

D7.5 Guideline for Health Data Access Bodies on linkage of health datasets

TEHDAS2 – Second Joint Action Towards the European Health Data Space

16 September 2026

Co-funded by the European Union

0 Document info

Disclaimer
Views and opinions expressed in this deliverable represent those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

0.1 Authors

Author(s)	Organisation
Pia Brinkmann	BfArM, Germany
Zdeněk Gütter	MZCR, Czech Republic
Lise Skovgaard Svingel	RM, Denmark
Hanna Tervonen	Findata, Finland
Tim Vlaar	HDH, France
Jannik Wheler	RM, Denmark
Minerva Alvarez Matas	MoH, ES
Justin Ansotte	HDA, Belgium
Gianluca Carlini	IRCCS-ISNB, Italy
Francesco Casadei	IRCCS-ISNB, Italy
Ana Martin Moreno	MoH, ES

0.2 Keywords

Keywords	TEHDAS2, Joint Action, Health Data, European Health Data Space

0.3 Document history

Date	Version	Editor	Change	Status
09/12/2024	0.1	Pia Brinkmann	Initial document creation	Draft
20/05/2025	0.2	All contributors	First draft	Draft
19/12/2025	0.3	All contributors	Second draft after initial EC feedback	Draft
13/02/2026	1.0	All contributors	Submission for consortium feedback before Milestone publication	Draft
24/03/2026	2.0	All contributors	Milestone submission for acceptance in PSG	Final

24/08/2026	3.0	All active contributors	Incorporating feedback from public consultation and submission for consortium feedback before Deliverable finalisation	Draft
14/09/2026	4.0	All active contributors	Incorporating last comments before PSG	Draft
16/09/2026	5.0	Pia Brinkmann	Last editing after PSG acceptance with minor changes	Final

Accepted in Project Steering Group on 16 September 2026.

Copyright Notice
Copyright © 2024 TEHDAS2 Consortium Partners. All rights reserved. For more information on the project, please see www.tehdas.eu.

Contents

1 Executive summary 5
1.1 Abbreviations 6
2 Introduction 7
2.1 Work Package 7 Introduction 8
2.1.1 Task Overview – Data linkage (T7.5) 8
3 Scope 10
4 Data linkage 12
4.1 Why, when, and by whom are data linked? 12
4.2 What data are being linked? 17
4.2.1 Electronic health data protected by intellectual property rights or trade secrets 21
4.3 How is data linked? 21
4.3.1 Which methods can be used? 21
4.3.2 Which method should be preferred? 25
4.3.3 Which are the requirements for each method? 26
4.3.4 How should data be prepared before linkage? 26
4.3.5 How to perform data linkage? 27
4.3.6 How to validate data linkage? 28
4.4 How is quality of linked datasets ensured? 30
4.4.1 Linkage quality 32
4.4.2 Techniques for linkage quality assessment 36
4.4.3 Recommendations related to linkage in the interest of ensuring its quality 38
5 Role-based access to linkage processes 41
5.1 Before data extraction for data linkage 42
5.2 After linkage feasibility is verified 44
5.3 Managing data linkage documentation 45
5.4 Destiny of the linked data 45
6 Penalties for misuse of (linked) data 45
7 Repeated and frequent applications to (linked) datasets 46
8 Open questions 47
Annex 1 – Methodology 50
Annex 2 – Public consultation summary 51
Annex 3 – User journey 52
Annex 4 – Glossary 54
Annex 5 – Data description template 57
Annex 6 – Data linkage scenarios 60
Annex 6.1 Data access application 60
Annex 6.2 Data request 65
Annex 7 – Excursion on linkage quality assessment methods 73
Annex 8 – Template for instructions for data users regarding upload of data they possess 75
Annex 9 – Use cases 77
9.1 Linking claims data and cancer registry data in Germany: Lessons learned from record linkage approaches 77
9.2 AI development using cross-country national health data 78
9.3 Data linkage in the case of ready-made datasets 81
9.4 Data linkage with a Trusted Third Party 82
Annex 10 – Data quality considerations 84

Annex 11 – Scenarios data linkage with TTP 87
Scenario 1: 1 data holder, no TTP 88
Scenario 2: 2+ data holders, no TTP 89
Scenario 3: 2+ data holders, with TTP 89
Scenario 3.1: Heterogeneous starting configurations 90
Scenario 4: 2+ data holders, with TTP (who only communicates with DHs) 92
Scenario 5: 2+ data holders, with TTP (PPRL using blind pseudonymisation) 93
Scenario 5.1: Blind pseudonymisation with heterogeneous starting configurations 94
Scenario 6: 2+ data holders, no TTP (with multi-party computation) 96
Methods comparisons 98

1 Executive summary

This guideline aims at providing practical and legal support for Health Data Access Bodies (HDABs) on linkage of health datasets. It helps answering several questions of data linkage in the EHDS:

- **Why, when and by whom can data be linked within the EHDS?** You can find procedures for data linkage, role descriptions and their legal references (section Why, when, and by whom are data linked?). More detailed scenarios are presented in Annex 6 – Data linkage scenarios for data access applications (Annex 6.1 Data access application) and data requests (Annex 6.2 Data request).
- **What data are being linked?** Here, we also refer to Article 51 from the EHDS Regulation (section 4.2).
- **How is data linked?** Including the methods to be used, which method should be preferred, what are the requirements for each method, how the data should be prepared before linkage and how to validate the linkage (section 4.3).
- **How is quality of linked datasets ensured?** In addition, we elaborate on linkage quality assessment and provide recommendations (section 4.4).

Moreover, you can find information regarding role-based access to linkage processes and regarding repeated and frequent applications to (linked) datasets in the guideline.

Finally, we added as much practical guidance as possible in the **Annexes**, covering a wide range of topics:

- Annex 5 provides a data description template that linking organisations could use in communication with the data user. Optional elements may be added reporting information on data quality.
- Annex 8 is a template with suggestions for instructions for data users regarding the upload of the data they possess.
- Annex 9 contains three use cases addressing (1) pitfalls (2) AI development using cross-country national health data and (3) data linkage in case of ready-made datasets and (4) data linkage with a trusted third party (TTP).
- Annex 10 elaborates on data quality for direct and indirect linkage activities.
- Annex 11 illustrates possible data linkage scenarios including TTPs.

1.1 Abbreviations

Term	Abbreviation
Advanced Encryption Standard 256 bits Galois/Counter Mode	A256GCM (or AES256GCM)
Data holder	DH
Data user	DU
European data protection board	EDPB
European Medicines Agency	EMA
European Patient Identity	EUPID
European Rare Disease Registry Infrastructure	ERDRI
Regulation (EU) 2025/327	EHDS
Electronic health record	EHR
European Union	EU
Regulation (EU) 2016/679	GDPR
Health data access body	HDAB
Keyed-hash message authentication code	HMAC
Personal identifier	ID
Privacy preserving record linkage	PPRL
Rivest–Shamir–Adleman cryptosystem with optimal asymmetric encryption padding 256-bit	RSA-OAEP-256
Secure Hash Algorithm 256-bit	SHA256
Secure multi-party computation	SMPC
Secure Privacy-preserving Identity management in Distributed Environments for Research	SPIDER
Secure processing environment	SPE
Trusted health data holder	TDH
Second joint action towards the European health data space	TEHDAS2
Trusted third party	TTP
Work package 7	WP7

2 Introduction

Advancing health data use in the European Health Union

As part of the European Health Union, the European Union (EU) is advancing the use of health data for secondary purposes, including research, innovation and policymaking. Smooth and secure access to data will drive the development of new treatments and medicines and optimise resource utilisation – all with the overarching goal of improving the health of citizens across Europe.

TEHDAS2, the second joint action Towards the European Health Data Space, represents a significant step forward in this vision. The project will develop guidelines and technical specifications to facilitate smooth cross-border use of health data, and support data holders, data users and the new Health Data Access Bodies (HDABs) in fulfilling their responsibilities and obligations outlined in the European Health Data Space (EHDS) regulation.

TEHDAS2 focuses on several critical aspects of health data use, including:

- Data discovery: findability and availability of health data, ensuring it is accessible for secondary purposes.
- Data access: developing harmonised access procedures and establishing standardised approaches for granting data access across member states.
- Secure processing environments (SPEs): defining technical specifications for environments where sensitive health data can be processed safely.
- Citizen-centric obligations: providing guidance on fulfilling obligations to citizens, such as communicating significant research findings that impact their health, informing them about research outcomes and ensuring transparency in how their data is used.
- Collaboration models: developing guidance on collaboration and guidelines on fees and penalties as well as third country and international access to data.

TEHDAS2 will contribute to harmonised implementation of the EHDS Regulation through the concrete guidelines and technical specifications. Some of these documents and resources will also provide input to implementing acts of the Regulation. Hence, the joint action will increase the preparedness for the EHDS implementation and lead to better coordination of member states' joint efforts towards the secondary use of health data, while also reducing fragmentation in policies and practices related to secondary use.

This document should be understood as an expert opinion and guidance document developed within the TEHDAS2 framework, reflecting technical and expert input from the project partners. It is not legally binding and does not constitute a formal guideline or technical specification under the European Health Data Space.

This document does not represent the position of the European Commission.

Legally binding and enforceable requirements under the European Health Data Space are laid down in Regulation (EU) 2025/327 and, where applicable, in Implementing Acts adopted by the European Commission, within the limits of the empowerments provided by the Regulation.

2.1 Work Package 7 Introduction

The work performed in Work package 7 (WP7) addresses “Safe and secure processing” of electronic health data within the EHDS infrastructure. The goal is to enable secure processing of EU citizen’s electronic health data while fostering a secure, interoperable, and efficient health data ecosystem. The output of this work package consists of guidelines and technical specifications that shall inform further decisions and technical frameworks to set up the EHDS.

The results of WP7 are distributed across five tasks. Task 7.1 provides guidance to users about their duties and responsibilities when analysing data in an SPE. Next, guidelines for data minimisation and de-identification give directions on how to address the challenges of health data minimisation, pseudonymisation, anonymisation and the generation of synthetic data (task 7.2 includes sub-tasks: 7.2.1, 7.2.2, 7.2.3 & 7.2.4). Specifications for the implementation of a common IT infrastructure (task 7.3) shall help member states to connect to the EHDS ecosystem. To ensure interoperability, common security requirements applicable to all SPEs are defined in addition to functional and technical services that should be part of all SPEs (task 7.4). Lastly, information about data linkage techniques and possibilities of quality control of linked data are collected (task 7.5).

Here is an overview of the documents that are part of WP7:

1. Guidelines for data users on how to use data in a secure processing environment (task 7.1);
2. Guidelines for Health Data Access Bodies on data minimisation, pseudonymisation, anonymisation and synthetic data (task 7.2);
3. Technical specifications for Health Data Access Bodies on the implementation of the common IT infrastructure (task 7.3);
4. Technical specifications for Health Data Access Bodies on the implementation of secure processing environments (task 7.4);
5. Guidelines for Health Data Access Bodies on linkage of health datasets (task 7.5).

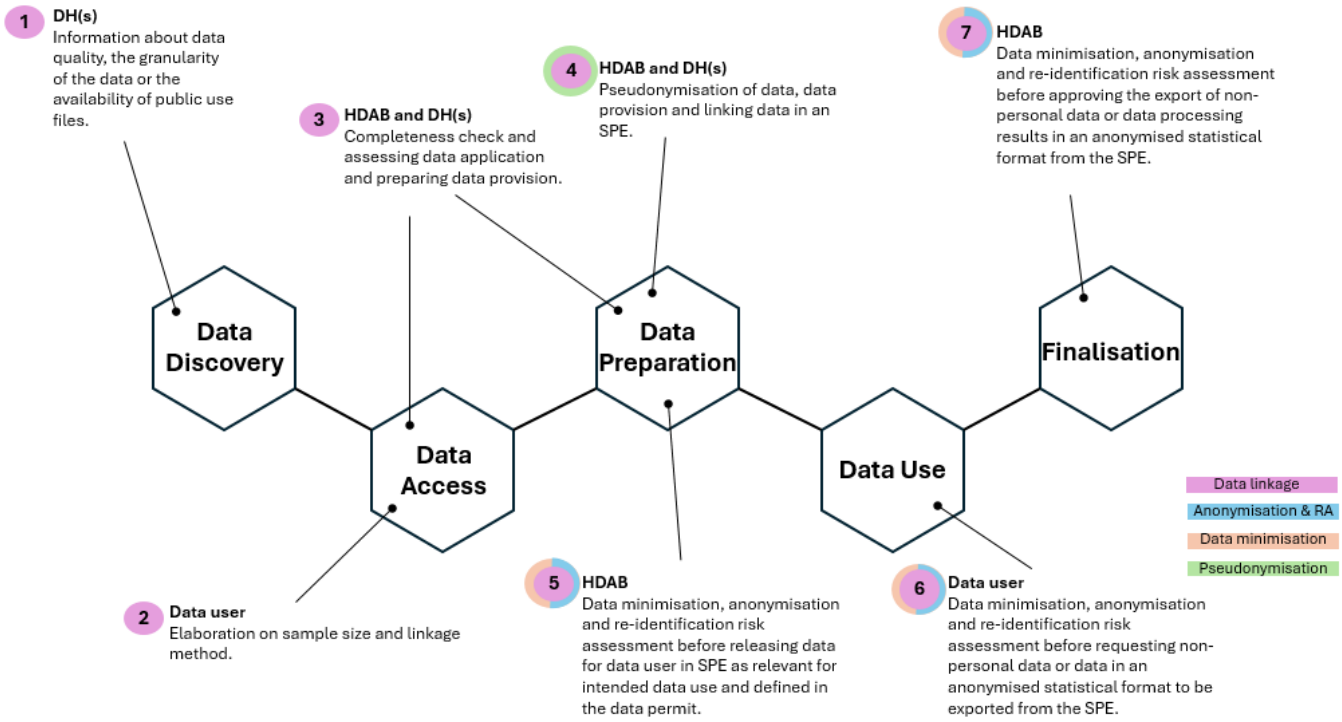
2.1.1 Task Overview – Data linkage (T7.5)

Linkage of personal data is governed by the provisions laid down in the General Data Protection Regulation (GDPR). Data linkage can be subsumed under the broader term data processing, as defined in Article 4(2) of the GDPR. Non-personal data that are outside of the scope of GDPR but could also be subject of data linkage.

To begin the work on this task, a survey was conducted, which served as orientation (see more information in Annex 1). This survey was conducted to receive a better impression of the interests and needs of stakeholders for data linkage topics under the EHDS across Member States but was not systematically analysed. In the following, some impressions from the survey were discussed and work started on the chapters of the guideline. The outline was finalised in close alignment with the EC (DG SANTE) and contents were added by the contributors and discussed in bi-weekly task meetings. Once a first full draft was finalised, every contributor was invited to provide comments. These comments were incorporated and then submitted as Milestone. After the public consultation of this draft, received comments were clustered and evaluated before they were incorporated. Comments from the public consultation also prompted adjustments to the document to

make it more understandable and usable for the relevant participants in data linking processes, which resulted in the final Deliverable (Annex 2).

Figure 1 User journey with colours indicating data processing aspects such as data minimisation, pseudonymisation, anonymisation, and data linkage.



Abbreviation: DH: Data holder; HDAB: Health Data Access Body; RA: re-identification risk assessment; SPE: secure processing environment.

Figure 1 gives an overview of the different stages along the user journey (see Annex 3 – User journey) where aspects of data linkage are relevant. When data users want to use linked data, the information available during the data discovery stage may inform about data quality, but also about specific information such as the granularity of the data or certain variable types. Then, to get access to the linked data, data users shall provide as detailed information as possible in their data application, irrespective of whether they submit a data request or apply for a data permit. When the application has been approved, the data preparation stage starts. To perform the data linkage, the data controllers and/or processors (section Why, when, and by whom are data linked?) need to know which data they should link (section What data are being linked?) and choose the appropriate method for data linkage, dependent on the data, the national law and the permit (section

How is data linked?). Moreover, data users and HDABs should be informed on data quality measures (section How is quality of linked datasets ensured?) and about actions taken to prevent misuse of linked datasets (section 6 Role-based access to linkage processes).

3 Scope

In scope:

- This deliverable focuses on linkage of records from two or more health datasets at the individual level, including cases where records are matched through direct or indirect identifiers under pseudonymised conditions, as per Article 66(1) and Article 68(11) of the EHDS regulation.
- The guidance provided in this deliverable applies to linkage performed as part of data applications access and data request applications under the EHDS Regulation.
- This deliverable also serves to support implementation under Article 67(2)(f) of the EHDS Regulation, assuming linkage is always carried out by authorised personnel under HDAB supervision.
- It describes which EHDS actors (e.g., linkage organisations) may be involved in performing data linkage, depending on the data sources and the national setup, including HDABs, health data holders, trusted health data holders (TDHs) and trusted third parties (TTPs).
- Moreover, considerations on linking methods, linkage quality including accuracy and completeness of linked datasets are described. Linkage quality is also considered in the context of overall dataset quality and its dimensions.
- It analyses the relationship between data quality and data linking quality, presents linkage quality parameter definitions and summarises methods for quantification of linkage accuracy and completeness, assessment of linkage quality and provides relevant recommendations for data holders and HDABs.
- The target audience include HDABs, data holders and TDHs, as they are the organisations that can perform data linkage within the EHDS ecosystem. Moreover, this guideline may also assist data users in understanding the linkage process and formulating appropriate requests or applications.
- This guideline aims at supporting HDABs supervising or performing data linkage within the EHDS.

Out-of-scope:

- Linking data across member states is not in the focus of this deliverable. Scenarios with multi-country applications are very relevant but should be addressed in future guidelines, as this topic involves other governance structures as well. Moreover, linking data across European data spaces as foreseen in Recital 80 of the EHDS regulation is out-of-scope for this guideline.
- Research infrastructures (e.g., the European Research Infrastructure Consortium [ERIC]) may be involved in data preparation or request formulation under national procedures. However, their operational roles in linkage are not addressed in detail in this guideline and may be clarified in future implementation guidance or discussions within the EHDS Board.

- Impact of data linkage on utility of the linked datasets. This deliverable merely introduces the issue of utility and the impacts of linking on it, without elaborating on the relevant operational details.

4 Data linkage

Data linkage refers to combining data from different data sources that relate to the same entity (e.g., individual, institution) to create a more comprehensive dataset (see Glossary).

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

In the EHDS, linkage of data can be requested for all secondary use purposes under Article 53 of the EHDS Regulation, although not all types of health data from all health data categories are intended for all secondary use purposes (please refer to D5.2 Guideline for Health Data Access Bodies on allowed purposes and prohibited secondary use according to EHDS Reg.).

Data linkage is performed to strengthen different dimensions of the dataset, ultimately to the benefit of patients and society through improved prevention, more effective and safer treatment, and better quality and equity of health services. As such, there are three main reasons to perform data linkage[1]:

- To enable longitudinal research
  - Study populations may be followed over time and across different health-related events to investigate disease aetiology and prognosis.
- To enrich data sources
  - Diverse data sources may be integrated to create rich datasets including both health data and data on factors impacting on health, such as socioeconomic, environmental and behavioural determinants of health.
- To generate population-level insights
  - Large sample sizes and hard-to-reach populations may be investigated, e.g., to inform policy makers, develop new drugs, evaluate interventions, enable health economics and outcomes research, assess inequalities, and study rare diseases and vulnerable populations.

Data linkage must be conducted before the dataset is made available to the health data user in an SPE for their processing in accordance with the data permit issued to them by the HDAB (Articles 61(1), 68(11), and 73(1)). This deliverable concerns linkage carried out in response to a data access application or data request. Some datasets are, however, linked on a standing basis as part of their maintenance, independently of any individual application; in such cases the linkage described here concerns only any further linkage required to produce the permitted dataset. The possibility for linking datasets held by data users themselves with data obtained via the EHDS remains subject to national implementation and technical safeguards. Such datasets, referred to in this deliverable as the data user's "own data", are datasets that the data user lawfully holds outside the EHDS

[1] Harron, K., Doidge, J. C., & Goldstein, H. (2020). Assessing data linkage quality in cohort studies. *Annals of Human Biology*, *47*(2), 218–226. https://doi.org/10.1080/03014460.2020.1742379

and that are described in the data access application or data request. Where the data user is also a health data holder for the datasets concerned, the linkage is still carried out by the linking organisation, and the separation between the roles of data holder and data user is maintained. This deliverable outlines data linkage scenarios (see also Annex 6 – Data linkage scenarios) involving datasets in the data user's possession while assuming linkage is always carried out by authorised personnel under HDAB supervision, and that data users do not perform data linkage directly and are not granted access to identifiers in the SPE. Clarification on exceptional scenarios may be considered in further implementing guidance. These exceptional scenarios concern only the technical and organisational arrangements for incorporating datasets that data users already possess, which are uploaded in accordance with the instructions provided by the HDAB and linked by the linking organisation; they do not permit data users to perform the linkage themselves or to access direct identifiers.



In the EHDS, data linkage is performed by a linking organisation, i.e., HDAB, data holder, or trusted data holder. The data user may propose a linkage plan in their data application. A trusted third party may support the linkage, for example by generating the pseudonyms used to match records, without itself becoming the linking

Health data users may propose a data linkage plan in their data access application or data request form (form sections 6 and 7, as described in D6.2 Guideline for data users on good application and access practice, sections 6.4.6 and 6.4.7). However, in these instances, the final linkage process is defined and carried out by the HDAB or TDH under HDAB supervision in an SPE (D6.2 Guideline for data users on good application and access practice, section 8.6). Where the proposed linkage plan is not feasible, for example because the required variables are unavailable or of insufficient quality, or because it would entail a disproportionate risk of re-identification, the linking organisation adapts the plan or informs the data user, who may amend or withdraw the application. In this context, HDAB supervision means that the HDAB approves the linkage as part of the data permit, determines which organisation performs it, and remains responsible for the linked dataset being made available to the data user in accordance with the permit.

The expected timeframes from data application to data access can be found in section 8 in D6.2 Guideline for data users on good application and access practice. For a detailed overview please inspect Figure 7 of D6.2. To shortly summarise: The HDAB has 3 months to issue a permit after receiving a complete data access application (Art 68(4)). The DH has 3 months to prepare the data (Art. 60(2)). The HDAB has 2 months to provide the data in the SPE (Art. 68(7)). If a trusted third party is involved in the process of data provision, the time allocated for their services falls within the timeframe of the HDAB. Please note, that extensions are possible and there is an accelerated procedure (Art 68(6)) for e.g., public sector bodies and Union institutions. In addition, shorter timeframes apply for TDH (Art. 72(5)) and for data requests, the response should be provided to the user within 3 months (Art. 69(4)). Hence, for the data acquisition and data linkage 3 months (extension of 3 months possible) and for the data provision, 2 months (possible extension) are allocated (see also Figure 2).

Which organisation that is responsible for the process of data linkage in any given scenario depends on where the permitted datasets are held. As such, there are different procedures

for data linkage involving different organisations, including HDABs, data holders, and TDHs (from here on forward referred to collectively as the “linking organisation"), as outlined in Figure 2 and Table 1. The three main data linkage processes involving these different linking organisations are outlined in the following text and in Figure 2 below. Please note that, for simplicity, only data linkage scenarios involving data access applications are outlined in this section (please refer to Annex 6.2 Data request for data linkage scenarios involving data requests) and that all scenarios outlined below and in Annex 6.1 Data access application assume that a data permit is granted.

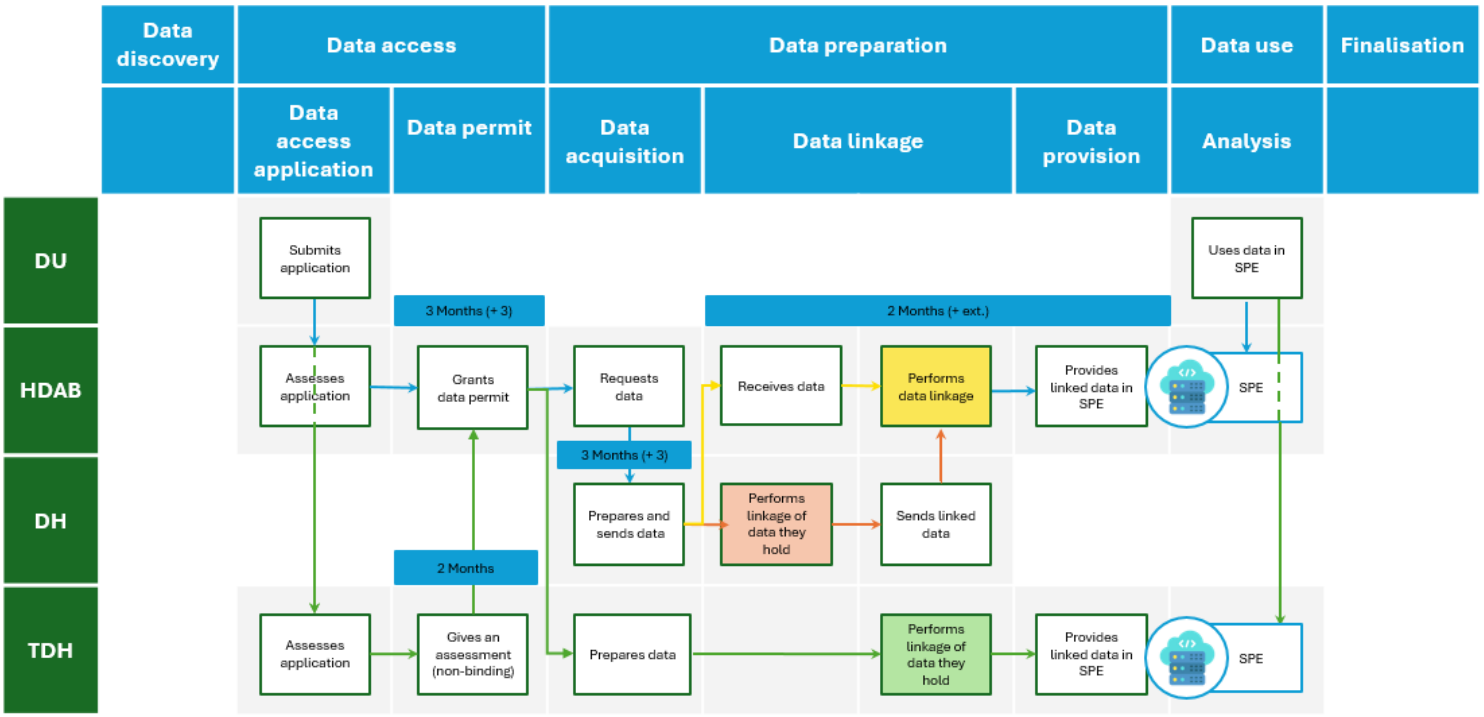
- **HDAB performing data linkage:**
    - Receipt of data application: The HDAB receives a data application submitted by the data applicant (Article 57(1)(a)).
    - Assessment of data application: The HDAB assesses the data application and provides a data permit to the data applicant (Article 57(1)(a)).
    - Request of data:
        - In scenarios where the permitted datasets are held by different (trusted) data holders, the HDAB requests secure delivery of the permitted datasets from the relevant (trusted) data holders (Articles 57(1)(a) and 68(7)).
        - In scenarios where a data user applies for access to data via the EHDS (held by one or multiple (trusted) data holders) and intends to link that data with datasets in their own possession (Article 67(2)(f)), the HDAB should provide them with instructions on how to ensure secure data transfer (see Annex 6 and D6.2 Guideline for data users on good application and access practice, section 6.4.7)
    - Receipt of data: The HDAB receives the datasets from the (trusted) data holder(s) and potentially the data user (Article 57(1)(b)).
    - Data linkage: The HDAB performs data linkage (Article 57(1)(b); see also M6.2 Draft guideline for data users on good application practice for data access and requests, section 8.6.
    - Data provision: The linked dataset is made available to the data user in the SPE in accordance with the data permit issued by the HDAB (Article 68(7) and Article 73(2)).
- **Data holder performing data linkage, under HDAB supervision**:
    - Receipt of data request: In scenarios where all permitted datasets are held by one single data holder, the data holder may receive a request for a linked dataset from the HDAB.
    - Data linkage: The data holder performs the requested data linkage of the datasets they hold.

- Data transfer: The data holder transfers the linked dataset to the HDAB (Article 60(1) and (2)) to be made available for the data user in the SPE in accordance with the data permit issued the HDAB (Article 68(7) and Article 73(2)).

- **Trusted health data holder performing data linkage, under HDAB supervision:**
  - Receipt of data application: In scenarios where datasets applied for are held by a TDH, the TDH receives the data application from the HDAB to which is has been submitted by the data applicant (Articles 67(2)(f) and 72; see also D6.2 Guideline for data users on good application and access practice, section8.3).
  - Assessment of data application: The TDH provides their assessment of the data application to the HDAB for further evaluation (Article 72).
  - Receipt of data request: The TDH receives a request for provision of a linked dataset from the HDAB (Article 72(6)).
  - Receipt of data: In data linkage scenarios where data user has been permitted linkage of data held by the TDH with data in their own possession, the TDH receives the dataset from the data user (Article 57(1)(b) and Article 72(6)).
  - Data linkage: The TDH performs the requested data linkage of the datasets (Article 57(1)(b)).
  - Data provision: The linked dataset is made available to the data user in the SPE provided by the TDH (Articles 57(1)(a)(i) and 72(6)).

- **Other entities and data linkage**
  - The EHDS Regulation recognises the utility of data linkage for public sector bodies where they obtain data in the circumstances referred to in Article 15, point (a) or (b), of Regulation (EU) 2023/2854 (exceptional need to use data), and enables the HDAB to provide technical support to process those data or combining them with other data for joint analysis (Article 57 (4), EHDS Reg.).
  - Member states may in their national law establish health data intermediation entities to carry out the duties of certain categories of health data holders (Article 50(3) EHDS Reg.) to reduce the administrative burden on some data holders. It is logical that data linkage of datasets held by one data holder may also be entrusted to the relevant health data intermediation entity as a kind of data processing. An intermediation entity could also be a trusted third party. The details of such scenarios are not elaborated on in this deliverable.

Figure 2. Procedures for data linkage in the HealthData@EU with indication of the entity responsible for data linkage and relating to the European Health Data Space user journey.



The colours yellow (HDAB), orange (data holder), and green (trusted data holder) indicate the organisation performing data linkage. *Abbreviations:* DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Table 1. Summary of the linkage scenarios, their linking organisations and references to the EHDS Regulation.

Linkage scenario	Who performs data linkage?	EHDS Regulation	Data linkage scenario (see Annex 6.1)	Comment
Linking data from multiple (trusted) data holders	HDAB	Article 57(1)(a) and (b) Article 68(7) Article 73(2)	1.1	
Linking data from one or multiple (trusted) data holders with data provided by data user	HDAB	Article 57(1)(a) and (b) Article 67(2)(f) Article 68(7) Article 73(2)	1.2	

Linking data from one data holder	DH	Article 60(1) and (2) Article 68(7) Article 73(2)	1.3	
Linking data from one trusted data holder	TDH	Article 72	1.4	
Linking data from one trusted data holder with data provided by data user	TDH	Article 67(2)(f) Article 72	1.5	
Linking data between the EHDS and other data spaces	HDAB	Recital 80 Article 75(9) and (11)		Detailed guidance out of scope for the present guideline

Abbreviations: EHDS: European Health Data Space; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder

4.2 What data are being linked?

The EHDS Reg. Article 51 'Minimum categories of electronic health data for secondary use' specifies categories of electronic health data that health data holders shall make available for secondary use. These datasets may be linked by HDABs, DHs or TDHs, in line with the Articles 57, 60, 68 and 69. These data categories are presented in Table 2. For data categories f, g, i and q, member states may introduce stricter measures and additional safeguards at national level. These measures or safeguards should be clearly communicated and documented in the metadata. For example, the dcat::qualifiedRelation property could be helpful here.

The scope of the health data for linkage comprises datasets in the metadata catalogue. Datasets contain personal data, anonymised, aggregated data, administrative data, various contextual data (e.g., location, time periods, weather, outbreak of the epidemic), medicinal product- and medical device- related data. Datasets provided by health data users themselves may also be linked in an SPE (Article 67(2)(f), if appropriate safeguards are in place. Such cases require approval by the relevant HDAB.

To clarify the content of the various health data categories in terms of their sensitivity and potential linkage strategy, data categories have been divided into 1) individual-/ patient-level data and 2) non-individual- / aggregate-level data (Table 2). Some data categories can include both individual- and non-individual-level data. Linkage strategies differ accordingly. Individual-/ patient-level data can be linked at the level of an individual person. Non-individual-level data can be linked across datasets sharing a common contextual variable, for example, a geographic area/ region code, Anatomical Therapeutic Chemical Classification (ATC) code, hospital/ health care centre ID, or time period. Non-individual-level data can also be used without linking, for example, as background information.

To enable high-quality data linkage, variable(s) used for linkage should be available in standardised formats in all datasets that are being linked. For example, if data is linked using personal identity codes, these codes should be recorded in the same format in each dataset. Please note, that the same logic should be applied for pseudonyms. Variables

used for linkage should be recorded with high-quality and -coverage to support successful linkage. Feasibility and quality of linkage depend on the availability of common identifiers or variables that allow records to be matched reliably as well as their completeness and representativeness. Datasets may require mapping, transformation, and harmonisation before linkage can be performed effectively.

Comprehensive metadata is important for ensuring linkage quality. Metadata should include clear and consistent variable definitions, data source and collection context, provenance, coding standards used (e.g., ICD-10, ATC codes), temporal and geographic coverage and representativeness, geographical granularity, known limitations or biases, and quality information to support robust and traceable linkage (see D5.1 Guideline for data holders on data description). Metadata using common standards is especially important for understanding the quality, availability, and characteristics, including linkability, of linking variables.

When possible, the use of common data model mapping and international standards, such as Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) or ISO standards (e.g., ISO 13606), can facilitate high-quality data linkage and use. Consistent coding systems, terminologies, and metadata standards are not only needed for successful data linkage, but they also support meaningful, reliable and interoperable secondary use of data. In addition to high-quality technical linkage, retaining clinical meaningfulness, ensuring semantic harmonisation and consistency of the data as well as understanding the context of data collection are essential for successful data use. It should be noted that even data with standard coding systems (such as ICD-10) may hold a risk of bias if the purpose and / or method for data collection or reporting differ between datasets. In addition, there can be interoperability challenges arising from differences in coding systems, data models, and care settings.

The type of data being linked can have impact on linkage activities and considerations. Examples of linkage scenarios that may require special consideration include (please, note, the list of examples or considerations are not exhaustive):

- Imaging data: Linkage could be done using metadata, e.g., personal identity codes that identify individuals.
- Device-generated data: The need for pre-processing and cleaning as well as potential intellectual property rights should be assessed carefully before linkage can be performed.
- Genetic and genomic data: Privacy and re-identification risk considerations are especially relevant when linking genetic and genomic data.
- Non-health sector data: The availability and comparativeness of linking variables should be assessed properly before linkage can be performed. In addition, data permit considerations are needed (i.e., which organisation is responsible of granting the permit and how information is communicated between different organisations).

For more examples, see Annex 9 which presents three different practical use cases of data linkage.

Both individual-level and aggregate-level data can be linked. In both cases, variables used for linkage should be available in standardised format to ensure high-quality data linkage.

In addition to data categories presented in Table 2, member states may also provide additional data categories available for secondary use as stated in their national law (EHDS Reg., Article 51).

Table 2. Minimum categories of electronic health data for secondary use (EHDS Regulation, Article 51) [1] and their division into individual- / patient-level and non-individual- / aggregate-level data [2].

Data categories		**Individual/ patient level data**	**Non-individual level/ aggregate level data**
a)	electronic health data from EHRs;	x	
b)	data on factors impacting on health, including socio-economic, environmental and behavioural determinants of health;	x	x
c)	aggregated data on healthcare needs, resources allocated to healthcare, the provision of and access to healthcare, healthcare expenditure and financing;		x
d)	data on pathogens that impact human health;	x	x
e)	healthcare-related administrative data, including on dispensations, reimbursement claims and reimbursements;	x	x
f)	human genetic, epigenomic and genomic data;	x	x
g)	other human molecular data such as proteomic, transcriptomic, metabolomic, lipidomic and other omic data;	x	x
h)	personal electronic health data automatically generated through medical devices;	x	
i)	data from wellness applications;	x	x
j)	data on professional status, and on the specialisation and institution of health professionals involved in the treatment of a natural person;	x	

k)	data from population-based health data registries such as public health registries;	x	x
l)	data from medical registries and mortality registries;	x	x
m)	data from clinical trials, clinical studies, clinical investigations and performance studies subject to Regulation (EU) No 536/2014, Regulation (EU) 2024/1938 of the European Parliament and of the Council34, Regulation (EU) 2017/745 and Regulation (EU) 2017/746;	x	
n)	other health data from medical devices;	x	
o)	data from registries for medicinal products and medical devices;	x	x
p)	data from research cohorts, questionnaires and surveys related to health, after the first publication of the related results;	x	
q)	health data from biobanks and associated databases.	x	x

[1] Minimum categories included in the table reflect the categories in the EHDS Regulation, Article 51.
[2] Division shows the most common way of using data. All individual-level data can be aggregated. Division may not be fully applicable to all data categories in all member states.

Data used for linkage are described in the data access application (see D6.3, Annex 4, 6.1.5 Description of the data needed). The applicant must describe in the application how data from different sources will be linked. The applicant must also describe in the application whether other data will be combined and, if yes, how the linkage will be conducted (see D6.3, Annex, 6.1.6. Other data to be combined).

Data linkage refers to combining data from different sources that relate to the same entity (e.g., individual, institution) to create a more comprehensive dataset (see Glossary). When expanding a study population by amending the data extraction criteria and by doing this, combining data from different entities using the same variables, data linkage activities are required to identify whether newly retrieved records correspond to entities already included, thereby avoiding duplication and misclassification. Data linkage is often a method used to achieve data enrichment. For more information concerning data enrichment, see M5.4 Draft guideline for data enrichment (or the final deliverable, which is currently not published yet).

Linking several different datasets can increase the risk of re-identification. For more information on data protection, re-identification risk, data minimisation and variables with privacy risk, see D7.2 Guideline for health data access bodies on data minimisation, pseudonymisation, anonymisation and synthetic data.

4.2.1 Electronic health data protected by intellectual property rights or trade secrets

Table 2 does not elaborate on data containing information or content protected by intellectual property rights or trade secrets. In accordance with Article 62(2) and Recital 60 of the EHDS Regulation, linkage of datasets that may contain content protected by intellectual property rights or trade secrets (e.g., clinical trial data, product supply data) must be assessed on a case-by-case basis. In such cases, appropriate legal, organisational and technical measures must be put in place to ensure that linkage does not result in disclosure or reverse engineering of protected information. The organisation performing linkage requires information about the intellectual property rights or trade secrets related to the data under linkage, e.g., from the HDAB. For more detailed discussion of intellectual property rights and trade secrets, see D5.2 Guideline for Health Data Access Bodies on minimum categories and limitations on the reuse of health data (Section 9). In addition, FORTIFY consortium project will develop legal, governance and technical solutions for enabling secondary use of health data while protecting intellectual property rights and trade secrets (for more information, see Home - FORTIFY).

4.3 How is data linked?

Considering various methods of data linking described in section 4.3.1, there are two broad categories of linkage methods: deterministic (rule-based) methods and probabilistic (score-based) methods, including techniques that make use of advanced machine learning algorithms[2]. These approaches have similar implications for linkage error and analysis. The aim of each approach is to classify record pairs according to their true match status (whether they belong to the same person or entity), and they achieve this principally using information derived from matching identifiers (variables) that records have in common, such as names and dates of birth. The term ‘match’ refers to the true relationship between a pair of records, term ‘link’ refers to their derived or assumed classification, and ‘agreement’ to describe their similarity in terms of matching variables. Linkage variables must be variables that are common to the datasets being linked.

These definitions are relevant for understanding the mechanics of linkage algorithms, but in operational terms, HDABs and (trusted) data holders should document linkage procedures, thresholds, and error rates to ensure accountability and reproducibility.

4.3.1 Which methods can be used?

Data linkage can be performed with one or multiple methods, including:

- **Direct linkage**: Matches records using exact identifiers (i.e., linkage variables), effective when high-quality unique identifiers are available.
- **Deterministic indirect linkage**: Matches records using indirect identifiers, e.g., date of birth, sex, admission date, city, postal code and hospital ID. Uses stepwise decision rules, including full agreement across fields, partial agreement across a limited number of fields, etc. and expert arbitration for duplicates.
- **Probabilistic linkage**: Probabilistic record linkage is a method used to link together records that lack unique identifiers. In the absence of a unique identifier, it is possible to

[2] Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

use a combination of individually non-unique variables such as name, gender and date of birth to identify individuals. Probabilistic linkage assigns a match score based on field agreements, using:

- Weighting: Assigning importance to fields based on uniqueness.
- Probabilistic models: Methods like the Fellegi-Sunter model classify records as matches, non-matches, or potential matches. Advanced machine learning algorithms may be used e.g., to estimate the statistical parameters for probabilistic record linkage without human review, when human intervention is not possible or practical[3,4]. It is to note that where machine learning algorithms are used to support linkage, their configuration should be documented, explainable, and auditable by the HDAB or the (trusted) data holder responsible for linkage, and in line with Article 67(2)(f), where relevant.



Highlighting linking claims data and data from cancer registries in Germany, while elaborating on challenges and opportunities when linking data using non-deterministic linkage methods. Please look at Annex 9 – Use cases.

Cross-border record linkage (i.e., linking datasets from different countries) may be challenging, as it requires that study subjects are registered in databases across different countries. In practice, many national databases are limited to a single country's population. In addition, it may be difficult to figure out if entries in both datasets relate to the same person. However, cross-border federated analyses based on different populations using the same variables may be beneficial in various cases. Recital 80 of the EHDS Regulation mentions cross-border linkage, but the EHDS Regulation does not provide a harmonised mechanism to perform it. However, such linkage may be considered in specific bilateral or multilateral contexts, provided that legal and interoperability conditions are met. Cross-border use cases might more often rely on data combination processes, such as concatenation, than data linkage.

Note: Although there is no obligatory harmonisation of data that should be implemented by data holders or one common standard for data formats and semantics required by the EHDS Regulation in the context of secondary use of health data, it is clear that the use of common standards have positive impact even on the ability to perform cross-border data linkage. Efforts to adopt EEHRxF for primary use of health data are moving in this direction. However, this relates to only limited number of health data categories. Other initiatives that aim to create data standardisation framework on the international level are therefore welcome. An example of such initiative is the Data standardisation strategy developed by the European Medicines Agency (EMA), which aims, among other things, to enable quicker uptake of international data standards across the EU with an explicit intention to enable

[3] Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Annu Symp Proc. 2003;2003:259-63. PMID: 14728174; PMCID: PMC1479910.

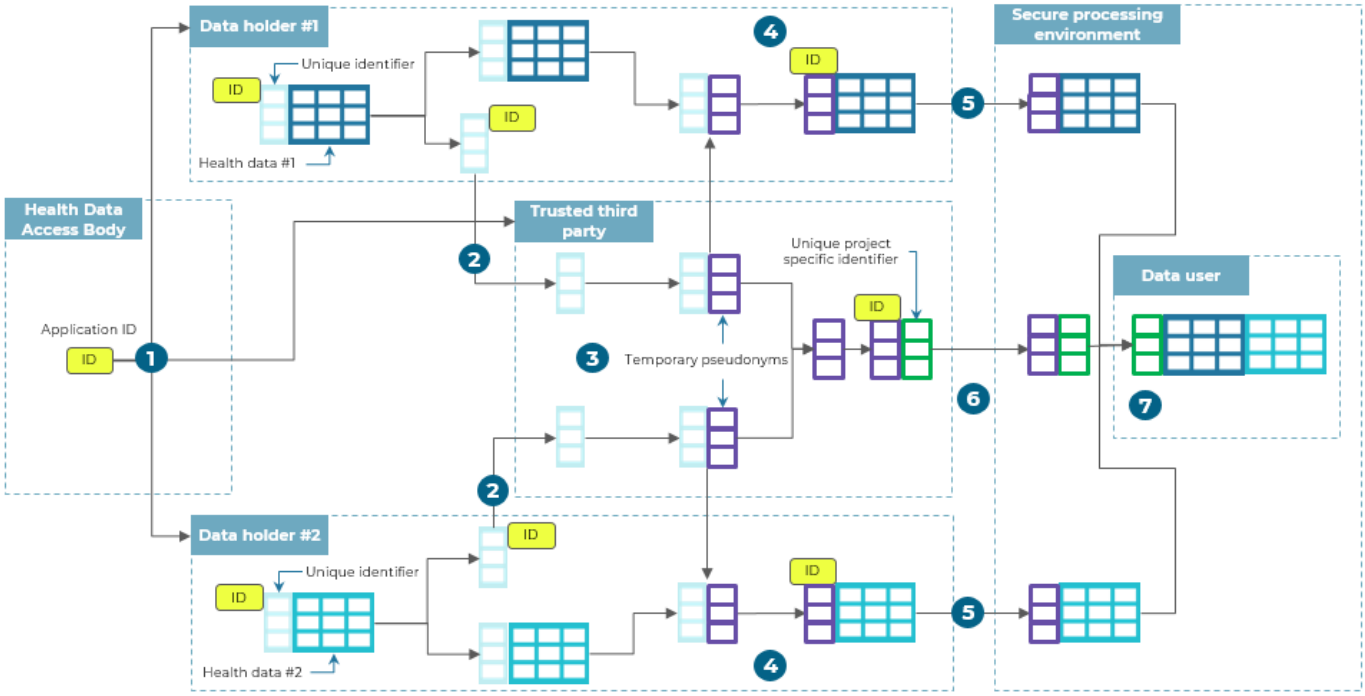
[4] Mayer A, Stockdale M. Developing standard tools for data linkage: February 2021. Office for National Statistics. Available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/developingstandardtoolsfordatalinkagefebruary2021

data linkage. Similarly, European Rare Disease Registry Infrastructure (ERDRI) aims at the harmonisation of used data elements and data structures in existing and upcoming rare disease registries. For this purpose, a set of common data elements for rare diseases registration has been developed.

Safeguards can be introduced by, for example, performing privacy preserving record linkage (PPRL) methods to ensure that record linkage is performed without revealing personally identifiable information[5]. In the context of the EHDS, a TTP may perform the pseudonymisation process before datasets being sent by data holders to the SPE (see Figure 3). Please find an overview of possible linkage scenarios using TTPs in Annex 11 – Scenarios data linkage with TTP. Note: this overview is not exhaustive. Linkage can be carried out through cryptographic techniques such as secure multi-party computation (SMPC)[6]or hashing-based methods to match records without exposing personal identifiers. Multi-party computation without the use of a TTP is currently being investigated within the Belgian public sector, for example[7].

[5] https://doi.org/10.1016/j.is.2012.11.005
[6] Laud, P., Pankova, A. Privacy-preserving record linkage in large databases using secure multiparty computation. BMC Med Genomics 11 (Suppl 4), 84 (2018). https://doi.org/10.1186/s12920-018-0400-8
[7] Verslype, K., & De Decker, B. (2025). Privacy-By-Design in the Belgian Public Sector: Pseudonymising and Joining Personal Data Fragmented Over Multiple Organizations. In *Public Governance and Emerging Technologies: Values, Trust, and Regulatory Compliance* (pp. 103-129). Cham: Springer Nature Switzerland. https://library.oapen.org/bitstream/handle/20.500.12657/100827/9783031847486.pdf?sequence=1#page=105

Figure 3. Example of a privacy preserving record linkage process through a trusted third party. 1: HDAB sends application ID to both data holders and TTP. 2: Data holders transfer strictly necessary identifiers to the TTP (e.g. social security number) with added application ID. 3: TTP generates temporary pseudonyms through identical generation process and sends them to the data holders, restricted to common patients, with the application ID. 4: Data holders replace unique identifiers by the temporary pseudonyms in their medical data. 5: Data holders send data with temporary pseudonyms and application ID to the SPE. 6: TTP generates unique project specific identifier and sends it to the SPE. 7: Pseudonymised data are linked inside the SPE by the linkage organisation and temporary pseudonyms are replaced by the unique project specific identifier before giving access to the data user.



It is worth noting that several tools are available facilitating PPRL by generating pseudonyms for patients and allowing linking patients' data across different data holders without revealing their identities:

- The European Patient Identity Management (EUPID) system is a privacy-preserving identity management service designed by the Austrian Institute of Technology. Regarding the process, first, personal identifiers (name, date of birth, sex) are transformed into cryptographic pseudonyms by the data holder and transmitted securely to the EUPID service. Then, the system determines whether the individual already exists in its identity management system or not. If so, it links the record to the existing identity, otherwise, a new identity is created. EUPID finally generates context-specific pseudonyms that are unique to each data-sharing context but can be translated through the central identity management system when authorised.
- The European Rare Disease Registry Infrastructure (ERDRI) provides a quite similar pseudonymisation tool called SPIDER (Secure Privacy-preserving Identity management in Distributed Environments for Research). However, as opposed to EUPID, the pseudonym in the identity management system (SPIDER ID), once established, is shared across different registries.

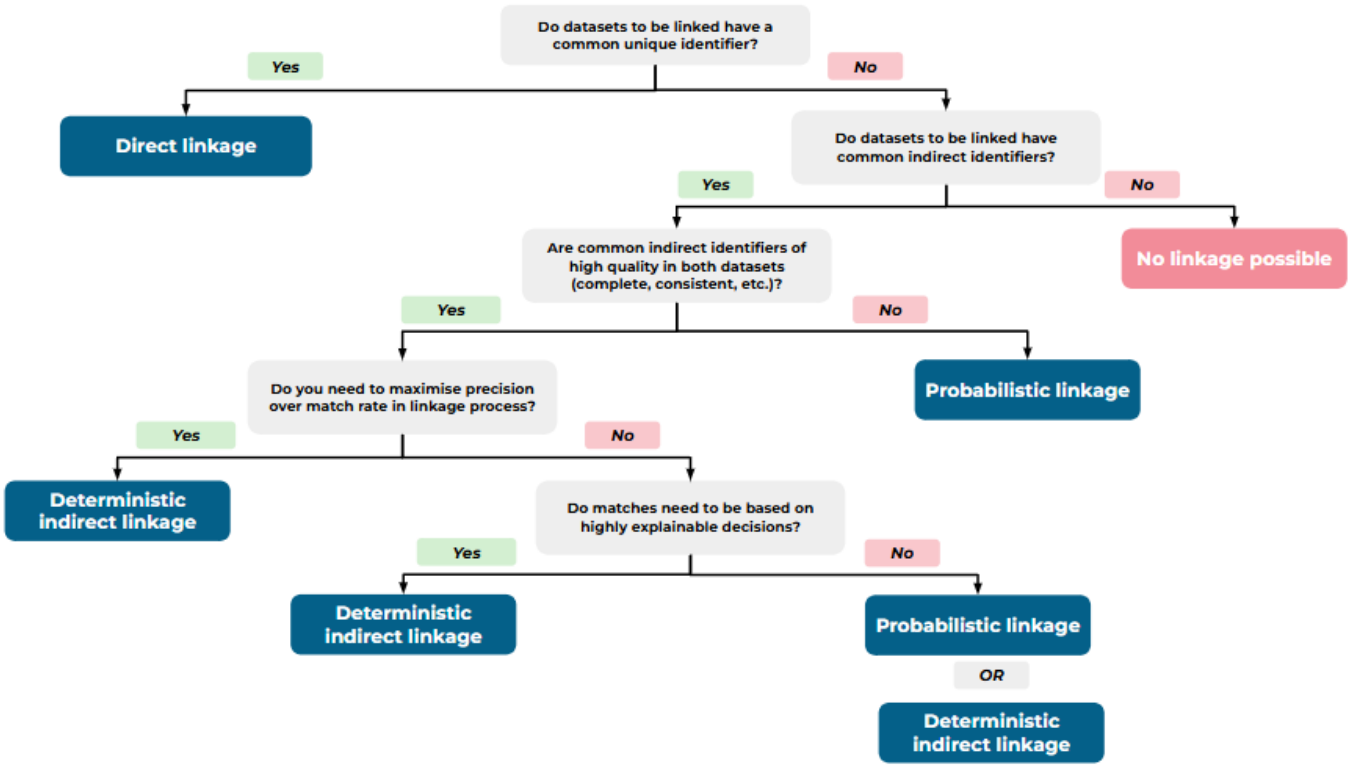
Please note that difficulties may arise using these tools when the linkage information, typically the unique identifiers, are already pseudonymised.

4.3.2 Which method should be preferred?

In some cases, data users may propose a preferred linkage method, especially when they have knowledge of the datasets. However, the final decision on the method rests with the HDAB or the entity that performs the linkage (i.e., the linking organisation), who must ensure that linkage is performed in compliance with legal and technical safeguards as part of their monitoring and supervisory tasks (Article 63(1)).

If feasible, direct linkage should generally be preferred because it can provide highly accurate and reproducible results when high-quality, consistent identifiers are available in both datasets. However, the performance of any linkage method depends on the quality and completeness of the identifiers and data available. Where identifiers are incomplete, inconsistent or unavailable, deterministic indirect or probabilistic linkage methods may provide more appropriate results. Figure 4 provides a decision tree.

Figure 4. Decision tree for choosing linkage method to apply.



Different linkage methods may be combined if relevant. For example, for the same dataset, direct linkage can be used for individuals with unique identifiers and deterministic indirect linkage for individuals without unique identifiers.

In some cases, deterministic indirect linkage and/or probabilistic linkage may be applied not only when unique identifiers are unavailable, but also as part of a two-step linkage strategy based on an *a priori* unique identifiers. For example, when first name, last name, and date of birth are treated as unique identifiers, the process may first (1) search for exact matches through direct linkage and, if none are found, (2) search for potential matches through deterministic indirect or probabilistic linkage to account for typographical errors or other discrepancies.

PPRL should be preferred where technically feasible, as it complies with GDPR principle of data minimisation (PPRL allows linkage without exposing the raw identifiers) and European Data Protection Board guidelines regarding pseudonymisation (PPRL ensures that the "pseudonymisation secrets" (e.g., cryptographic keys or salts) used to generate the pseudonyms are kept separate and secure).

4.3.3 Which are the requirements for each method?

If individual-level record linkage is needed, datasets to be linked must contain records referring to the same individuals, identifiable through shared linkage variables.
Direct linkage requires common unique identifiers in each dataset. These identifiers should ideally meet the following criteria:

- Uniqueness: Each identifier should refer to a single individual (or other unique entity) across datasets.
- Exclusivity: No individual (or other unique entities) should be associated with more than one identifier.
- Completeness: Identifiers should be available (i.e., not missing) for all relevant records in each dataset.
- Identical encoding: The unique identifiers used should be encoded the same way in each database (or standardised if not and that standard is known to the linking organisation).

Deterministic indirect linkage and probabilistic linkage require common fields used for matching to meet the following criteria:

- Comparability: Information has to be standardised and coded the same way or in formats that can be directly compared across datasets.
- Usability: Variables should be clearly identifiable and sufficiently reliable in each dataset.
- Discriminative power: The combined linkage variables should distinguish individuals effectively within and across datasets.

4.3.4 How should data be prepared before linkage?



When applicable, direct linkage should be preferred over other linkage methods, as it relies on exact matches and generally yields better results.

Where data are to be pseudonymised before being made available for secondary use, the linkage keys must be constructed in a way that ensures both data protection and technical linkability. One method is to generate project-specific pseudocodes using cryptographic hashing, while ensuring that the same method and salt are applied consistently across

datasets to be linked. Keyed hash functions based on methods such as HMAC-SHA256, A256GCM, RSA-OAEP-256, etc. can be used as they comply with security standards and produce standardised 256-bit pseudonyms, making it easier to verify that the pseudocodes have been effectively generated. For each project, specifically generated pseudocodes should be used for identifiers to avoid unauthorised linkage between datasets of different projects. The EUPID tool could be used to simplify this linkage process and reduce the burden on HDABs as it allows to link datasets from different sources without the need for a pre-established common pseudonym.

Deterministic indirect linkage and probabilistic linkage require data standardisation: Ensure consistent coding across datasets to be linked, for example, by harmonising data using a common data model.

4.3.5 How to perform data linkage?

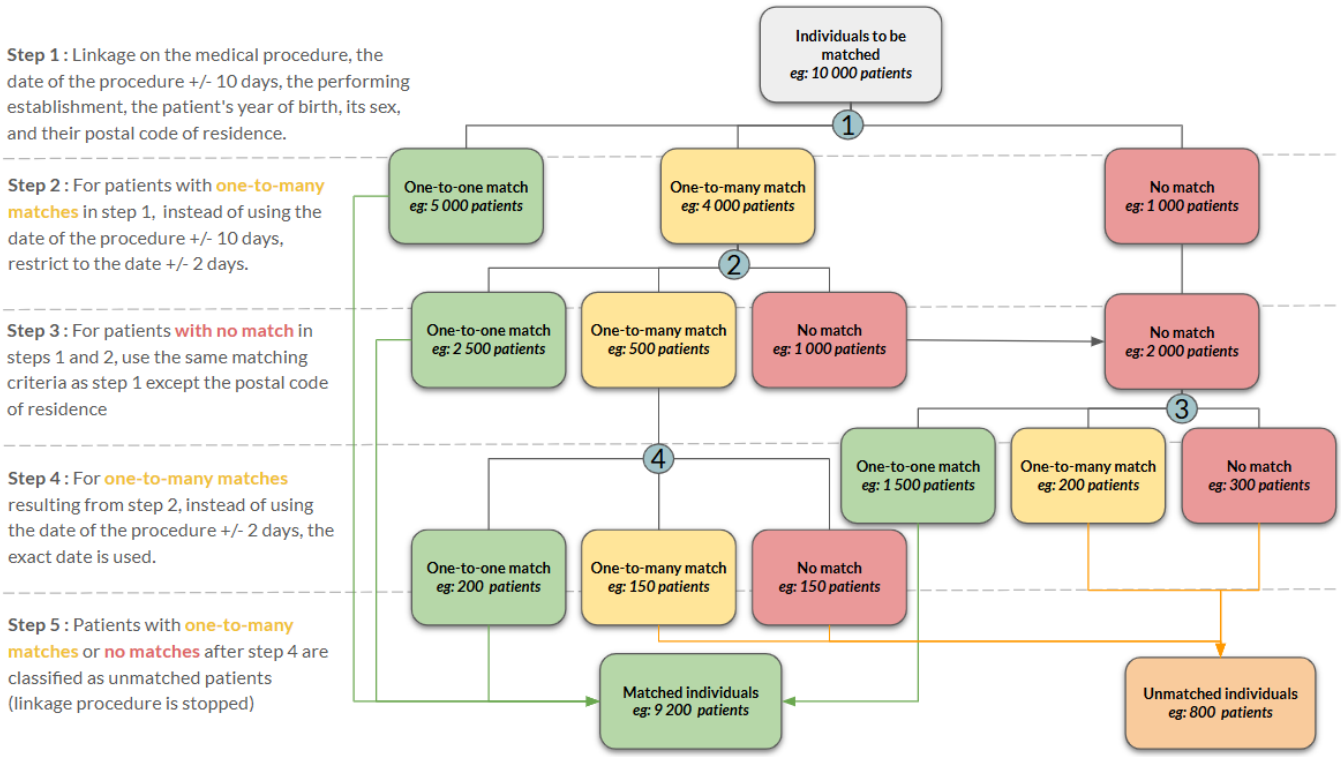
For direct linkage, perform a join on the unique identifiers (beforehand pseudonymised through the same method) across datasets. This links individuals with same unique identifiers across all datasets. The join method (inner join, full join, etc.) determines if individuals that are not part of all to be linked datasets will be included in the study cohort. Typically, only data for individuals present in the intersection of all datasets is made available, though this is ultimately determined by the defined target group.

For deterministic indirect linkage, execute multiple merges using different combinations of matching variables. The outcomes at each step may include:

- One-to-one match: One record in dataset A matches one in dataset B
- One-to-many match: One record in A matches several in B (insufficiently specific variables)
- Many-to-one match: Several records in A match one in B (e.g., multiple entries for the same person)
- No match: No corresponding record found due to missing or non-aligned variables

In deterministic indirect linkage, a decision tree can be used to guide and show the matching process, as illustrated in Figure 5.

Figure 5. Illustrative example of a decision tree supporting deterministic indirect linkage. Such tools help document and visualise rule-based linkage procedures in a structured, auditable format. The logic and outcomes of the decision tree may be shared with the data user as part of the data permit documentation or via a data description annex (see Annex 5 – Data description template). This contributes to transparency and allows users to assess the reliability and limitations of the linkage process.



For probabilistic linkage, apply matching models using the specified variables. The result will include matched individuals with associated plausibility scores.

It is recommended that the HDAB or data holder inform the applicant of the linkage rules applied, particularly when a decision tree or similar logic is used. Where appropriate, they should also engage in dialogue with the user to clarify the implications of the linkage logic for the intended analysis.

4.3.6 How to validate data linkage?

Validating data linkage involves evaluating the quality of the overall linkage process. This evaluation should focus on characteristics critical to data users, typically including match rate (the proportion of records that were linked), sensitivity (the proportion of true links that were

detected), or precision (the proportion of detected links that were true). Section 4.4.1 provides more detail about linkage quality parameters.

This guideline presents several methods and techniques for assessment of linkage quality and its statistical presentation. However, it is the responsibility of each linking organisation to choose and offer methods that align with their operational capabilities and with the specific requirements of the data linkage task. Professional data linkage offer should include at least essential parts of the validation as described in this section as a baseline.

- Assessing of match quality – linkage accuracy and completeness. Match quality refers to how accurately the linkage process classifies records among the datasets according to their true relationship (e.g., same individual vs. different individuals). This assessment is essential to ensure the fit-for-purpose of the linked dataset (see also section 4.4.1). HDABs should evaluate:
    - Match rates: Evaluate linkage – matched (true match – false match) and not matched (true non-match - missed match), and duplicate rates (see also section 4.4.1 which provides formulas for statistical presentation of the assessments hat can be used if for more demanding projects, where the data user is not satisfied with the mere number of matches and is interested in statistical parameters that consider, for example, the range of linked datasets). Section 4.4 also elaborates on the relationship between quality of data and linkage quality.
    - Plausibility scores: Review thresholds and edge cases (applies for: deterministic indirect linkage and probabilistic linkage).
    - Agreement on common fields: Compare overlapping field values.

- Perform manual validation:
    - Sampling: Manually review a subset of matched pairs.
    - Expert review: Involve domain experts for complex cases.
    - Manual review should be documented, and – where possible – anonymised examples retained as part of the validation trail should be provided.

- Perform data consistency checks:
    - Temporal consistency: Ensure chronological logic (e.g., birth date precedes registration date).
    - Field consistency: Ensure stable values across linked records.
    - No illogical duplicates: Prevent conflicts within linked records.
    - Post-merging data integrity: To maintain data integrity after merging datasets via data linkage correspondence, implement checks to prevent duplicate rows resulting from 1:many or many:many relationships and to avoid the generation of null or empty fields.

- Check for overlinkage or underlinkage:
    - Overlinkage (false positives/ false match): One record is incorrectly linked to multiple distinct individuals.
    - Underlinkage (false negatives/ missed match): Matching records belonging to the same individual are not linked.

- Validate unmatched records: Analyse missing identifiers or unique field values.

- Perform statistical validation: Compare demographics (e.g., age, sex, geolocation) of linked datasets.

- Cross-check across linkage methods: If different linkage methods are combined, compare linkage results to resolve discrepancies. For example, if we have a dataset composed of individuals with incomplete fields on the unique identifier. We may first perform direct linkage for those with a unique identifier. Then we may perform deterministic indirect linkage or probabilistic linkage for all individuals (with and without a unique identifier). For those with a unique identifier – already matched previously during direct linkage – we can then compare if the direct linkage (golden standard) yields the same match as deterministic indirect linkage or probabilistic linkage method and evaluate the correctness of these two latter methods.



Especially for indirect and probabilistic linkage, the HDAB should carry out several validation steps to ensure quality of the linkage process. Among the methods mentioned above, assessing match quality, performing data consistency checks and checking for overlinkage or underlinkage should be mandatory.

4.4 How is quality of linked datasets ensured?

Linked datasets are an important resource for data users, but linkage error can lead to biased, worthless[8] or incorrect results. To illustrating linkage mechanisms, the datasets are assumed to be internally consistent and to contain at least some linkage identifiers. Section 4.3.6 discuses match quality, which represents several dimensions including accuracy and completeness of linkage and more details also in statistical terms are provided in section 4.4.1 below. The resulting quality and usability of data created by a particular data linking method depends on:

- The quality of original datasets with records that are to be linked;
- The standardisation of direct or indirect identifiers that are used for the linkage;
- The due diligence in carrying out the linkage.

Depending on the reasons for linkage (see section 4.1), there are various implications of missed matches and false matches[9].

In general, if the purpose for linkage is to define a project or study population, linkage error can lead to erroneous exclusion or inclusion from the study population (i.e., through missed or false matches where linkage provides information on inclusion/exclusion criteria).

[8] Moore CL, Amin J, Gidding HF, Law MG (2014) A New Method for Assessing How Sensitivity and Specificity of Linkage Studies Affects Estimation. PLoS ONE 9(7): e103690. https://doi.org/10.1371/journal.pone.0103690

[9] Katie L Harron, James C Doidge, Hannah E Knight, Ruth E Gilbert, Harvey Goldstein, David A Cromwell, Jan H van der Meulen, A guide to evaluating linkage quality for the analysis of linked data, *International Journal of Epidemiology*, Volume 46, Issue 5, October 2017, Pages 1699–1710, https://doi.org/10.1093/ije/dyx177

If the purpose of linkage is to define a variable of interest, or to provide information on additional variables, false matches can lead to misclassification or measurement error in any of the variables captured through linkage (i.e., if the wrong records are linked together).

Linkage error can also result in double-counting (when one individual's records are counted multiple times, due to missed matches), or undercounting (when records for multiple individuals are counted as one, due to false matches)[10].

As outlined above, quality of the linked data is influenced by the quality of the original dataset, the quality of the metadata in the data catalogue, the methodology used for linkage, the availability and types of linkage identifier (variables) in the datasets. Types of identifiers mean a single variable or a range of variable values.

There are several mechanisms by which linkage errors can bias analyses based on linked data[1]. The strategy used for data linkage influence bias in linked data and choice of the strategy is usually determined by the purpose of the data linkage. One typical linking strategy is based on a basic (spine) dataset to which a new file(s) can be linked. Another strategy is sequential linking of datasets based on variables that match in adjacent linked datasets[1]. In sequential linkage, datasets are linked in a stepwise fashion where each linkage depends on successful matching in the previous step. Errors spread further and affect the quality of the resulting linked data file. Quality and utility of datasets that were selected for linkage cannot be changed in the normal workflow of the HDAB or the data holder. The availability and accuracy of the linkage identifier in the datasets is also given. The linkage identifier is specified by the data user in the linkage plan, however, communication should be allowed between the data user and the linking organisation or data holder to clarify the specification of the identifier for the linking, especially when the identifier is not a personal identifier (ID) but another variable. This comes into consideration, for example, if the identifier is a specific disease and data holders use different standards for coding for classification of diseases.

It should be noted that linking organisations cannot guarantee or commit to specific values of sensitivity, specificity or predictive values for any given linkage. Estimates may be provided as indicative ranges, based on the known properties of the datasets and identifiers. The improvement of data quality parameters may exceptionally, in individual cases, be carried out after an additional dialogue between the data user and the HDAB or TDH, e.g., in case the selected data files do not have a quality label or are found to be of insufficient quality for the purposes of data processing in the issued permit. In parallel, data cleaning tasks may be carried out by the HDAB or TDH to perform deterministic indirect linkage. It cannot be assumed that each data holder will be able to perform these specific tasks. This would go beyond the tasks set out in the EHDS Regulation and further implications need to be considered as the data will still be considered to be made available to the health data holders.

[10] Harron, K., Doidge, J. C., & Goldstein, H. (2020). Assessing data linkage quality in cohort studies. *Annals of Human Biology*, *47*(2), 218–226. https://doi.org/10.1080/03014460.2020.1742379

4.4.1 Linkage quality

Data linkage quality refers to the accuracy, completeness, and reliability of the record-linkage process, measured by how effectively true matches are identified and false matches are avoided.

Linkage quality of given datasets comprises these dimensions:

- **Accuracy (Precision):** Proportion of linked records that are truly correct matches.
- **Completeness (Recall/Sensitivity):** Proportion of all true matches that are successfully linked.
- **False Match Rate:** Percentage of linked records that should not have been linked.
- **Missed Match Rate:** Percentage of true matches that were not linked.

Note: In the EHDS. The data user selects the datasets to be linked, and therefore it is beyond the competence of the linking organisations to evaluate whether these selected datasets are suitable for the data user's project, e.g., in terms of proportional representation of different groups, or bias.

The level of linkage error is dependent on the quality and completeness of the identifying data available within a dataset and can occur irrespective of the linkage methods employed. However, careful data cleaning, where possible, and linkage design can help reduce the likelihood of errors, and linkage strategies can be designed to minimise false matches or missed matches (or to strike a balance between the two), depending on the aims of the project[11].

Table 3 quantifies the result after linking against the correct relationship of data for a single entity (in two records). This case assumes the linking of two datasets, containing m and n records, respectively.

Table 3. Linkage quality matrix for data that should relate to a given entity (e.g., individual) in case of linking two datasets[1].

	Match status	
	Match (pair from the same entity)	Non-match (pair from different entities)
Assigned link status		
Link	True match a	False match b
Non-link	Missed match c	True non-match d

In two identical datasets of size n, a perfect linkage procedure would give:

$a = n$ and $d = n(n - 1)$.

Various parameters of linkage quality can then be calculated:

[11] Harron, K., Doidge, J. C., & Goldstein, H. (2020). Assessing data linkage quality in cohort studies. *Annals of Human Biology*, *47*(2), 218–226. https://doi.org/10.1080/03014460.2020.1742379

The **sensitivity** (or **recall,** also known as true predictive rate) is the proportion of true matching record pairs that are classified as matches = a/(a + c).

The **specificity** is the probability of not detecting an event if the event is truly absent and is equal to = d/(b + d).

The **positive predictive value** (or **accuracy, precision**) is proportion of compared record pairs classified as matches that are true matches = a/(a + b).

The **negative predictive value** = d/(c + d).

The **false match rate** is the proportion of linked records that should not have been linked

= b/(a + b).

The **missed match rate** is the proportion true matches that were not linked = c/(a + c)**.**

Error rate (a + d)/(a + b + c + d), the proportion correctly classified record pairs).

The values of all parameters in the case of ideally perfect linking are equal to 1.

In data linkage, the most popular measures are accuracy (precision) and sensitivity (recall). If cases where they can be calculated, they are also recommended as a minimum for use by the HDAB or data holders for reporting linkage quality.

The main reason for the popularity of precision and sensitivity is that record linkage is commonly a very unbalanced classification problem. If $m \times n$ record pairs are being compared across two databases, and assuming there are no duplicates in each database (i.e. no two records in a single dataset refer to the same real-world entity), then the maximum number of true matches will be $a = min(m, n)$. This number (a) grows linearly with the size of the databases being linked, while the comparison space, $m \times n$, grows quadratically.

Note: The function $min(m, n)$ compares two input arguments, m and n, and returns the smaller of the two.

Combination of sensitivity and precision provides the F-measure, which is a formula that enables deriving a single number to measure linkage quality, the F-measure (the harmonic mean of precision and sensitivity):

F = (2 x precision x sensitivity) / (precision + sensitivity), or $2a/(2a + b + c)$

The value of F is high if both sensitivity and precision are high.

The importance of these parameters in a concrete case depends on the project or analysis that the user intends to perform with the data. The F-measure is equivalent to using different performance criteria for different linkage methods. There may be also other parameters calculated but they are outside of the scope of this guideline. It should be stressed that the values a,b,c,d are frequently not known in the practice so that the linking organisation (HDAB, TDH, DH) cannot commit to a certain value of any of these parameters. If the datasets that are to be linked are accompanied by information that there may be some errors in datasets, in e.g., ID's, the data user should be informed about it. This is valid even for datasets without errors in all other variables, in their classifications

(e.g., all patients are with diabetes type 1, but just their ID's may not be consistent across several datasets).

It is therefore useful for the entity that performs the linkage to know as much information as possible about known inconsistencies in the datasets for linkage purposes.

Depending on the quality requirements of the data user requesting data linkage, the linking organisation can then respond to the data user with useful information that may have impact on the planned intended use of the electronic health data, even in the case of demanding quality requirements.



Data holders may be requested to provide answers to the following questions about the datasets that should be linked:

- Are there possible/known errors in the dataset?
- If so, which variables are affected?
- Can the relevant error rate be estimated (quantified)?

Quality of data and linkage quality

It is necessary to understand the position of linkage quality in the context of overall data quality, which can be expressed by the quality label of the original datasets from which the data records are to be linked (Art. 78, EHDS Regulation).

Accuracy, coherence, completeness, consistency, precision and validity of data in a dataset, are the most linkage quality-relevant dimensions among all the dimensions for data quality as proposed by QUANTUM project in its deliverable 1.1 "Specification of the data sets' quality and utility label". And the data accuracy may be expressed through statistical information on accuracy at variable and/or individual level. Even not fully accurate or complete dataset may still obtain quality label because there are several other quality-related dimensions whose positive values/weights may allow assignment of a quality label. Moreover, QUANTUM proposes five dataset quality levels, so it is difficult to estimate the actual conditions for flawless linking even in a dataset with a quality label of a certain level. Therefore, it cannot automatically be assumed that linking data files labelled for quality will result in the highest quality of linkage. However, it is advisable that the linkage quality-relevant values or measures of data quality dimensions detected or calculated by the data holder in the process of determining the data quality level for labelling following the Article 78 of the EHDS Regulation are maintained and made available for projects involving data linkage that data users may wish to ask for. The linking organisation can also use these parameters when providing an estimate of the linking accuracy (see section 4.4.3).

Note: There may be a question of data users about what happens with the quality label of original datasets after performed linkage. There is no single and simple answer to that. It can be estimated that if the linkage quality-relevant dimensions had higher levels in the original datasets, the resulted linked data could represent also adequate data quality. Nevertheless, specific projects require specific, as precise as possible information about

where the errors are in the resulting file, and a general quality label assigned to the original datasets can no longer provide information for that.

Art 78(4) of the EHDS Regulation addresses a case in which the HDAB has reason to believe that a data quality and utility label might be inaccurate. The HDAB shall then assess whether the dataset covered by the label meets the quality requirements that are part of the elements of the data quality and utility label referred to in Art. 78(3) of the EHDS Regulation. If the dataset does not meet the quality requirements, or of there are systematic discrepancies in quality elements, the HDAB shall revoke the label (Art. 78(4), EHDS Reg.).

Another relevant topic is utility of the dataset. The QUANTUM project (e.g., its deliverable D1.2 Specification for the assessment of data holders´ maturity) recognises potential utility (typically information that describes the dataset in a standard manner, such as source, provenance or lineage, content, distribution, etc.), which designates fit-for-use of a dataset. In contrast, fit-for-purpose refers to the actual utility of an accessed dataset according to the objective of the health data request. The information necessary to evaluate fit-for-purpose depends on the user's experience. Datasets that have undergone some processing (e.g., data cleaning) to make them more suitable for more accurate data linkage may have significant value for further use or even for other data users. That is why these refined datasets may be recommended for inclusion in the data catalogues (see note in section 4.4.3).

If we resume to linkage quality and real-world linkage examples, it is important whether the linking organisation knows about linkage errors/inaccuracies of the original datasets or not.

Recognised linkage errors. Errors that can be inferred from metadata or known data limitations (e.g., temporal or geographical mismatch).

Unrecognised linkage errors. Errors that cannot be predicted from the description of the datasets that are to be linked and result from the degree of accuracy of the used linkage method and availability of unambiguous identifiers.

Variables used for linkage can be diverse, and biases have different kinds of impact on the resulting data intended for processing. It is important to understand the structure of the linkage to be performed and acknowledge the impact of the number of individuals or parameters of interest in the datasets on the objective of the planned data analysis. For example, linkage of data on a small number of subjects with a rare disease with other data using their IDs may be worthless if the IDs are incorrectly assigned to relevant individuals, or if an ID is missing in a linked dataset and probabilistic method of linkage is applied. This is important in cases where data on relatives are to be linked (e.g., mothers and their babies) and the IDs are different or not available.

A simple approach to understand the impact of linkage error in analysis is to consider the best- and worst-case scenarios: how much of each type of linkage error could there be, and how strongly might the error be correlated with parameters of interest? This type of quantitative bias analysis can be sufficient to demonstrate the sensitivity of results to the range of plausible assumptions that could be made about linkage error. The linking organisation may provide estimated values of sensitivity and/or specificity for a requested data linkage, but the actual consideration of whether the whole exercise with available

datasets will comply with the given project requirements for the quality of data is on the data applicant.

It shall be stressed that quality of data in a dataset, e.g. expressed by quality and utility label, may differ from resulting quality of all linked data. It's more apparent in cases when the relevant datasets were composed by different entities. Competent entities should establish mechanisms to systematically query data holders about known dataset limitations, to better inform linkage planning and support transparency for data users.

4.4.2 Techniques for linkage quality assessment

Uncertainty in data linkage, particularly with administrative data, is generally inescapable[12]. Datasets with small number of records can be assessed directly, manually (true match – false match) and we can get directly the values listed in Table 3 above and calculate parameters (e.g. sensitivity) as agreed with the data user.

Every conceivable technique for linkage quality assessment in larger datasets is either a partial measure (it only identifies some errors) or indirect (it only estimates the rate or distribution of errors).

While it is rarely possible to measure of larger datasets linkage quality in full, a combination of direct and indirect methods can provide meaningful estimates of linkage quality, particularly when supported by metadata, reference statistics, or clerical review, which is a manual human check of data matches.

[1]More advanced analysis of the impact of linkage errors on the given project of the data user is possible, but it is outside of the scope of this guideline.

Note: Advanced quality assessment of data linkage can also provide estimates of the rates of missed links and false links, and moreover:

- estimates of how these errors influence clustering, and
- estimates of how errors vary according to any variables of interest for a given analysis[13].

Specific techniques for linkage quality assessment usable for larger datasets are summarised in Table 4.

Details about these techniques can be found in[14]. In Annex 7 – Excursion on linkage quality assessment methods are shortened descriptions of the three techniques that require knowledge of identifiers, which is most likely case for (trusted) data holders or HDABs.

---

[12] Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

[13] Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

[14] Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

Table 4. Techniques for linkage quality assessment also found in Doidge et al (2024).

Technique	Required inputs	Potential outputs
Techniques requiring access to identifiers:		
Training data (gold standard)	Identifiers	Rates of missed links and false links and/or distribution of error rates. The training data can be used to estimate rates and distributions of linkage errors in the remaining records in which values for all identifiers are not complete. The training data are assumed to be representative of the joint distribution of matching variable quality and any analysis variables, which, however, may not be always the reality.
Clerical review	Identifiers and/or supplementary matching data.	Human estimation of match status leading to estimated rates of false links and distribution of false links.
Negative controls	A set of records not expected to link that can be submitted to the linkage procedure.	Rates of false links.
Techniques not requiring access to identifiers:		
Unlikely or implausible links	Links excluded by authorised personnel performing data linkage during quality assurance; and/or records with multiple candidate links even when only one is possible; and/or payload data.	Distribution of false links, rate of false links.
Analysis of matching variable quality	Record-level or aggregate indicators of matching variable quality.	Identification of unlinkable records. Distribution of missed links and/or likely missed links.
Comparison of linked vs unlinked records	Unlinked records, or aggregate characteristics of unlinked records, when all records in one or both files are expected to have matches.	Distribution of missed links Rate of missed links when expected match rate = 100%, given rate of false links can also be estimated.
Positive controls	Unlinked records for a subset of records expected to have matches.	Distribution of missed links Rate of missed links, when rate of false links can be estimated.
Comparison of linked data to external reference statistics	Statistics derived from another representative dataset for observable	Rate of missed links. Rate of false links Distribution of missed links. Distribution of false links.

	characteristics of the linked data.	

Note: 'Payload' data mentioned in *Unlikely or implausible links* technique are variables used for analysis by the data user but not always for linkage[15]. The term therefore refers to variables intended for use in the project for analysis and may also serve secondarily to evaluate linkage quality. The term 'rate' is used very loosely in this table to refer to any measurement of error frequency or distribution.

Selected techniques should provide at least some values necessary for calculation of the parameters listed in section 4.4.1, (e.g., precision, sensitivity, specificity) and they can then be communicated to the data user. Also, the F-parameter may be calculated as a single number characterising linkage quality.

As also stated in section 4.4.3, quality assessment is an optional activity for linking organisations and may only be offered if they control the relevant processes, have the necessary information about the linked data sets, intend to improve their services, and there is interest in this information from data users.

4.4.3 Recommendations related to linkage in the interest of ensuring its quality

While the functional separation between data users and data processors (i.e., HDABs or (trusted) data holders) ensures confidentiality and complies with legal safeguards, it may also limit feedback loops that are beneficial for ensuring high-quality linkage. Within the linking organisation, internal organisational measures – such as segregation of duties with structured collaboration – can help reconcile privacy requirements with technical accuracy.

Practices that can be recommended to help establish a more cohesive system that improves the extraction of value(s) from linked data are summarised below. Linking organisations are encouraged to (based on[16]):

- Prior to linkage and where possible, engage with data users of the linked data to understand their requirements regarding data linkage quality and the potential implications of both missed links and false links for the intended application. Use the relevant template to collect this information (Annex 5 – Data description template), or provide users with the template in Annex 8 – Template for instructions for data users regarding upload of data they possess.
- Be transparent about the approach taken to data linkage, including data cleaning and possible harmonisation, assessment of agreement between records, use of clustering algorithms and quality assurance. Relevant information including fees and terms should be communicated to the data users in time.
- Where feasible, obtain additional information about the datasets for possible linkage quality estimation from the data holder(s). Information about common error rates in

[15] Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

[16] Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

IDs or other variables in given kind of datasets is useful; these parameters are often known to data holders with some accuracy. It may help to estimate quality of planned data linkage.

- If a dataset is provided by the data user (e.g., for linkage with authorised datasets), the HDAB must ensure that appropriate legal and technical safeguards are in place. The user should be required to provide clear documentation on variable formats (see Annex 8 – Template for instructions for data users regarding upload of data they possess), data structure, and quality parameters.
- Accept (and inform the data user) that linkage error is generally inevitable and that no single approach to linkage quality assessment is likely to be sufficient for measuring it. Consider also possible impact of data minimisation on the quality of linkage.
- Find, together with the data user, a balance between the quality needs of the intended application, the quality of the data available for linkage, and the resources available to support the linkage process.
- Where appropriate, provide feedback to data holders on identified linkage errors and other findings from the linkage quality assessment to support development of the dataset and linkage algorithm.
- Report on the quality of the linkage to data users.

To what extent quality assessment can be implemented depends on what information is available to linking organisation (HDAB or DH, TDH) and what was also agreed with the data user.

At a minimum, linkage outputs provided by a linking organisation should include:

- Detailed information about the linkage process.
- Record-level information about matching variable quality where possible (e.g., an indicator of whether each matching variable was missing, invalid or valid).
- Link-level information about match quality.
- Aggregate information about any identified linkage errors (for example, links excluded during quality assurance). If linkage quality assessment was performed, a report on this.

A simplified template for a linked dataset description that linking organisations may use is added to Annex 5 – Data description template. This template also includes optional elements to report linkage quality assessments.

Wherever possible, **linkage outputs provided by the linking organisation should also include**:

- Uncertain links.
- Record-level information about any identified linkage errors.

If data users have sufficient information about the possible linkage errors from the HDAB (or from the TDH), they can appropriately address linkage errors in the analyses conducted and/or take them into account when interpreting the analysis results.

Recommendations for linking organisations with respect to linkage quality:

I. Linking quality estimation

Requests by data applicants may require a review of the quality of planned linking, including what is known about the data files in terms of accuracy, coherence, completeness, consistency, correctness, and validity of the variable used for linking. The HDAB or data holder can also offer estimation of linkage quality after initial review of the data user application, particularly if probabilistic linkage should be used. These are activities for which the HDAB, or possibly the data holder, can prepare offers including pricing. As data holders have the best knowledge of their data, they should be able to cooperate with the HDAB in providing relevant information to data users that may have specific impact on linkage quality.

As to linkage quality parameters, it is left on the HDAB or data holder to choose suitable ones to express expected quality of the future linkage. Linking organisations may also conclude that to achieve linking quality adequate for a given project, it will be appropriate to improve the data quality of the original datasets. It is strongly recommended to agree on these optional activities, their scope, and price with the data applicant. Regarding estimating the quality of linkage, linkage organisations should only offer services for which they are qualified, have the capacity, and have sufficient information about the datasets.

II. **Communication of the linkage quality estimation to the applicant**

Estimated values of linkage quality can be presented to the data applicants either using the statistical parameters or absolute numbers so that they are aware of the possible results and can consider the usefulness of such a linkage in a timely manner.

When the HDAB or the data holder performed linkage quality estimation, in case the estimated linkage quality is clearly below the expectations of health data user, three main options for further progress in a given case:

- the data user withdraws the application,
- the data user accepts possible lower quality of data linking results,
- or the data user accepts extra processing of the data (e.g., cleaning), in case such improvement of the datasets in the interest of higher quality of the result is feasible for the HDAB, TDH, data holder.

It is recommended that after the HDAB or the data holder makes an estimate of the linkage quality, the resulting values (e.g., at least overlinkage, underlinkage, or precision, recall) should be communicated to the data user as soon as possible. This communication should occur before HDAB sends the user a preliminary invoice for processing the request. Limitations of this activity should be clearly communicated to the data user, particularly that some incorrect matches may still occur and that the linking process itself may still introduce unexpected errors, especially in the case of deterministic indirect linking (probabilistic linking is considered not recommended for common requests). Please also have a look at Annex 10 – Data quality considerations.

Data users can withdraw the application and then possibly submit modified application if they are not satisfied with the estimated linkage quality, price or with the proposed timelines for making corrections in the datasets that HDAB or the data holder could perform before linking. Art. 62 (5) of the EHDS Regulation applies in case of withdrawal of the application,

and the health data applicant shall only be charged the costs that have already been incurred.

III. **Reporting of linkage quality parameters**

After the datasets are linked, the linking organisation should provide the data user with basic information about the quality of the linkage through selected linkage quality parameters, possible reduction in the number of records due to efforts to achieve accurate linkage, and other relevant information that may be useful for the data user's planned analysis (see Annex 5 – Data description template). Techniques for linkage quality assessment listed in Table 4 can be used. The EHDS Regulation does not impose any obligation to assess the quality of the linkage or the quality of the resulting linked data. With regard to reporting the quality of linkage, linkage organisations should only offer services for which they are qualified and have sufficient information about the datasets.

Note: If the extra processing of catalogued datasets aimed at improvement of the data for linking lead to the creation of new version of the dataset with higher quality degree, it should, depending on conditions on the side of the data holder, be properly described by metadata and made available for future use.

5 Role-based access to linkage processes

To perform successful data linkage with the necessary safeguards in place, while adhering to the EHDS Regulation, role-based access measures are helpful. The following section elaborates on recommendations on how to address and implement some role-based access measures.

Below, you may find an overview of the role-based accountability along the EHDS user journey. Please note that multiple actors may be involved in each procedure along the EHDS user journey. Listed are only the actors accountable according to the EHDS Regulation. It is recommended to make the detailed steps needed to follow the procedures listed in the table below as transparent as possible to support transparency and non-discrimination. Please also consult Annex 11 – Scenarios data linkage with TTP.

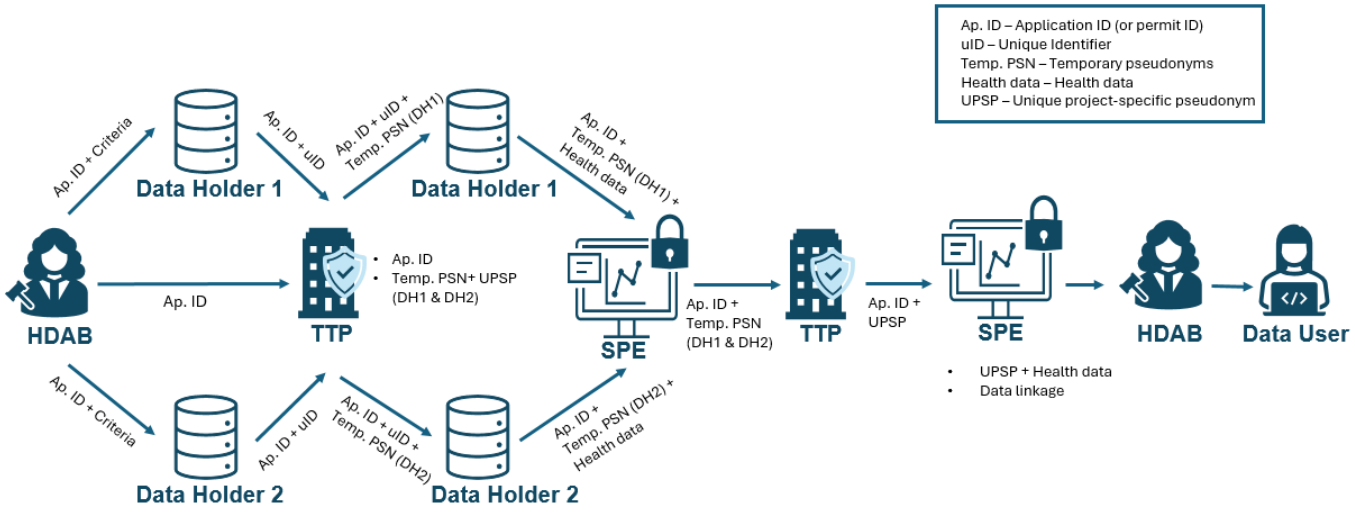
Table 5. Procedures necessary for data linkage under the EHDS with their respective accountabilities listed. Note that HDABs or TDH can have a TTP or contact to other intermediate entities that function as a TTP.

Procedure	Accountability	EHDS Regulation
Receipt of data application at HDAB *Optional*: Forward application to TDH	HDAB HDAB	Art. 57(1)(a) Art. 72(3)
Assessment of data application	HDAB TDH	Art. 57(1)(a) Art. 72(4)
Decision on data permit Proposal for decision	HDAB TDH	Art. 57(1)(a) Art. 72(5)
Issuing data permit	HDAB	Art. 57(1)(a)
Request of data from DH	HDAB	Art. 57(1)(a)(iii)/Art. 68(7)
Sending data	DU	Art. 60(1)
Receipt of data	HDAB	Art. 57(1)(b)

Data linkage	HDAB TDH	Art. 57(1)(b) Art. 72(6)
Data provision to data user	HDAB	Art. 68(7)/Art. 73(2)

In the following, an overview of how data linkage involving a trusted third party can look like under the EHDS is presented.

Figure 6. Depiction of an exemplary workflow for data linkage under the EHDS when two data holders and one HDAB are involved. More complex scenarios are possible, please see this as a simplified overview. Certain steps along the workflow may vary. The trusted third party (TTP) does not need to be a separate trust centre but is here represented in its operational role. Meaning, a TTP can be part of an HDAB, a trusted data holder or an intermediation entity, if the role-separation is guaranteed, legally sound, and appropriate safeguards are in place. Please note that there are repeated occurrences of the same actors referred to in the figure (i.e., the same HDAB is involved in the beginning and towards the end of the workflow).



5.1 Before data extraction for data linkage

To prevent low utility and low cell count linkages, the actors involved in the linkage process should communicate before data linkage. To the best of our current knowledge, no standardised ways of communication are in place on how HDABs communicate with each other. We reinforce the importance of structured and transparent communication, aligned with the clear description of the data user's needs (including linkage) (Art. 61(1)), that HDABs can request additional information from users or data holders (Art. 57(1)(a)(ii)) and the coordination of data preparation performed by HDABs (including linkage) (Art. 67(2)(f). Pleaser also refer to TEHDAS2 D6.3 and D6.1.

It should also be noted that the data minimisation principle (Article 5(1)(c), GDPR; Article 66 EHDS Reg.) applies to all steps along the user journey (see Annex 3 – User journey) and can be monitored by the HDAB (Article 57(1)(a)(ii), EHDS Reg.).

There may be many ways to preserve role-based safeguards when linking data in the data preparation stage. In the following, we are listing some recommendations:

- Health data and the IDs (e.g., identifying data, such as unique identifiers) must be separated and handled separately.
- Trusted third parties, or the section within the HDAB that performs the pseudonym handling, may communicate with each other to enable linkage. If the HDAB performs the pseudonymisation and data linkage, organisational segregation between the personnel who has access to the identifiers and those who assess the output (for example) should be in place (Art. 55(3)).
- The linking organisation coordinates with the data holder(s), who assesses – based on in-/exclusion criteria from the application – the eligible population.
- It is recommended that the HDAB shares with other possible parties involved in data linking only the criteria absolutely necessary for the project in question. This should prevent gradual knowledge accumulation on characteristics of persons in the data outside the original mandate of the data holder and in breach of the data minimisation principle.
- Data extraction can be parallel or sequential. The general recommendation is to apply parallel data extraction, while sequential extraction procedures might be applicable in certain cases (i.e., when one data holder already holds part of the information from another data holder, whose data should be linked).
- It is recommended that the data holder first communicates the size of the eligible population to the HDAB before performing data linkage. This might be useful to assess linkage feasibility, and the HDAB can in turn communicate the size of the population to the data user, who can decide to advance with the project or retract the application. However, according to the EHDS Regulation, the HDAB grants access to the health data when approving the permit (Art. 68(1) & Art 68(2)), before the feasibility of the linkage is assessed as data can only be extracted after a permit has been issued (Art. 68(7)).
- No individual-level data are transferred to the SPE until the HDAB confirms that linkage meets legal and statistical thresholds based on the data permit
- Following approval by the HDAB, the data can be transferred to the SPE, linked, and re-identification potential has to be assessed before the user is granted access to the SPE.

It is acknowledged that flexibility must be preserved to account for divergent national legal frameworks and operational maturity. If, under unforeseen circumstances, the cell count is too small (i.e., based on the sample size specified in the application, or due to the HDABs risk assessment), linking of such data could lead to easier re-identification and the HDAB could proceed in line with Art. 68(3) and the data user may optionally be advised to reapply with a modified version that does not trigger such risky situations. It is strongly recommended to balance the trade-off between stronger protection against re-identification and data minimisation.

5.2 After linkage feasibility is verified

The following recommendations address steps that occur after it has been verified that the data to be provided are useful and sufficient for the applicant to answer their research question. In other words, the intersection between the to-be-linked datasets is adequate for the research purposes of the data user.

Here are some recommendations after it has been decided what role-based safeguards are in place between the different data holders:

- The organisational and technical safeguards for (sensitive) data processing shall be in place (here we also refer to Deliverable 7.4).
- The personnel performing the data linkage must be trained and authorised. There might be extra safety procedures in place, for example reasonable criminal record checks. Those checks should be context-dependent, performed with caution, proportionally and aligned with national law.
- Re-pseudonymisation should be performed by the linking organisation immediately after the linkage operation and prior to any access by the data user. It must be ensured that these pseudonyms are project-specific and non-linkable across different data permits. This ensures study-specific segregation and limits re-identification risks.
- Logs and audits for monitoring linkage activities.

Logging helps to identify and avoid errors, disruptions, and capacity bottlenecks at an early stage. Audit logs help organisations to document an historical record activity for compliance purposes and other business policy enforcement. Audit logs create an audit trail, which can be inspected when needed.

We recommend that linkage activities are logged and audited to facilitate the understanding of the HDAB and to be able to retrospectively follow up on the steps taken during the data preparation. Audit trails should be tamper-resistant, retained for a legally defined period (see Article 73(1)(e) EHDS Reg.), and periodically reviewed by the HDAB or a designated oversight body. Logs should cover access, linkage operations, and downstream queries on linked datasets.

However, how linkage procedures might be logged or audited remains to be discussed. This is especially relevant when deterministic linkage activities are performed that contain direct and indirect identifiers. It is recommended that logs for linkage activities include as minimum information: the actors involved in the linkage, a timestamp, the datasets involved, the system where linkage was performed, the linkage method, the variables/identifiers used for linkage, and if performed, validation or quality assurance methods performed. As the recommended audit trail includes information on the actors and timestamps, it is considered personal data and falls under GDPR.

Data linkage should occur in an SPE with no access for data users, prior to making the data available for secondary use. Data linkage could also be performed in an SPE where the relevant actors have access to (i.e., HDAB, data holder(s), TDH(s)). To enforce technical and organisational security measures, it is recommended here to adhere to the recommendations from Deliverable 7.4. Please note that the same governance and security mechanisms hold for SPEs with and without data user access.

Logs and audit trails will be generated when data users get access to an SPE (EHDS Reg. Article 73 (1)(e)), Article 73 (3)). The logs will be kept for the period necessary to verify and audit all processing activities in that environment (EHDS Reg. Article 73 (1)(e)). The frequency of audits is not defined in the EHDS Regulation, it states that audits of SPEs should be performed on “a regular basis” (Article 73(3)).

5.3 Managing data linkage documentation

Documentation of the processes related to data linkage, the linkage methods used and any linkage quality assessment should be communicated to the data user upon data access or provision and maintained by the linking organisation for the same period as the application information is retained. Please consult Annex 5 – Data description template. Other ways of linkage documentation might be using metadata.

The data user should be informed about the information that may have impact on the purpose of processing the data and its quality in accordance with the data permit or approved health data request. Relevant information from the data linking documentation should be attached to the quote and invoice or provided in the SPE.

Data linkage documentation intended for data users should be provided through the linking organisation.

5.4 Destiny of the linked data

After the linking organisation has placed the linked data either in the SPE for the data user to analyse in line with the permit or handed over in an anonymised statistical format to the data user, as a new specific dataset, there may be question whether this new dataset may be useful also for other users in the future. According to the EHDS Regulation, Article 68(12), the electronic health data within the secure processing environment shall be deleted within six months of the expiry of the data permit. At the request of the health data user, the formula for the creation of the requested dataset may be stored by the health data access body.

The EHDS Regulation therefore does not pave the way for a harmonised EU method of recycling this newly created data set for others. However, data linkage can be considered in conjunction with data enrichment, which may be possible at national level, and if so, then governed by national law.

6 Penalties for misuse of (linked) data

The legal foundation for the supervisory and sanctioning responsibilities of HDABs, also in terms of penalties for misuse of health data can be found in Articles 63 and 64 of the EHDS Regulation. The explicit use case of penalties for misuse of linked data is not addressed in the EHDS Regulation but equally applies to linked health datasets. Therefore, to prevent repetition, please refer to the Deliverable of task 4.1 in TEHDAS2.

7 Repeated and frequent applications to (linked) datasets

It is likely that data users repeatedly request linked data from data holders. Please also refer to D6.2 Guideline for data users on good application and access practice (e.g., section 6.4.7) and to D6.3 Guidelines for health data access bodies on the procedures and formats for data access.

It should be specified whether the data user would like to:

(1) exclusively extract new information (e.g., annual updates),

or

(2) update the previously extracted (and possibly linked) data, while adding information (e.g., for the current year).

Both options are possible under the EHDS but must be submitted as new applications. This aims at reducing complexity and avoid ambiguity. For these new applications, simplified procedures might be in place. This is also to prevent re-identification, as adding data to an already approved permit will require a new re-identification risk assessment and data minimisation assessment.

There are three key aspects that the linking organisation should be aware of and inform the data applicants about repeated applications. First, data users must be made aware that repeated data (access) requests for the same linked datasets may result in different outcomes. This might occur due to changes in the underlying data, which can stem from the opt-out of individuals or from updates to registry entries. When the underlying data is changed, a bias can be introduced to the datasets, which future users should be made aware of. It is recommended that the HDAB or the DH manage the version control of these dynamically changing datasets to support reproducibility.

Second, the linking organisation should provide clear procedural guidance to data holders and data users for handling repeated requests, while distinguishing between incremental updates and full re-extractions.



For repeated requests simplified procedures could be established at the linking organisation, as the application has already been assessed.

To facilitate the processing of repeated requests, which could decrease the administrative burden for the HDAB and the data holder(s), and reduce costs for the data user, it should be clearly stated in the application:

- what data should be extracted,
- the data extraction dates,
- what actors are involved,
- and to which already approved request or permit the current request pertains.

Third, for requests that update the previously extracted data while adding information, it is recommended to refrain from probabilistic linking methods. To further enhance the stability of the linkage processes, it should be ensured that the variables used to perform the linkage (i.e., ID's and/or indirect identifying variables) are identical to the pervious application, stable and available. In addition, documenting the linkage parameters, including parameters help to facilitate future linkages and is strongly recommended.

It should be noted that for data access applications, the data relating to one data permit is provided in one SPE instance.

Additional important aspects to consider are:

- **Amendments**. They are possible under the EHDS Reg. (Article 68(13). However, they refer to minor changes, such as a change in the research personnel or a slight extension of the data retention period. A change to data categories, purposes or risk profile does not fall under the concept of amendments in the EHDS Regulation. Please also see D6.3.

- **Frequent data access or frequent data updates**. Under the EHDS, future and frequent data extractions are possible for public sector bodies and Union institutions, bodies, offices and agencies (Article 67(5), but not for other data users. In other words, public sector bodies and Union institutions may request recurring access covering new data extractions over time (Article 67(5), if these cover the same scope, for the same purposes and same use. In addition, they can, request access data for future periods (e.g., annual updates).

8 Open questions

The **data minimisation** principle (Article 5(1)(c), GDPR and 66(1), EHDS Reg.) is part of every step along the user journey and its technical implementation for an application process under the EHDS may render data linkage more challenging. The EHDS Regulation explicitly allows HDABs to process data where strictly necessary for data preparation, including linkage (Article 67(2)(f), EHDS Reg.). This controlled processing must occur before data access is granted and be subject to documentation and safeguards. Nevertheless, it remains challenging to determine the right amount of granularity during data preparation, especially, when multiple data sources are involved.

Another challenging aspect is the treatment of **quasi-identifiers**, which are sometimes needed for successful linkage, but they can also lead to revealing identifying information. To mitigate this, clear variable descriptions should be available to the linking organisation (see the recommendations from Annexes 4 or 7).

Moreover, standardised practices and **efficient communication** schemes should be in place. The EHDS Regulation supports structured communication between data users and HDABs, including clarification of linkage requirements (Art. 61(1)) and the possibility for HDABs to request further information from data users or holders (Art. 57(1)(a)(ii)). While direct contact between data users and data holders is not foreseen, HDABs may serve as intermediaries to ensure appropriate interpretation of technical and legal conditions. It remains challenging to implement and connect the numerous actors in an efficient way and in a timely manner.

Related to this is the **structural separation between linking organisations** (i.e., data holders, TDHs, HDABs) **and data users**, which is required under EHDS Articles 60 and 73. While it can limit interactive collaboration, it serves to uphold confidentiality and safeguard against re-identification. Internal coordination within HDABs and controlled metadata sharing may mitigate this limitation.

Annex number	Annex title
1	Methodology
2	Public consultation summary
3	User journey
4	Glossary
5	Data description template
6	Data linkage scenarios
7	Excursion on linkage quality assessment methods
8	Template for instructions for data users regarding upload of data they possess
9	Use cases
10	Data quality considerations
11	Scenarios data linkage with TTP

Annex 1 – Methodology

Survey development:
In the preparatory phase, thematic brainstorming sessions were performed internally, followed by drafting the first version of the survey questions. After an internal feedback loop, feedback was provided by the EC and the result was further distributed to the major and minor contributors to provide comments. The final version was implemented in an online survey tool (i.e., LimeSurvey), published and distributed to the whole TEHDAS2 consortium and further to maximise outreach.

Demographic information from the survey:
A total of 64 responses were recorded. Incomplete responses and multiple entries by one Institution were cleaned, which resulted in 27 (full: 24, partial 3) contributions. The top three personal fields of expertise were: Data management (12/44%), Project management (11/41%) and Data science (10/37%). The most frequent represented countries were Italy (3), Belgium (3), Germany (3), Sweden (2), and Spain (2). The top three roles that were represented were: Project lead (7/26%), Project coordination (7/26%) and Team lead (6/22%). Among the respondents, there were data holders (12/44%), HDABs (12/44%), data users (5/19%), trusted third parties (2/7%) and others (4/15%). The data these respondents are (planning) to make accessible are tables (22/81%), relational databases (18/67%), unstructured data (13/48%), imaging data (9/33%), genomic data (7/26%), bio-sample data (5/19%), and other (3/11%).

Survey results for 5.2 What data is being linked?
According to the survey, types of data that are or will be made available via health data access application included tables (81%), relational databases (67%), unstructured data (48%), imaging data (33%), genomic data (26%), bio-sample data (19%) and other data (11%). Majority of respondents (67%, n=18) came from institutions that currently link electronic health data from different datasets and 52% (n=14) linked data from different data holders.

In terms of types of data that are being currently linked, respondents mentioned a wide variety of health data, including administrative data, clinical data, electronic health records data, (population-based) health data registries, other (disease specific) health data or medical device registries, health data from medical devices and digital health applications, research data (e.g., surveys), social/ socio-economic data, mortality data, genomic data, residency data and data on health care professionals.
Six respondents indicated that there are specific data types that are prohibited from being linked. Reasons for not allowing the linkage of specific data types included anonymisation and minimisation aspects as well as administrative reasons, such as not being able to link data from different regions.

Survey results for 5.3 How is data linked?
According to the survey, the most frequently used unique identifiers for direct linkage across countries are:

- Social Security (or Health Insurance) numbers
- National Identification numbers

Annex 2 – Public consultation summary

A draft version of this document was in public consultation between the 5th of May and 28th of June 2026. This document was commented in total 81 times. One response was deleted, as it was a duplicate. Some respondents have also responded both from data holder's and data user's perspective. The responses came from 20 different countries from the EU countries and the European Economic Area countries. Responses from Eastern and Southern European countries and international organisations were largely missing. The respondents were primarily from three main types of organisations, listed in order of prevalence: public organisations, academic/research organisations, and private organisations.

The comments were screened and labelled (labels: relevant and implement; relevant but will not be implemented; relevant but not feasible/postponed; required discussion; irrelevant/out of scope). The comments from the *Generic Feedback* were sorted to the sections they pertained to. Based on the comments, each section was adapted, which is described in more detail below.

We added information on timeframes, other entities and data linkage, interoperability, a decision tree for choosing the linkage method to apply (Figure 4). More concrete linkage scenarios, including a trusted third party, were added (see Figure 3). Information on timeframes and on the accountability of each actor was added (see Table 5). A section for optional data quality assessment was added to Annex 5 – Data description template. More information on interoperability was added, as well as sections on other linking organisations or considerations regarding possible reuse of linked datasets. Finally, Annex 11 – Scenarios data linkage with TTP was added, showcasing possible linkage scenarios including TTPs.

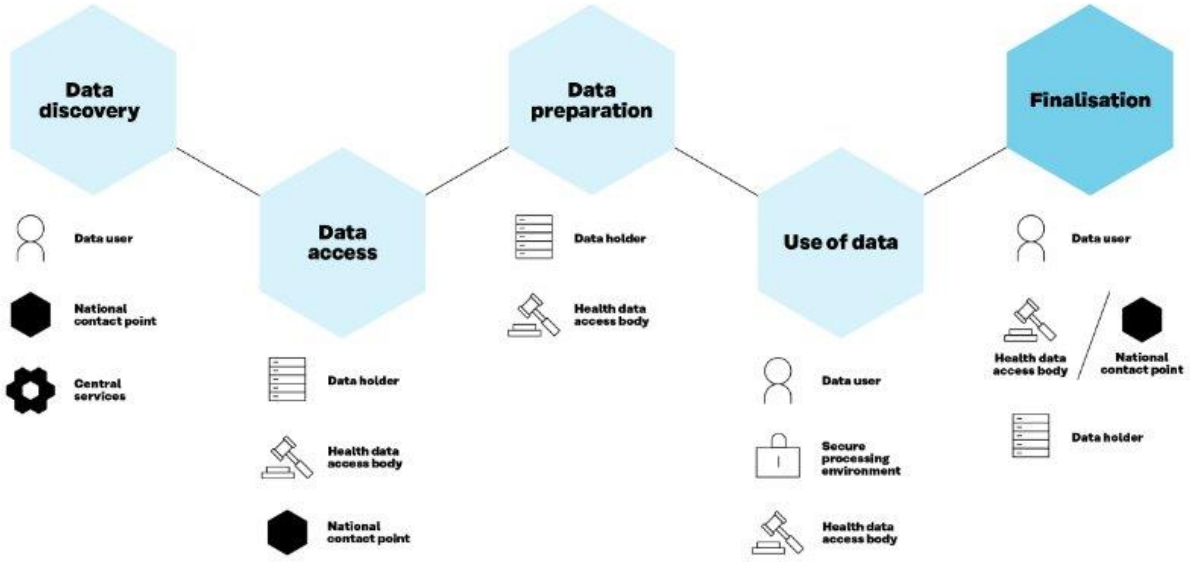
Cross-border linkage or multi-country applications are out-of-scope for this guideline, as scenarios involving multiple member states are still to be defined. It will be possible, however, to use the EU SPE for future applications that involve data from several member states.

It was mentioned multiple times that standardised procedures for data linkage are currently lacking, combined with standardised ways of communication.

Annex 3 – User journey

When a data user[17] applies for electronic health data for secondary use purposes, such as research and innovation activities, education, and policymaking, within the European Health Data Space (EHDS), the user journey consists of several stages (see Figure 1). Access for certain purposes (public or occupational health, policy-making and regulatory activities, and statistics) is reserved for public sector bodies and Union institutions (see Chapter IV, Art. 53(1) and 53(2)).

Figure 1: EHDS user journey consists of five main phases: data discovery, data access, data preparation, use of data and finalisation.



Data discovery

Before being able to use the data, the user needs to investigate whether the data needed is available, and whether it is available in the necessary format for the secondary use purpose. This phase is called data discovery. Datasets available in the EU can be found in a metadata catalogue at https://qa.data.health.europa.eu/. Once the data discovery is completed, the user can begin the process of applying for the data.

Data access

In the data access phase, the user fills in and submits a dedicated and standardised data access application form or a data request to a health data access body (HDAB)[18]. The user must complete the information required in the form, upload necessary documents, and provide justifications as needed.

Data access application form is used when the user seeks to use personal level data.
Data request is for cases when the user wants to apply for anonymised statistical data.

[17] Data user = a person using electronic health data for a secondary use purpose
[18] Health data access body (HDAB) = the authority responsible for assessing the information provided by the data user who applies for electronic health data for a secondary use purpose

Data preparation

During this phase, the data holder(s)[19] deliver(s) the necessary data to the HDAB, which starts to prepare the data for secondary use. Techniques for pseudonymisation, anonymisation, generalisation, suppression, and randomisation of personal data are employed. The data minimisation principle (as per the GDPR) must be respected to ensure privacy.

Use of data

In this phase, the user performs analyses based on the received data for the purpose defined in the application phase. Analysing personal level data must be performed in a secure processing environment[20]. The duration of this phase is specified in the Regulation (Art 68(12)).

Finalisation

This last phase of the user journey concerns data user's duties regarding analysis outcomes derived from secondary use of data. Data user must publish the results of secondary use of health data within 18 months of the completion of the data processing in a secure processing environment or of receiving the requested health data. The results should be provided in an anonymous format. The data user must inform the health data access body of the results. In addition, the data user must mention in the output that the results have been obtained by using data in the framework of the EHDS.

[19] Data holder = Any natural or legal person, public authority or other body in the healthcare or the care sectors that has the right or obligation to provide electronic health data for secondary use purposes or the ability to make such data available (see more EHDS Regulation Art. 2 (1t)).
[20] Secure processing environment = an environment with strong technical and security safeguards in which the data user can process personal level electronic health data

Annex 4 – Glossary

Term	Description
Agreement	A description of similarity in terms of ***matching*** variables.
Data combination	The process of bringing together data from multiple ***datasets*** that can be processed pursuant to one or multiple data permit(s) or data request(s) (Regulation (EU) 2015/327 (EHDS) Articles 57, 68, 69) or other legal basis (such as consent or permits based on other legislation than EHDS). ***Data linkage*** can be part of this process.
Data extraction	Data extraction is the process of retrieving data from its source dataset. Structured data extraction involves extracting data from datasets that are already organised in predefined formats. Unstructured data extraction pertains to extracting data from databases handling unstructured formats such as PDFs, images, or free text. There may be one or more different data sources from which data extraction may be required.
Data linkage	The process of combining data/ ***health records*** from different ***datasets*** from several sources on one ***entity***, ***topic*** or data subject (based on ISO 5127:2017, 3.1.11.12, extended). This can be done using unique identifiers, probabilistic methods, or a combination of techniques.
Data linkage quality	The degree to which records from two or more datasets are correctly matched to the same entity.
Data minimisation	A principle mandating to only collect, store and process personal data in a manner that is adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed. (GDPR Article 5(1)(c)) Access is only provided to electronic health data that is "adequate, relevant and limited to what is necessary in relation to the purpose of processing indicated in the health data access application by the health data user and in line with the data permit issues pursuant to Article 68." (EHDS Regulation, Article 66(1)) Data minimisation applies to all stages of the data lifecycle.
Data permit	An "administrative decision issued by a health data access body to process certain electronic health data specified in the decision for specific secondary use purposes, based on conditions laid down in Chapter IV of this Regulation"; (Regulation (EU) 2025/327 (EHDS), Article 2(2)(v))
Dataset	A "structured collection of electronic health data. Regulation (EU) 2025/327 (EHDS), Article 2(2)(w))
Entity	Is "something capable of being uniquely identified. Note 1 to entry: *Entities* include material objects, electronic representations of content, abstract items (such as times, places),

	parties (human and corporate), as well as anything else that can be identified uniquely. Note 2 to entry: A defined fragment of an *entity* is itself an entity." (ISO 5127:2017(en), 3.1.13.27)
Health data access body (HDAB)	Member state-designated authority that facilitates the secondary use of electronic health data. HDABs assess the information provided by the health data applicant and decide on health data requests and access applications, authorise and issue data permits, obtain data from data holders and make data available in Secure Processing Environments. HDABs systematically track the data request and data access applications received and the data permits issued. As per Article 58 of the EHDS, HDABs are required to publicly list information on the data permits issued (Regulation (EU) 2025/327 (EHDS), Article 55 and Recital 52).
Health data holder	Any person, organisation or public body involved in healthcare, care services, health-related products, wellness apps or health(care) research, that has the right to process data for health care provision or for public health purposes, reimbursement, research, policy making, official statistics or patient safety. This includes, for example, hospitals, insurers, research institutes and EU institutions. For a more detailed definition: EHDS Regulation, Article 2(2) point (t).
Health data intermediation entity	A legal person that may be established by national law for the purpose of fulfilling the obligations of certain categories of health data holders and that is able to process, make available, register, provide, restrict access to and exchange electronic health data for secondary use provided by health data holders. (EHDS Regulation, Article 50 (3) and Recital 59)
Health data user	A "natural or legal person, including Union institutions, bodies, offices or agencies, which has been granted lawful access to electronic health data for secondary use pursuant to a data permit, a health data request approval or an access approval by an authorised participant in HealthData@EU;" (Regulation (EU) 2025/327 (EHDS), Article 2(2)(u))
Health record	A "data repository regarding the health and care" (ISO/TS 16551:2025(en), 3.9, shortened) of a data subject.
Link	The derived or assumed classification between ***health records***.
Linking organisation	The organisation performing data linkage (HDAB, data holder, trusted data holder).
Match	A true relationship between ***health records***.
Match quality	Based on ISO/IEC 19794-14:2022(en), 3.4.12, but with broader scope: The "level of ***agreement*** between different ***health records***.
Overlinkage (false positives/false match)	When one ***health record*** is incorrectly ***linked*** to multiple data subjects.
Pseudonymisation	The processing of personal data in such a way that the "data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to

	ensure that the personal data are not attributed to an identified or identifiable natural person". (Regulation (EU) 2016/679 (GDPR) Article 4(5))
Re-identification risk	The risk of a successful re-identification attack (ISO/IEC 20889:2018(en), 3.33), which describes an action performed on de-identified data by an attacker with the purpose of ***re-identification*** (ISO/IEC 20889:2018(en), 3.32).
Secure processing environment (SPE)	An environment in which access to electronic health data can be provided in following a data permit. A secure processing environment is subject to technical and organisational measures and security and interoperability requirements. Specifically allowing access to only those persons listed in the permit, as well as user authentication, authorisation, restricted data handling, logging and the compliance monitoring of respective security measures. (EHDS Regulation, Article 73)
Topic	An "***entity*** (3.1.13.27) used as a subject of a work (3.2.1.07)." (ISO 5127:2017(en), 3.2.1.17)
Trusted health data holder (TDH)	Member State designated health data holder for whom a simplified procedure can be followed for the issuance of data permits. Trusted health data holders leverage their expertise on the data they hold to assist the health data access body by providing assessments of data requests or access applications. Once data permits are authorised, these trusted data holders provide the data within a secure processing environment that they manage. (EHDS Regulation, Article 72 and Recital 76)
Trusted third party (TTP)	A ***pseudonymisation entity*** which is independent from the data user, data holder and possibly the HDAB, that processes identifiers into pseudonyms. (ENISA, Pseudonymisation techniques and best practices, p. 10, *modified*). The TTP needs only to know the identifiers of the data subjects on the basis of which it will compute the pseudonyms, and no other data (EDPB Guideline 01/2025, §126).
Underlinkage (false negatives/missed match)	When ***matching*** health records belonging to the same data subjects are not ***linked***.

Annex 5 – Data description template

Disclaimer

This template is a suggestion and open to adaptation. It might be too simplistic for different complex linkage scenarios or not feasible for other linkage scenarios.

Permit decision diary number:

Linking organisation:

Contacts regarding data: *Contact information*

Date when linkage was performed (possibly with version nr.):

Data description

- Description of the study / target population:
- Number of unique persons in the data: N =
- Direct identifying variable names:
- Description of data processing and editing before data linkage:
- Data linkage method (e.g., share information from the decision tree with the user):
- Variable(s) used for data linkage:
- Aggregate information about any identified linkage errors:
- Logs:
- Used data format:

Included datasets

Name of the data file	Number of rows in the data	Number of columns (variables) in the data	Number of persons in the data
Data holder 1			
Data holder 2			

Reviewing the data

Once you have received the material, please review it as soon as possible. If you find any omissions or errors, please send a message to *(contact information)*.

Optional: Correction or improvement of data in datasets by linking organisation

Date when the data correction or improvement was performed:

The following variables for linking were reviewed and corrected/improved for the quality of the linkage of the dataset to be increased:

Variable(s):

The following elements of data quality of the variables in datasets were addressed (one or more):

- Accuracy, Coherence, Completeness, Consistency, Correctness, Validity

Data files	Element of the data quality improved	Method used (manual, an AI tool)	Expected resulting degree of the quality of the of the variable(s) in %
Data holder 1 (Name of the data file)			
Variable 1			
Variable 2			
Data holder 2 (Name of the data file)			
Variable 1			
Variable 2			

Optional: Data linkage quality reporting template

Contacts regarding linkage quality assessment and dataset improvement:

Date when the linkage quality assessment was performed:

Technique used for assessment of linkage quality:

Data quality assessment	**Estimates**
Match rate: • **True match (a)** • **False match (b)** • **Missed match (c)** • **True non-match (d)**	
Sensitivity	

Specificity	
Positive predictive value	
Negative predictive value	
False match rate	
Missed match rate	
Error rate	
F-measure	
Plausibility scores (in case deterministic indirect linkage or probabilistic linkage)	
Agreement on common fields (if it was checked and the result)	
Reasons for unmatched records (e.g., due to missing identifiers or unique field values)	
Results of consistency checks	
Statistical parameters of linkage quality (select those that were calculated)	
Statistical validation of demographics in linked datasets	
Cross-checks across linkage methods (if more linkage methods were used)	

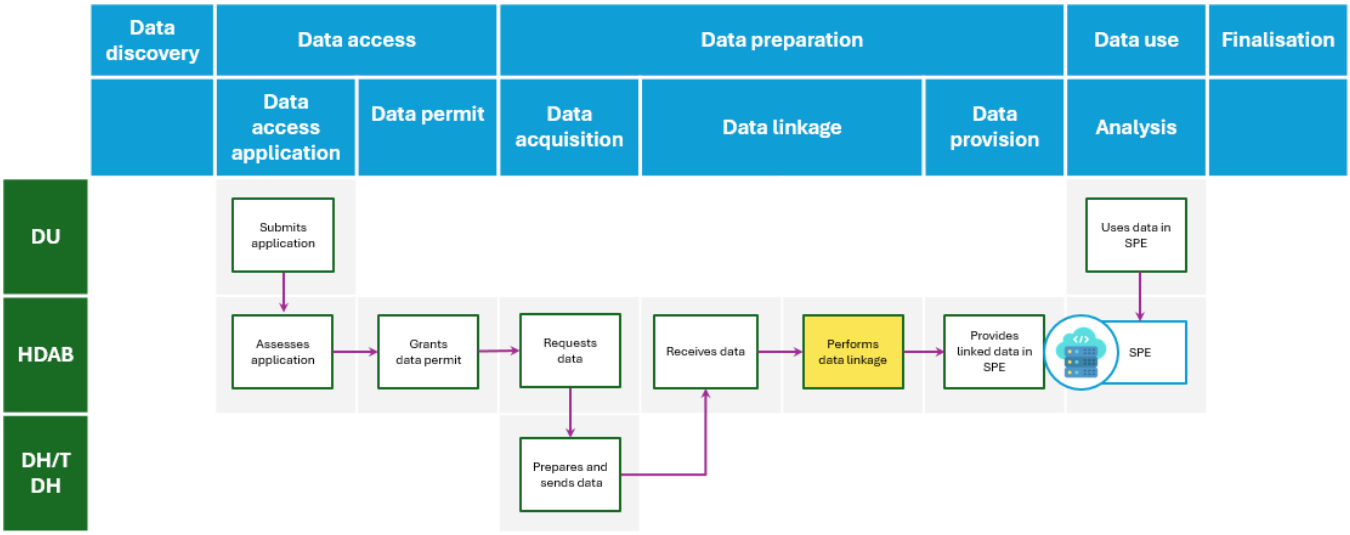
Known limitations:

Notes:

Annex 6 – Data linkage scenarios

Annex 6.1 Data access application

Scenario 1.1 – Linkage by HDAB of data from one or multiple (trusted) data holders (data access application)

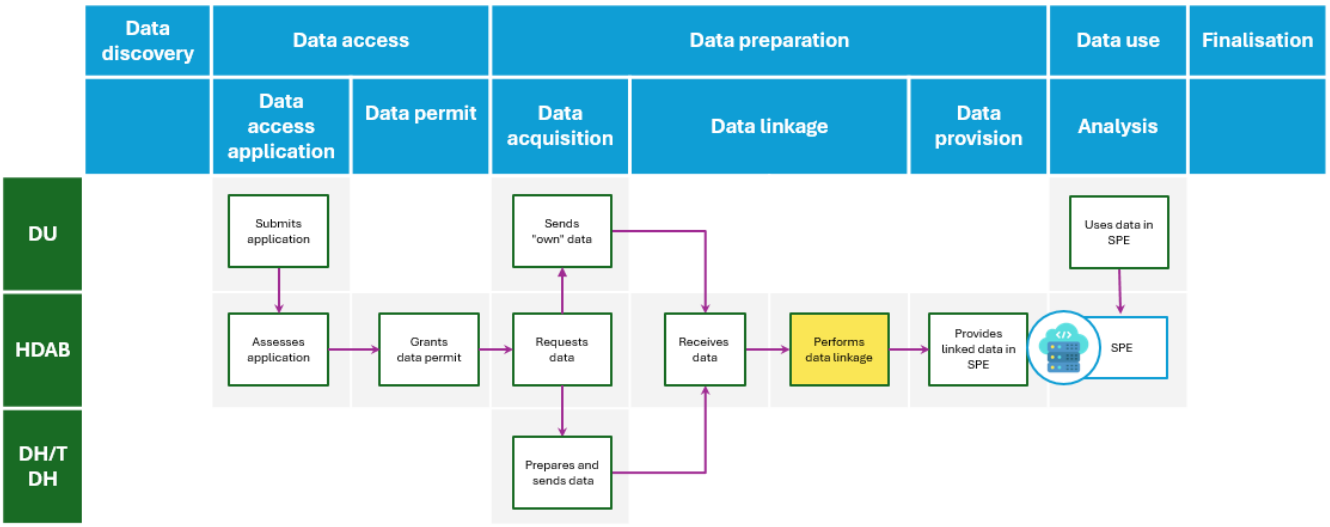


Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data application: The HDAB receives a data application submitted by the data applicant (Article 57(1)(a)).
- Assessment of data application: The HDAB assesses the data application and provides a data permit to the data applicant (Article 57(1)(a)).
- Request of data: The HDAB requests secure delivery of the permitted datasets from the relevant health data holder(s) (Article 57(1)(a) and Article 68(7)).
- Receipt of data: The HDAB receives the datasets from the data holder(s) and potentially the data users (Article 57(1)(b)).
- Data linkage: The HDAB performs data linkage (Article 57(1)(b); see also M6.2 Draft guideline for data users on good application practice for data access and requests, chapter 8.6.
- Data provision: The linked dataset is made available to the data user in the SPE provided by the HDAB (Article 68(7) and Article 73(2)).

Scenario 1.2 – Linkage by HDAB of data from (trusted) data holder(s) and data user (data access application)

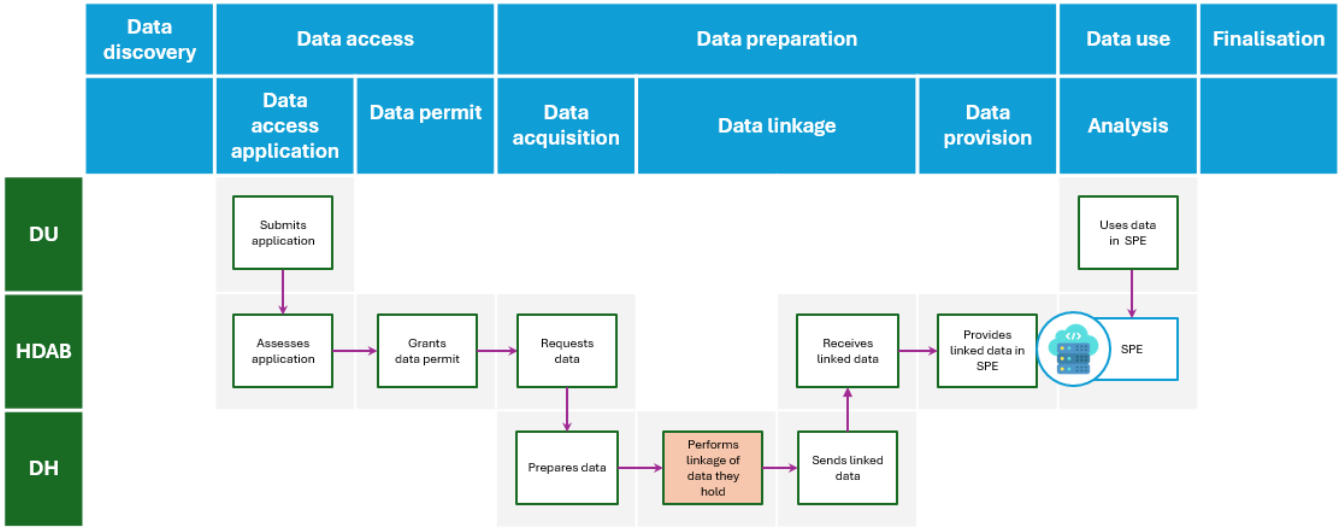


Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data application: The HDAB receives a data application, including description of the dataset in the data applicant's possession, submitted by the data applicant (Article 57(1)(a)).
- Assessment of data application: The HDAB assesses the data application and provides a data permit to the data applicant (Article 57(1)(a)).
- Request of data:
    - The HDAB requests secure delivery of the permitted datasets from the relevant health data holder(s) (Article 57(1)(a) and Article 68(7)).
    - The HDAB provides the data users with instructions on how to deliver the data in their own possession securely to the HDAB (Article 67(2)(f)); see also D6.2 Guideline for data users on good application and access practice, chapter 6.4.7).
- Receipt of data: The HDAB receives the datasets from the data holder(s) and the data users (Article 57(1)(b)).
- Data linkage: The HDAB performs data linkage (Article 57(1)(b); see also M6.2 Draft guideline for data users on good application practice for data access and requests, chapter 8.6.
- Data provision: The linked dataset is made available to the data user in the SPE provided by the HDAB (Article 68(7) and Article 73(2)).

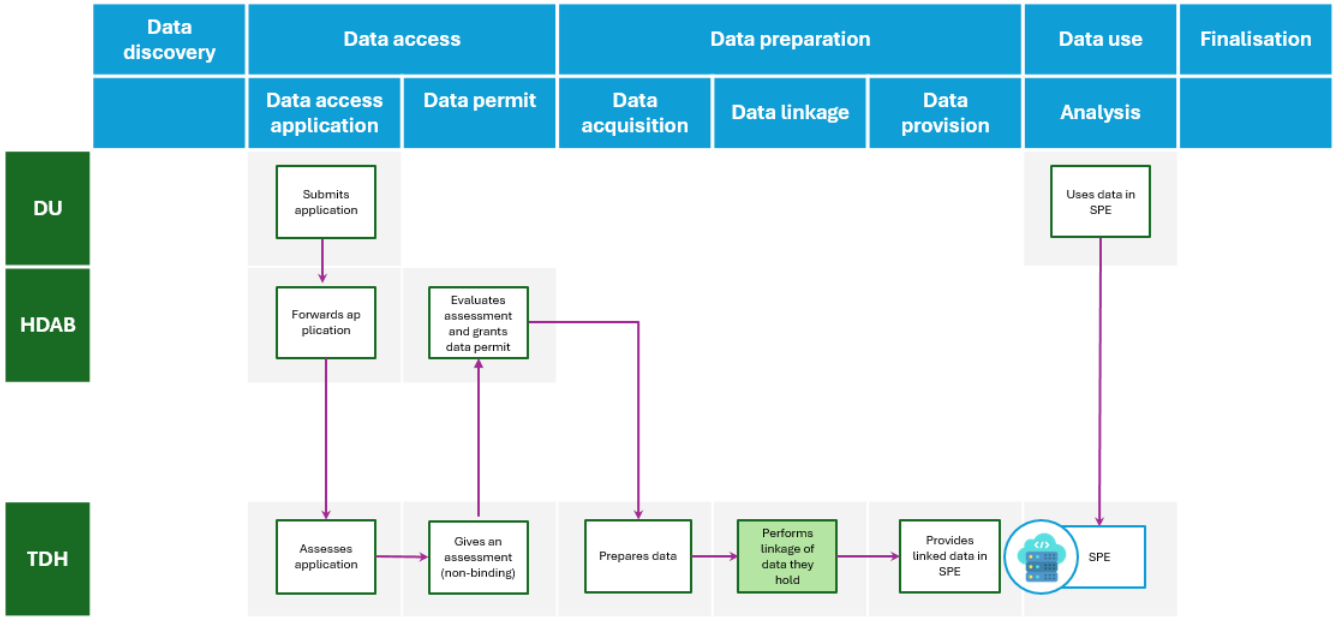
Scenario 1.3 – Linkage by data holder of data from data holder (data access application)



Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

- Procedure: Receipt of data application: In scenarios where all permitted datasets are held by one single health data holder, the data holder may receive a request for a linked dataset from the HDAB.
- Data linkage: The data holder performs the requested data linkage of the datasets they hold.
- Data transfer: The data holder transfers the linked dataset to the HDAB (Article 60(1) and (2)) to be made available for the data user in the SPE provided by the HDAB (Article 68(7) and Article 73(2)).

Scenario 1.4 – Linkage by trusted data health holder of data from a trusted health data holder (data access application)

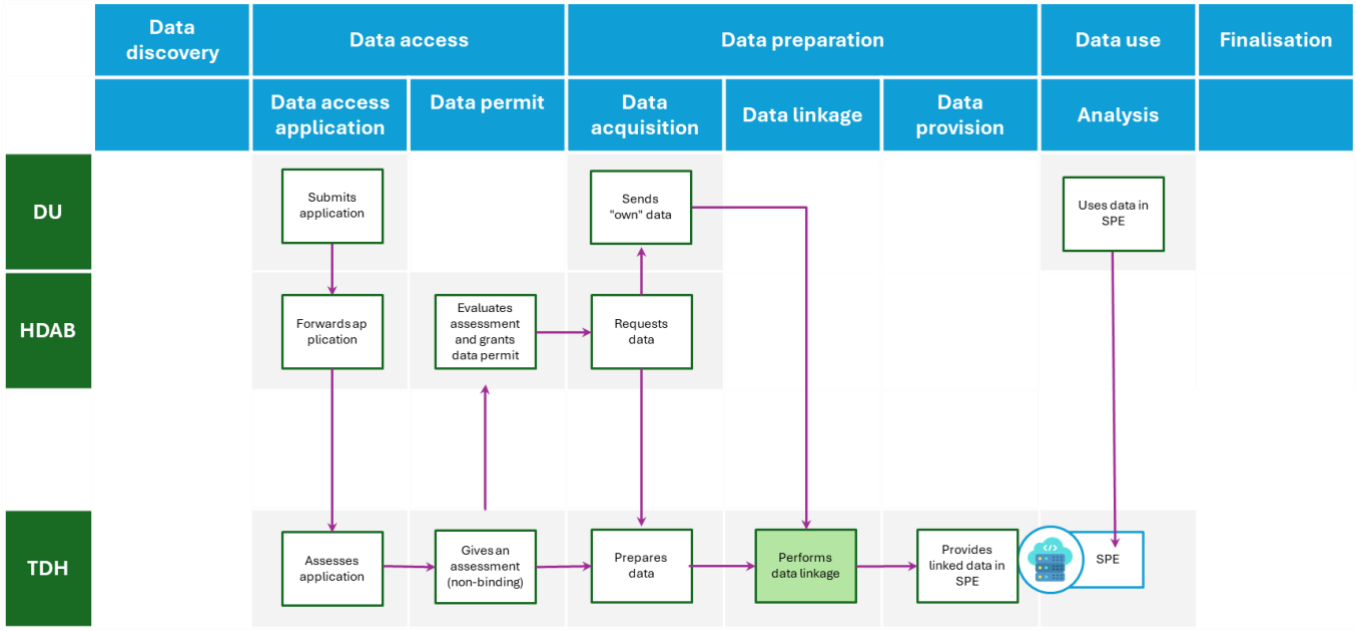


Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data application: In scenarios where datasets applied for are held by a trusted health data holder, the trusted health data holder receives the data application from the HDAB to which is has been submitted by the data applicant (Article 72; see also D6.2 Guideline for data users on good application and access practice, chapter 8.3).
- Assessment of data application: The trusted health data holder provides their assessment of the data application to the HDAB for further evaluation (Article 72).
- Receipt of data request: The trusted health data holder receives a request for provision of a linked dataset from the HDAB (Article 72(6)).
- Data linkage: The trusted health data holder performs the requested data linkage of the datasets (Article 57(1)(b)).
- Data provision: The linked dataset is made available to the data user in the SPE provided by the trusted health data holder (Articles 57(1)(a)(i) and Article 72(6)).

Scenario 1.5 – Linkage by trusted health data holder of a data from trusted health data holder and data user, under HDAB supervision (data access application)



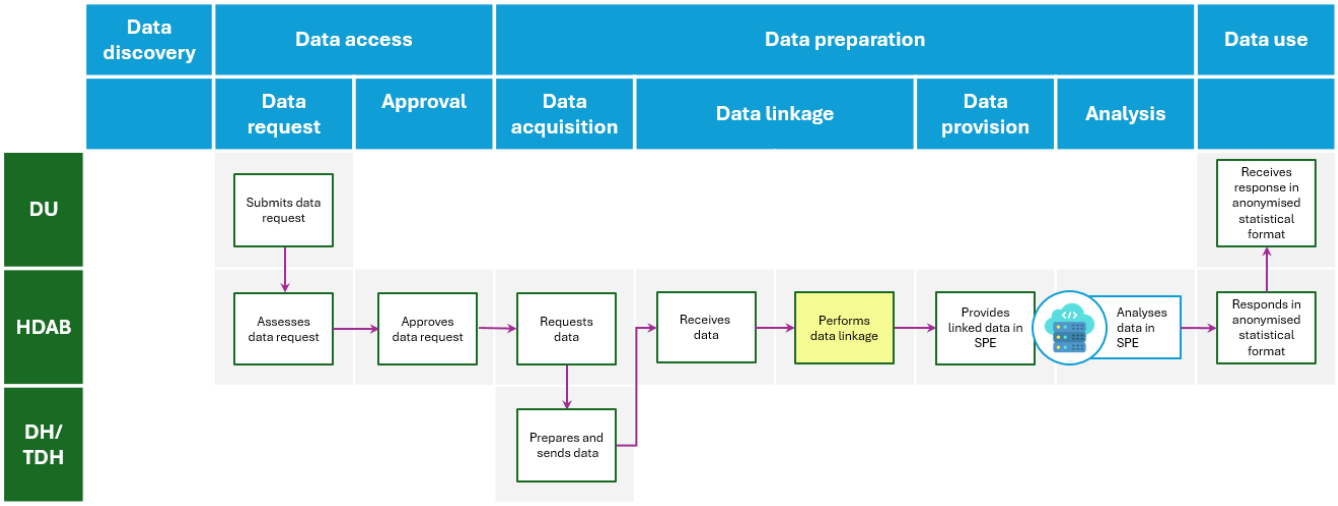
Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data application: In scenarios where datasets applied for are held by a trusted health data holder, the trusted health data holder receives the data application, including description of the dataset in the data applicant's possession, from the HDAB to which is has been submitted by the data applicant (Article 67(2)(f) and Article 72; see also D6.2 Guideline for data users on good application and access practice, chapter 8.3).
- Assessment of data application: The trusted health data holder provides their assessment of the data application to the HDAB for further evaluation (Article 72).
- Receipt of data request: The trusted health data holder receives a request for provision of a linked dataset from the HDAB (Article 72(6)).
- Receipt of data: The trusted health data holder receives the dataset in the data user's possession from the data user (Article 57(1)(b) and Article 72(6)).
- Data linkage: The trusted health data holder performs the requested data linkage of the datasets (Article 57(1)(b) and Article 72(6)).
- Data provision: The linked dataset is made available to the data user in the SPE provided by the trusted health data holder (Articles 57(1)(a)(i) and Article 72(6)).

Annex 6.2 Data request

Scenario 2.1 – Linkage by HDAB of data from one or multiple data holders (data request)

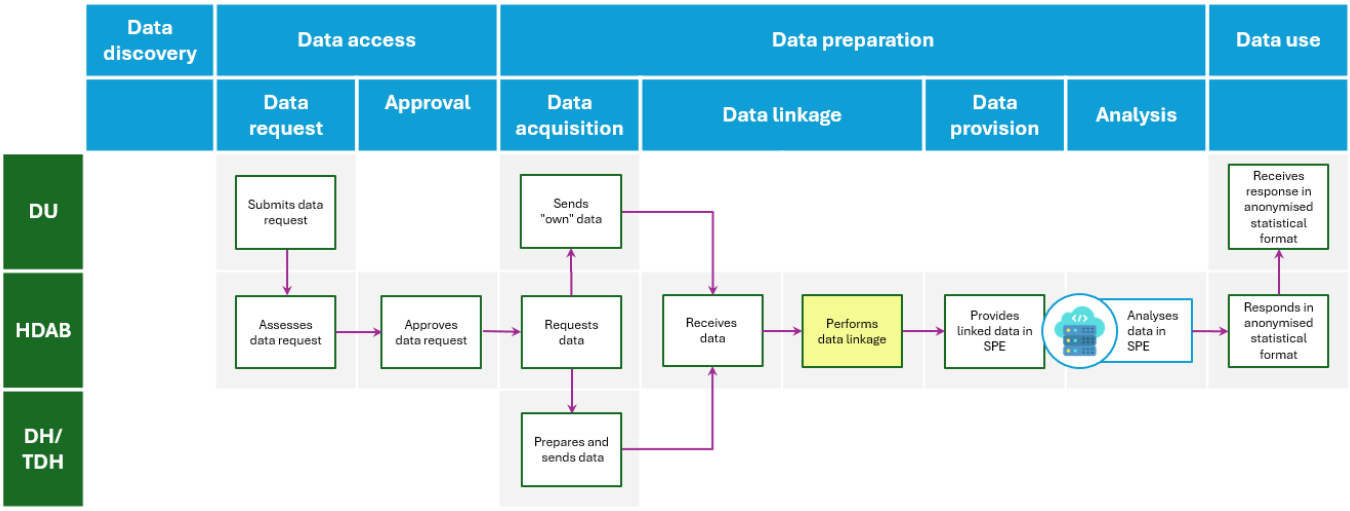


Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data request: The HDAB receives a data request submitted by the data applicant (Article 69(1) and (2)).
- Assessment of data request: The HDAB assesses the data request and provides an approval of the data request to the data applicant (Article 57(1)(a) and Article 69(3) and (4)).
- Request of data: The HDAB requests secure delivery of the permitted datasets from the relevant health data holder(s) (Article 57(1)(a) and Article 60(1)).
- Receipt of data: The HDAB receives the datasets from the data holder(s) (Article 57(1)(b) and Article 60(2)).
- Data linkage: The HDAB performs data linkage (Article 57(1)(b); see also M6.2 Draft guideline for data users on good application practice for data access and requests, chapter 8.6.
- Data provision: The linked dataset is made available to HDAB authorised personnel in the SPE provided by the HDAB (Article 57(1)(b)).
- Data analysis: The authorised personnel perform data analysis in the HDAB internal technical solution for secure data processing.
- Finalisation: The HDAB provides the data user with a response to the data request in an anonymised statistical format (Article 69(4)).

Scenario 2.2 – Linkage by HDAB of data from data holder(s) and data user (data request)



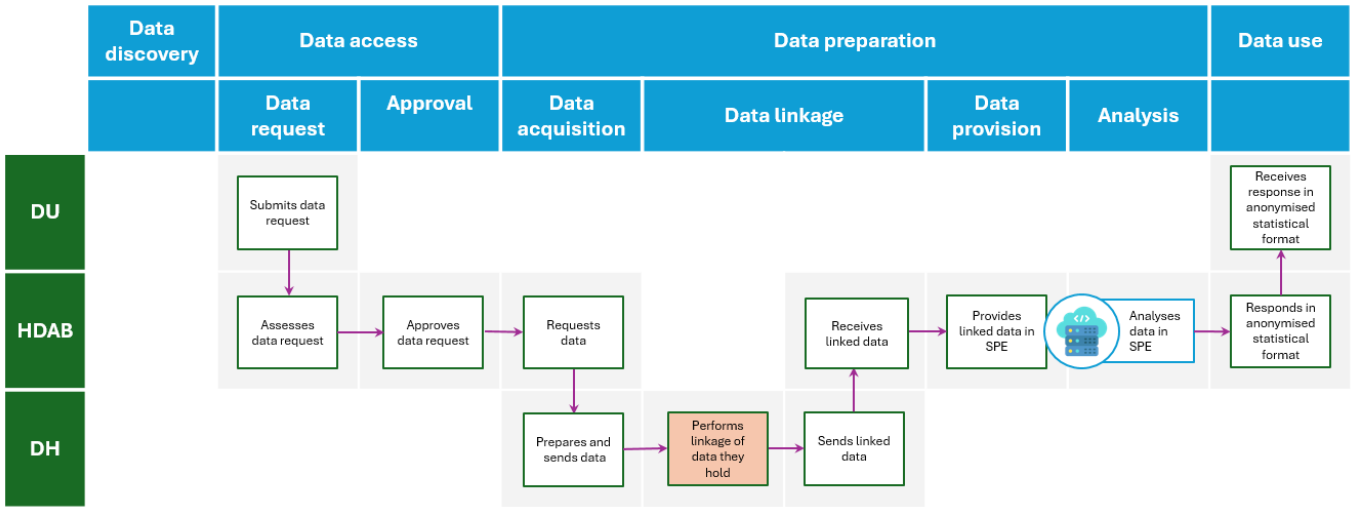
Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data request: The HDAB receives a data request, including description of the dataset in the data applicant's possession, submitted by the data applicant (Article 69(1) and (2))).
- Assessment of data application: The HDAB assesses the data request and provides an approval of the data request to the data applicant (Article 57(1)(a) and Article 69(3) and (4)).
- Request of data:
    - The HDAB requests secure delivery of the permitted datasets from the relevant health data holder(s) (Article 57(1)(a) and Article 60(1)).
    - The HDAB provides the data users with instructions on how to deliver the data in their own possession securely to the HDAB (Article 67(2)(f)); see also D6.2 Guideline for data users on good application and access practice, chapter 6.4.7).
- Receipt of data: The HDAB receives the datasets from the data holder(s) and the data users (Article 57(1)(b) and Article 60(2)).
- Data linkage: The HDAB performs data linkage (Article 57(1)(b); see also M6.2 Draft guideline for data users on good application practice for data access and requests, chapter 8.6.
- Data provision: The linked dataset is made available to HDAB authorised personnel in the SPE provided by the HDAB (Article 57(1)(b)).

- Data analysis: The authorised personnel perform data analysis in the HDAB internal technical solution for secure processing.
- Finalisation: The HDAB provides the data user with a response to the data request in an anonymised statistical format (Article 69(4)).

Scenario 2.3 – Linkage by data holder of data from data holder (data request)

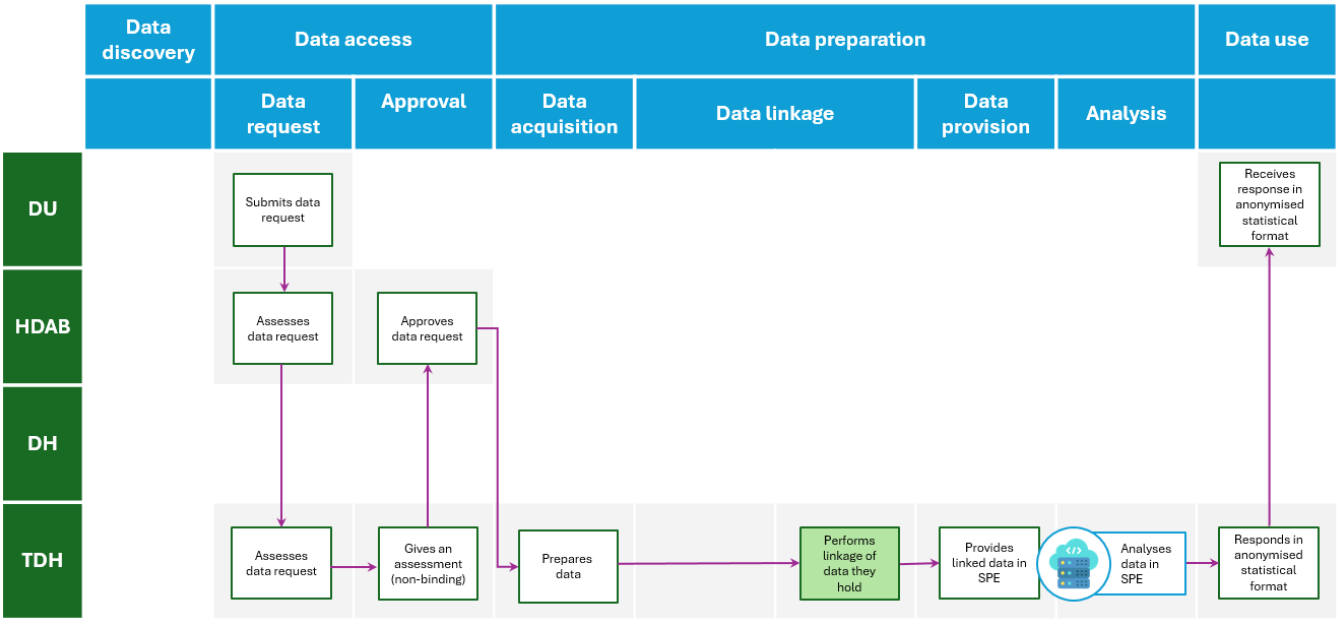


Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data request: In scenarios where all permitted datasets are held by one single health data holder, the data holder may receive a request for a linked dataset from the HDAB (Article 60(1)).
- Data linkage: The data holder performs the requested data linkage of the datasets they hold.
- Data transfer: The data holder transfers the linked dataset to the HDAB (Article 60(2)).
- Data provision: The linked dataset is made available to HDAB authorised personnel in the SPE provided by the HDAB (Article 57(1)(b) and Article 60(2)).
- Data analysis: The authorised personnel perform data analysis in the HDAB internal technical solution for secure data processing.
- Finalisation: The HDAB provides the data user with a response to the data request in an anonymised statistical format (Article 69(4)).

Scenario 2.4 – Linkage by trusted health data holder of data from trusted health data holder (data request)



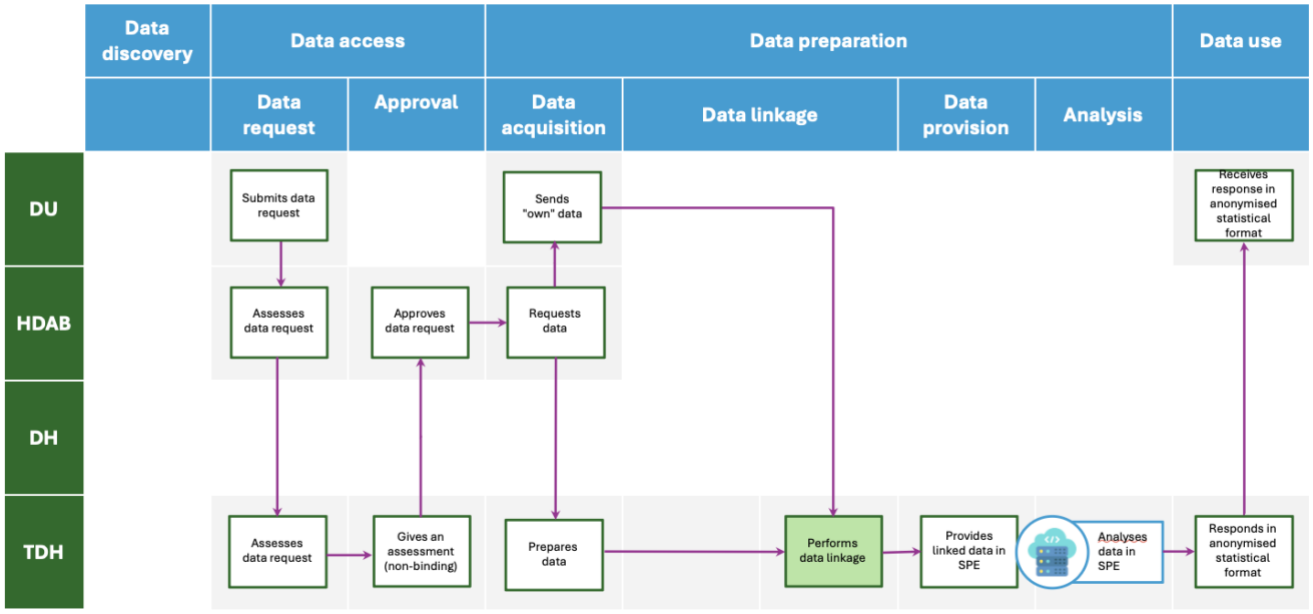
Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data request: In scenarios where the date request pertains to datasets held by a trusted health data holder, the trusted health data holder receives the data request from the HDAB to which is has been submitted by the data applicant (Article 67(2)(f) and Article 72; see also D6.2 Guideline for data users on good application and access practice, chapter 8.3).
- Assessment of data request: The trusted health data holder provides their assessment of the data request to the HDAB for further evaluation (Article 72).
- Receipt of data request: The trusted health data holder receives a request for provision of a response to the data request from the HDAB (Article 72(6)).
- Data linkage: The trusted health data holder performs the requested data linkage of the datasets (Article 57(1)(b)).
- Data provision: The linked dataset is made available to the trusted health data holder authorised personnel in the SPE provided by the trusted health data holder (Article 57(1)(b)).
- Data analysis: The authorised personnel perform data analysis in the trusted health data holder internal technical solution for secure data processing.

- Finalisation: The trusted health data holder provides the data user with a response to the data request in an anonymised statistical format.

Scenario 2.5 – Linkage by trusted health data holder of data from trusted health data holder and data user (data request)



Abbreviations: DU: Data user; HDAB: Health Data Access Body; DH: Data Holder; TDH: Trusted Data Holder.

Procedure:

- Receipt of data request: In scenarios where date request pertains to datasets held by a trusted health data holder, the trusted health data holder receives the data request, including description of the dataset in the data applicant's possession, from the HDAB to which is has been submitted by the data applicant (Articles 67(2)(f) and Article 72; see also D6.2 Guideline for data users on good application and access practice, chapter 8.3).
- Assessment of data request: The trusted health data holder provides their assessment of the data request to the HDAB for further evaluation (Article 72).
- Receipt of data request: The trusted health data holder receives a request for provision of a response to the data request from the HDAB (Article 72(6)).
- Receipt of data: The trusted health data holder receives the dataset in the data user's possession from the data user (Article 57(1)(b)).
- Data linkage: The trusted health data holder performs the requested data linkage of the datasets (Article 57(1)(b)).
- Data provision: The linked dataset is made available to the trusted health data holder authorised personnel in the SPE provided by the trusted health data holder (Article 57(1)(b)).

- Data analysis: The authorised personnel perform data analysis in the trusted health data holder internal technical solution for secure data processing.
- Finalisation: The trusted health data holder provides the data user with a response to the data request in an anonymised statistical format.

Annex 7 – Excursion on linkage quality assessment methods

Explanation of three data linkage assessments methods that require access to identifiers is taken from Doidge et al. (2024) and amended. Description of methods that do not require access to identifiers can be found in Doidge et al. (2024).

1. The gold standard technique – **training data** (a subset of data where the true match status is known) – is useful in situations, when relevant training data are available, where the true match status of records is known. Such data is only ever available for a subset of records (or else linkage would not be required). Training data is sometimes derived from a subset of records that have additional information available for matching and can be used to estimate rates and distributions of linkage errors in the remaining records, or from a subsample of records that have been manually reviewed or otherwise determined to be matches (or non-matches), or from a representative synthetic dataset (e.g. generated through simulating data). Gold standard datasets allow us to identify where errors have occurred in the linkage; where there was failure in linking records that should have been linked (missed matches); or where we have linked together records belonging to different entities (false matches). Gold standard data should be linked in the same way as the requested data (for comparison), so in case of linking on personal IDs it shall be performed by authorised personnel that has access to the identifying data; thus, involvement of the linking organisation is required.

   When gold standard data are used *to train linkage algorithms*, the set are assumed to be representative in terms of the joint distribution of errors and natural discrepancies in the values of matching variables. When gold standard data are used to assess linkage quality *for analysis of linked data*, the set are assumed to be representative in terms of the joint distribution of matching variable quality *and any analysis variables*. In practice, training data often struggle to meet these ideals and the extent to which they do is generally untestable.

2. Another approach that is commonly used also to generate a gold standard is **clerical review** (human decision-making about link status). Other than being resource-intensive, the main limitation of clerical review is the data available to support it. Humans can only outperform linkage algorithms if provided with sufficient or supplementary matching data (e.g. original records, or additional variables). If the matching data are insufficient (e.g., missing) then neither human nor algorithm will be able to accurately classify match status. The sheer number of non-matching record pairs is another factor; nearly always limiting clerical review to pairs linked by an algorithm or those classified as likely candidates (e.g. having match weights just below the threshold for acceptance). Matching pairs that have substantial missing data or substantially inconsistent data will be thoroughly hidden in the proverbial haystack of non-matches and never submitted for review. For these reasons, clerical review is more useful for estimating precision (by identifying false links) than for estimating recall (identifying missed links).

3. **Negative controls** technique is based on records that are known not to have matches but are intentionally submitted to the linkage process to assess the rate of false links. Time and date dimensions often create implausible combinations of records. This approach to estimating specificity relies on being able to identify a set of non-matching records that is representative in terms of matching variable quality and requires a high degree of confidence in the quality of timestamps or other variables that imply implausibility. Provided that the information governance framework supports the use of negative controls (i.e., records that may be only

indirectly relevant to an intended analysis), negative control records can be submitted alongside the rest of records to be linked.

Required inputs and potential outputs for various techniques are presented in Table 4.

Reference: Doidge, J., Christen, P., & Harron, K. (2024). *Quality assessment in data linkage*. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

Annex 8 – Template for instructions for data users regarding upload of data they possess

Direct linkage

- Pre-conditions:
  - Data permit/request must be approved
  - A unique identifier (e.g., social security nr) must be available in the datasets to be linked
- Direct linkage can be performed on the unique identifier
- Afterwards the documentation and quality checks are shared with the data user

Indirect linkage

- Pre-conditions:
  - Data permit/request must be approved
  - Collected information about the datasets to be linked must be shared with the linking organisation
    - Information included:
      - The name of the datasets (i.e., sources) to be linked;
      - The file information (see Table 5);
      - The variable information (see Table 5).

Table 5. Information about files and variables that should be provided by the data user to the entity that performs the linkage when the data user wants to bring their own data.

Question	Options/examples
File information	
What is the name of the source this file is stemming from?	Can be a name or a link to a platform with publicly available information
What is the format?	CSV, PNG, JSON, XML, DCM, JPEG, TIF, MHA, Other (specify), etc.
A short description	Diagnostic table, DICOM results, graphical analysis of air pollution, etc.
Does the file name need pseudonymisation? Please choose.	suppression, masking, generalisation, n/a
Variable information (can be adapted, but strongly recommended for primary linking variables)	
What is the file name this variable is in?	Table_A.csv
What is the variable name?	Exam_type, centre_region, image_height, image_width, subject_age, subject_sex etc.
Is this variable a primary linking variable?	True, False
What is the description of the variable?	Type of exam, Region of the medical centre, Height of the image in pixels (only for 2D pictures), Age of the subject, etc.
Give an example of a value.	A, Normandie, 250000000, 48, etc.
What is the variable type?	String, long string, character, boolean, date, date string, integer, float, unstructured text, n/a, etc.
Category of sensitive variable information.	Address of birth, address of residence, email address, IP address, date of consultation, start/end of the consultation, date of death, date of birth,

	initials, initials of the health care professional, social security number, name of the treating hospital, name, last name, name and last name of the treating health care professional, ID of the health care professional, number of the medical record, sex, phone number, URL (professional website, user profile, …), any other ID linked to this individual, etc.
If there are variables with sensitive information, what are the pseudonymisation rules?	Delete, mask, generalise date (JJJJ/MM), generalise geolocation (state, region), do not generalise, DICOM (tag as empty, absent, masked, ...), n/a, etc.
Is the justification to use this variable valid?	Yes, no, needs justification

- Topics that should be discussed or clarified before the data linkage procedure:
  - Possibly a short presentation of the project;
  - List of primary linking variables;
  - List of secondary linking variables;
  - Questions (for example):
    - How reliable is the information provided for linkage?
    - Are there extreme cases that hold an increased re-identification risk? How should they be treated?
    - Can the same subject appear under several identifiers?
    - Etc.
  - Steps to prepare the data for the linkage:
    - Select only the variables used for matching;
    - Unify subject identifiers in case of multiple occurrences of the same subject (if possible);
    - Remove outliers;
    - Quantify missing values and define how they are considered in the linking process;
    - Document data & data quality;
    - Set up table in csv format.
  - Clarification on the indirect matching procedure:
    - 1) Data holder carries out the quality control of the linking table and provides any necessary documentation.
    - 2) Transfer the linking table to the entity that performs the linkage.
    - 3) The project lead decides on the steps and the variables to be used to perform the indirect linkage.
    - 4) Inspection of the intermediate results and reporting to the data user (if needed multiple iterations of steps 3 and 4 are needed).
    - 5) Finalisation of the linkage process; unique matches are kept.

Annex 9 – Use cases

9.1 Linking claims data and cancer registry data in Germany: Lessons learned from record linkage approaches

Use case 1

Objective
To investigate in- and out-patient services, diagnoses and the utilisation of drugs, claims data can be very informative. However, claims data lack information on tumor grades or laboratory tests, which can be found in cancer registries.

Background
In Germany, there are 16 cancer registries (i.e., one per state, one for kids, while the cancer registry for Berlin and Brandenburg is combined). Moreover, there is a central cancer registry where the other cancer registries must regularly report to. In the past, cancer registries could document unique identifiers, such as the health insurance number for administrative purposes, but sharing them with other cancer registers fell under strict legal and data protection specified in the respective federal cancer registry laws. Therefore, they often had to rely on quasi-identifiers for individual level record linkage. Current and future frameworks, however, enable and support the usage of unique identifiers, and there are also methods available to enhance the successful linkage rate when no unique identifier is available.

Evidence from two German studies can be found below.

Study 1: Record linkage using indirect identifiers
In an exemplary study Lendle and colleagues assessed multiple linkage methods (i.e., 7 in total) when linking claims data from a database with data from two cancer registries (i.e., Bremen and Lower Saxony) and elaborated on linkage failures for golden standard links (i.e., naive deterministic 1:1 matching). The study used the variables birth year, sex, area of residence, date of diagnosis, ICD10-GM code, and vital status to link individuals with colorectal cancer. Due to data privacy, only a randomly selected 50% sample could be provided by the cancer registry of Lower Saxony, which complicated implementing prediction models.

From 641 cases with colorectal cancer in the claims database and 42,781 cases from the cancer registries, 199 golden standard links were found, of which 106 cases had to be eliminated due to missing or nonconforming data. The method with the highest F*-measure, recall and precision values was a gradient boosting algorithm.
8% of the individuals in the claims database and 33% from the two cancer registries could not be uniquely identified based on the selected indirect-identifiers.
Reference
Lendle et al. (2025), https://pubmed.ncbi.nlm.nih.gov/40130753/

Study 2: Privacy-preserving linkage using trusted third parties
The DigiNet study by Kästner et al. implemented a privacy-preserving linkage infrastructure using a trusted third party to connect claims data from several statutory health insurance funds with cancer registry data. The study demonstrated that secure linkage across multiple institutions is technically feasible while maintaining data protection requirements and enabling high-quality linkage for oncological research.

The results indicated that linking the data from three cancer registries in Germany with one of three statutory health insurances that were participating, while applying a privacy-preserving record linkage (PPRL), led to more than 90% successfully linked records.

Reference
Kästner et al. (2025), https://link.springer.com/article/10.1007/s00432-025-06384-7

Conclusion
Record linkage based on indirect identifiers remains a valuable methodological alternative when unique identifiers are unavailable and has proven useful in several research settings. However, it is inherently affected by incomplete matches and linkage uncertainty. Please be aware that future linkage scenarios that involve claims data and/or data from other (cancer) registries in Germany will be able to use unique identifiers in collaboration with a trusted third party.

9.2 AI development using cross-country national health data

Use case 2

Chronic Kidney Disease (CKD) Prediction Model – a SHAIPED use case exploring the transportability of an existing AI/ML model on different cohorts from various countries.

Objective
The aim of this use case, conducted as part of the SHAIPED project, is to test the transportability and adaptability of an advanced artificial intelligence (AI)/machine learning (ML) model across multiple European member states. The prediction model, KDPredict, was originally developed and internally tested in Canada and externally tested in Scotland and Denmark (https://doi.org/10.1136/bmj-2023-078063) to predict the 5-year risk of kidney failure and death (all-cause mortality) in patients with chronic kidney disease (CKD) stage G3b-4, based on readily obtainable clinical variables.
The use case aims to examine the challenges of transferring a prediction model based on AI/ML between EU member states and to develop and evaluate the capacity of Health Data Access Bodies (HDABs) within the HealthData@EU infrastructure. The use case is crucial in demonstrating AI model transportability between Secure Processing Environments (SPEs), providing validated, predictive AI models that support chronic disease management across diverse populations in Europe.

Project stakeholders

- Denmark (coordinator)
  - Research team: Aarhus University Hospital
  - SPE provided by the Danish Health Data Authority (DHDA, Danish HDAB equivalent)
- Finland
  - Research team: THL – Finnish Institute for Health and Welfare
  - SPE provided by THL and Findata (Finnish HDAB equivalent; the coordinating HDAB)

- France
  - Research team: Clinical Epidemiology Team of the Centre for Epidemiology and Population Health, Inserm unity 1018
  - SPE provided by Health Data Hub (French HDAB equivalent)

Study populations

- Denmark and Finland
  - Cohorts of patients with newly documented CKD stage 3b or 4
    - Identified in population-based laboratory registries
- France
  - Cohort of patients with well-established CKD diagnosis, stages 3a –4, under nephrology care
    - Identified and recruited from a nationally representative sample of 40 nephrology facilities as part of the Chronic Kidney Disease-Renal Epidemiology and Information Network (CKD-REIN) cohort study (https://doi.org/10.1093/ndt/gft388) and hosted by France Cohortes (a service unity of the Inserm).

Data requirements

The use case requires various types of data, including:

- demographics (age and sex),
- vital status,
- laboratory tests (outpatient creatinine and albuminuria measurements),
- comorbidities (diabetes, cardiovascular disease, cancer, and chronic lung disease),
- kidney failure (long-term kidney replacement therapy, kidney transplantation, and based on outpatient creatinine measurements).

Data linkage

The use case will not involve any linking of individual-level data between the participating countries. Instead, the analyses will be federated, i.e., the datasets will be processed and analysed in the SPE of each respective HDAB, and only aggregated results will be transferred to the coordinating HDAB in Denmark.

The linkage process within each country comprises:

- Denmark
  - The study population will be identified using the laboratory database (https://doi.org/10.2147/clep.s380840) and linked with other registries (https://doi.org/10.2147/clep.s91125, https://doi.org/10.1093/ije/dyw213, https://doi.org/10.1007/s10654-014-9930-3) in the DHDA SPE to identify predictors and outcomes.
  - Linkage method: Direct linkage with records matching using a pseudonymised linkage variable (Data Protection Policy - The Danish Health Data Authority, Sundhedsdatastyrelsen: Pseudonymiseringsprincipper for sundhedsdata til statistikproduktion-v1.0.docx [in Danish]).
    - Linkage variable: The Civil Personal Register (CPR) number (https://doi.org/10.1007/s10654-014-9930-3), a unique identifier included in all Danish administrative and medical registers and

databases (https://doi.org/10.2147/clep.s179083, TEHDAS Denmark country visit factsheet).

- Finland
  - The study population will be identified using the laboratory database and linked with other registries held by within THL's SPE to identify predictors and outcomes.
  - Linkage method: Direct linkage with records matching using a pseudonymised linkage variable.
    - Linkage variable: THL pseudonymised linkage variable (TEHDAS Finland country visit factsheet).
- France

The study population has already been identified and enrolled in the CKD-REIN cohort study (https://doi.org/10.1093/ndt/gft388). Selected variables from the study database will be transferred to the SPE of the Health Data Hub by Secure File Transfer Protocol (SFTP).Within the CKD-REIN cohort framework, a passive follow-up of participants has been implemented through record linkage with national registries and databaseshttps://doi.org/10.1093/ndt/gfi198https://frdata.org/docs/pdf/DatapaperFHMD_04_04_2022.pdf. No additional data linkage is required for the present use case.

- Linkage methods:
  - Linkage with the national kidney replacement therapy registries (https://doi.org/10.1016/j.nephro.2022.01.004) and vital status are based on a direct, deterministic approach relying on personal data. The percentage of successful linkage to either database is 100% and 95%, respectively.
    - Linkage variables: full name, birthdate, sex, and address.

https://doi.org/10.1093/ndt/gft388https://doi.org/10.1186/s12882-020-1692-4https://doi.org/10.1016/j.respe.2017.05.004https://doi.org/10.2196/36711

Statistical methods

Each participating country will be responsible for the initial data management, ensuring the dataset follows a common structure. Each country will then apply a common data model to standardise variables and allow for comparisons across datasets.

Initially, each country will validate the original KDpredict model by applying the matrix of predicted risk estimates to the respective cohorts. Prediction accuracy, calibration, and discrimination will be assessed.

Following, an AI model based on a super-learner algorithm will be deployed within the SPE of each of the contributing HDABs. The model will integrate multiple ML models, including the predictor variables, to select the best-performing one for each dataset.

The performance of the models will be examined and aggregated results will be compared at the coordinating HDAB.

Reference

Chronic Kidney Disease (CKD) Prediction Model - Shaiped

9.3 Data linkage in the case of ready-made datasets

Use case 3

Findata's ready-made datasets

Background
Finnish Social and Health Data Permit Authority Findata provides ready-made datasets that are pre-compiled and pre-processed datasets which can be applied for by submitting a data permit application to Findata. Ready-made datasets are available quickly and easily without the need of data extraction by original holders. Findata is the data holder of the ready-made datasets. Findata offers ready-made datasets on specific themes. Currently, Findata provides two options: a dataset collection based on registry data from the FinRegistry research project, and a COVID-19-themed ready-made dataset.

Data and study population
Findata's FinRegistry ready-made dataset is based on registry data collected in the FinRegistry research project (Home | FinRegistry). Dataset includes people who lived in Finland 1.1.2010 and their spouses, children, parents and siblings. The dataset includes information from the following sources, insofar as the data is covered by the Act on the Secondary Use of Health and Social Data:

- Digital and Population Data Services Agency
- Cancer Registry
- Finnish Centre for Pensions
- Kanta Services
- Kela (Social Insurance Institution)
- Finnish Institute for Health and Welfare
- Statistics Finland

The COVID-19 dataset includes people who fell ill with COVID-19 in the HUS (Helsinki University Hospital) area in 2020–2021. The target group is formed based on the Infectious Disease Register. The COVID-19 dataset includes information from the following sources:

- Finnish Institute of Health and Welfare
- Kanta Services
- Kela (Social Insurance Institution)
- Finnish Medicines Agency
- Statistics Finland

Data linkage and preprocessing
Data in the ready-made datasets are linked using the Finnish personal identification code which is a unique identifier. Data are linked using direct linkage method. Datasets are tailored and pseudonymised separately for every permit holder. Ready-made datasets can also be linked with other data.

Individuals, who have objected to the processing of their data according to the Article 21 of the GDPR, will be removed from the ready-made datasets. These removals are done at regular intervals and every time when data are extracted and provided to the secure processing environment (SPE) for the analyses conducted by permit holders. Therefore, study populations are constantly changing because of individuals using their right to object the processing of their data.

Reference
Ready-made datasets - Findata

9.4 Data linkage with a Trusted Third Party

Use case 4

Deterministic linkage using a Trusted Third Party in Belgium: example from HISLink

Objective

A Trusted Third Party (TTP) can act as a neutral intermediary that enables deterministic linkage while ensuring that no single actor has access to both raw identifiers and the pseudonyms.

Background

The HISlink project in Belgium joined data from the Belgian Health Interview Survey (BHIS) and from the Belgian Compulsory Health Insurance (BCHI). It provides an operational example of this approach, based on a strict separation between linkage and analysis roles.

Linkage process

1. Selection of individuals: survey participants are identified using a stable personal identifier (the Belgian National Register Number)
2. Secure transfer to the TTP: identifiers are encrypted and transmitted to the TTP through controlled and audited channels
3. Generation of common pseudonyms: the TTP replaces the original identifier with a project-specific pseudonym.
    - The same person receives the same pseudonym across all datasets
    - Pseudonyms are unique to the project and cannot be reused anywhere else
4. Dataset linkage: the entity responsible for the linkage procedure receives datasets where identifiers are removed and records are labelled only with the project pseudonym. Because all datasets now share the same pseudonym, records can be linked deterministically without disclosing the original identifiers to the linking or analysing parties
5. Access for research: researchers access the linked, pseudonymised dataset in a secure environment.

Data protection and governance aspects

Using a TTP prevents any actor from holding both identifiers and attributes while operationalising GDPR principles such as data minimisation, confidentiality and purpose limitation. It also helps building trust among data holders and citizens. A centralised TTP facilitates the monitoring, correction and re-execution of linkage procedures when linkage quality issues are detected.

Practical considerations

Some practical considerations from the HISlink project:

- TTP governance and legal mandate must be clearly defined
- Pseudonyms should be project-specific and time-limited
- Linkage processes are operationally complex and require early planning
- Researchers should be informed about linkage quality and potential biases

Recent developments

More recently, the Blinded Pseudonymisation REST service provided by the Belgian eHealth platform further strengthens and simplifies the procedure as it removes the need for successive identifiers by blinding the pseudonymisation operation from the TTP and encrypting the pseudonym from the data holder.
Reference

Linkage of Health Interview Survey Data with Health Insurance Data | sciensano.be

Pseudonymisation & Anonymisation | Platform eHealth

Annex 10 – Data quality considerations

Indirect linkage: Data quality assessments are handled on a case-by-case basis. Sometimes variables with common information can be compared between two databases for the linked patients, but this depends on which variables were available and used for the linkage process. Standardised methods are not yet available as it is highly dependent on the external data that needs to be linked.

For example, the Centre for Health Record Linkage (CHeReL) in Australia uses the software ChoiceMaker to convert linkage weights to probabilities ranging from 0 to 1, while 1 is a definite match and 0 a definite non-match. They suggest a procedure where linkage starts with default cut-offs of:

Upper cut-off p=0.75
Lower cut-off p=0.25

Then, these cut-offs are adjusted until the upper cut-off (i.e., the false positive rate) is below 5 per 1,000, or 0,5% as well as the lower cut-off (i.e., the false negative rate) is below 5 per 1,000[21].

Direct linkage (e.g., on social security numbers): Duplicates are assessed in the data and compared with the pseudonyms that are available. In the following, three scenarios are provided as examples. It should be noted that more complex scenarios might arise when the number of datasets that should be linked increases. Moreover, the data user should be able to indicate their preference on how to deal with the duplicates.

- Variable 1: Cohort_Number
- Variable 2: Pseudonym_1 (e.g., pseudonym created from the social security number)
- Variable 3: Pseudonym_2 (e.g., pseudonym created from the social security number in addition to date of birth and sex)

1. Examination of the data provided when investigating pairs of Pseudonym_1 and Pseudonym_2 result in [insert number here] duplicates in Cohort_Number.

Cohort_Number	Pseudonym_1	Pseudonym_2
24323	PSA1	ANO1
34234	PSA1	ANO1
23352	PSA2	ANO2
33445	PSA2	ANO2
00231	PSA3	ANO3
00005	PSA3	ANO3
…	…	…

- This suggests that different Cohort_Numbers point to the same individual in source data.

To-do: Align with data user if the Cohort_Numbers should be grouped into a single identifier.

[21] https://www.cherel.org.au/quality-assurance

2. Examination of the data provided when investigating duplications of Pseudonym_2 pointing towards different individuals in the sample cohort resulted in [insert number here] duplicates in Chohort_Number.

Cohort_Number	Pseudonym_1	Pseudonym_2
24323	PSA1	ANO1
34234	PSA2	ANO1
23352	PSA3	ANO1
33445	PSA4	ANO1
00231	PSA5	ANO2
00005	PSA6	ANO2
…	…	…

- This suggests that different Cohort_Numbers point to the same individual in source data if relied on the Pseudonym_2.

To-do: Align with data user how to proceed.

Note: The same may occur for duplicates in Pseudonym_1 that point to the same individual but are assigned different Cohort_Numbers. [Adapt this template if this is the case.]

3. Examination of the variable Cohort_Number pointing to pairs of Pseudonym_1 and Pseudonym_2 results in [insert number here] different Cohort_Numbers.

Cohort_Number	Pseudonym_1	Pseudonym_2
24323	PSA1	ANO1
24323	PSA2	ANO2
33445	PSA3	ANO3
33445	PSA4	ANO4
00005	PSA5	ANO5
00005	PSA6	ANO6
…	…	…

- This suggests that identical Cohort_Numbers point to different individuals in source data.

To-do: Align with data user how to proceed.

Note: In some cases, it is for example possible that the corresponding Pseudonym_2 values are not matching, while the Pseudonym_1 values match. In other words, the pseudonymisation of Pseudonym_1 is compatible, but not for Pseudonym_2. This could stem from differences in the pseudonymisation processes of the trusted third party (TTP) (or, possibly multiple TTPs) that create the pseudonyms from the social security number. Another option is missing data on the components of Pseudonym_2 (date of birth, sex). Depending on the application and the organisational structure (i.e., what

data sources are involved, who performs the linkage, involvement of one or more TTPs), more scenarios can emerge.

Annex 11 – Scenarios data linkage with TTP

The scenarios presented in this document provide a starting point for discussing possible data linkage and pseudonymisation approaches in the context of the SPE. They illustrate different configurations with a particular focus on data minimisation, separation of information and the distribution of trust between actors.

These scenarios should not be considered exhaustive or as a definitive list of possible solutions. Other architectures, combinations of the approaches presented here, or alternative technical solutions may be considered. The objective is to provide a common basis for discussion and comparison, rather than to prescribe a specific implementation. In comparing the scenarios, particular attention should be given to which actors have access to identifying information and health data, and whether these informational powers remain separated. Implementation and operational complexity are also considered but should be assessed alongside (and not independently from) the data protection and security benefits provided by each architecture.

Separation of informational powers is considered effective when no single actor, acting within its normal technical capabilities and access rights, can independently associate identifying information with the corresponding health data beyond what is strictly necessary for its assigned role. Such an assessment should therefore consider not only the direct access to identifiers and health data, but also whether an actor can:

- Correlate or link records across datasets
- Access pseudonymisation, encryption or other cryptographic secrets
- Convert pseudonyms in a way that restores linkability
- Initiate or perform re-identification without involvement or authorisation from the required actors
- Combine information obtained through separate operations, systems, or stages of the process to reconstruct the relationship between an individual and their health data.

The purpose of introducing a trusted third party in the following scenarios is to reduce the informational power of each actor involved in the data linkage process by separating access to identifiers, pseudonyms, and the associated data. Additional pseudonymisation measures are intentionally abstracted from the scenarios. These may include, for example, introducing multiple layers of pseudonymisation, separating the pseudonymisation secrets, or applying additional transformations to make unauthorised re-identification more difficult and limit the impact of a data leakage. The need for such additional layers depends on the level of trust placed in the different actors, the threat model, and the technical and organisational measures surrounding the processing. At the same time, additional pseudonymiastion layers should remain proportionate to the identified risk and operational requirements. Over-complexifying the process may render the system unsuitable for operational use or make the reversal of pseudonymisation difficult in legitimate and authorised scenarios where re-identification is required. The diagrams therefore focus on the distribution of information and responsibilities between actors, rather than on all cyptographic and security measures that would be required in a concrete implementation.

In the scenarios, the concept of a “data preparation environment” is introduced. It is intentionally kept abstract and should be understood as the environment in which data preparation activities required for a data permit take place, such as data minimisation, anonymisation and pseudonymisation (see D7.2). These activities may be performed by the HDAB, a (trusted) data holder, or more generally by the entity responsible for the relevant data preparation tasks. The data

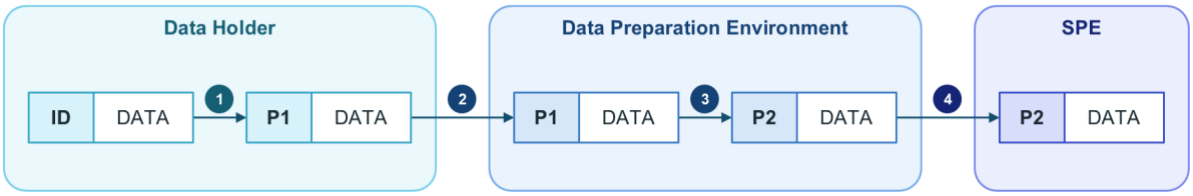
preparation environment therefore does not necessarily correspond to a single technical component or physical environment. It could, for example, be implemented as a dedicated workspace using technology similar to the secure processing environment, or encompass other trusted components involved in preparing and transferring the data, including secure data transfer pipelines.

For the purpose of illustrating the scenarios, the following assumptions are made:

A. We are working with tabular data.
B. Pseudonyms are generated per approved data request or access application.
C. Where a scenario requires returning to the original identifier, the relevant actor (the DH or a TTP) is capable of depseudonymising the existing pseudonym.
D. Where identifying information and health data follow separate flows, DHs have the ability to send only the identifier or pseudonym required for the linkage operation, rather than the whole dataset. A mechanism exists to subsequently associate the resulting pseudonym with the corresponding data, for example using row numbering.

"Who sees what?" refers to information available in clear/usable form to each actor. Cryptographically protected intermediate values may still transit through an actor.

Scenario 1: 1 data holder, no TTP

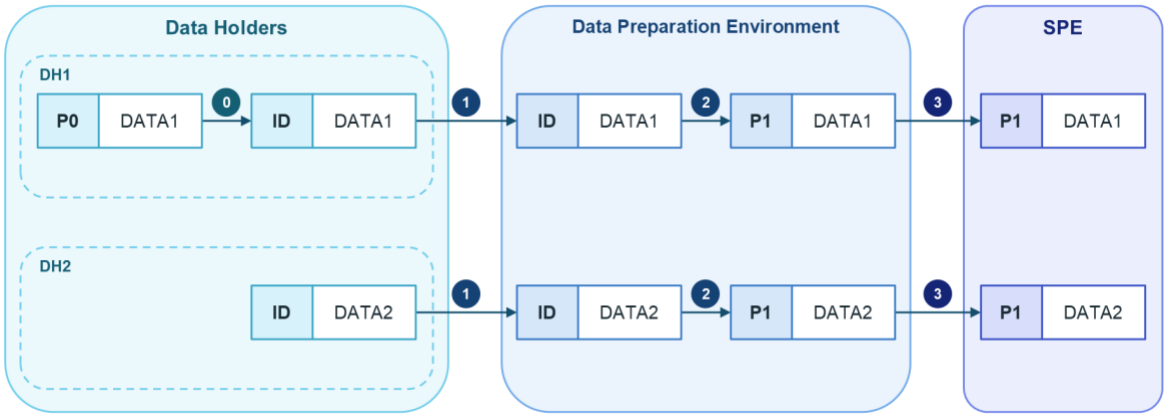


1. **Pseudonymisation at the data holder**: the data holder replaces the original identifier ID with a first pseudonym P1. DATA remains unchanged.
2. **Transfer to the data preparation environment**: the data holder sends [P1 | DATA] to the data preparation environment.
3. **Second pseudonymisation in the data preparation environment**: P1 replaced with a new pseudonym P2.
4. **Transfer to SPE**: [P2 | DATA] transferred to the SPE.

Who sees what?

	Data Holder	**Data Preparation Environment**	**SPE**
Identifiers	ID + P1	P1 + P2	P2
Data	DATA	DATA	DATA

Scenario 2: 2+ data holders, no TTP

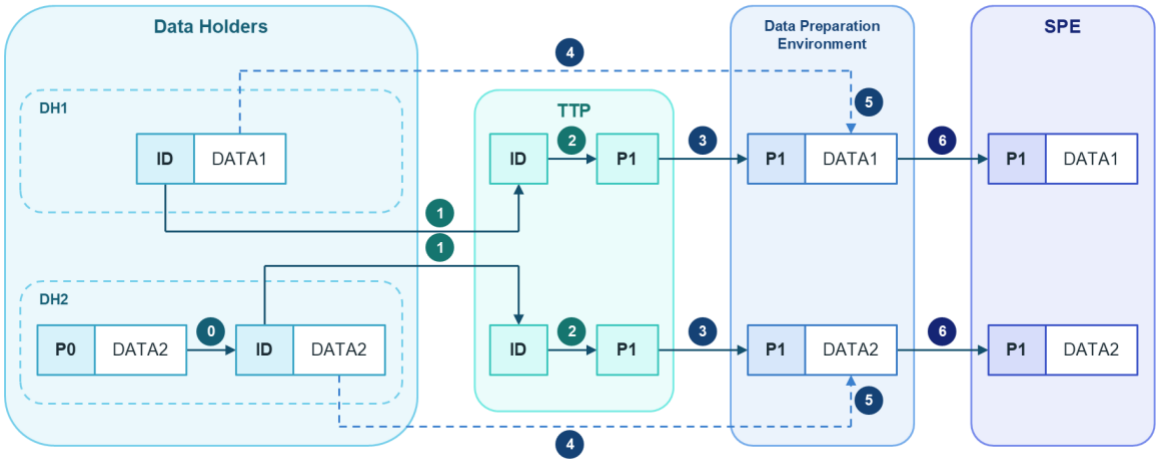


0. **Depseudonymisation at source where necessary:** if a data holder initially holds the identifier under an existing pseudonym (P0), it first converts it back to the original ID.
1. **Transfer to data preparation environment:** each data holder sends its dataset containing the original identifier to data preparation environment.
2. **Pseudonymisation in the data preparation environment:** the original ID is replaced with a pseudonym P1.
3. **Transfer to SPE:** [P1 | DATA] transferred to the SPE.

Who sees what?

	DH1	**DH2**	**Data Preparation Environment**	**SPE**
Identifiers	P0 + ID	ID	ID + P1	P1
Data	DATA1	DATA2	DATA1 + DATA2	DATA1 + DATA2

Scenario 3: 2+ data holders, with TTP



0. **Depseudonymisation at source where necessary:** if a data holder initially holds the identifier under an existing pseudonym (P0), it first converts it back to the original ID.
1. **Transfer of identifiers to TTP:** each data holder sends the original ID to the TTP. The associated data remains at the data holder.
2. **Pseudonymisation by TTP:** the TTP replaces the original ID with a pseudonym P1.
3. **Transfer of pseudonyms to the data preparation environment:** the TTP sends P1 to the data preparation environment.
4. **Transfer of data to the data preparation environment:** each data holder sends its corresponding DATA directly to the data preparation environment, separately from the identifier flow.
5. **Recombination in the data preparation environment:** P1, received from the TTP, is associated with the corresponding DATA received from the data holder.
6. **Transfer to SPE:** [P1 | DATA1] and [P1 | DATA2] transferred to the SPE.

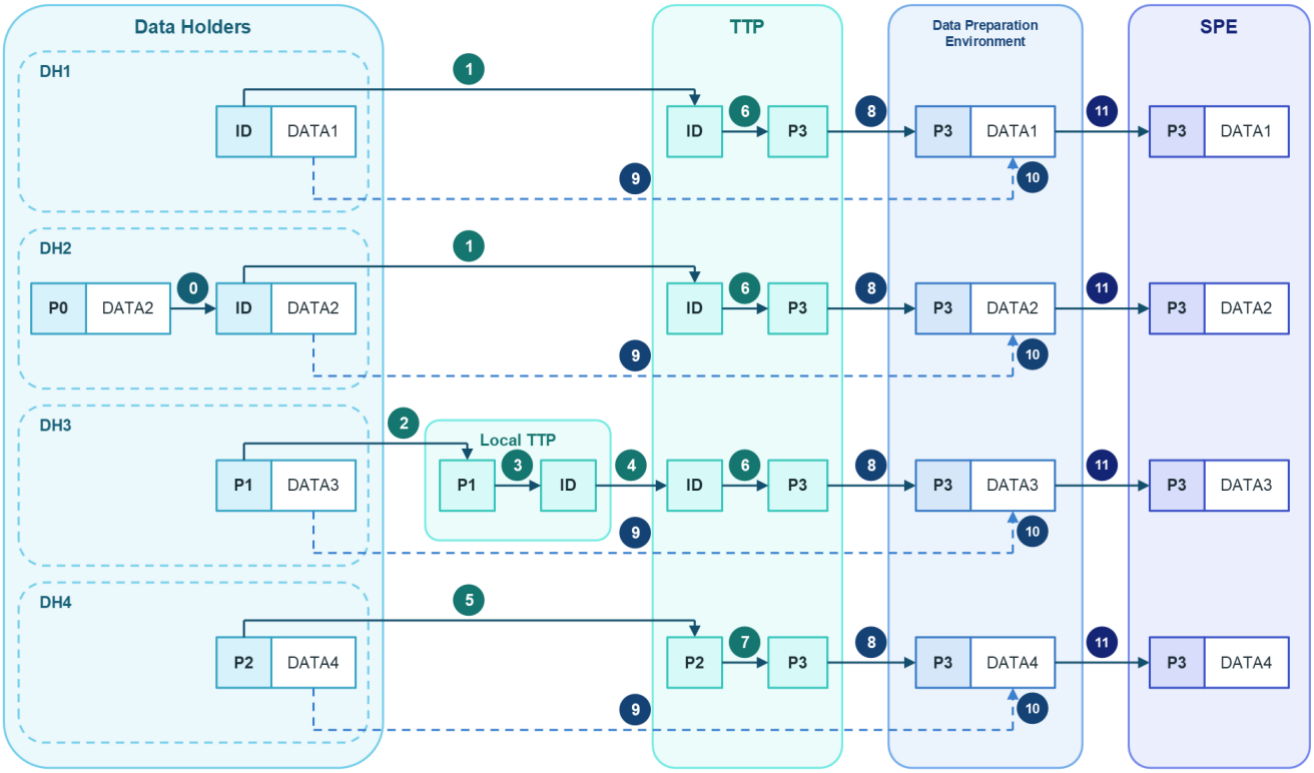
Who sees what?

	DH1	DH2	TTP	Data Preparation Environment	SPE
Identifiers	ID	P0 + ID	ID + P1	P1	P1
Data	DATA1	DATA2	—	DATA1 + DATA2	DATA1 + DATA2

Scenario 3.1: Heterogeneous starting configurations

This scenario extends Scenario 3 to account for the fact that data holders may not all start with the same type of identifier. Depending on their existing pseudonymisation setup, different paths are therefore used to obtain the common data access application pseudonym. The four starting configurations are:

- DH1 — Original ID: the data holder already holds the original ID and sends it directly to the TTP.
- DH2 — Locally reversible P0: the data older starts with P0 but can depseudonymise it locally to recover the original ID before interacting with the TTP.
- DH3 — P1 managed by a Local TTP: the data holder cannot itself recover the original ID. P1 is sent to its Local TTP, which converts it back to ID and communicates it to the central TTP.
- DH4 — P2 managed by the central TTP: the existing P2 was generated by the same TTP used for the data linkage. Depending on the cryptographic capabilities of the pseudonymisation system, the TTP can either convert P2 directly to P3 without recovering the original ID, or depseudonymise P2 to the original ID and then repseudonomise it as P3 in the new data access application-specific pseudonymisation space.



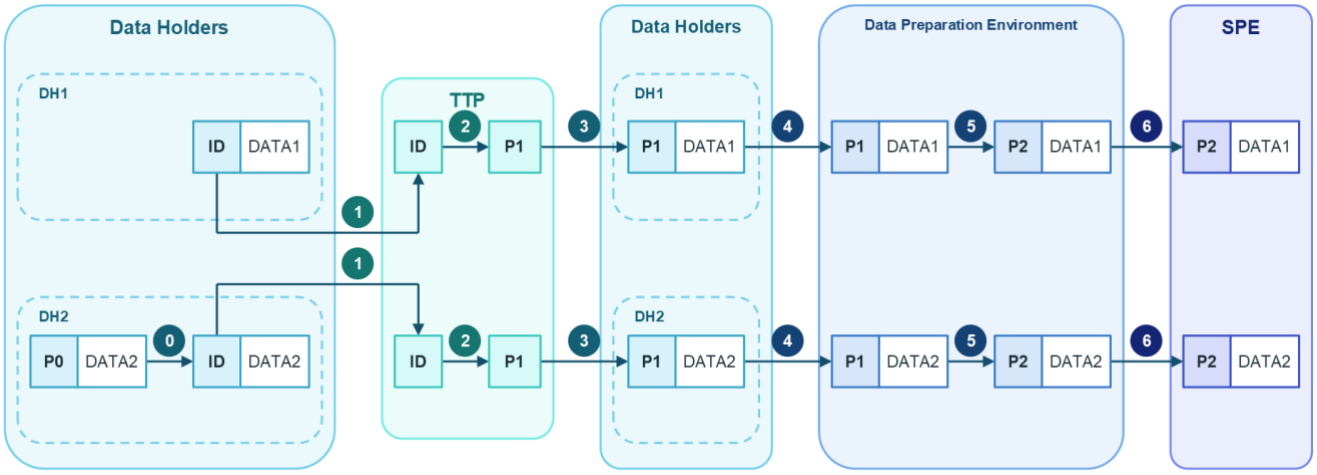
0. **Local depseudonymisation by DH2:** DH2 converts its existing pseudonym P0 back to the original ID.
1. **Transfer of IDs to TTP:** DH1 and DH2 send their original IDs to the central TTP.
2. **Transfer to local TTP:** DH3 sends its existing pseudonym P1 to its Local TTP.
3. **Depseudonymisation by Local TTP:** the Local TTP converts P1 back to the original ID.
4. **Transfer between TTPs:** the Local TTP sends the resulting ID to the central TTP.
5. **Transfer of P2 to TTP:** DH4 sends its existing pseudonym P2 directly to the central TTP.
6. **Pseudonymisation from ID:** for DH1, DH2 and DH3, the central TTP converts the original ID into the common data access application pseudonym P3.
7. **Transcoding from P2:** for DH4, the central TTP converts P2 directly into the common data access application pseudonym P3.
8. **Transfer of pseudonyms to the data preparation environment:** the central TTP sends P3 to the data preparation environment for each data holder.
9. **Transfer of data to the data preparation environment:** each data holder sends its corresponding DATA directly to the data preparation environment, separately from the identifier flow.
10. **Recombination in the data preparation environment:** P3, received from the TTP, is associated with the corresponding DATA received from the data holder.

11. **Transfer to SPE:** [P3 | DATA1], [P3 | DATA2], [P3 | DATA3] and [P3 | DATA4] transferred to the SPE.

	DH1	DH2	DH3	DH4	Local TTP	TTP	Data Preparation Environment	SPE
Identifiers	ID	P0 + ID	P1	P2	P1 + ID	ID + P2 + P3	P3	P3
Data	DATA1	DATA2	DATA3	DATA4	—	—	DATA1 + DATA2 + DATA3 + DATA4	DATA1 + DATA2 + DATA3 + DATA4

Scenario 4: 2+ data holders, with TTP (who only communicates with DHs)

This scenario explores an alternative TTP-based architecture in which the TTP communicates only with the data holders and does not interact directly with the data preparation environment. The TTP pseudonymises the identifiers and returns the resulting pseudonyms to the Data Holders, allowing them to associate P1 with their data before transfer. Compared with Scenario 3, this shifts the recombination step from the data preparation environment to the data holders and exposes P1 to the data holders. It is therefore included primarily as an alternative organisational configuration rather than an equivalent solution in terms of separation of informational powers.



0. **Depseudonymisation at source where necessary:** if a data holder initially holds the identifier under an existing pseudonym (P0), it first converts it back to the original ID.
1. **Transfer of identifiers to TTP:** each data holder sends the original ID to the TTP. The associated data remains at the data holder.
2. **Pseudonymisation by TTP:** the TTP replaces the original ID with a pseudonym P1.

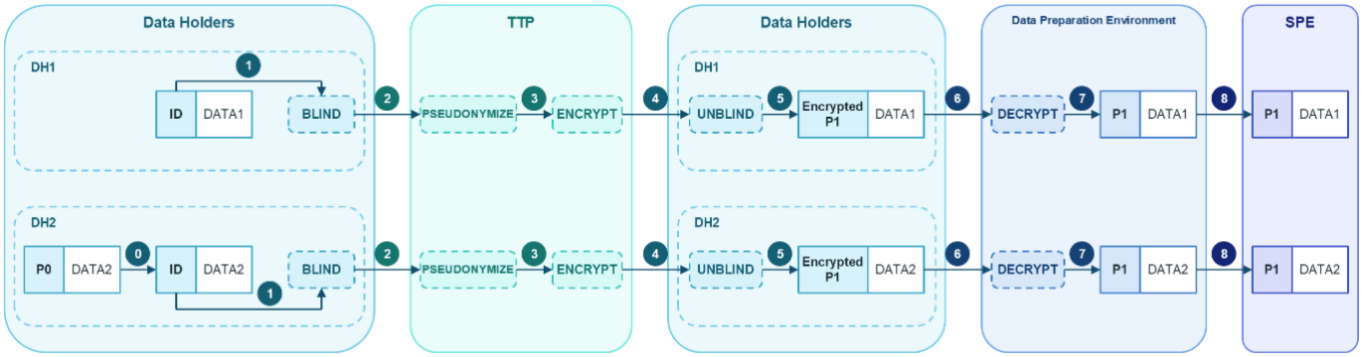
3. **Return of pseudonyms to data holders:** the TTP sends P1 back to the corresponding data holder.
4. **Transfer to the data preparation environment:** each data holder sends its dataset [P1 | DATA] to the data preparation environment.
5. **Pseudonymisation in the data preparation environment:** P1 is replaced with a second pseudonym P2.
6. **Transfer to SPE:** [P2 | DATA1] and [P2 | DATA2] transferred to the SPE.

Who sees what?

	DH1	DH2	TTP	Data Preparation Environment	SPE
Identifiers	ID + P1	P0 + ID + P1	ID + P1	P1 + P2	P2
Data	DATA1	DATA2	—	DATA1 + DATA2	DATA1 + DATA2

Scenario 5: 2+ data holders, with TTP (PPRL using blind pseudonymisation)

This scenario illustrates a Privacy-Preserving Record Linkage (PPRL) approach based on the Blind Pseudonymisation service developed by the Belgian eHealth Platform[22][23][24][25]. The approach enables identifiers to be pseudonymised while preventing the pseudonymisation service from accessing either the clear identifier or the resulting pseudonym, thereby supporting separation of informational powers between the actors. The TTP uses a project-specific pseudonymisation key to generate P1 and a separate symmetric key, shared with the data preparation environment, to encrypt the blinded P1. The data holder never receives either key: it only removes the blinding layer, using its own key.



[22] https://www.ehealth.fgov.be/ehealthplatform/fr/service-pseudonymisation-anonymisation
[23] Verslype, K., De Decker, B. (2025). Privacy-By-Design in the Belgian Public Sector. In: Goossens, J., Keymolen, E., Stanojević, A. (eds) Public Governance and Emerging Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-84748-6_6
[24] https://www.smalsresearch.be/basisprincipes-voor-een-moderne-pseudonimiseringsdienst-2/
[25] https://www.smalsresearch.be/kruisen-van-persoonsgegevens-met-ehealths-blinde-pseudonimiseringsdienst/

0. **Depseudonymisation at source where necessary:** if a data holder initially holds the identifier under an existing pseudonym (P0), it first converts it back to the original ID.
1. **Blinding at the data holder:** each data holder blinds the original ID before sending it outside its environment.
2. **Transfer to TTP and pseudonymisation:** the blinded identifier is sent to the TTP, which pseudonymises it without having access to the original ID or generated pseudonym.
3. **Encryption by TTP:** the TTP encrypts the resulting blinded pseudonym.
4. **Return to data holder:** the encrypted blinded pseudonym is sent back to the corresponding data holder.
5. **Unblinding at the data holder:** the data holder removes the blinding while the pseudonym remains encrypted and associates the resulting encrypted P1 with its DATA.
6. **Transfer to the data preparation environment:** each data holder sends [Encrypted P1 | DATA] to the data preparation environment.
7. **Decryption in the data preparation environment:** the encrypted pseudonym is decrypted, resulting in [P1 | DATA].
8. **Transfer to SPE:** [P1 | DATA1] and [P1 | DATA2] transferred to the SPE.

Who sees what?

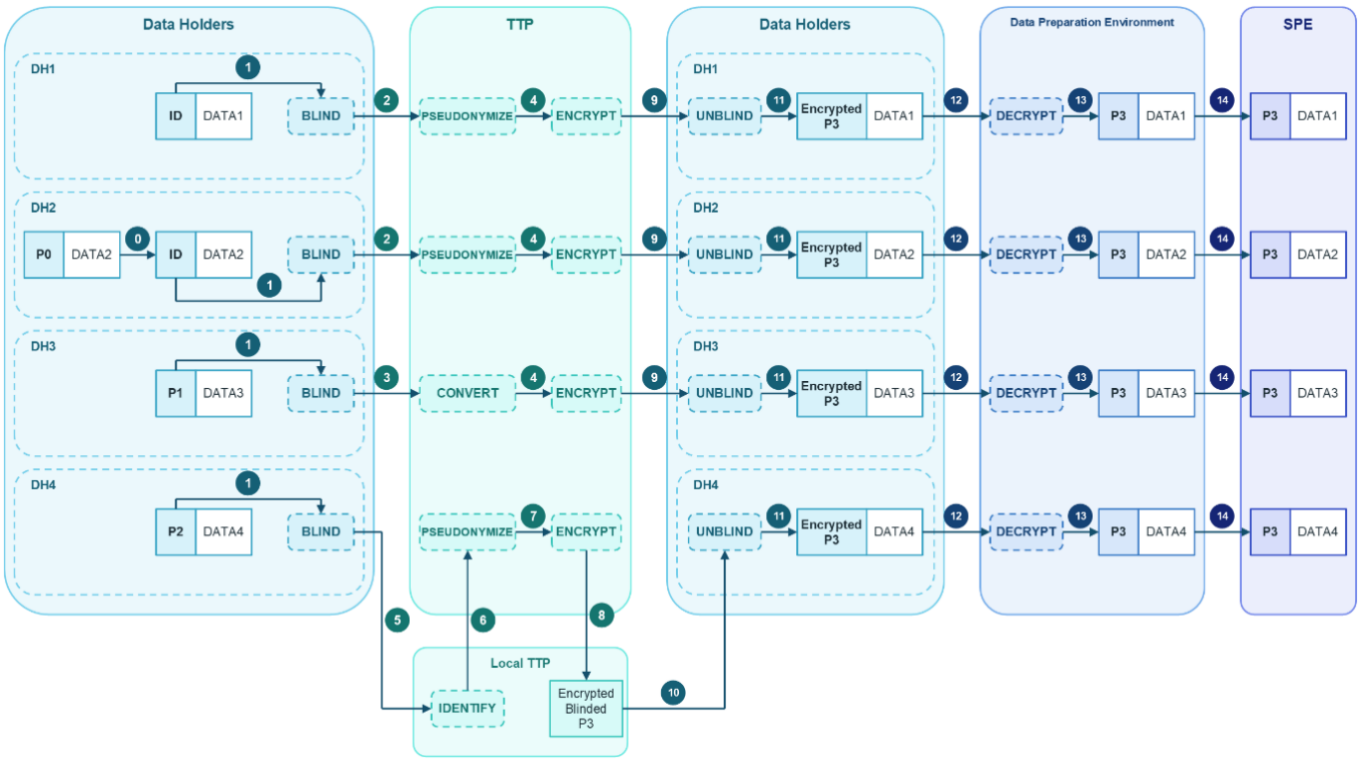
	DH1	DH2	TTP	Data Preparation Environment	SPE
Identifiers	ID	ID + P0	—	P1	P1
Data	DATA1	DATA2	—	DATA1 + DATA2	DATA1 + DATA2

Scenario 5.1: Blind pseudonymisation with heterogeneous starting configurations

This scenario extends blind pseudonymisation to account for different starting configurations. Depending on how the identifier is initially represented and which actor controls the corresponding pseudonymisation, different paths are used to obtain the common data access application pseudonym. The four starting configurations are:

- **DH1 — Original ID:** the data holder holds the original identifier.
- **DH2 — Locally reversible P0:** the data holder starts with P0 but can locally recover the original ID.
- **DH3 — P1 managed by the linkage TTP:** P1 was generated by the same TTP used for the linkage, allowing it to convert P1 to P3.
- **DH4 — P2 managed by a Local TTP:** P2 was generated by another TTP. DH4 interacts only with its Local TTP, while the Local TTP and linkage TTP cooperate to obtain P3.

In all four configurations, the objective is to produce the same data access application pseudonym **P3**, while limiting the disclosure of identifiers to the actors involved in the conversion and keeping DATA outside the TTPs.



0. **Depseudonymisation at source where necessary:** if a data holder initially holds the identifier under an existing pseudonym (P0), it first converts it back to the original ID
1. **Preparation and blinding at the data holder:** the starting identifier/pseudonym is prepared where necessary and blinded. For DH2, P0 is first locally converted back to ID. DH1 and DH2 therefore blind ID, while DH3 and DH4 blind P1 and P2 respectively.
2. **Blind pseudonymisation for DH1 and DH2:** the blinded IDs are sent to the linkage TTP, which pseudonymises them towards P3 without receiving the original IDs in clear.
3. **Blind conversion for DH3:** the blinded P1 is sent to the linkage TTP, which converts it towards P3 without receiving P1 in clear. Conversion transforms an existing pseudonym into another, without first reverting to the original identifier.
4. **Encryption by the linkage TTP:** the resulting blinded P3 values for DH1, DH2 and DH3 are encrypted.
5. **Transfer to Local TTP:** DH4 sends its blinded P2 to its Local TTP.
6. **Identification and transfer between TTPs:** the local TTP identifies the blinded P2 within its pseudonymisation domain and sends a blinded representation to the linkage TTP. The clear identifier is not exchanged between the TTPs.
7. **Pseudonymisation towards P3 for DH4:** the linkage TTP pseudonymises it into a blinded P3. From the linkage TTP's perspective, it received a blinded ID from the local TTP, just like it does from DH1 and DH2.
8. **Encryption for DH4:** the resulting blinded P3 is encrypted and returned through the Local TTP flow.
9. **Return to data holders:** the encrypted blinded P3 values for DH1, DH2 and DH3 are returned to their respective data holders.
10. **Return to DH4:** the encrypted blinded P3 for DH4 is returned via its Local TTP. From DH4's perspective, it interacts only with the Local TTP.

11. **Unblinding at the data holders:** each data holder removes the blinding while P3 remains encrypted and associates the resulting encrypted P3 with its DATA.
12. **Transfer to the data preparation environment:** each data holder sends [Encrypted P3 | DATA] to the data preparation environment.
13. **Decryption in the data preparation environment:** P3 is decrypted, resulting in [P3 | DATA].
14. **Transfer to SPE:** [P3 | DATA1], [P3 | DATA2], [P3 | DATA3] and [P3 | DATA4] transferred to the SPE.

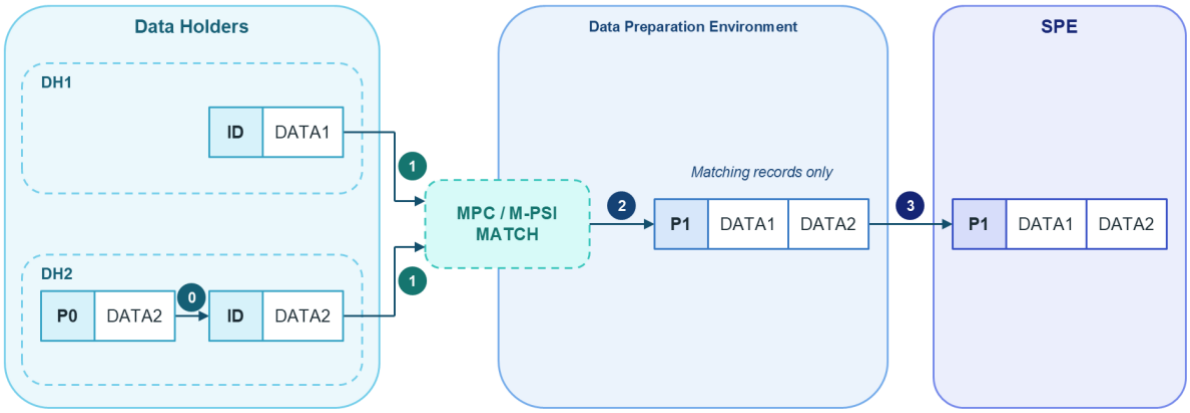
Who sees what?

	DH1	DH2	DH3	DH4	Local TTP	TTP	Data Preparation Environment	SPE
Identifiers	ID	P0 + ID	P1	P2	—	—	P3	P3
Data	DATA1	DATA2	DATA3	DATA4	—	—	DATA1 + DATA2 + DATA3 + DATA4	DATA1 + DATA2 + DATA3 + DATA4

Scenario 6: 2+ data holders, no TTP (with multi-party computation[26][27])

At each data holder, the linkage protocol produces a pseudonym and a corresponding secret key for each local record. These can then be handed locally to an encryption component or script, which uses the key to encrypt the associated data before it leaves the data holder. The key is not transferred together with the encrypted payload. Instead, as part of the linkage protocol, the data preparation environment can reconstruct the corresponding decryption key only for records that are part of the intersection. It can therefore decrypt and combine the data for matching records, while data associated with non-matching records remains encrypted. The exact technical mechanism used to hand the key from the linkage component to the local encryption component is implementation-specific and is abstracted in this scenario.

[26] https://www.smalsresearch.be/innovationsmals-croisement-des-donnees-a-caractere-personnel-dans-le-respect-de-la-vie-privee/
[27] https://doi.org/10.48550/arXiv.2512.08558



0. **Depseudonymisation at source where necessary:** if a data holder initially holds the identifier under an existing pseudonym (P0), it first converts it back to the original ID.
1. **MPC / M-PSI matching:** the data holders jointly execute the MPC / M-PSI protocol using their IDs. This identifies the records common to both Data Holders without either data holder disclosing its full set of IDs to the other. No data is shared across DHs.
2. **Transfer of matching records to the data preparation environment:** for matching records only, the protocol produces a common pseudonym P1 and combines the corresponding data into [P1 | DATA1 | DATA2], which is transferred to the data preparation environment.
3. **Transfer to SPE:** [P1 | DATA1 | DATA2] transferred to the SPE.

Who sees what?

	DH1	DH2	Data Preparation Environment	SPE
Identifiers	ID	P0 + ID	P1	P1
Data	DATA1	DATA2	DATA1 + DATA2 *(matching records only)*	DATA1 + DATA2 *(matching records only)*

Methods comparisons

Scenario	TTP	TTP sees ID	DPE sees ID	ID / DATA split	DH complexity	Flow complexity	Crypto complexity	Key mgmt.	Pros	Cons
1 — 1 DH, no TTP	No	N/A	No	No	Low	Low	Low	Limited	—	Only applicable to single-DH linkage
2 — 2+ DH, no TTP	No	N/A	Yes	No	Low	Low	Low	Limited	—	High level of trust placed in DPE
3 — TTP → DPE	Yes	Yes	No	Yes	Medium	Medium	Low	Limited	Limited burden on DHs	Recombination of pseudonyms and DATA required in DPE
4 — TTP ↔ DH	Yes	Yes	No	Yes	Medium	High	Low	Limited	Recombination occurs at DHs, before transfer to DPE	Additional interactions with DHs; possible collisions between DHs
5 — Blind pseudonymisation	Yes	No	No	Yes	High	High	High	High	No actor has access to both clear ID and resulting pseudonym	Requires support for blind pseudonymisation protocol
6 — MPC / M-PSI	No	N/A	No	Yes	Very High	Very High	Very High	High	Only matching records need to be disclosed	Implementation depends on selected MPC/M-PSI protocol