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 *

Netherlands

2. Type of responder *

Other
non profit foundation

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

If yes: On behalf of how many organisations?

No

4. Sector *

Other
health data infrastructure

5. Organisation size *

Small to medium enterprise (10–249 employees)

6. Professional role / function

No answers

Quality

7. Is the document easy to understand? *

3

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

2

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

2

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.

The implementability of the guidelines constitutes a significant concern. It is essential that, at the outset, the workflow of the “data journey” can at least be executed end-to-end in a minimal form. This should then gradually evolve from a basic process into an increasingly comprehensive one. Accordingly, it is not advisable to attempt to regulate and operationalise all details from the very beginning through implementing acts. Instead, a learning-by-doing approach should be adopted.At the same time, a critical challenge lies in preventing the creation of excessive discretionary space for Member States to establish divergent national legislative arrangements (as illustrated by the shortcomings experienced under the GDPR). Therefore, particularly in the initial phase, a learning and growth perspective should be applied, combined with clearly defined, directly implementable requirements per phase regarding the HOW, which do not allow for differing interpretations or implementations across Member States.

Each guideline should be designed to incorporate an iterative learning and development trajectory. In this context, implementation should commence with a limited scope, including a reduced set of mandatory fields and/or requirements, while, at the same time, establishing predefined timelines at which additional obligations will be introduced, without predetermining in advance the specific nature of those future obligations. Each implementation phase should serve as a basis for evaluation, on the grounds of which decisions shall be taken as to the introduction of further mandatory elements in subsequent phases.

Each guideline should include cross-references to other relevant guidelines and should also contain explicit references to Chapters 2 and 3 of the EHDS. At present, the overall coherence between the guidelines is insufficiently ensured, and the timelines and processes for input and consultation meetings are, in some cases, (too) short. Moreover, there are significant differences in the level of detail, with some guidelines being highly high-level, while others are detailed and technical. In certain instances, the substantive content is also not fully aligned across the documents. The inclusion of an overall reader’s guide for the full set of documents would improve accessibility and usability; in addition, attention should be given to ensuring consistent terminology and a uniform roles and responsibilities model.

In the development of the TEHDAS2 documentation, which describes the full process of data application, data discovery and related procedures, it has been observed that the documentation is largely drafted from the perspective of a centralised analysis environment. Insufficient clarity is provided regarding the concept of federated analytics, its implications (for example, the requirement that permits must also be capable of being ‘controlled’ by data stations), as well as the respective advantages and disadvantages of centralised versus federated approaches.

The guidelines set out in Chapter 4 are largely drafted from a cohort- and research-

oriented perspective. Since the data in question predominantly originate from the healthcare sector, the guidelines should also be formulated from a clinical care perspective. Accordingly, in addition to research expertise, the appropriate (clinical) expertise should be duly involved in the development of the substantive guidelines, including expertise relating to Chapters 2 and 3 of the EHDS.

The document needs a (thorough) revision given the recent judgement of the Court of Justice of the EU in the case of EDPS vs SRB. The authors already stated that the current version does not include this ruling, but the next version needs to take this into account.

Ther document refers a lot to other documents and at least one of them is not yet available (M7.5) which makes reading hard(er). The document cannot be used without the others, let alone be reviewed without some of the others.
Nowhere is it described how the results of minimization and pseudonymization/anonymization are monitored across all Member States to ensure they are "comparable/standardized"

Recommendation: this aspect must be an integral part of M7.2 (and also the other consulted documents)

Related to this observation: Regarding linkage and pseudonymization techniques used: nowhere is it stated that if multiple TTPs and trusted data holders are involved in an application, they must apply harmonized/standardized "encryption"

Available (best practice) privacy enhancing technologies (PETs) (including MPC) are lacking, while they offer possibilities to "shield data and still query/analyse data federatedly".
Recommendation: include those technologies.
At various pages the term “research objectives” is used.
Recommendation: rephrase to “secondary use objectives” (or at start of document state that with research objectives all secondary use objectives are meant).

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? *

Other
health data infrastructure, direct member HDAB-NL program

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.

No answers

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.

Genetic data is named as one of the relevant data types (in 2.1.1 Data types). While the definition used here is fine, it should be noted that genetic information can also be deduced from other types of omics data (e.g. proteomics) and that these data types therefore are potentially also sensitive.

In the same paragraph (2.1.1) Concerning the definition of imaging data (2.1.1, p.6):

- “Medical imaging data refers to digital representations of visual information captured for clinical purposes”

Why the term medical and clinical here and not for the other data types? I think the terms are redundant, as the scope of the EHDS is Health Data (“gezondheidsgegevens” in Dutch). What falls under Health Data, is/should be defined clearly in the EHDS.

Suggestion: Imaging data refers to digital representations of anatomical or physiological information acquired through diverse imaging modalities, such as X-ray, CT, MRI, ultrasound, nuclear imaging, as well as photographs.

- Mentions of pathology and microscopy are missing, potentiall angiography as well.

- 'Raw data' might be confusing. It would be preferred to use 'pixel data' to indicate it concerns the image itself

And a more general concern: will the HDAB have enough specialised knowledge to accurately evaluate the measures taken for data minimsation, pseudonimisaion, anonimisation?

At various pages confidentiality, integrity, and availability are often described as being part of the GDPR. It is however also part of information security regulations like NIS2 and other ISO standards like 27001. Security and privacy are different aspects and therefore should be, when necessary to prevent confusion, explicitly stated as such.

Page 20: The main steps that HDABs should put into effect towards the goal of issuing data permits and requests are as follows… At bullit 3 it is stated "or propose a changeover to a data request (e.g., if only aggregated data in a statistical format are appropriate)" This should be the first check: is permit okay or change to data request. Why: this check/change is minimisation to the max.

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.

In the summary it is stated: "This principle (Editor: data minimisation) applies throughout the entire lifecycle of the data, from when it is first collected and prepared by the health data holder, to when it is assessed by a health data access body (HDAB), and finally, during its use and processing by the health data user. "

the publication/archiving of results seems to be lacking. The export out of the SPE needs data minimisation/pseudo/etc. too.

Recommendation: Include this explicitly in Executive Summary, like it is done in Figure 1, and paragraph 3.1-table 1.

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)?

2

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)?

No answers

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

No answers

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.

(quasi-identifier): Related to the definition of quasi-identifiers: Is this the right way to define quasi-identifiers? Isn’t it more handy to define them as parameters that are both seriously reducing the population and knowable for an individual, so that it is possible for a known individual to be matched to it, by the data user?

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.

On the domain-specific considerations for data minimisation for imaging data (3.4.1, p.22):

- Missing a mention about data aggregation here in general, not only specific for imaging data. However, aggregation might pose additional burdons on data holders. Who is doing that?

- This section is written quite directive. It should be emphasized that the suggested procedures COULD be used for data minimisation.

All procedures have the risk of reducing quality and usability of the requested data, and should therefore be taken in careful consideration with the data user to make sure the data is still of sufficient quality for the intended use.

- Next to digital blurring, also masking should be named as a potential procedure.

- Some concerns around tooling: verified tools should be used to make sure the anonimisation is reliable and does not create artifacts. This is especially important when combining data from different sources.

- For imaging data, EUCAIM is working on a list of verified tools for dataholders for theses purposes.

On the 'What' dimension, concerning genetic data; it is stated that 'genetic data are classified as sensitive personal data under Article 9(1) (GDPR), and special safeguards apply to them.'. Last part might be rephrased to 'special safeguards may apply to them'.

On the 'How' dimension, concerning domain-specific considerations for genetic data (p. 22): Additionally to filtering for specific variables, it is also possible to filter for specific regions of the genome (e.g. give access to only a limited set of chromosomes or smaller), or to reduce resolution (e.g. identify that there are variants to a gene, but removing information about what variation exactly).

“Who”

Page 15: "An estimate of the expected cohort size should be provided by the applicant in the application phase, along with the inclusion and exclusion criteria used to derive this estimate from the data assets of interest. This information serves to support the HDAB in

assessing the proportionality and feasibility of the requested data provision. For instance, the applicant may state: 'We apply inclusion criteria A, B, and C to dataset X, which contains 120,000 individuals. Based on prior studies, we estimate that 5,000 individuals will match our criteria.”

Feedback: This estimation is only possible if it is clear how many individuals a dataset contains and this is therefore stated in the metadata of the dataset.

Page 24: Closing remarks "This section has outlined how HDABs, data holders and data applicants/users can operationalise this principle across the data lifecycle — from application, through assessment, to permit issuance and finally to data usage."

Recommendation: The step of exporting data/results from an SPE or publishing results, such as in an article with an accompanying dataset, seems to be lacking in “this principle” (data minimization). This text should indicate how "publication of data and results" is handled with respect to the applicable principles. It's quite possible that the requirements, for example, regarding traceability, are much stricter than necessary for analysis within the SPE.

Part 3: Pseudonymisation

19. Are the described purposes and goals for processing pseudonymised data

within the EHDS clearly articulated and comprehensive?

2

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

In the summary it is stated: "This principle (Editor: data minimisation) applies throughout the entire lifecycle of the data, from when it is first collected and prepared by the health data holder, to when it is assessed by a health data access body (HDAB), and finally, during its use and processing by the health data user. "

the publication/archiving of results seems to be lacking. The export out of the SPE needs data minimisation/pseudo/etc. too.

Recommendation: Include this explicitly in Executive Summary, like it is done in Figure 1, and paragraph 3.1-table 1.

At page 3 Figure 1: Why is pseudonimisation at “Data use” and “Finalisation” excluded? Once results including data are exported/published, 1) one wants/needs to be able to enrich data(sets) or make parts publicly available, 2) if this exported/published data(set) is reused and incidental findings need to be reported back one needs pseudonimisation otherwise one cannot fulfil subjects wishes/control over their data

Page 4: "Pseudonymisation (see section 6) should be performed as early as possible and re-pseudonymisation must occur before data provision in a SPE (see section 6.3 for more detail)."

Feedback: Pseudonymization should be performed as early as possible. However, the term "as early as possible" can be misinterpreted. It's not about being technically as early as possible, but rather appropriate for the data request. Early pseudonymization can complicate data linking.

Page 6: "Genetic data refers to information derived from an individual's DNA, representing their genetic makeup. It includes sequences of nucleotides (A, T, C, G), genetic variants, mutations, and structural variations that influence traits, disease susceptibility, and responses to treatments."

Feedback: It may be useful to clarify whether this also includes data types that are biologically relevant but not directly derived from DNA, such as epigenetics, proteomics, or metabolomics data.

Page 22: "Variables that are not linked to study objectives and cannot be considered as control/confounder variables should be excluded. If different variables that point to similar kind of information is available at data holder’s side, these variables should be evaluated as alternatives in terms of utility and risks."

Feedback: The text assumes that researchers always know in advance which variables are confounders. In reality, confounders are not always known before the analysis and can only be identified during the study. It is unclear who determines what a confounder can be.

Page 22: "Technical metadata should be removed too, if not explicitly required."

Feedback: It can often be necessary to maintain technical metadata, for example, to know which pipeline to use. There's a risk: who decides whether something is "explicitly required"?

Page 22: "Filtering for relevant variables (i.e., if the analysis concerns only some mutations - e.g., BRCA1 for breast cancer - only those variants can be preserved from the entire genome)."

Feedback: Filtering for relevant variables supports data minimization. However, the reliability of this filtering depends on the tools and pipelines used to access these variables (especially with genomics data). It is important to clarify how this is ensured.

20. Does the guideline provide adequate detail and recommendations on the

practical implementation of pseudonymisation transformations?

2

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

At page 8 Data minimization it is stated that “Data minimisation may include: reducing the volume of data made available.

Recommendation:

Specify what is meant with reducing the volume. Since it could mean various things, like total data size in bytes, number of records, number of columns, …

Next, it's important to emphasize that sometimes sufficient data is precisely what's needed to ensure anonymity. The current wording might give the impression that less data is always better, while anonymization (e.g., k-anonymity) may require sufficient records to prevent traceability.

Page 9: "Under Article 68 (EHDS), HDABs are responsible for deciding whether anonymised or pseudonymised data should be used (see Article 68(1)(c)). Regardless of the legal status of the data, data minimisation applies to all phases of data user’s journey and the actors involved (data holders, HDABs, data users), including within and beyond the SPE. In fact, data minimisation also applies after the data is made available, including during processing by the data user, during analysis and results export."

Feedback: It's good to be explicit about how data minimisation applies to data archiving!

Data minimisation: on page 10 for the first time: “a changeover from a data permit to a data request (Article 68(3), EHDS)” is mentioned

Recommendation: This should be explicitly be stated as the first step in data minimisation: is a permit needed!

Table 1. Role-based involvement for data minimisation in the EHDS. The list below addresses data access applications, while similarities for data requests hold. For simplicity reasons, trusted data holders are left out from this table.

This raises questions like is it just to keep the layout of the table simple or is the role of trusted data holders adding “complexity”.

Recommendation: at least address what they mean with “for simplicity reasons”. Make it

explicit and next state what the role of other entities like “authorised participants” and “health data intermediation entities” will be with respect to data minimization.

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)?

2

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

Feedback D1: The part about Pseudonimisation addresses data linkage several times and refers to M7.5 which is currently not yet available. We find it difficult to judge without the other guideline being available.

In addition, wondering if there will be designated place in the metadata to place information about pseudonimisaion, anonimisation etc. Or will this have to be placed in the general documentation?

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.

It is stated that the document is written before this judgement was published. This document should therefore be re-evaluated in light of the judgement.

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?

2

Please elaborate:
Max. 5000 characters.

Concerning anonimisation

- The relevant approaches are mentioned here for anonimisation of imaging data, but are missing in the section about data minimisation. It is not clear that this is specified elsewhere in the document.

- Some concerns around tooling: verified tools should be used to make sure the anonimisation is reliable and does not create artifacts. This is especially important when combining data from different sources.

- For imaging data, EUCAIM is working on a list of verified tools for dataholders for theses purposes.

Concerning Tooling (7.5.5, p 44)

This section discusses tools that should be included in the SPE. However, some measures should be taken before the data enters the SPE. There should be verified tools for data holders to use before the data is sent to the SPE.

For imaging data, EUCAIM is working on a list of verified tools for dataholders for theses purposes.

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?

2

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.

No answers

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?

No answers