The risks of modifying or using different data minimization methods across and within studies, as well as the risks associated with a lack of harmonization, should be acknowledged.
The guideline does not clearly articulate the data holder’s responsibilities for minimising data during primary collection versus retaining full datasets for clinical, administrative, or legal reasons. Practical criteria for deciding what should be collected vs. excluded, and how these choices interact with secondary-use obligations under the EHDS should be included. In the case of data collection in clinical trials, investigators and sponsors should apply protocol-level minimization. The sponsor/CRO (data holders) should minimize the data before transfer, with additional filters applied by SPEs.
While HDABs are instructed to evaluate whether the requested data is proportionate, the guideline does not define how or by what criteria this should be done. The tools (e.g., whether removing variables, aggregating data, or substituting synthetic/anonymized datasets) for data minimization should be clarified as should who performs these transformations.
During data processing, the guidance should clarify who performs data minimization (the HDAB, the data holder before transfer, or the SPE operator), which transformations (pseudonymization, generalization, suppression) are suggested and why, and whether minimization should be static or adjusted iteratively as the data user’s analysis evolves.
At the results export stage, SPEs should not release the results until appropriate minimization checks have been performed to avoid any residual identifier leaks. Who is responsible, what standards or thresholds, and how HDABs should document and/or audit should be clarified.

17. How practical are the recommendations for identifying and managing direct and indirect/quasi-identifiers in line with data minimisation principles, particularly regarding the trade-off between reducing privacy risks and maintaining data utility?

2

Please provide examples of challenges or alternative approaches for managing indirect/quasi-identifiers:
Max. 5000 characters.

While the EHDS provides a useful foundation, the recommendations would benefit from greater operational clarity to support the consistent implementation of data minimization, thereby reducing re-identification risk and preserving analytical utility. HDABs and data holders would benefit from more concrete criteria or decision frameworks for determining when suppression, generalization, or perturbation are appropriate, and how to document these choices. Similarly, the process for identifying quasi-identifiers remains high-level, leaving ambiguity about how comprehensive such assessments must be, what tools or methodologies are expected, and how domain-specific risks should be evaluated—especially for rare diseases and high-dimensional clinical and genomic data. Notably, various indirect identifiers, such as rare disease status, certain biomarker profiles, and site locations, among others, are necessary for scientific validity but also raise the risk of re-identification. For example, anonymizing data by aggregating the trial site geography has the potential to obscure safety signals linked to environmental or certain population-specific factors.
With longitudinal data, minimizing dates by converting to age groups or time intervals could compromise the ability to detect temporal patterns in adverse events. It is virtually impossible to fully anonymize genomic/omic data without attenuating utility. To address these issues, it would be beneficial to support dynamic risk assessment tools rather than a "one size fits all" approach. Similarly, the degree of minimalization can be adapted to the purpose in a more context-based fashion. The guideline should offer operational tools for stripping identifiers embedded in unstructured data (notes, images, linked files), which is a major practical challenge. Quasi-identifiers (e.g., age, postal code, rare diagnoses, dates of procedures) are the main source of re-identification risk in health datasets, and the guidance could offer a standardized list of common quasi-identifiers, tools to quantify residual re-identification risks, and a decision framework for what minimization tools to utilize (e.g., suppression, aggregation, tokenization). Additional guidance on linking multiple datasets and various types of data (including imaging, genomic data, and others) would be welcome. The roles and responsibilities of the data holder, HDAB, SPE, and others in data minimization before and after transfer should be clarified.
Without clearer thresholds, exemplars, and risk–utility evaluation steps, there is a risk of both over-minimization, which hinders research value, and inconsistent practices that may expose individuals to undue privacy risks.

18. Does the detailed examination of the five dimensions of data provision ("Who," "What," "When," "Where," "How") provide sufficient guidance for data

users and HDABs in preparing and assessing data permit/request applications to ensure data minimisation?

2

Are there any dimensions that require more elaboration or specific examples?
Max. 5000 characters.

The structured examination of the five dimensions of data provision (“Who,” “What,” “When,” “Where,” “How”) is a helpful organising framework, but the current draft stops short of offering sufficiently actionable guidance for HDABs and data users preparing or reviewing data permit applications. While the dimensions prompt applicants to reflect on the necessity and proportionality of requested variables, time ranges, populations, and data transfer methods, they provide limited detail on how these considerations should translate into concrete decisions about data minimisation. For example, it remains unclear what level of justification is expected for expanding a cohort (“Who”), extending follow-up periods (“When”), or requesting higher granularity (“What”). Similarly, HDABs would benefit from clearer evaluation criteria, templates, or examples illustrating how to assess whether a request meets the ‘minimum necessary’ standard. Without such operational guidance, the framework risks being applied inconsistently, with variable interpretations of sufficiency and proportionality across jurisdictions and projects.
Expanding the examples within the 5 dimensions would make the guidance more directly applicable and more specifically actionable for HDABs and data users preparing or reviewing data permit applications and for trial sponsors and data users who are managing sensitive participant-level data. dimensions of data provision (“Who,” “What,” “When,” “Where,” “How”) provide limited detail on how these considerations should translate into concrete decisions. Specifically:

- Who: Criteria for assessing whether the requested population necessary or could be narrower (e.g., clinical phenotyping thresholds, inclusion/exclusion logic, justification for age bands or geography), and examples of acceptable and overly broad cohort definitions. Additional clarify on whether “data users” includes regulators, auditors, or only secondary researchers. Further, in the clinical trial space, there are multiple categories of stakeholders such as sponsors, CROs, investigators, regulators to consider.
- What: Additional examples on what constitutes “essential” trial data versus excessive is needed. Provision of evaluation methods for granularity requirements (e.g., date → month → year), and examples of permissible reductions without compromising analytical validity.
- When: Additional guidance on what minimization is for longitudinal trials would be beneficial (at interim analyses vs. final datasets), as would guidance on appropriate observation windows, minimum look-back periods, and justification requirements for long-term follow-up.
- Where: How minimization requirements differ if/when processing occurs in different Member States or under centralized SPEs should be explicitly addressed. How spatial granularity (e.g., address vs. zip code vs. county vs. region) should be assessed in relationship to its analytic necessity compared to the risk of reidentification.
- How: Concrete expectations for secure processing arrangements, including criteria for safe environments, acceptable pseudonymization schemes, and documentation standards for minimization decisions. Providing practical templates, such as sample

minimization plans and workflows for data reduction before secondary use, would be helpful in addition to the conceptual descriptions provided.

Part 3: Pseudonymisation

19. Are the described purposes and goals for processing pseudonymised data within the EHDS clearly articulated and comprehensive?

3

Are there any additional purposes or challenges of pseudonymisation that should be highlighted?
Max. 5000 characters.

Purposes
- Enabling longitudinal linkage: Pseudonymization supports longitudinal, multi-episode, and/or multi-source linkage while decreasing risk of reidentification; however, the guideline could clarify how to prevent unnecessary linkability.
- Supporting tiered access models: Pseudonymization enables low-detail preview datasets prior to access of more complete research datasets.
- Facilitating cross-domain interoperability: Pseudonymization supports linkage across between health, care, administrative, or other datasets
- Providing auditability and accountability: Pseudonymized supports opt outs, error correction, and audits and monitoring.

Challenges
- Risk of residual identifiability through high-dimensional data: Even with pseudonyms, certain data (e.g., clinical, genomic, imaging, or free-text) remain identifiable.
- Key security and potential misuse: Risks arise from key custody, a lack of sufficient re-identification controls, and unanticipated re-linking.
- Interoperability vs. privacy tension: Using consistent pseudonymization methods across datasets improves linkage but increases re-identification risk; using inconsistent methods protects privacy but impedes research.
- Re-identification risk inflation over time: As auxiliary datasets grow and IT/AI advances,, pseudonymized data may become re-identifiable. Guidance detailing expectations for periodic risk reassessment (who, how, how often) should be provided.
- Pseudonymization quality: Quality varies and depends on algorithms, hashing, AES, and implementation. Explicit technical expectations and minimum standards is needed.
- Cross-border differences in legal interpretation: Pseudonymisation may be treated differently across Member States under GDPR and the EHDS. Clarifying how the guideline expects HDABs to address divergent national practices could reduce uncertainty.

Expanding the current guideline to include additional clinical trial-specific information and considerations would strengthen the pseudonymization content. Ensuring ethical and regulatory continuity would strengthen participant rights; this is also a requirement included in other European regulations. Regarding the longitudinal aspects of clinical trial data, this section of the guidance could be further developed to support the temporal dimension across trial phases. Because pseudonymization involves multiple stakeholders who interact with the data at different points, a more clearly defined role for each stakeholder would be beneficial.

20. Does the guideline provide adequate detail and recommendations on the practical implementation of pseudonymisation transformations?

3

What are the main practical challenges you foresee in implementing these recommendations, and what further guidance would be helpful?
Max. 5000 characters.

The guideline would benefit from additional detailed and practical recommendations for implementation in real-world HDAB and data-holder environments. The text leaves substantial uncertainty about required technical parameters (e.g., acceptable hashing algorithms, key-management practices, handling of quasi-identifiers, treatment of high-dimensional data (e.g., genomics, imaging, free text)). The guideline could also enhance its specificity in recommendations about operational processes, including who should perform pseudonymization at each stage, how responsibilities should be segregated, and how to document and audit transformations. Without these, data holders, HDABs, and data users may adopt different approaches that vary significantly in both privacy protections and data utility.
With respect to clinical research, guidance to mitigate several practical and logistical challenges would be helpful:

- Establishing and coordinating a consistent pseudonymization process with data that are first acquired by the clinician or clinical investigator and are then transferred and processed by other entities, including CROs, sponsors, and HDABs.
- Managing pseudonymization for trials that utilize multiple types of data (e.g., imaging, free text, genomic), all of which require different techniques.
- Managing the linkage of datasets with other systems such as disease registries, EHRs, etc.,
- Establishing pseudonymization service centers for smaller clinical research entities or academic institutions who do not possess that sophisticated expertise in-house.

21. s the guidance on pseudonymisation across the different phases of the EHDS user journey (data discovery, access application, data preparation, data

processing, and finalisation) clear and actionable for relevant actors (data holders, HDABs, data users)?

3

Are there any stages where the responsibilities or procedures related to pseudonymisation need further clarification?
Max. 5000 characters.

Within the helpful phased structure laid out in the guidance, the level of detail remains insufficient to make expectations actionable for HDABs, data holders, and data users. The following may provide additional clarity, with particular reference to clinical trials:

- Data Discovery: Determination if pseudonymization occurs before dataset registration or only after access approval, and the application of pseudonymization to metadata or preview datasets (leaving uncertainty about acceptable granularity and disclosure-control thresholds.) With pediatric and rare disease populations, more stringent methods to guard against indirect identification (rare pediatric conditions, neonatal studies, rare disease biomarkers).
- Data Access Application: Identification of who holds responsibility-- sponsor/data controller or Health Data Access Body (HDAB)—for verifying pseudonymization; clarify what information applicants must provide about anticipated pseudonymization needs and who (data holder vs. HDAB) is responsible for assessing feasibility; with pediatric populations, ensure access committees include pediatric ethics experts
- Data Preparation: Operational requirements for the pseudonymization workflow—such as key separation, encryption standards, role segregation, or handling complex quasi-identifiers. Clearly lay out the steps that define the data-flow mapping for the data controllers; with pediatric populations, ensure clear responsibility for maintaining re-identification capability if/when future re-consent is needed
- Data Processing / Analysis: Specify what analytical operations are allowed, the minimum controls necessary, and the documentation required within secure processing environments, and provide concrete examples. In clinical trials, what standards, controls, and documentation are necessary when using pseudonymized trial data; with pediatric populations, ensure advanced oversight with pediatric data sets with regard to cell size thresholds.
- Finalization / Result Disclosure: Provide further details for how pseudonymized trial data outputs can be archived, further shared, or converted to an anonymized form; provide guidance on the interaction of pseudonymization with output checking, especially with respect to inadvertent re-identification through reporting; with pediatric populations, include additional guidance for how pediatric pseudonymized data will be converted to anonymized or archived data in situations where further follow up is not anticipated

More explicit mappings of responsibilities, technical norms, and step-by-step process expectations across each phase would materially improve consistency and implementation clarity. A responsibility table by phase would serve to provide the logistical details necessary for use and implementation.

22. Are there further ambiguities in the pseudonymisation section that should be addressed regarding the recent judgement of the Court of Justice of the EU in the case EDPS vs SRB (C-413/23 P)?

Max. 5000 characters.

The pseudonymization section provides a solid conceptual foundation and is generally aligned with Court of Justice of the EU in Case C-413/23 P (EDPS v. SRB). To support operational clarity, the guideline should:

1. Introduce a contextual risk–identifiability assessment framework for pseudonymized data transfers — including factors to consider, how to weigh them, and examples.
2. Define minimum organizational, technical, and contractual safeguards (key management, access controls, encryption, role separation) that reduce re identification risk, referencing the “means reasonably likely to be used” standard.
3. Clarify who is responsible at each stage (controller/data holder, HDAB, data user) for assessing identifiability and maintaining documentation.
4. Provide template documentation/audit records for decisions about identifiability, pseudonymization, and anonymization status.
5. Offer sector-specific guidance, especially for health data, on when pseudonymization may be insufficient (e.g., genomics, rare diseases) and additional privacy-enhancing methods should be used.
6. Require transparency toward data subjects, reflecting that pseudonymized data may be transferred to third parties and possibly treated as non-personal depending on context.

Specifically, further guidance will help prevent overly conservative or lax pseudonymization, leading to either unnecessary data utility loss or unintended re-identification risk; cross-Member State and cross-project inconsistency, which limits utility; misunderstanding the identifiability of data (and therefore requirements for consent, reporting, etc.); and compliance problems.

With respect to clinical trials, and from a clinical-trialist’s viewpoint this section could be improved with:

- More detailed operational guidance for multi-actor pseudonymization
- A responsibility matrix aligned with the phases previously mentioned

Important areas to address include:

- Identifiability of data (re-identification) depends on each data user’s perspective and is therefore not absolute; for example, for pediatric participants, identification may occur through family, school or other community health care settings
- Clarity on the likelihood of re-identification over time (e.g., cases of long-term follow-up)
- Cross-border cohorts that may involve different regulations (e.g., reporting on ethnicity, ages of majority).

Part 4: Anonymisation and synthetic data generation

23. Does the guideline adequately describe how anonymisation and synthetic data generation can be applied within the EHDS?

3

Please elaborate:
Max. 5000 characters.

The terms are defined separately and conceptually in the guidance. Further clarification, as discussed in prior sections for both anonymization and synthetic data would be helpful.

Regarding anonymization involving genomic data, high-dimensional health datasets, and rare diseases, among other applications (e.g., clinical trials and complex trial designs), “anonymization” is not always possible. What is the level of acceptable risk of re-identification that is considered anonymized versus pseudonymized? When would moving to synthetic data be important, and when would that also be impossible (e.g., genomic data)? A decision tree tool could help users determine which specific category applies to a particular trial situation. Specific techniques (e.g., differential privacy, aggregation thresholds), risk assessment methods, and documentation standards for demonstrating compliance with GDPR Recital 26 and EHDS requirements would be helpful.

Similarly, the section on synthetic data describes the potential to preserve utility while mitigating privacy risks, but does not provide practical guidance on generation methods, validation metrics, or privacy safeguards. It would benefit from recommendations for evaluating fidelity, balancing privacy and utility, and documenting the provenance of synthetic data. Finally, although synthetic data is presented as a separate category, the criteria that render it safe and trustworthy for clinical research—quality assurance, representativeness, validation—could be more strongly emphasized in the guidance.

24. How clear and applicable are the proposed use cases (Table 2) and the high-level architecture (Figure 6) for implementing anonymisation, synthetic data generation, and privacy risk assessment within the EHDS framework?

3

Do you have additional examples of use cases where anonymisation or synthetic data might be relevant?
Max. 5000 characters.

No answers

Are there specific use cases or architectural components that require more detailed explanation or examples?
Max. 5000 characters.

The following cases/types of trials would benefit from additional detailed explanation and examples:

- In pediatric and rare disease studies (small study populations) even aggregated output may be identifying. Additional concrete case studies and methods such data suppression or controlled access would be helpful.
- Longitudinal safety data: When age bands are used rather than dates, the ability to detect temporal safety signals is diminished. Additional, nuanced techniques would be helpful.
- Genomic data: The utility of individual-level genomic sequences data is essentially destroyed when existing anonymization techniques are applied. This point should be more clearly stated in the guidance.
- Operational and implementation detail are lacking; tools such as checklists, exemplar workflow diagrams would help in practical application of the theoretical materials offered in the guidance.

25. How effectively do the guidelines address the requirements for documentation of anonymisation/synthetic data generation, privacy risk assessment, and tooling recommendations for supporting these processes?

2

What specific suggestions do you have for improving these areas?

The guideline recognizes the trade-offs but does not provide operational guidance for deciding the appropriate, risk-proportionate balance of data utility and data minimization. For example, in clinical trials, statistical power and subgroup analyses may depend on retention of age and sex data. A structured risk-benefit assessment framework would help users decide the right balance in practice.

The guideline should provide standardized templates, required content, or guidance on how to demonstrate compliance for audits or cross-border review. For privacy risk assessment, it acknowledges the need to evaluate residual identifiability and re-identification risk, yet it does not specify methodologies, thresholds, or quantitative approaches that are appropriate for different types of health datasets, nor does it clarify how often assessments should be repeated as data or auxiliary information evolve. The guideline mentions the use of statistical, technical, or software tools to support anonymization or synthetic data generation but provides no recommendations on validated methods, minimum capabilities, or interoperability standards, leaving HDABs and data users to select approaches independently. Without concrete guidance on documentation standards, risk assessment processes, and tools, there is a risk of inconsistent practices, potential gaps in compliance, and variable protection of data subjects across EHDS implementations.