# Summary page:
# TEHDAS2 public consultation on Draft guideline on linkage of health datasets

This consultation has 4 pages and 22 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
Foundation working on better health for citizens and patients by reusing health and life science data with an integrated health data infrastructure for research, policy and innovation

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

If yes: On behalf of how many organisations?

No

## 4. Sector *

Other
Foundation working on better health for citizens and patients by reusing health and life science data with an integrated health data infrastructure for research, policy and innovation

5. Organisation size *

Small to medium enterprise (10–249 employees)

6. Professional role / function

Program manager

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.

We note that the guidelines have largely been developed and described as separate documents by different authors (which is understandable), but as a result, there is limited insight into how they relate to each other within the broader process.

Better horizontal alignment between the guidelines is needed to support a clear and coherent data journey. This would help stakeholders understand how the various components connect, ensure consistency within the framework, and facilitate practical implementation throughout the entire data lifecycle.

This guideline does not mention any direct consequence for any of the digital business capabilities of the HealthData@EU blueprint, e.g. catalogue, DAAMS, SPEs.

Linking back (re-identification) to a "natural person" if there is a basis for doing so, such as reporting an incidental finding, is not mentioned anywhere in this guideline (not even in the "out of scope" section), nor is the return of enriched data to the original data holder.

**The following questions are specific for TEHDAS2 Draft guideline on linkage of health datasets**

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

Other (please specify in text field)
Foundation working on better health for citizens and patients by reusing health and life science data with an integrated health data infrastructure for research, policy and innovation

**Why, when, and by whom are data linked?**

## 12. On a scale of 1–5, how clear is section 4.1?

(1 = unclear, 2 = rather unclear, 3 = neutral, 4 = rather clear, 5 = very clear)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | ○ | ○ | ◉ | ○ | ○ |

13. Do you have comments for section 4.1 that have not yet been addressed in the generic feedback? The comments can address potential improvements, missing aspects or inconsistencies.

Max. 5000 characters.

Throughout the entire process, particularly in 4.1 but also in the scenarios in Annex 5, nothing at all is indicated regarding when data minimization can/must take place (reference is made to many EHDS articles but not to Article 66).

Since this section states the "when", it should be explained via extending the scenario's when in the process data linking takes place and when data minimisation and purpose limitation. Most likely the data minimisation and purpose limitation will take place after data linkage but before the linked dataset is made available for the data user in the SPE. If data minimisation and purpose limitation is done before data linking, linking might be not possible at all, might have higher error rates, etc.

Include article 66 inside figure 2, most optimal "when" will be as last step before the data user gets access to the linked data in the SPE; in Table 1 on the linkage scenario's; and in Annex 5.1 all scenario's

The roles of authorised participant (except once in Annex 3 under the description of the term Health Data User) and health data intermediation entities within data linkage are lacking throughout this guideline.

The importance of establishing a clear division of responsibilities for data linkage should be added.

**What data are being linked?**

14. On a scale of 1–5, how clear is section 4.2?

(1 = unclear, 2 = rather unclear, 3 = neutral, 4 = rather clear, 5 = very clear)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | ○ | ◉ | ○ | ○ | ○ |

15. Do you have comments for section 4.2 that have not yet been addressed in the generic feedback? The comments can address potential improvements, missing aspects or inconsistencies.

Max. 5000 characters.

Section 4.2 states that: "The applicant must describe in the application how data from different sources will be linked." To be able to do this as an applicant, you need knowledge of the datasets involved (including unique identifiers or variables that can be used) for linking, as well as the various linkage methods that are available for the specific datasets. There is nothing about this in the guideline (and presumably nothing in the dataset description articles of the EHDS).

**How is data linked?**

16. On a scale of 1–5, how clear is section 4.3?

(1 = unclear, 2 = rather unclear, 3 = neutral, 4 = rather clear, 5 = very clear)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ◉ | ○ | ○ |

17. Do you have comments for section 4.3 that have not yet been addressed in the generic feedback? The comments can address potential improvements, missing aspects or inconsistencies.

Max. 5000 characters.

This paragraph should provide based on various linking methods guidance in when data minimisation should take place. Especially applying minimisation before linking with deterministic indirect or probabilistic methods might result in lower linking quality.

SPIDER is mentioned in 4.3.1. and 4.3.4. but the document does not reference the EUPID. Authors must consider if mentioning SPIDER to also mention EUPID as one of the best practices (suggested reference: https://services.eupid.eu/).

**How is quality and accuracy of linked datasets ensured?**

18. On a scale of 1–5, how clear is section 4.4?

(1 = unclear, 2 = rather unclear, 3 = neutral, 4 = rather clear, 5 = very clear)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ◉ | ○ | ○ |

19. Do you have comments for section 4.4 that have not yet been addressed in the generic feedback? The comments can address potential improvements, missing aspects or inconsistencies.

Max. 5000 characters.

No answers

**Role based access to linkage processes**

20. On a scale of 1–5, how clear is section 5?

(1 = unclear, 2 = rather unclear, 3 = neutral, 4 = rather clear, 5 = very clear)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ◉ | ○ | ○ | ○ |

21. Do you have comments for section 5 that have not yet been addressed in the generic feedback? The comments can address potential improvements, missing aspects or inconsistencies.

Max. 5000 characters.

Authors must respect/adhere to NIS2 and other laws/regulations and state that organisations must apply strict authorisation principles (like ”least privilege”).

Text annotation: pag 33: 5. ”role-based access measures are helpful. The following section elaborates on recommendations on how to address and implement some rule-based access measures” -> ” role-based access measures are helpful. The following section elaborates on recommendations on how to address and implement some role-based access measures”.

22. Do you think there are other topics or questions in the context of data linkage in the EHDS ecosystem for secondary use of health data that should be added or clarified?

Max. 5000 characters.

It is not clear in the main text nor in the appendices (particularly the scenarios) how the linking organisation (the organization performing the data linkage: HDAB, DH, or TDH) will maintain the "key table(s)" in the event that multiple datasets from different data holders (and the data user) need to be linked, such that the data user does not have access to them, but the linking organisation can access them in case of emergency, or when enriched data must be returned to the original data holders so that the correct persons can be re-identified, and also in the event of incidental findings, the identification of the persons can take place so that the responsible persons can be contacted (in the finalisation step).

The roles of authorised participant (except once in Annex 3 under the description of the term Health Data User) and health data intermediation entities within data linkage are lacking throughout this guideline.

Annex 4: Not all data is tabular. Authors should include in the various sections provide insights on how to link with non-tabular data! (e.g. what if you want to link clinical data to DICOM-images?)

It is important to know what minimum information is required to properly request a data linkage, and that this often demands specialized knowledge and (personal) communication with the data holder(s). As well as which fields in application forms are required with respect to data linkage to streamline preliminary communication as much as possible, in the interest of an efficient data linkage process.

The document lacks information on how citizens will be informed about data linkages involving their data?

EHDS data can also be combined with other datasets; how is data linking handled in that case?