M7.2 Draft guideline on data minimisation, pseudonymisation, anonymisation and synthetic data

TEHDAS2 – Second Joint Action Towards the European Health Data Space

05 September 2025

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
Luca Augello	ARIA/RL, Italy
Petr Holub	MU, Czech Republic
Jaakko Lähteenmäki	VTT, Finland
Eva Anjo	Serviços Partilhados do Ministério da Saúde (SPMS), Portugal
Victor Leandre-Chevalier	Health Data Hub, France
Peter van Meerendonk	Nictiz, The Netherlands
Farzaneh Michaud	Health Data Hub, France
Juha Pajula	VTT, Finland
Henriette Rau	Robert Koch Institute, Germany
Lea Rizzuto	Health Data Hub, France
Lise Skovgaard Svingel	RM, Denmark
Hanna Tervonen	Findata, Finland
Steven Wolter	BfArM, Germany

0.2 Keywords

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

0.3 Document history

Date	Version	Editor	Change	Status
08/04/2025	0.1	Pia Brinkmann	Initial document creation	Draft
12/05/2025	0.2	All contributors	First draft	Draft
01/07/2025	0.3	Luca Augello, Pia Brinkmann, Jaakko Lähteenmäki, Petr Holub	Draft for consortium feedback	Draft
05/09/2025	1.0	Luca Augello, Pia Brinkmann, Jaakko Lähteenmäki	Document to be submitted for public consultation	Final

Accepted in Project Steering Group on 12 September 2025.

Contents

Abbreviations

Term	Abbreviation
Computed tomography	CT
Digital imaging and communications in medicine	DICOM
Data model	DM
Differential privacy	DP
Electrocardiograms	ECG
Electrodermal activity	EDA
European data protection board	EDPB
Electroencephalograms	EEG
Regulation (EU) 2025/327	EHDS
Electronic health record	EHR
Electromyograms	EMG
European Union	EU
Generative adversarial network	GAN
Regulation (EU) 2016/679	GDPR
Health data access body	HDAB
Intellectual property	IP
Large language models	LLM
Machine learning	ML
Magnetic resonance imaging	MRI
Named entity recognition	NER
Positron emission tomography	PET
Public use file	PUF
Secure processing environment	SPE
Trusted health data holder	TDH
Second joint action towards the European health data space	TEHDAS2
Trade secret	TS
Trusted third party	TTP
Variational autoencoders	VAE
Work package 7	WP7
Zone improvement plan	ZIP code

1 Executive summary

This guideline focuses on processing electronic health data within the European health data space (EHDS), by detailing methods for **data minimisation**, **pseudonymisation**, **anonymisation**, and **synthetic data generation**. The goal is to create a secure, interoperable, and efficient health data ecosystem for secondary use in compliance with the EHDS and General Data Protection (GDPR) regulations, which means using health data beyond direct patient care.

One foundational principle for handling health data is **data minimisation**. This means that only the **minimum amount of personal health data** that is adequate, relevant, and limited to what is necessary for a specific purpose should be processed. This principle applies throughout the entire lifecycle of the data, from when it is first collected and prepared by the health data holder, to when it is assessed by a health data access body (HDAB), and finally, during its use and processing by the health data user. Data minimisation can involve reducing the volume of data, limiting its detail (granularity), making sensitive information less specific, or restricting geographical or temporal scopes. In addition, it can be applied in five dimensions (“Who”, “What”, “When”, “Where”, “How”). This helps to significantly reduce risks related to confidentiality, integrity, and availability of data.

Pseudonymisation is another crucial technique that helps protect personal data while keeping data useful for analysis. It involves replacing direct identifying information (for example names or social security numbers) with new identifiers called pseudonyms. Pseudonymised data is one data format HDABs may permit access to, if the re-identification risk is justified and appropriately mitigated (see Article 66(3), Regulation (EU) 2025/327 (EHDS)). The information needed to link these pseudonyms back to the original individuals is kept entirely separate and secure. Pseudonymisation is particularly valuable because it allows for the **linkage of different health datasets** related to the same entity, even if they come from various sources or countries, without revealing the individual's direct identity. This is vital for comprehensive research. It also supports the **rights of data subjects**, such as the ability to opt-out of data use for future projects, or to be informed of significant findings related to their health data. The HDAB plays a key role in defining and overseeing the pseudonymisation process.

Finally, **anonymisation** and **synthetic data generation** offer strong privacy protection, often used when data or analysis results are intended to be exported or made publicly available. **Anonymisation** (Regulation (EU) 2016/679 Recital 26 (GDPR), Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, WP216, adopted on 10 April 2014) transforms original personal data so that it no longer relates to an identified or identifiable person, meaning the individual cannot be re-identified by any reasonable means. **Synthetic data generation**, on the other hand, creates entirely new, artificial datasets that mimic the statistical properties and relationships found in original data, without containing any actual personal information. While distinct, both methods require the HDAB to establish similar processes for evaluating **data quality**, performing **privacy risk assessments** (looking at risks of re-identification and inferring sensitive information), and implementing **disclosure controls**. All such activities must be thoroughly **documented** to ensure transparency and accountability. It should be noted that the EHDS does not impose legal obligations regarding synthetic data generation, but HDABs may support its use via evaluation frameworks, as part of enabling responsible data access.

2 Introduction

Advancing health data use in the European Health Union

As part of its work on 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 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.

- 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 environment (SPE): 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 Regulation (EU) 2025/327 (EHDS) 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.

The work performed in work package 7 (WP7) addresses "Safe and secure processing" of electronic health data within the HealthData@EU infrastructure. The goal is to enable secure processing of EU citizens' electronic health data for secondary purposes while fostering a secure, interoperable, and efficient health data ecosystem. The output of this work package consists of guidelines and technical specifications that shall further inform 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 a SPE. Next, guidelines for data minimisation, pseudonymisation, anonymisation and synthetic data generation give guidance on how to address these topics and their challenges (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

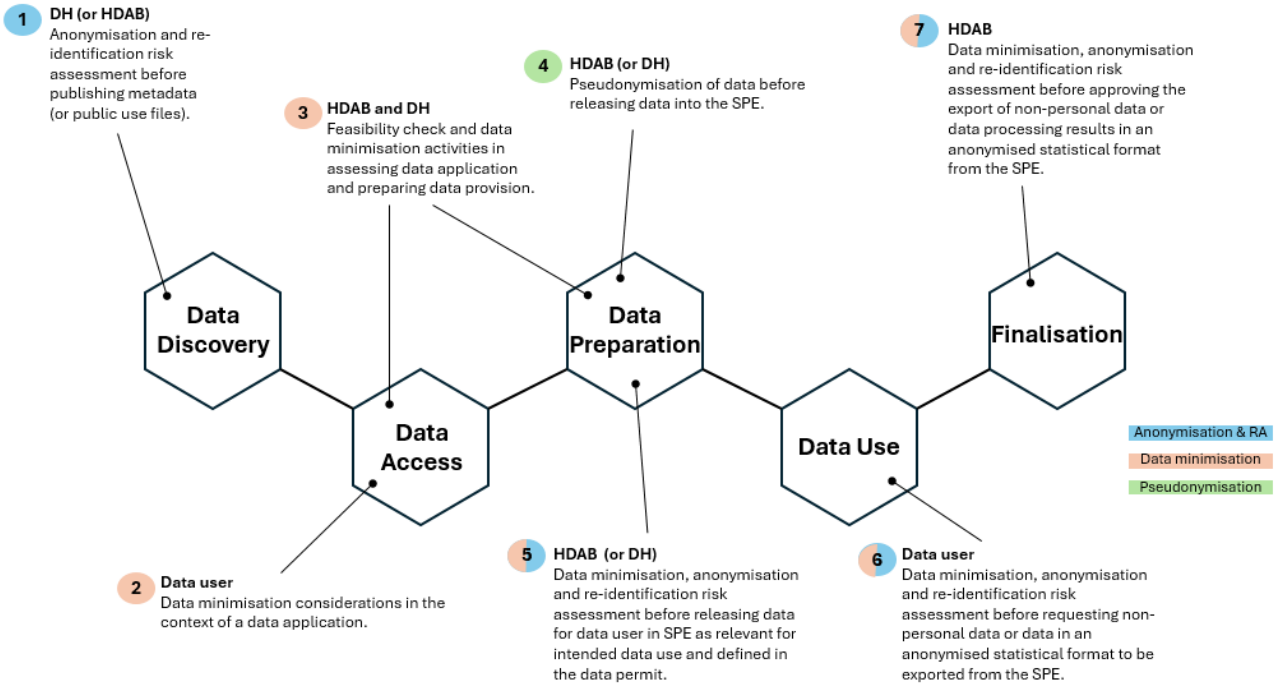
infrastructure (task 7.3) shall help member states to connect to the EHDS ecosystem. To ensure interoperability, common security requirements applicable to all SPEs are defined in addition to functional and technical services that should be part of all SPEs (task 7.4). Lastly, information about data 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:

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

2.1 Overview of the content and scope of this guideline

Figure 1. The EHDS user journey depicting data minimisation, pseudonymisation and anonymisation.



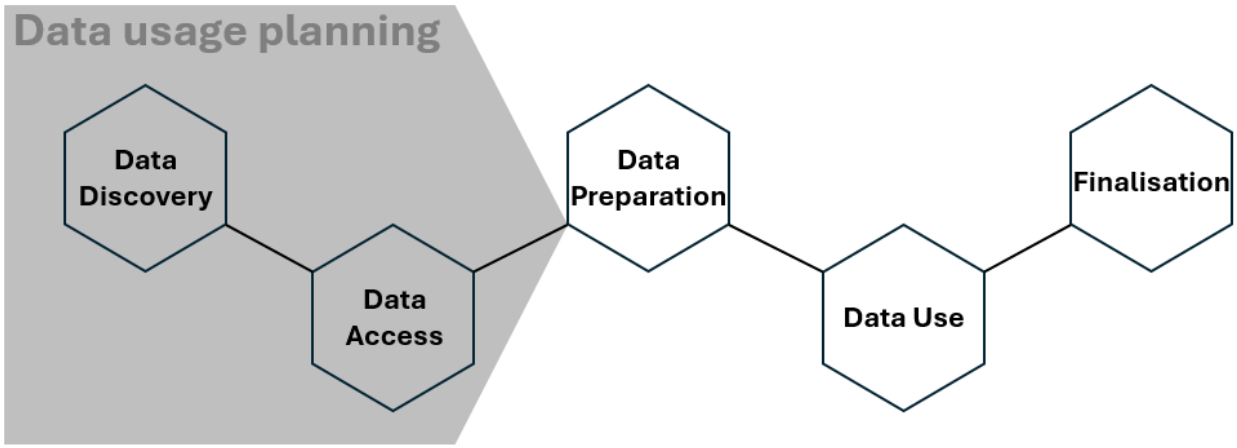
Abbreviation: DH: Data holder; HDAB: Health Data Access Body; RA: re-identification risk assessment; SPE: secure processing environment.

Figure 1 provides an overview of the key phases of the EHDS user journey where data minimisation, anonymisation and re-identification risk assessment, and pseudonymisation are relevant. Several phases involve multiple of these aspects simultaneously. For simplicity,

trusted health data holders, intermediation entities, and trusted third parties have been excluded from the figure.

The scope of this guideline focuses on the phases after the data discovery phase (see Figure 1 and Annex 1: User journey) and covers various aspects of health data minimisation and de-identification.

Figure 2. The EHDS user journey divided into before and after the actual data processing takes place.



Along the user journey, the first phases include *data usage planning* (see Figure 2). This planning involves specifying an approach on how to perform data minimisation, specific requirements for pseudonymisation, anonymisation or synthetic data generation before the data is actually being processed. The data usage planning ends when the health data request (Article 69, EHDS) or data access application (Article 67, EHDS) is approved and data preparation begins.

Regarding data minimisation (see section 5), this guideline elaborates on the dimensions of data minimisation for the use of secondary health data. This includes limiting the amount, type, and granularity of data during data preparation. While a first assessment must occur prior to the approval of the data access application or data request (Article 66(1), EHDS), the obligation to minimise data continues throughout all processing activities (i.e., data access, data preparation, data use and finalisation phases). It applies equally to data holders, HDABs, and data users.

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

Furthermore, it should be noted that anonymisation (see section 7) is not a binary nor permanent status. Previously anonymised data may, in the future, cease to meet the conditions for anonymity due to technological advances or the ability to combine multiple datasets. Under Recital 26 of the Regulation (EU) 2016/679 (GDPR), data are considered anonymised only if the data subject is not identifiable by any means reasonably likely to be used, taking into account objective factors such as cost, time, and available technologies. Therefore, the concept of anonymised data in this guideline refers to data that has been processed through anonymisation techniques that meet this “reasonableness”.

Anonymisation and synthetic data generation are related but distinct concepts. While they may involve similar techniques, anonymisation implies a transformation of real data to prevent re-identification, whereas synthetic data generation creates artificial data based on models or distributions. Synthetic data is not necessarily anonymised. The EDPB stated that the documentation of synthetic data generation should include the model's theoretical resistance to re-identification techniques (§58e) and meet the purpose and data minimisation principles (§64) (EDPB Opinion of the Board (Art. 64), Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models[1]). In other words, synthetic data may fall under GDPR if individuals can still be re-identified with reasonable effort. Therefore, it is necessary to demonstrate the resistance to re-identification.

Next, it should be noted that oftentimes, this guideline left out trusted data holders for simplicity reasons.

It is recognised that different types of SPE architectures may be implemented across Member States, depending on technical and organisational choices. For example, there may be isolated components dedicated to data preprocessing by the HDAB, and other components enabling data processing by authorised users. In this guideline, the term SPE (Article 2(1)(c), EHDS) refers to the complete set of environments or components that, taken together, meet the requirements set out in Article 73 of the EHDS regulation. These include strict control over data access, processing, export, logging, and compliance with data protection, intellectual property, and confidentiality obligations.

Lastly, it should be mentioned that this document was written before the judgement of the Court of Justice of the EU in the case EDPS vs SRB C-413/23 P was published on the 4th of September 2025.

2.1.1 Data types

Under the EHDS regulation, all categories of data processed must be subject to appropriate minimisation, pseudonymisation or anonymisation measures depending on purpose and context (Article 66(3), EHDS). Similarly, the generation of synthetic data may be relevant across all data types. In the following, a high-level categorisation of relevant datatypes is presented below (please see D5.1 Guideline on data description, section 5.3 for details on classification on EHDS categories):

- **Structured data** refers to healthcare or health-related information that is stored according to a predefined schema typically in tabular form, where each row represents an individual person or an observation, and columns correspond to specific attributes such as demographics, diagnoses, treatments, lab results, and service usage. Structured health data can be cross-sectional or longitudinal. It may

[1] https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf (Adopted 17 December 2024).

comply with a healthcare standard, such as HL7 FHIR, but can also exist in non-standardised formats.

- **Medical imaging data** refers to digital representations of visual information captured for clinical purposes. This includes traditional scans such as X-rays, CT scans, MRI, ultrasound, and PET, as well as clinical photographs—such as images of skin, wounds, or surgical sites. These data are typically stored in standardised formats like DICOM (Digital Imaging and Communications in Medicine), and may include raw image files, metadata (e.g., patient identifiers, imaging parameters), and annotations used for diagnosis, monitoring, or research. Image metadata may include direct identifiers (e.g., patient names or IDs in DICOM headers).
- **Bio-signal data** refers to measurable signals derived from a biological source, reflecting underlying physiological or biological processes recorded from the human body. These signals include electrocardiograms (ECG) for heart activity, electroencephalograms (EEG) for brain activity, electromyograms (EMG) for muscle function, electrodermal activity (EDA) for skin conductance, audiological signals (e.g., tympanometry), and other measurements such as temperature, blood pressure, respiration rate, and oxygen saturation.
- **Genetic data** refers to information derived from an individual's DNA, representing their genetic makeup. It includes sequences of nucleotides (A, T, C, G), genetic variants, mutations, and structural variations that influence traits, disease susceptibility, and responses to treatments.
- **Textual data** refers to unstructured or semi-structured information recorded in free-text form within electronic health records (EHRs) or other information systems that include clinical notes, discharge summaries, pathology reports, patient histories and other relevant information. This data often includes physician observations, diagnostic assessments, treatment plans, and patient-reported symptoms, but can also be non-clinical data such as interviews and questionnaire responses.
- **Multimodal data** refers to information that combines different types of media, such as text, images, graphs, audio, and video, in defined formats. This data is typically provided on forms that are filled out by devices used in testing, or manually by people (usually patients) undergoing testing for diagnostic or screening purposes, and the completed forms are digitised. This data includes audiograms, dementia memory tests, cognitive tests, and other tests whose results include media other than text.

2.1.2 Terminological notes

- In this guideline, references to activities conducted by the HDAB may also encompass activities carried out by a processor acting on behalf of the HDAB, such as the operator of a SPE. This is consistent with Article 28 (GDPR) and Article 73 of the EHDS regulation. In such cases, the HDAB remains responsible for ensuring full compliance with the applicable legal framework and must implement appropriate contractual and technical safeguards to govern the processing by the processor.
- The term *health data user* refers to an organisation or natural person who has been lawfully granted access to electronic health data for secondary purposes in the

context of the EHDS regulation (in line with Article 2(2)(u), EHDS). In this deliverable, *health data user* may also refer to the individual staff member acting on behalf of the organisation processing the data as employee of the data user organisation, provided there is no ambiguity. Please note that instead of the defined terms *health data user* or *health data holder*, shortened versions are also used, i.e., *data user* or *data holder*.

3 Data minimisation

Data minimisation is a fundamental principle under Article 5(1)(c) of the GDPR, requiring that personal data be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"[2].

In the context of secondary use of personal data as governed by Chapter IV of the EHDS Regulation, data minimisation and purpose limitation must be applied throughout the entire data lifecycle including:

- during data collection and preparation by the data holder,
- when assessing the data request or data access application (by the HDAB and potentially the trusted data holder),
- during the use and processing of data by the data user, including export of results from the SPE.

Purpose limitation implies that HDABs should assess their admissibility and negotiate with the data applicant whenever a refinement or restriction of the scope of the purposes is needed. This should be documented clearly and reflected in the data permit conditions under Article 68(3) EHDS.

Data minimisation may include:

- reducing the **volume of data** made available,
- limiting the **granularity** and **sensitivity** of variables,
- restricting **temporal or geographical scopes** where appropriate.

It forms part of the broader risk mitigation strategy that includes pseudonymisation, anonymisation, and organisational safeguards. It also supports compliance with other regulatory requirements, such as:

- the protection of intellectual property and trade secrets (Article 52, EHDS), when relevant.
- the exclusion of public interest risks from permitted uses (Articles 68(2) & 69(3), EHDS).

Within their risks evaluation strategy towards data provision, HDABs and data users must consider not only re-identification risk but also inference attacks. Attribute inference is described as the process of inferring unknown attributes of known individuals from the data. Membership inference refer to inferring the presence of known individuals in the data (see section 7.5.5). Attacks on attribute inference, group membership inference, linkage attacks should be considered during the risk evaluation before data provision (see ISO/IEC 29100:2024 , Recital 49 (GDPR)).

While minimisation is not an active measure of enhancing data security control, it significantly reduces exposure related to:

- confidentiality (unauthorised access),

[2] See Article 5(1)(c) (GDPR).

- integrity (unauthorised alterations), and
- availability (data not accessible to authorised parties).

Hence, data minimisation indirectly supports data confidentiality, integrity and availability, by limiting the amount of the to be exposed data. However, the GDPR triad (confidentiality, integrity, availability) must be clearly distinguished from broader data protection risks and traditional IT security domains (e.g., authenticity in Recital 49, GDPR).

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

If the HDAB provides a response to a data request, data made available to data user will be non-personal aggregated data, i.e. data in an anonymised statistical format. Synthetic data may also be used during the data processing of the request and used as basis to create data in an anonymised statistical format. Even in that case, HDABs must pay attention to which relevant data elements should be provided to data users, based on their stated purposes, that should not be exceeded. More information and insights on the procedures and techniques to produce aggregated data within data requests are offered in section 7, i.e., Anonymisation chapter.

Regarding data permits, data minimisation procedures and techniques that can be used by HDABs and data holders will be further described in this paragraph and should be considered by applicants in the data access phase of the EHDS user journey (Figure 1).

3.1 When should data minimisation be performed?

Controllership. A data controller is a person or organisation that determines the purposes and means of processing personal data, as regulated by the GDPR. According to Article 4(7) (GDPR), the data controller is responsible for ensuring that data processing complies with the principles set out in the regulation, including data minimisation. The EHDS regulation specifies (Article 74) the controllership rules applicable to data processing workflows within the EHDS ecosystem. Data holders are deemed controllers for the initial processing and provision of data to the HDABs. This condition may occur when pseudonymised or anonymised data are foreseen in the data permit, or when HDAB will perform data aggregation of personal data to be offered in an anonymous statistical format. HDABs are deemed controllers for processing personal data within the scope of their tasks indicated by the EHDS regulation (Article 57, EHDS), having received data from data holders, for activities like data linkage, additional pseudonymisation that may be needed, additional risk reduction or anonymisation, and upload to the SPE. Data users are deemed controllers for data that following a data permit issuing, has been made available to them in a SPE. Data users' scope of controllership is then limited to the permitted purposes within the regulated framework they have access to (see also D7.1, TEHDAS2).

Communication channels. Specific considerations on data minimisation should always be made by data applicants, because the EHDS directive mandate the HDAB to ensure that the

requested data are limited to what is necessary for the permitted purposes. HDABs will have to assess compliance with data minimisation principles of the data access application or the data request received, before issuing the data permit (see Article 68, EHDS) or approving the data request (see Article 69, EHDS). If compliance is not sufficiently demonstrated by the data applicant, some delays in the process, a changeover from a data permit to a data request (Article 68(3), EHDS) or even a rejection may occur. For that reason, it is important to establish an effective communication channel between the data applicant and the HDAB, that could resolve doubts or controversial points specific to data minimisation early in the process (see also D6.2, TEHDAS2).

Another communication channel may be needed between HDAB and data holders, to clarify some aspects of data minimisation, especially when the application is incomplete (see D6.2 Guideline for data users on good applications and access practice, section 8.5) to manage the amount and the granularity of information requested by the data applicant.

Further ahead in the user journey, data users shall take data minimisation principles into consideration during the use of data in the SPE and when requesting data exports from the SPE (see more in section 7 of this document).

Summing up all those considerations, data minimisation shall be applied by different actors and most importantly during the phases presented in Table 1. See also Figure 1 for a more general overview.

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

Actors	User journey phase	Task	Activities
Data Applicant	Data access	Data access application preparation and submission	Considering and documenting in the application relevant data minimisation actions on tables, variables, granularity levels, quasi-identifiers.
HDAB (involving data applicant(s) and data holder(s) whenever relevant)	Data access	Data access application evaluation	Engaging with the data user to assess whether data minimisation has been adequately demonstrated, and—if needed—formulating adjustments to ensure compliance with Articles 66–68 EHDS.
HDAB	Data preparation	Data preparation before granting access to data in the SPE	Applying and verifying what is stated in the data permit issued.
Data Holders	Data preparation	Data preparation pursuant to a data permit	Applying and verifying what is stated in the data permit issued.

Data User	Data use	Preparation and request for result export from SPE	Proposing the structure of a result export to the HDAB.
HDAB	Finalisation	Approval of result export from SPE	Risk assessment before approving the export of results from the SPE.

The GDPR emphasises the necessity to manage data minimisation activities also at the data holder's side before any data permit has been issued (i.e., at the data collection phase and when pre-processing data to be shared). However, those activities are out of scope for this document.

3.2 Direct identifiers and quasi-identifiers

Within the context of human subjects' research, **direct identifiers** are variables that point explicitly to specific individuals and are sufficient to identify the data subject either alone or in their mutual combination (e.g., names, national id numbers, social security numbers, telephone numbers, email addresses, fingerprints). As widely discussed in the EDPB guideline on Pseudonymisation[3], effectively pseudonymised data do not contain direct identifiers as they are stored separately as part of the "additional information". Pseudonymised data provisions, however, in some cases may contain variables whose combination is sufficient to attribute at least part of the pseudonymised data to identifiable data subjects. Those attributes are called quasi-identifiers (or indirect identifiers[4]). Some examples of quasi-identifiers are age, date of birth, sex, ethnicity, education, employment status, marital status, income, place of residence or work/study, or a sequence of hospital visit dates. In the context of employee data, relevant quasi-identifiers may include structural role, number of working hours, and length of service.

Quasi-identifiers may be considered as critical variables that increase the assessed privacy risk (including the re-identification risk) and should be removed from the data provision, unless they play a crucial role for data usage. When they play a crucial role, only strictly necessary quasi-identifiers should be retained, and their risk mitigated, in line with the GDPR's data minimisation principle. Appropriate risk mitigation techniques may be applied to reduce the likelihood of re-identification, such as generalisation, suppression, or randomisation.

In the GDPR (Recital 26), directly or indirectly identifiable data is data that can identify a natural person by reference to an identifier or more factors (Article 4(1), GDPR).

Reducing re-identification risk typically require some kind of transformation in the original datasets. Generalisation is the most used method and involves reducing the level of detail (e.g., using age brackets instead of exact age). Suppression involves removing or hiding specific cell values. Randomisation[5], which introduces noise to diminish predictability like any other forms of non-deterministic methods should be carefully calibrated because they may significantly reduce data utility. Non-deterministic techniques like noise injection are more

[3] See EDPB Guidelines 01/2025.
[4] See EDPB Guidelines 01/2025.
[5] See EDPB Guidelines 01/2025.

commonly used in anonymisation than in pseudonymisation contexts, especially under EHDS which prioritises data fidelity. See more in the next chapters.

Some examples are as follows:

- An example of generalisation is when income data are grouped into ranges (e.g., €20,000-€30,000) to prevent identification.
- An example of suppression is when the values of cells that represent a small number of contributing entities are suppressed, for instance the extreme values in a lab test are removed. For aggregated data, an example of suppression is removing the number of patients discharged from a small hospital for a low-incidence disease.
- An example of randomisation is altering the number of working hours by adding or subtracting a small random value to/from each entry.

Although a case-by-case approach may be adequate, some recommendations and general principles to be respected will be further discussed in the following paragraphs.

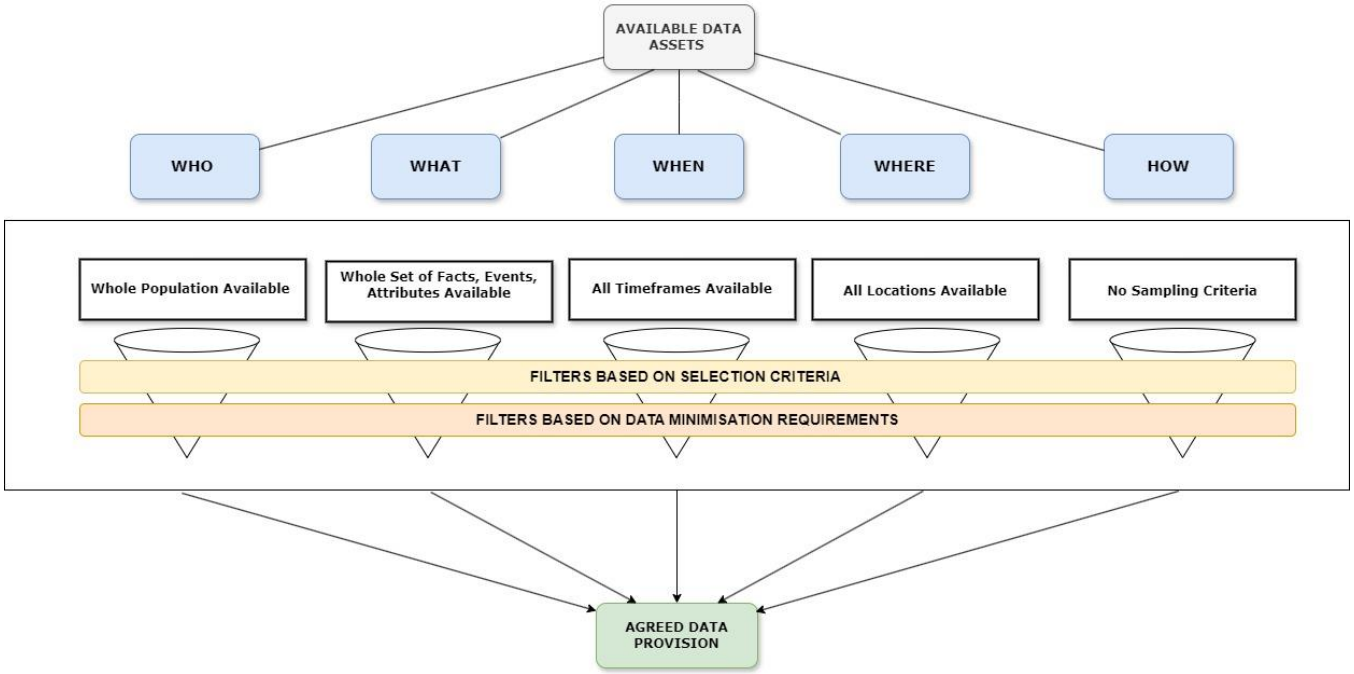
In the EHDS, when data access to a pseudonymised or anonymised dataset is provided, the HDAB should ensure to the maximum extent possible that natural persons cannot be re-identified (Recital 72, EHDS). In addition, the data user should refrain from any attempt to re-identify natural persons from the dataset (Recital 72, EHDS). Therefore, it should be clear that ensuring adequate protection against re-identification is a shared responsibility between the HDAB and the data user.

The EHDS also states that HDABs should apply tested state-of-the-art techniques that ensure that data processing preserves the privacy of the information contained in the data, including methods such as generalisation, suppression or randomisation (Recital 65, EHDS).

3.3 The dimensions of data provision

Whenever a data permit is issued by an HDAB, meaning the HDAB has determined that access to individual-level data (pseudonymised or anonymised) is justified and appropriate based on the purpose and privacy risk assessment, a data provision with individual-level data will be offered to the users in an SPE. It is possible to identify some specific dimensions of such data provision, as described in Figure 3. It should be noted that access to data via a data permit is not necessarily equivalent to granular pseudonymised data, but can be restricted, depending on proportionality and risk mitigation principles.

Figure 3. Data provision dimensions.



During the data discovery phase (Figure 1), the data applicants would check data assets that are available from specific data holders of their interest and would select some of them that can work well in addressing a limited number of predefined research questions (or endeavours) and objectives. Information from available data assets can be distinguished for practical reasons into the following five dimensions:

- **Who** - Which individuals are relevant for the study purposes (Study population), from the population;
- **What** - What information is relevant for the study purposes (Research variables), from the whole set of facts, events, attributes available;
- **When** - Which timeframes are relevant for the research (Research timeframe), from the whole available periods of time;
- **Where** - Which specific locations, e.g., place of residence, work or health assistance, are relevant for the study purposes (Research geographical perimeter), from the whole geographical distribution available;
- **How** - What extraction methods are used to get samples of interest (Research extraction methods), from the entire available sets of information.

Please also see M6.3 Guideline for Health Data Access Bodies on the procedures and formats for a data access application template (Annex 5), where the five dimensions are also used to help applicants adhering to the data minimisation principle.

The cross-classification of multiple variables is the real issue to be tackled. In a dataset containing a medium to high number of variables, a large proportion of records will be unique, due to the cross-classification of their values. Moreover, in the case of quasi-identifiers, it could also occur a risk of singling out (which corresponds to the possibility to isolate some or

all records which identify an individual in a dataset) due to combinations of different personal information (see Opinion 05/2014 on Anonymisation Techniques).

Based on the study objectives, applicants should identify the necessary information and apply inclusion and exclusion criteria accordingly. This process may reduce the data requested across the five data provision dimensions. Data minimisation principles and data protection risk management should inform the exclusion of information not directly relevant to the study or that presents additional privacy risks. Applicants must justify their data selection decisions across all dimensions.

The result of those activities, that imply a mediation between the study objectives and the data minimisation requirements, which may require negotiation between the applicant and the HDAB (or even the data holder in particular cases, see more in the next paragraph), will end up defining data provision characteristics, and the specificity of each piece of information required by the applicant, in all five dimensions of data provision.

A general consideration that applies to all those dimensions is that it must be stated in each data permit (Art. 68(10), EHDS) which are "the *categories,* ***specification*** and *format* of the electronic health data to be accessed". Some specifications are consequently considered mandatory in the data access application form and need to be justified by the applicants as required for reaching the study objectives and yet compliant with the data minimisation principles. The access application form, as well as the data request template can be inspected in Annexes 5 and 6 of the M6.3 guideline.

Practical strategies to enhance the data provision may include the followings:

- It is possible to provide a certain percentage of the data during data provision (e.g., 1 million of data subjects, or 10 years of data) that allow data users to build and test their analysis scripts and elaborate on intermediate results, then executing the final scripts on the complete dataset applied for.
- It is possible to provide pre-defined granularity levels by the data holder that the applicant can consider.

Those ways of working should be made clear as early as possible, so that applicants may know this possible chance in advance and be aware of this already during the data discovery phase.

In the following paragraphs, special attention will be given to each data provision dimension.

3.3.1 The "Who" dimension

Explicit study population definition support both data minimisation and purpose limitation. Specification of data subjects included in the study population is mandatory when individual-level pseudonymised or anonymised data are requested by the data applicants. Applicants should clearly define their study population, and the relevant inclusion and exclusion criteria

from the whole population included in the datasets offered by the data holders[6]. This decision will affect which subjects will be included in the data provision. Moreover, eligibility criteria are also critical for enabling the HDAB to assess necessity and proportionality of the requested sample. Common inclusion and exclusion criteria normally act jointly on demographic characteristics of subjects and on study-specific variables of interest like diseases, exposures, treatments, and comorbidities. All eligibility criteria specified by the data applicant would define the study population.

An estimate of the expected cohort size should be provided by the applicant in the application phase, along with the inclusion and exclusion criteria used to derive this estimate from the data assets of interest. This information serves to support the HDAB in assessing the proportionality and feasibility of the requested data provision. For instance, the applicant may state: 'We apply inclusion criteria A, B, and C to dataset X, which contains 120,000 individuals. Based on prior studies, we estimate that 5,000 individuals will match our criteria.' The HDAB may then verify this estimation against the structure of the dataset and, if needed, request clarification or propose adjustments (e.g., stratified sampling, changes in inclusion logic). Such a check does not imply that the HDAB re-derives the full cohort itself but helps identify inconsistencies or misalignments between the stated study population and the available data.

It should be noted that most demographic variables typically fall into the category of quasi-identifiers and hence require some remarks as follows:

- ***Date of birth*** and ***Date of death*** are strong quasi-identifiers and should be avoided whenever possible. They can be substituted with other relevant variables like year of birth, age group or age at death. For age-sensitive studies year/month of birth, year/month of death, or even days at hospital discharge for children might be relevant.
- ***Age*** is a quasi-identifier that requires attention. Age should only be used as an exact value when it is indispensable for the research question and cannot be replaced by age categories, i.e., generalised, without loss of analytical validity. In some cases, where knowing the exact age is fundamental for the analysis, it may, from a data minimisation perspective (to reduce privacy risk), be relevant to categorise the extreme values where the number of subjects reduces considerably (e.g., ages from 0 to 5 and above 85 are grouped together).
- ***Nationality, ethnicity, education, profession, employment status, marital status, income, state of vulnerability*** (persons under guardianship, trusteeship) are quasi-identifiers that require attention and should be used only when strictly relevant, taking the precaution of standardising, classifying and/or grouping their possible values into the lowest level of detail that is useful to meet the research objectives.

Please also note that some of the variables listed above (e.g., ethnicity, health status) represent sensitive data under Article 9(1) (GDPR), which requires an additional layer of legal and ethical assessment.

[6] see also TEHDAS2 D6.2 Guideline for data users on good application and access practice.

Moreover, **combinations of demographic variables**, like “date of birth plus sex plus postcode” should be avoided because they significantly increase the re-identification and privacy risk of the data subject.

3.3.2 The “What” dimension

Specification of datasets, tables and data elements needed to pursue the aim and topic of the application is mandatory. During the data discovery phase, applicants should invest some effort to understand what is available from a specific data holder of their interest, and which information is relevant for their study objectives. In accordance with Article 5(1)(b) of the GDPR, personal data must not be made available for undefined or overly broad secondary use purposes. Instead, applicants should request only the minimum data that is necessary to accomplish the clearly defined research objectives specified in their data access application. Variables to be requested can be divided in two broad categories: can be divided in two broad categories:

- **Study-specific variables:** the ones that are functional to answer research questions and address study objectives, and that may be linked with patients, diseases, characteristics of diseases, treatments, exposures, comorbidities, disabilities and so on.
- **Control & confounding variables**: the ones that are needed not to address the study objectives, but to enhance the validity of the research, giving the researcher the ability to control external and extraneous influences on the observed outcomes.

Data applicants are strongly encouraged to explicitly mention, during the application phase, to which of these categories each requested variable refers to. Separating variables in these two categories may help HDABs assess the relevance of the data application and its data minimisation compliance. In some cases, a transformation in the type, format, level of detail or coding of a variable made available by a data holder may be asked by the applicant in the data access application form or suggested by the HDAB for data minimisation compliance. In such cases it is important to define who will be in charge of applying those transformations, that virtually can be done both at data holder’s or HDAB’s side and could influence the production and costs of data provision (see more at paragraph 5.4).

The ”What” data provision dimension consists of a list of objects and data elements that are strictly useful for the expected data usage, and do not fall in any other dimension like “who”, “when”, “where”, or “how”. The “what” dimension can sometimes manage direct and/or indirect identifiers. Overall, the higher the number of indirect identifiers requested by applicants, the higher the risk of re-identification. For more clarification on the EHDS categories for secondary use see section 5.3 in D5.1 Guideline on data description. They require some general remarks:

- ***Medical images*** produced by healthcare procedures may reveal unique body features of individuals with specific diseases (e.g., head and neck cancers). This uniqueness may contribute to increase the risk of re-identification of data subject’s identity. Depending on the case, those images could be considered as direct-identifiers or quasi-identifiers and should be treated consequently. It is important to stress that evaluations depend on the case and the re-identification risk.

- ***Genetic data*** require special attention because depending on the research may represent direct identifiers or quasi-identifiers. Moreover, genetic data are classified as sensitive personal data under Article 9(1) (GDPR), and special safeguards apply to them.
- ***Rare disease codes*** can be considered quasi-identifiers. Therefore, when researching rare diseases, attention must be posed to remove or generalise from data provision other quasi-identifiers. Practically, the higher the number of indirect identifiers requested by applicants, the higher the risk of re-identification, and consequently, the more importance should be given to risk reduction (e.g., anonymisation, differential privacy techniques) and risk management procedures.
- ***Unstructured data*** always needs particular attention, because attributes selection is difficult to manage. Thus, unstructured data potentially includes more background information on facts and people being investigated than what is strictly necessary to data usage purposes. Also, the activities that may be performed to remove direct and quasi-identifiers (e.g., searching and removing or obfuscating them) are less deterministic compared to the ones performed for structured data.

3.3.3 The "When" dimension

Specification of the data provision timeframe is mandatory. It may regard years, or months, or days, and both *absolute* and *relative* timeframes are possible. An absolute timeframe specification may specify a given calendar period, e.g., from the year 2020 to 2025. A relative timeframe specification may be relevant for longitudinal studies, where the timeframe can be aligned with the patient inclusion period (e.g., +/- 1 year after/before hospitalisation). Specifying the timeframe does not mean that it cannot be the maximum available, it only stresses out the necessity to align to the study purposes/objectives, and so a clear justification of the timespan is required. Applicants are expected to provide considerations on the granularity of time-specific variables also, because the granularity of temporal data (especially when requesting day-level information or information on even shorter timespans) is directly linked to privacy risk and must be justified accordingly. For instance, if the data applicant is requesting information on the incidence of a disease for a selected population, it should be clear enough if the year of incidence (instead of the month or the exact date) is relevant for the study objectives.

Quasi-identifiers that fall under the "When" category of data provision dimension require some remarks:

- ***Dates*** are often critical variables, especially when combined with other information. Some consideration on dates linked to patient's demographics have already been discussed in paragraph 5.3.1. Other relevant dates available in a data provision may be time-specific ones. For instance, hospital admission and discharge dates or treatments dates. Those dates may point at a specific episode of care in the life of a person, that linked with other information may reveal data subjects' personal information or their identity. Therefore, absolute dates are variables that should be

avoided when not essential and replaced by relative timespans (e.g., time since diagnosis, age at event), which may often be safer and sufficient for most analyses. Possibly, year/month of hospital admission and length of stay may be considered valid substitutes of hospital admission and discharge dates. discharge dates.

These considerations align with the GDPR's overarching principles of confidentiality (Article 5(1)(f)) and with recommendations under EDPB Guidelines 01/2025 for controlling re-identification risk in datasets containing high-dimensional or temporally specific variables.

3.3.4 The "Where" dimension

Specification of geographic constraints applicable to the study population should be indicated by the applicant in the data access application form submitted to the HDAB (see Annexes 5 and 6 of Guideline M6.3, TEHDAS2). The geographical dimension can directly affect the amplitude of data provision.

Selecting a large research population may be beneficial for the research, especially when machine learning models are going to be used but may also increase the need to control for confounding variables when geographical differences imply differences in the health determinants and/or in the healthcare service characteristics and consumptions. Choosing a very large population may also increase the privacy risks, as it increases the probability of providing records with unique characteristics that may be correlated with identifiable data subjects and/or that may be linked with external available sources of information (i.e., using geotags or geocoding[7]). A very small population, on the other hand, may also increase the privacy risks because their demographic and geographic variables may easily point out toward a data subject that is known to be included in the dataset. In both cases, a general guidance can be to pay close attention on the variable selection. Next to population size, the granularity of the location data is of great importance and should always be generalised by default unless granular data is clearly justified and proportionate.

Geographic attributes are generally easily generalisable without reducing the utility of data provided. Examples include:

- ***Place of residence, of work, of health assistance***, being critical variables, especially when combined with other information, should be provided at the highest possible level, like a ZIP code with only 2 digits instead of 5, or the region, or the local health authority. On the other end, data projects that are very specific in terms of geographic scope (for instance, comparing the catchment area of the city hospitals for specific procedures) should be critically assessed and should offer very controlled variables (for instance, ZIP codes could be provided with 4 or 5 digits but most of the other demographic variables would be omitted or generalised, or otherwise specific variables like "distance from workplace" or "distance from residence" could be calculated and offered to data user).

[7] See DOI: 10.3389/fsoc.2022.910111.

- ***General Practitioner*** associated with the data subject should always be pseudonymised and may give a strong local attribution if linked with other information included in the data provision like general practitioner's visits.
- ***Health Facilities*** linked to data subjects' episodes of care, having a known geographic location, may sometimes require attention and may suggest using some form of obfuscation (i.e., substituting identifying facilities codes with sequential numbers).

3.3.5 The "How" dimension

Specification of how the general population available from a data holder should be sampled and/or managed to obtain the final data provision may be included in the data access application form.

In some cases, data applicant may ask to apply specific data cleansing methods to the identified target population to increase data quality or data analysis (remove redundant records, apply standardization procedures to specific columns, apply specific rules to treat missing information etc..). However, these methods may conveniently be applied once data provision has been issued by data users themselves.

In other circumstances, modifying the consistency of the whole population available at data holder side may be needed for data minimisation necessities (i.e, removing outliers, aggregating information for low-frequency groups when they do not influence the study objectives).

Sampling procedures that produce statistical samples from the identified target population are also possible (i.e., the applicant want to conduct a quicker preliminary study on available data before starting the main study).

In any of those situations it is required that:

- The principle of data protection by design and by default (Article 25, GDPR) should guide sampling decisions, particularly in how subsets of the population are chosen and filtered.
- Sampling methods must preserve representativeness and avoid introducing bias.
- Data applicants should be transparent on methods to be applied in the data application form.

3.4 Data minimisation steps towards issuing a data permit

The data minimisation principle applies at all phases of health data access application.

One of the primary responsibilities of HDABs, as illustrated by Article 57 (EHDS), is to issue data permits, based on specific criteria, as expressed in Article 68 (EHDS). During the data access application assessment phase, HDABs are legally responsible for ensuring that the scope and granularity of requested data are limited to what is necessary. Note that trusted data holders may also perform the first application assessment.

This section outlines the main procedural steps HDABs should implement to ensure that only data that are necessary for the stated study purposes are made available, and that personal data are minimised at every phase of the permit process.

The main steps that HDABs should put into effect towards the goal of **issuing data permits and requests** are as follows[8]:

1. **Completeness and relevance check**: Review the data access application to ensure that all required information is present, and that each requested element is clearly justified, relevant to the research objectives, and aligned with the data minimisation principle.to ensure that all required information is present, and that each requested element is clearly justified, relevant to the research objectives, and aligned with the data minimisation principle.
2. **Data access application assessment**: The HDAB should examine the content of the application in detail; identify critical elements that may pose minimisation challenges; clarify applicant choices and goals where needed and initiate communication with the data applicant and/or the data holder to resolve ambiguities and reduce unnecessary data demands.
3. **Decision and outcome**: Based on the assessment, the HDAB should either issue the data permit, request specific revisions to ensure compliance with the minimisation principle, or propose a changeover to a data request (e.g., if only aggregated data in a statistical format are appropriate).

3.4.1 Data Access Application Assessment

In evaluating a data access application (after a successful data access application completeness check, see Annex 7 in M6.3), HDABs must ensure that all requested data are strictly necessary for the stated data usage objectives and comply with the principles of data minimisation and purpose limitation (Articles 66 & 68, EHDS; Article 5(1)(b–c) GDPR). This assessment consists of two layers:

Layer 1. HDABs verify whether the data access application justifies access to the data applied for. The HDAB examines whether the applicant can get access to anonymised or pseudonymised data (Article 66).. They also examine (possibly with the help of the data holder) whether risks related to public interest, intellectual property, or trade secrets are implicated (Articles 52, 68(2), and 69(3), EHDS).

Layer 2. Once these general conditions are considered, a more **detailed examination** is conducted. This includes reviewing the proposed data provision across all five dimensions (“Who”, “What”, “When”, “Where”, and “How”) to ensure that:

- each data element is relevant and justified for the study purpose;
- the granularity and sensitivity of the data are proportionate;

[8] TEHDAS2 D6.3 Guideline for Health Data Access Bodies on the procedures and formats for data access.

- unnecessary or overly detailed data are excluded or transformed.

Applicants should take a proactive role in anticipating potential data minimisation concerns, and structure their requests accordingly. HDABs may engage in dialogue with applicants or data holders to refine the request and ensure regulatory compliance.

The assessment of a data access application involves both **general** and **domain-specific considerations**. The following section first outlines cross-cutting criteria that apply to all applications, regardless of data type or research domain. It then provides tailored recommendations for specific types of data, such as structured datasets, images, genomic datasets, and unstructured data, which require additional attention due to their unique privacy or minimisation implications (as outlined in section 6.3.).

The following considerations also apply in the case of data permit amendments (for instance upon the user's request for additional variables or changes in scope) which may occur after the initial permit is granted. The HDAB must reassess the relevance and proportionality of any such modifications before approval and remains responsible for data minimisation and necessity assessment when such modifications are requested.

General considerations specific to data minimisation and purpose limitation that should be made during the detailed examination of a data application assessment include:

- **Revising** the specificity of data usage objectives. Significantly different objectives should be managed with different data applications and data provisions.
- **Revising** the appropriateness of the study population to answer data usage objectives, in terms of number and characteristics of the data subjects for which personal data are expected to be included in the data provision.
- **Verifying** that no direct identifiers are asked to be provided, and that quasi-identifiers' inclusion is based on solid motivations and pertinent risk reduction safeguards. (Please note that in the case of data linkage, direct identifiers may be necessary to perform the linkage).
- **Evaluating** the legitimacy, proportionality, and granularity of quasi-identifiers and the potential increases in re-identification risk that may arise from their combination.
- **Exploring** all five dimensions of the expected data provision to check if justification for inclusion/exclusion of all variables is available and appropriate.
- **Evaluating** the impact of cross-classification of multiple variables: for instance, in a dataset containing a high number of variables, a large proportion of records can result to be unique, due to the cross-classification of their values. In these cases, there is a trade-off to be managed between the limitation of re-identification risk and the retention of utility for the expected data provision.

Domain-specific considerations on data minimisation and purpose limitation that should be made during the detailed assessment of a data access application may include:

For **datasets** and **data elements**:

- **Limiting** the **tables** requested to what is strictly necessary to the data usage objectives, in terms of tables that contain study-specific and/or control/confounder variables only.
- **Limiting** the number of **columns (variables)** to what is necessary to data usage objectives. For each variable that is not directly justified as related to data usage objectives, HDAB will ask the applicant for a revision of the application.
- **Confirming or changing the provision of columns (variables).** The use of variables should be assessed considering data minimisation and purpose limitation issues as well as data usage objectives. Variables that are not linked to study objectives and cannot be considered as control/confounder variables should be excluded. If different variables that point to similar kind of information is available at data holder's side, these variables should be evaluated as alternatives in terms of utility and risks. HDABs could propose data transformations for the purpose of minimisation (e.g., replacing detailed dates with derived indicators like hospital length of stay). This activity is typically limited to predefined variables and options already made available through data catalogues. Unless otherwise agreed with the data holder, custom derivations are not required.
- **Confirming or changing the specificity of information** (i.e., the structure and granularity of values in the columns). HDABs will confirm data applicant's choices or offer provision of more aggregated information (e.g., variables with categorised data or at a higher geographical level) or diluted information (e.g., suppression of row or, randomisation). In publishing data catalogues, data holders may pro-actively anticipate the available levels of specificity for each variable provided, to allow data minimisation considerations during the data discovery phase.
- **Confirming or changing cross-linkage between tables**. Only tables that are explicitly intended to be linked should be cross-linked within the data provision, to keep re-identification risk controlled.

For **images**:

- **Limiting metadata** to what is strictly necessary to the research objectives. Metadata should not include direct identifiers. Technical metadata should be removed too, if not explicitly required.
- **Digital blurring** of specific image areas (i.e., where personal identifiers or quasi-identifiers are present).
- **Reducing** the **resolution** of the image whenever is relevant, to what is pertinent to data usage objectives.

For **genomic datasets**:

- **Limiting metadata** to what is strictly necessary to data usage objectives. Technical metadata should be removed too, if not explicitly required.
- **Filtering** for relevant **variables** (i.e., if the analysis concerns only some mutations - e.g., BRCA1 for breast cancer - only those variants can be preserved from the entire genome).

- **Preparing files** in the format most suitable for analysis. Raw genomic data is generally avoided unless clearly necessary and explicitly justified.

For **unstructured data** and **qualitative research:**

- **Evaluating** the data collection approach and its sensitivity to minimise personal data. It might be helpful to use sample data from data holders.
- **Pre-processing** text files to remove direct identifiers, quasi-identifiers and data attributes that are described in the application as not linked to data usage purposes. For example, this may mean using regular expressions ("regex") to locate identifiers in free text, and to base these regular expressions on public databases such as a list of common surnames, or the exhaustive list of European communes or ZIP codes. The length and format of the text snippet to be extracted should be specified (ideally per variable) (e.g., detecting sequences of 13 digits in free text to identify a social security number).

Having assessed in detail the structure and the content of the application, HDABs will end up with different possibilities:

1. to issue a data permit;
2. to refuse the data access application;
3. to propose a revision of some critical aspects or gaps that need to be addressed to accept the application (see Article 68(3));

The first two cases are straightforward: HDAB will take its decision and document it. Refusing the data access application may lead to a new data request submission by the data applicant, when HDAB propose aggregated data instead of individual-level data. The third case is indeed the more complex scenario to be managed by the HDAB and may require communication with the data applicant and data holders. After the detailed analysis of the application, the HDAB might have a list of the gaps that need to be addressed to accept the application, for instance:

- requested objects, tables and variables match the study objectives, quasi-identifiers are justified and proportionate, and safeguards to reduce re-identification risk are in place, yet some of them or their cross-combination is considered critical and need to be managed further (**Gap 1**).
- some of the requested objects, tables or variables do not seem to match the study objectives or require clarification from the applicant to understand if they are strictly necessary and/or they can be offered with a lower level of detail (**Gap 2**).

These gaps may be resolved by direct communication with the applicant and/or data holders. Whenever clarification from the applicant is needed (i.e., Gap 2 of the previous bullet point), the HDAB will contact the data applicant and start a discussion aimed at closing the gap towards an acceptable data access application.

In any case, before issuing a data permit, the HDAB must always assess whether the provision of pseudonymised or anonymised individual-level data is justified, taking into account the purpose of processing and the risk of re-identification (Articles 68(6) & 70(2), EHDS). In particularly sensitive or complex cases, (i.e., Gap 1), the HDAB may request the data holder to apply quality metrics (e.g., k-anonymity, uniqueness analysis) to the expected datasets, to support this assessment. Nevertheless, the responsibility to determine whether the re-identification risk is acceptably low and to decide whether data may be made available at individual level remains with the HDAB.

The quality of the data access application may also affect the level of interaction needed between the HDAB, the applicant and the data holder:

Scenario A – Detailed application: The applicant specifies precise variables, granularity levels, and richly justifies the use of quasi-identifiers. The HDAB validates alignment with minimisation principles, and the data holder can provide a cost estimate (see TEHDAS2 M4.1 for guidelines on fees related to the EHDS) based on well-defined parameters.

Scenario B – Vague application: The applicant requests broad tables or domains with insufficient detail. The HDAB must act as a facilitator, requesting clarification from the applicant and inviting the data holder to suggest a concrete variable set and associated minimisation techniques. The data holder may also propose specific transformations and provide a cost estimate based on the resulting dataset.

In each of the two scenarios, pursuing a collaborative approach allows the HDAB to make a legally sound, proportionate and documented data permit decision. It also reduces the likelihood that the data holder would later need to request modifications due to unmitigated re-identification risk or unresolved minimisation issues.

3.5 Closing remarks

Data minimisation is a foundational principle of lawful and proportionate secondary use under the EHDS regulation. This section has outlined how HDABs, data holders and data applicants/users can operationalise this principle across the data lifecycle — from application, through assessment, to permit issuance and finally to data usage. Each dimension of data provision has been reviewed in terms of minimisation logic, risk management, and justification requirements.

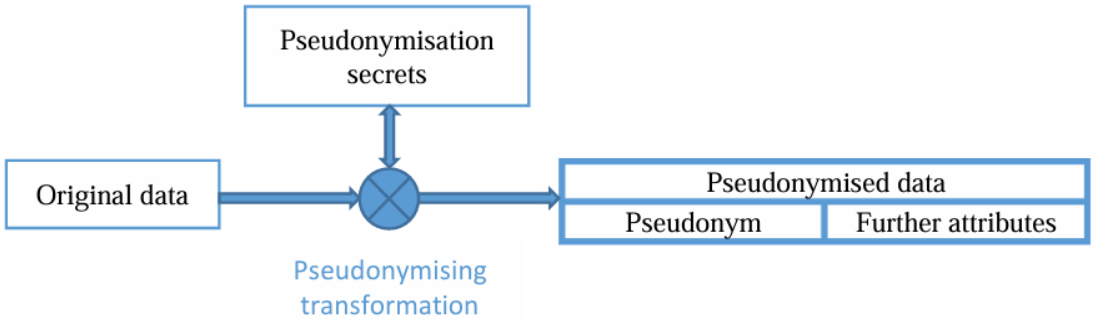
With the procedural clarifications and practical examples provided, this section supports HDABs in implementing a robust, transparent and compliant data minimisation process. It also provides a clear framework for dialogue between actors involved in the data access process.

The principles and methods described here should remain applicable not only to initial applications, but also to any subsequent amendments, ensuring continuity in the application of the minimisation principle.

4 Pseudonymisation

Pseudonymisation is defined in Art. 4(5) (GDPR) as “the processing of personal data in such a manner that the personal 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.” (EDPB-G, §16).

Figure 4. Depiction of the pseudonymising transformation taken from the EDPB Guideline 01/2025.



The purpose of pseudonymisation is balancing between privacy protection of data subjects, retaining data fidelity and quality, and the ability of data subjects to exercise their rights. In the context of this section, we will use terminology depicted in Figure 4, consistent with EDBP Guideline 01/2025. *Pseudonymisation* is a transformation, in which *direct identifiers are replaced by new identifiers called pseudonyms*.[9] The mapping of pseudonyms to the direct identifiers they replace is performed using *pseudonymisation secrets* which have to be kept separately from the *pseudonymised data* with appropriate technical and organisational measures. Pseudonymisation secrets constitute the *additional information* described by GDPR (Article 4(5)), which is necessary to map the data to data subjects. Note that we use the term *additional information* in this specific meaning in this section.

4.1 The purpose of processing pseudonymised data within the EHDS

The overarching purpose of pseudonymised data within the EHDS is to enable the provision of personal data for secondary use by data users, pursuant to a data permit. Processing of pseudonymised data within the EHDS must comply with both the EHDS regulation and the GDPR. Pseudonymisation enables lawful secondary use in a SPE as part of a data permit granted by an HDAB, under Article 68 (EHDS).

The following goals of processing pseudonymised data are considered within the EHDS, when a data applicant files a data access application (i.e., applies to get a data permit):

- Ensuring high data fidelity while applying strict data minimisation according to the approved research purpose (Article 66(1), EHDS; Article 5(1)(c), GDPR);

[9] See §83 EDPB Guidelines 01/2025.

- Enabling linkage of data records across datasets from one or more data holders, under the supervision of the HDAB, while managing re-identification risks (Article 70(2), EHDS);
- Supporting the implementation of data subject rights, such as opt-out processes that need close collaboration between the pseudonymisation entity (Article 71, EHDS), and significant findings (Recital 67 & Article 58(3), EHDS) where reversible pseudonymisation allows traceability when required by law.

4.2 The concept of pseudonymisation in the context of the EHDS

Pseudonymised data can often be usefully analysed since, in large part, the information content of the original data can still be evaluated (i.e., fidelity of data is preserved). Moreover, if consistent pseudonymisation is used, the insertion of pseudonyms enables the linkage of various records of pseudonymised data relating to the same person without the need to use additional information.[10]

Processing pseudonymised data preserves fidelity of data, as only direct identifiers are replaced by pseudonyms and the data is only subject to data minimisation for the purpose defined by the data applicant in the data access application. To make data available for secondary use in the EHDS, pseudonymisation should be performed as early as possible, taking into account the purposes of the processing (Recital 72, EHDS). Data minimisation pays particular attention to quasi-identifiers, which in conjunction with other attributes imposes increased risk of indirect identification of data subject in the dataset.

Pseudonymisation supports linkage of data across different datasets available at a single data holder, or at multiple data holders within one or even multiple countries (i.e., when data subjects have health records in several Member States). As dataset linkage increases privacy risks (namely re-identification risk and attribute inference), its consequences need to be carefully assessed.

Moreover, it may also be possible to use additional information to link different sets of pseudonymised data whose linkage has not been planned at the outset, i.e., at the time the purposes and means for the processing of secondary use have been determined by the controller or controllers involved. Implementing such linkage should be performed only by persons specifically authorised for this purpose[11], such as the HDAB, for example. See TEHDAS2 M7.5 on data linkage for more details.

Linkage is very relevant in the context of the EHDS, as health data is often split across different data holders (e.g., different hospitals, registers, and biobanks) and different datasets (e.g., different sub-systems comprising the overall hospital information system, which may use different identifiers for the same patient – acting as pseudonyms).

- Linkage is done under the auspices of the HDAB in the EHDS, either by the HDAB itself or by the (trusted) data holders[12].

[10] See §31 EDPB Guidelines 01/2025.
[11] See §33 EDPB Guidelines 01/2025.
[12] See TEHDAS2 M7.5 Guideline for Health Data Access Bodies on linkage of health datasets.

For pseudonymisation to be effective, pseudonymised data must not contain direct identifiers (e.g., personal or social security identifiers) whenever those direct identifiers could be used in the pseudonymisation domain (i.e., the environment where the controller or processor wishes to prelude attribution of data to specific data subjects) to easily attribute the data to the data subjects. To this end, those direct identifiers are removed in the course of the pseudonymising transformation.

Table 2. Role-based involvement for pseudonymisation in the EHDS. The list below addresses data access applications, while similarities for data requests hold. For simplicity reasons, trusted data holders are left out from this table.

Actors	Use journey phase	Task	Activities
Data Holder	Data discovery	Dataset description	Ensuring sufficient information about the dataset (i.e., preferably including information about the dataset, sample distribution and personal data).
HDAB & Data Applicant	Data access	Data access application preparation	Agreement on properties and parameters of the pseudonymisation. Ensuring adequate utility of the dataset for the purpose defined by the applicant. This includes identification of direct identifiers to be removed from the dataset and replaced by the pseudonym.
HDAB (involving data applicant(s) and data holder(s) whenever relevant)	Data access	Data access application evaluation	The process of defining data minimisation parameters is informed by the risks of determined quasi-identifiers (indirect identifiers) during the data access application process.
HDAB	Data preparation	Pseudonymisation transformation pursuant to a data permit	Ensuring adequacy of the process.

Insofar as necessary for pseudonymisation to have the intended effect, it is complemented by the data minimisation process. In the context of pseudonymisation in the EHDS, which comes with complex combination of technical and organisational measures to minimise risks., The pseudonymisation transformation data should only use well-defined documented deterministic operations such as suppression or generalisation, but not other forms of data modification that may non-deterministically reduce fidelity of data.[13]

[13] Based on modified §84 EDPB Guidelines 01/2025.

- Note that this suggested range of techniques is narrower than the list in §84 of the EDPB Guidelines 01/2025. Non-deterministic operations disturb the original data to the point that the user can only have guarantees on statistical properties on the dataset level and not on the individual data values. Noise modulation techniques and other non-deterministic processes are suitable as a part of anonymisation techniques portfolio. Therefore, the choice of the pseudonymisation method may depend on whether data linkage is required for the purpose of processing and should be aligned between relevant involved parties before data access approval.

- In order to prevent unauthorised attribution of pseudonymised data, the pseudonymising transformation regularly involves secret data. The controller may choose these data prior or during the execution of the transformation. These data are often either cryptographic keys (for encryption), salt values for hashing algorithms (for one-way functions) or tables matching pseudonyms with the personal data they replace. This secret data will be stored as a part of pseudonymisation secrets.[14]
- When implementing consistent pseudonymisation, controllers need to define which sets of personal data will be pseudonymised consistently based on objectives of the pseudonymisation. For example, they may decide to pseudonymise all data they collect on the same day consistently allowing for the linkage of two data records pertaining to the same data subject and collected on the same day, but preventing linkage of records of data collected on different days.
    - In particular, three ways to arrange for controlled linkage of pseudonymised data are widely used: person, relationship, and transaction pseudonyms. Note, however, that other ways to segment the pseudonymised data are available and may be appropriate for the respective use case[15].

Generally, pseudonymisation can be reversible or irreversible, depending on whether the pseudonymisation secrets are kept (reversible) or discarded (irreversible). Irreversible pseudonymisation can be used in cases where re-identification is not legally or ethically required or desired. EHDS should rely on reversible pseudonymisation to support implementation of data subject rights such as return of information on significant findings. Also, if pseudonymised and not identifiable data is stored at the data holder, reversible pseudonymisation is needed to implement opt-out for future data access applications. Reversal of pseudonymisation requires cooperation of the entity storing pseudonymisation secrets (e.g., trusted third parties).

4.3 Pseudonymisation with respect to the different phases of the EHDS user journey: from data discovery, access, to data processing

Data discovery phase: Pseudonymisation does not come directly into the discovery phase of the user journey. There are, however, certain aspects related to pseudonymisation, which are recommended to be advertised as a part of the data catalogue using HealthDCAT-AP

[14] See §85 EDPB Guidelines 01/2025.
[15] See §115 EDPB Guidelines 01/2025.

(please consult D5.1, section 7.3.4 for a detailed overview of HealthDCAT-AP properties. For each dataset published by the data holder, it should be indicated:

- if the described original dataset dcat:hasPersonalData (see below);
- if the dataset is categorised as *personal electronic health data* (i.e., genetic data or data concerning health (Article 4 (13), (15), Directive 95/46/EC (General Data Protection Regulation) an open access subset must be provided to ensure meaningful use and interpretation of non-public datasets. Examples are anonymised or synthetic subsets of an original dataset. Information can be included as sample distribution in HealthDCAT-AP, using the dcat:sampledistribution property and accompanied by appropriate documentation (e.g., data dictionaries, codebooks, or a description of the anonymisation method). This approach ensures clarity and avoids mixing metadata related to the original dataset with that of its processed versions.
- if the described dataset is suitable for linkage activities by indicating the properties:
  - hasPersonalData: Indicates whether the dataset includes personal data, and if so, what type. This can support evaluation of linkage feasibility by identifying the presence of shared identifiers or quasi-identifiers (e.g., patient IDs, dates)
  - sampleDistribution: Can provide a structural view of the dataset through a data dictionary (e.g., CSVW) or an anonymised/synthetic dataset. This supports the evaluation of shared variables (e.g., time periods, population segments) that are essential for assessing linkage potential.
  - HealthDCAT-AP does not directly enable linkage but it helps data users and HDABs to evaluate compatibility between datasets for linkage planning. HealthDCAT-AP does provide data holders the possibility to inform data users that a dataset has already been successfully linked in the past with other datasets. All datasets are according DCAT-APinterconnected and the relationships are expressed using DCAT-AP property: dct:source. A related dataset from which the described dataset is derived.

Next to the information already specified in HealthDCAT-AP, it might be helpful to integrate additional information on the dataset regarding:

- pseudonymisation method: reversible vs. irreversible pseudonymisation used at data holder;
- linkage information: Can it be deterministically linked with other datasets from the same data holder, and if yes, with which datasets?
  - Can it be deterministically linked with other datasets from other data holders either specifically (e.g., oncology registries) or generally (e.g., any identifying patient records using the same citizen identifier)?;
  - Can be deterministically linked with other datasets from other data holders in other countries, and if yes, with which datasets either specifically (e.g., hospital records on rare disease patients with patient registers of European Reference Networks) or generally?;

Providing these descriptors will help manage expectations of data applicants during data discovery and streamlines the communication between data applicants and HDABs in the subsequent data access application phase.

Data access application phase: in this phase, an agreement needs to be reached between the data applicant and the HDAB (or TDH) on application of pseudonymisation and resulting data quality and fidelity. This is to ensure that the dataset to be released and paid for by the data applicant meets their requirements specified in the data access application and subsequently stated in the data permit (Article 68(10), EHDS) to ensure traceability and auditability of the agreement. This pertains to:

- When pseudonymised data can be provided to the data user into the SPE. If resources allow, the pseudonymising entity could provide more information on:
  - direct identifiers to be removed from the dataset and replaced by the pseudonym;
  - determined quasi-identifiers and their planned handling as a part of data minimisation.
  - pseudonymisation policy to be applied: deterministic pseudonymisation, or document-randomised pseudonymisation (i.e., randomly generated identifiers of each appearance), or fully randomised pseudonymisation (i.e., for any occurrence)[16]
- When data linking is to be implemented, the determination of direct identifiers and quasi-identifiers needs to consider the whole resulting linked dataset, since some combined data variables may become quasi-identifiers only after the linkage (e.g., in a simple hypothetical scenario, where dataset 1 contains day of birth, dataset 2 contains month of birth, dataset 3 contains year of birth). This may refer to data combination (i.e., when bringing together data from multiple datasets based on one or multiple data permits, or legal basis), but also data linkage (i.e., when bringing together datasets from several sources based on one topic or subject) (see Glossary).

Data preparation phase:

- Scenarios with or without linking data:
  - Identifying data arrives from data holder(s), pseudonymisation is implemented at HDAB(s) or at data holders when HDAB approved processes are in place.
  - Pseudonymised data arrives from data holder(s) and is passed on by the HDAB(s) to the data user without re-pseudonymisation. This is only possible if the data holder can implement per project pseudonym generation and uses a pseudonymisation algorithm approved by the HDAB.
  - Pseudonymised data arrives from data holder(s), data is re-pseudonymised by the HDAB(s) and passed on to the data user.
- Linking-specific scenarios:
  - Linkable pseudonymisation requested *across multiple data holder(s) in a single country*. Linkable pseudonyms are generated by the HDAB, or by a

[16] See p.13 Data Pseudonymisation: Advanced Techniques and Use Cases report by ENISA for more information.

trusted third party (TTP) contracted by the HDAB or defined by law and imposed on data holders.

- Linkable pseudonymisation *across multiple data holder(s) in multiple countries* (including rare diseases, or when cross-border registers are to be linked with other national data holder(s) data). Linkable pseudonyms are generated by a TTP contracted by all the involved HDAB and agreed by each HDAB on the data holder(s) in the specific country. An example of such a TTP is the SPIDER pseudonymisation tool operated by the European Commission[17].
- While there is no formal requirement under the EHDS regulation for such TTPs to be certified or audited under a specific scheme, they must implement appropriate technical and organisational measures in accordance with Article 32 of the GDPR and be subject to contractual oversight by the controllers (i.e., HDABs). The use of a common TTP should be based on mutual agreement between the HDABs and supported by documented responsibilities, safeguards, and auditability provisions.
- Instead of contracting with a TTP, linkable pseudonyms can be generated by an algorithm based on secure multi-party computing[18] when agreed by all the parties involved.
- Linkable pseudonymisation across one or more data holder(s) *and authorised participants*. Linkable pseudonyms are generated by a TTP contracted by all the involved HDABs and authorised participants, or by algorithms based on secure multi-party computing agreed by all the involved HDABs and authorised participants.

Data use phase:

- HDAB (and Trusted data holder where appropriate) is responsible for ensuring that pseudonymised or anonymised data is delivered into the SPE, possibly in collaboration with the data holder(s) (Articles 72(2), 73(2), EHDS);
- HDAB is responsible for imposing adequate technical measures for processing pseudonymised data in the SPE, including requirements on the SPE provider and the data user (requirements on the data user will be formulated in the data permit and become legally binding for the data user);
- HDAB is responsible for imposing adequate data protection-related organisational measures for processing pseudonymised data in the SPE, including requirements on the SPE provider and the data user (requirements on the data user will be formulated in the data permit and become legally binding for the data user).

Finalisation phase:

- Archival of pseudonymised data is possible based on the data permit (subject to Article 68(12), EHDS). Archival refers here to time-limited storage within the SPE for

[17] See https://eu-rd-platform.jrc.ec.europa.eu/spider/.
[18] See https://doi.org/10.1093/bioinformatics/btaa764, https://doi.org/10.1109/TIFS.2021.3114026.

reproducibility purposes, aligned with the validity of the data permit (as described in TEHDAS2 Deliverable 7.1) - possibly on a “cold storage”. Archival should not be interpreted as long-term or open-ended storage.

4.4 Pseudonymisation requirements

Pseudonymisation, risk assessment, and risk management needs to be considered in the overall context of the EHDS. The data access mechanism of the EHDS only allows processing of data in SPEs, which have their own technical and organisational safeguards and must have guarantees on the level of anonymisation of the results or any other output research object (e.g., AI model, software). These safeguards can be taken into consideration with managing risks of quasi-identifiers handling the data minimisation (i.e., if a pseudonymisation domain is in a SPE, data minimisation may accept higher risks in quasi-identifiers taking into account the specific purposes of the processing compared to a completely generic case of pseudonymisation in the pseudonymisation domain).

Recommended pseudonymisation techniques and best practices to achieve reversible pseudonymisation:

- Use of lookup tables:
  - Need to save the complete mapping to implement reversibility.
  - Examples of possible techniques: counters, random number generators.
- Use of cryptographical functions to generate pseudonyms:
  - Need to save function used, exact specification of inputs, value of salt (i.e., a randomly generated list of characters added to the data before creation of the pseudonym, the salt must be kept separate and secure).
  - Recommended to save also complete mappings between unique identifiers in the dataset and pseudonyms if technically feasible (e.g., due to capacity constraints) – to handle situations that input data used to generate the pseudonym is changed at source subsequently after pseudonymisation (e.g., correcting errors in inputs such as names or birth dates or national identifiers), where required.
  - Examples of possible techniques: message authentication code (MAC) (basically hash functions with additional secret key input – so called salt), keyed hash functions, symmetric encryption[19].
    - chosen schemes and their configuration (e.g., size of keyed hash functions) should consider post-quantum security requirements;
    - if the encryption has homomorphic properties, the impact needs to be specifically assessed in terms of risk-to-benefit ratio.
  - Given the architecture of the EHDS, where the process is controlled by the HDAB, and data holders are legally required to cooperate with the HDAB,

[19] See Data Pseudonymisation: Advanced Techniques and Use Cases report by ENISA for more information.

advanced scenarios such as chained pseudonym generation or pseudonyms with the proof of ownership are usually not necessary.

Any pseudonymisation secrets must be stored securely with adequate technical and organisational safeguards in accordance with Article 32 and Article 4(5) (GDPR) in a way, that reversibility of the pseudonymisation can only be implemented by the HDAB or a designated TTP and *not* by the data user (Article 66(3).

The recommended pseudonymisation policy, whether deterministic pseudonymisation, or document-randomised pseudonymisation, or fully randomised pseudonymisation,,[20] depends on the purpose and requirements of data processing. ENISA suggests choosing the pseudonymisation technique based on the identified risk and the identified or expected utilisation of the pseudonymised dataset. Random number generators and MACs are stronger encryption methods as they prevent exhaustive search, dictionary search and random search. However, the pseudonymisation entity may lean towards a combination of different methods due to practicality reasons. Regarding the pseudonymisation policies, fully randomised pseudonymisation offers the best protection, while hindering linkage. Therefore, document-randomised and deterministic methods are recommended, as they allow for data linkage[21]. To enhance data utility for the user, the pseudonymisation policy could be agreed upon between the data user and the HDAB in the data access application phase and stated in the data permit in accordance with (Article 68(10)(a), EHDS) to ensure traceability and auditability. In the context of the EHDS with all other technical and organisational safeguards in place, it is advisable to consider deterministic pseudonymisation, which gives the data user ability to recognise the same patient in the dataset. Only when this is not needed by the data user, one of the randomised policies can be used.

The use of irreversible pseudonymisation is discouraged within the context of the EHDS and should only be used for legacy situations where data is already irreversibly pseudonymised at the data holder. Irreversible pseudonymisation does not allow exercising data subject rights reliably, yet it does not bring the benefits of proper data anonymisation, since not all the attack vectors to be considered for anonymisation are necessarily considered for irreversible pseudonymisation.

- Pseudonyms MUST NOT be reused across different data permits, to minimise risks of intentional or unintentional linkage of data by the data user.
- For data requests, HDABs can repeatedly process datasets with the same pseudonyms, as the data does not leave the HDAB and only a response in an anonymised statistical format is shared with the data user.

4.5 Safeguarding pseudonymised data in the EHDS

- Separation must be implemented between pseudonymisation secrets and pseudonymised data. This includes implementing technical and organisational measures at the entity responsible for pseudonymisation (HDAB or TTP or data holder, depending on the scenario).

[20] See p.13 Data Pseudonymisation: Advanced Techniques and Use Cases.
[21] See p.13 Data Pseudonymisation: Advanced Techniques and Use Cases.

- Identification of quasi-identifiers and handing over information on them to the data minimisation process (see section 5) such as:
    - Minimisation of quasi-identifiers in the dataset as a part of data minimisation process – because during data minimisation, the purpose of processing and requirements on data quality are known (fitness for purpose), which helps to consider which quasi-identifiers need to be released and if any generalisation/suppression/noise generation can be applied to them
    - Provide health-data specific quasi-identifiers beyond what is in the EDPB guidelines (e.g., dates of visits in the hospital also become quickly identifying – possibly from the public health databases and health care registers)

4.6 Data subject rights within the EHDS

- Processing pseudonymised data means processing personal data – hence all the data subject rights apply (but see also Article 11, GDPR).
- Opt-out and significant findings propagation, as these scenarios require reversible pseudonymisation methods to be implemented.
- Opt-out only applies to the new projects. For ongoing project, the opt-out does not apply (see Article 71(3), EHDS).

5 Anonymisation and synthetic data generation

Anonymisation and synthetic data generation techniques can be applied throughout the EHDS data lifecycle to enhance the protection of personal data beyond what is achieved through pseudonymisation alone. Anonymisation and synthetic data generation are distinct approaches to protecting privacy. **Anonymisation** involves modifying or aggregating personal data so that it becomes impossible using reasonably likely means – to re-identify any individual or infer information about them. In contrast, **synthetic data generation** creates entirely new, artificial data that retains the statistical properties of the original dataset. Synthetic data is typically produced using statistical or AI-based models trained on real personal data. This guideline addresses synthetic data generation primarily from a data protection perspective, with the objective of ensuring that the resulting data is anonymous. Beyond data protection, synthetic data may also be generated within the SPE in order to augment datasets, for instance to enhance the performance of deep learning models.to enhance the performance of deep learning models. Although the methodologies for generating anonymised and synthetic data differ, both require the HDAB to establish similar processes and functional components. In particular, the same procedures and functionalities are needed for evaluating data quality metrics, conducting privacy risk assessments (covering both re-identification and inference risks), and enforcing disclosure control measures. Therefore, this chapter considers anonymisation and synthetic data generation in parallel. The main differences between the two approaches lie in the specific technical methods and tools used, which will be addressed separately.

5.1 Objectives

Data protection by using anonymisation and synthetic data generation techniques is especially needed when allowing the data user to export data processing results or data contents from SPEs, but they may also be useful for data users to develop analysis scripts within the SEP, before releasing data for use in the SPE or for enabling data users to test with *public use files* before applying for a data permit. Anonymisation is also essential when data is prepared for use under data request.

It is essential that the EHDS infrastructure provides appropriate guidance and support for assessing the quality of anonymised and synthetic data. From the HDAB's perspective, particular emphasis should be placed on privacy metrics and the assessment of privacy risks. Importantly, privacy risk in this context extends beyond the data subject to include other individuals, such as relatives and health care professionals.

The objectives of this chapter are to:

- Define the contexts (“use cases”) and high-level architecture where anonymisation, synthetic data generation and privacy risk assessment are carried out in the EHDS.
- Provide guidelines concerning the methodology and tools which should be deployed to ensure safe and efficient anonymisation, synthetic data generation and privacy risk assessment implementation.

5.2 Scope and assumptions

5.2.1 Assumptions about data

Foundation in real data. It is assumed that anonymisation and synthetic data generation are performed using real, individual-level data or results derived from such data. For instance, synthetic data generated solely from public aggregate statistics is anonymous by nature and, therefore, falls outside the scope of this deliverable.

Data utility and fidelity. While privacy protection is the primary concern from the HDAB's perspective, assessing the utility and fidelity of anonymised or synthetic data is essential to ensure the data remains suitable for its intended use.

Purpose of data use. It is assumed that data usage is driven by the permitted purposes defined in the EHDS regulation. Anonymisation or synthetic data generation are not standalone permitted purposes for data use. .

Evolving privacy risk. It is recognised that anonymisation and synthetic data generation almost always entail some residual privacy risk. These risks may increase over time as technologies and threat landscapes evolve. Therefore, anonymisation and synthetic data generation and the related privacy risk assessment should be understood as ongoing, adaptive processes that require regular review and maintenance.

Context-dependent risk tolerance. It is recognised that acceptable levels of privacy risk may vary case-by-case, with higher risk potentially tolerated for less sensitive data.

5.2.2 EHDS scope

Responsible actors. Anonymisation, synthetic data generation, and related risk assessments may be carried out by HDABs, (trusted) data holders, or data users, depending on the use case.

Scope within the EHDS. This deliverable focuses on anonymisation, synthetic data generation, and privacy risk assessment activities that support the responsibilities defined in the EHDS regulation. Activities conducted by data holders outside the EHDS infrastructure are not within the scope of this deliverable.

5.2.3 Limitations

Criteria for disclosure decision. Definition of the exact qualitative and quantitative criteria for disclosing anonymised or synthetic data for the data user is not in the scope of this deliverable as such definitions are heavily case-dependent.

Policy considerations. Policies related to granting permissions to data users to create and use anonymised or synthetic data—such as exporting data from the SPE—are outside the scope of this deliverable.

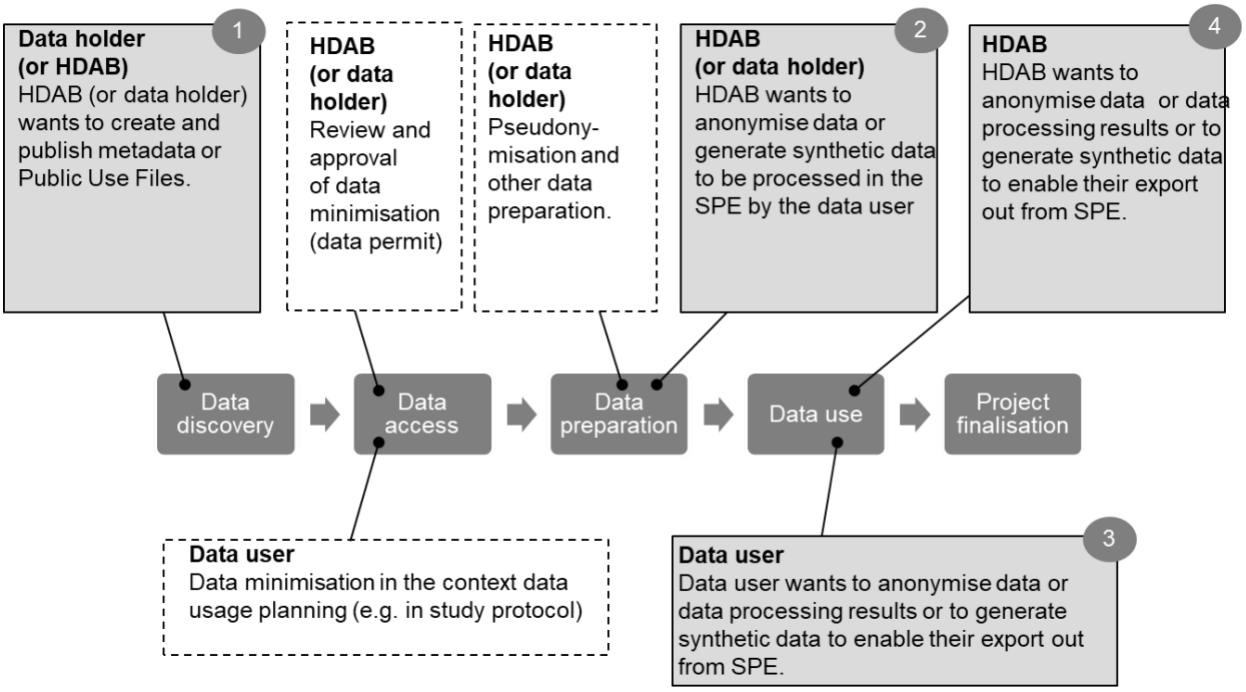
5.3 Use cases

The use cases for data anonymisation, synthetic data generation, and privacy risk assessment are listed in Table 3 and mapped to the EHDS user journey in Figure 1. Use case 1 covers regular EHDS activities that are not tied to a specific data permit. Use cases 2–4 are related to a specific data request or data permit, and the associated anonymisation and synthetic data generation measures must be defined and agreed during the data access application phase. This ensures that both the data applicant and the HDAB are aligned on the conditions for data access before the permit is granted and the data applicant commits to any data retrieval costs.

Table 3. Use cases for data anonymisation, synthetic data generation and privacy risk assessment.

Use case	Roles	Rationale
1. HDAB (or data holder) wants to create and publish metadata or public use files. (Article 57(1l) EHDS)	• HDAB (or data holder) anonymises data or generates synthetic data or metadata • HDAB (or data holder) assesses privacy risk of data or metadata intended to be published.	• Public use files are provided for testing purposes prior to data access application • All datasets must include metadata, which may entail a privacy risk if the dataset is small.
2. HDAB wants to anonymise data or generate synthetic data to be processed in the SPE by the data user (Article 73(1d) EHDS).	• HDAB (or data holder) anonymises data or generates synthetic data • HDAB assesses the privacy risk • HDAB optionally assesses the utility and/or fidelity of the anonymised or synthetic dataset.	• HDAB considers that data can only be released for processing by the data user in the SPE in an anonymised form due to the nature of the project • Evidence on utility/fidelity may be required to ensure that the anonymised or synthetic data sufficiently fulfils the needs of the data user.
3. Data user wants to anonymise data or data processing results or to generate synthetic data to enable their export from the SPE (Articles 61(4) and 73(2) EHDS).	• Data user anonymises datasets or data processing results or generates synthetic data. • Data user and HDAB assesses the privacy risk of the datasets or data processing results • HDAB approves the export of datasets or data processing results.	• Data user needs to export datasets or data processing results for scientific publication and other secondary use purposes.
4. HDAB wants to anonymise data or data processing results or to generate synthetic data to enable their export from the SPE(Articles 61(4) and 73(2) EHDS).	• HDAB anonymises data or generates synthetic for the data user. • HDAB assesses privacy risk • HDAB approves the export of data.	• Data user needs to export datasets for scientific publication and other secondary use purposes.

Figure 5. Anonymisation, synthetic data generation and privacy risks assessment use cases (1-4) mapped on the EHDS user journey. Data minimisation and pseudonymisation activities are covered in section 5 and 6, respectively, of this deliverable.

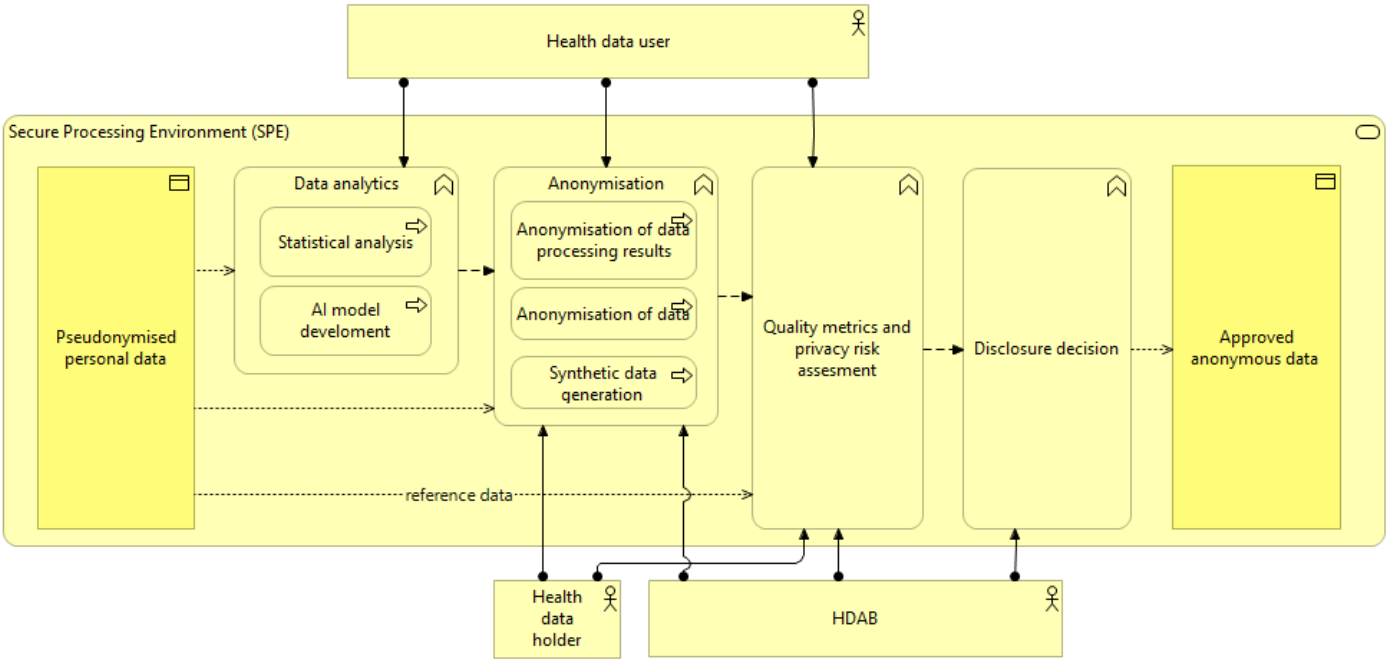


5.4 Architecture

The high-level architecture presented in Figure 6 defines the overall framework within which anonymisation, synthetic data generation, privacy risk assessment, and related activities are performed. These activities are mostly carried out by the HDAB and the data user as outlined in Table 3. The high-level architecture leaves freedom for the HDABs in implementing the detailed processes following existing models for disclosure control. The functions referred in the architecture are designated to take place within a SPE[22]. In this context, the term SPE is understood broadly, encompassing both the HDAB's infrastructure and any SPE provider infrastructure where data is processed in accordance with the EHDS regulation throughout the data lifecycle.

[22] TEHDAS2 milestone document M7.4

Figure 6. High-level architecture for ensuring safe disclosure of anonymised data, data processing results and synthetic data.



The high-level architecture is intended to be agnostic to the use cases and to the methods used for anonymisation or synthetic data generation. Therefore, it is applicable across all use cases and it allows the most appropriate methods to be selected for anonymisation and synthetic data generation.

The components of the architecture are:

Pseudonymised personal data (depending on the use case) can be:

- Use case 1: any individual-level data which the HDAB or data holder uses for generating public use files.
- Use case 2: data permit-specific data which the HDAB uses as basis for generating an anonymised or synthetic dataset to be released for the data user in the SPE.
- Use cases 3-4: data permit-specific data which the data user processes to create data processing results and which data user or HDAB uses as a basis for generating an anonymised or synthetic dataset.

Data analytics refers to the data processing carried out by the data user.

Anonymisation refers to anonymisation of original data[23] or data processing results or generation of synthetic data. The activities can be carried out by the data user, HDAB or data holder, depending on the use case.

[23] In this chapter “original data“ in most cases refer to pseudonymised data released for the data user under approved data permit.

Quality metrics & privacy risk assessment refers to the processes for calculating quality metrics and performing privacy risk assessment on the anonymised data or data processing results or synthetic data. The quality metrics calculation may require the use of original data as a reference. The activities can be carried out by the data user, HDAB or data holder, depending on the use case.

Disclosure decision refers to the processes where the HDAB makes the decision about disclosing the anonymised data or data processing results or synthetic data for the data user or to be used as a public use file.

Approved anonymous data refers to data (or data processing result) which has been deemed to be anonymous and disclosed for the data user to be used within the SPE (use case 2) or to be exported from the SPE (use cases 1, 3 and 4).

5.5 Guidelines

The high-level architecture depicted in Figure 6 outlines a disclosure control framework for the HDAB while allowing flexibility in the choice of methodologies, state-of-the-art technologies and tools. The following guidelines are intended to support the HDAB in establishing the capabilities needed for efficient generation of anonymised and synthetic data as well as for implementing controls that ensure the safe disclosure of anonymous data and processing results. The evaluation of quality metrics and assessment of privacy risk fall under the HDAB's responsibility. However, the metrics and risk assessment methods should also be known to and considered by data users when performing data anonymisation or generating synthetic data.

5.5.1 Documentation of anonymisation or synthetic data generation

Metadata and traceability. All anonymisation and synthetic data generation activities must be thoroughly documented. This documentation should be linked as metadata to the resulting anonymised or synthetic dataset, ensuring that essential information remains available to any future user of the data**.** Most essentially, the documentation should support the process of evaluating the privacy risk and in making the disclosure decision. The documentation will also be highly important when a future user of the data assesses the applicability of the data for a particular purpose.

Responsibility of documentation. All parties involved in anonymisation or synthetic data generation — as well as those responsible for subsequent steps such as calculating quality metrics, assessing privacy risk, and making disclosure decisions — must produce relevant documentation for the actions they perform. The HDAB is responsible for ensuring that appropriate documentation is created as part of the overall disclosure control process. This includes establishing suitable control functions and providing guidance documentation to all relevant stakeholders.

Documentation content. Where appropriate, the following elements are recommended for inclusion in the documentation:

- Metadata of the original dataset that was anonymised or used to generate synthetic data, including provenance information, dataset size, descriptive statistics and data quality and utility label (EHDS regulation, Article 78).
- Information about the anonymised or synthetic data or data processing results, including:
    - Dataset size and descriptive statistics;
    - Information about the anonymisation or synthetic data generation process, including tools, methodology and parameters used;
    - Specific measures applied to ensure privacy protection.
- Results of the assessment of quality metrics and privacy risk.
- Compliance statements and any remarks from the disclosure decision process.

Anonymity of documentation. The documentation shall not contain personal or identifying information, allowing it to be shared with the data user alongside the anonymisation results or synthetic data. Documentation must *not include any personal data or identifiers*, nor contain any reconstruction-relevant information that could compromise anonymity.

Documentation structure. Documentation shall be provided in a structured, standard format. Preferably, the documentation should be machine-readable to support the development of automatic and semi-automatic approval processes. Standard structure for anonymised or synthetic data documentation is currently not existing and should be developed.

5.5.2 Ensuring anonymity of data processing results

Ensuring anonymity of results. HDABs are responsible for ensuring the anonymity of data processing results intended to be exported from the SPE. The HDAB is *legally responsible* for ensuring the anonymity of any data processing results exported from the SPE, in line with Article 73(2) EHDS. Therefore, the HDAB shall establish well-defined criteria for privacy risk assessment, along with clear guidance for both its staff and data users to support them in fulfilling their respective responsibilities[24]. Findata's criteria and guidance[25] for producing anonymous results are a good example covering the following result types:

- descriptive analysis and indicators
- correlations and regression-type analysis
- graphs of different types
- images and other imaging materials
- results based on genome data
- machine learning models
- individual-level result materials
- synthetic data
- results of qualitative research.

[24] Guidelines for output checking.
[25] Producing anonymous results - Findata.

Besides the HDAB, the data user has a responsibility[26] in ensuring that the data processing results and outputs to be exported from the SPE are anonymous and can benefit from the criteria and guidance for privacy risk assessment.

Controlling privacy impacts of machine learning models. Large machine learning (ML) models can be difficult to assess directly for privacy risks[27]. When such assessment is not feasible, the data user shall mitigate privacy risks by training the model on an anonymised dataset or by applying relevant methods (such as differential privacy, DP) during the training process[28,29]. Detailed documentation of the ML development process, including the description of the measures taken to prevent privacy leakage from the ML model shall be included in the anonymisation documentation to support privacy risk assessment and disclosure decision-making.

5.5.3 Anonymisation of individual-level data

Anonymisation methods. HDABs should be prepared to apply commonly used anonymisation techniques and to assess the privacy risk of the resulting anonymised data. A variety of methods are used in healthcare data anonymisation, and the optimal approach depends on both the type of data and its intended use. As a baseline, all direct personal identifiers and pseudonyms must be removed from the dataset. In addition, the following methods are commonly applied:

Anonymisation methods should be selected based on a contextual risk assessment and not applied as fixed recipes. The risk of re-identification must be assessed using state-of-the-art techniques appropriate to the data type and use case.

- Common anonymisation approaches for tabular data are[30,31,32,33,34]:
    - Perturbation-based methods: noise addition, data shuffling, micro-aggregation, data swapping;
    - Generalisation-based methods: *k*-anonymity, *l*-diversity, *t*-closeness;
    - Suppression-based methods: direct suppression, redaction, top and bottom coding, obfuscation of identifiers;
    - Aggregation-based methods: grouping, binning, statistical summarisation
    - Encryption and tokenisation;
- Common anonymisation approaches for imaging and bio-signals data are[35]:

---

[26] EHDS Regulation Article 61(4).
[27] edpb_opinion_202428_ai-models_en.pdf.
[28] https://arxiv.org/pdf/2303.00654.
[29] https://findata.fi/en/services-and-instructions/producing-anonymous-results/#other-result-types .
[30] guide-to-basic-anonymisation-(updated-24-july-2024).pdf.
[31] Perturbation Methods for Protecting Data Privacy: A Review of Techniques and Applications.
[32] Opinion 05/2014 on Anonymisation Techniques.
[33] Anonymisation and Personal Data - Finnish Social Science Data Archive (FSD).
[34] ENISA, Data protection engineering, January 2022.
[35] Ministry of Social Affairs and Health (Finland), VN/23353/2022.

    - Metadata anonymisation, pixel redaction;
    - Image de-identification: face removal, surface rendering anonymisation, skull stripping;
    - Perturbation based methods.
- Common anonymisation approaches for free text are[36]:
    - Named entity recognition (NER) and rule-based filtering;
    - Text generalisation and masking;
    - Text perturbation and data synthesis.
- Common anonymisation methods for genetic data[37]:
    - Suppressing or removing identifiable variants;
    - Reducing resolution of genetic data.

Aggregated data. When a data user is granted access to data under a data request, the individual-level data shall be converted into anonymous statistical format in accordance with the EHDS regulation. The data must be aggregated over a sufficient number of individuals ensuring compliance with relevant k-anonymity criteria.

5.5.1 Synthetic data generation

Methods for synthetic data generation. HDABs should be prepared to apply commonly used synthetic data generation techniques and to assess the privacy risk of the resulting data. A variety of methods are used in generating synthetic data in healthcare, and the optimal approach depends on both the type of data and its intended use[38]. Synthetic data generation methods can be roughly divided into two groups:

- Statistical methods (e.g., gaussian multivariate models, generalized linear models, Bayesian networks, Markov models, decision trees) are used to learn multivariate distributions and relationships from real data, which are then sampled to generate synthetic data.
- Deep generative models—e.g., generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, transformer-based models, and large language models[39] (LLMs)—leverage neural networks to learn complex data distributions and generate synthetic samples.

Reducing privacy risk. Synthetic data may still qualify as personal data under GDPR if it can be linked back to individuals with reasonable effort. This must be considered when

---

[36] De-identification of Free Text Data containing Personal Health Information: A Scoping Review of Reviews | International Journal of Population Data Science.

[37] Computational tools for genomic data de-identification: facilitating data protection law compliance | Nature Communications.

[38] Synthetic data generation methods in healthcare: A review on open-source tools and methods - ScienceDirect.

[39] Large language models and synthetic health data: progress and prospects | JAMIA Open | Oxford Academic.

evaluating privacy risk. The following methods are considered useful for reducing the privacy risk of synthetic data. Their applicability depends on the specific context and type of data:

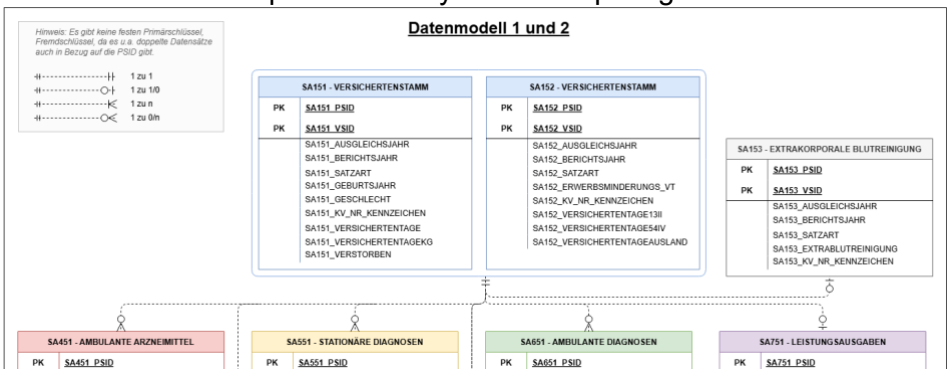
- Incorporating privacy-enhancing techniques into the synthetic data generation process. For example, applying differential privacy during model training can limit the influence of any individual record, thereby improving privacy guarantees without requiring extensive post-processing.
- Post processing of the synthetic data to mitigate residual risks:
    - *'Top and bottom coding'* of continuous variables to reduce the influence of extreme outliers that may resemble real individuals too closely.
    - *Applying distance metrics* (e.g., record-level closeness measures) to identify and remove synthetic data records that are too similar to real data records. While such assessments are typically part of the ex-post evaluation, incorporating them earlier in the generation pipeline can reduce then number of required iterations.
    - *Identifying and mitigating rare attribute combinations* that may pose disclosure risks. This includes techniques such as:
        - Application of *k*-anonymity-like checks to detect and manage unique and low-frequency combinations in synthetic datasets;
        - Detecting and removing decision tree leaves or other model structures that capture subgroups with very small counts.

Note: synthetic data may still qualify as personal data under GDPR if it can be linked back to individuals with reasonable effort. This must be considered when evaluating privacy risk.

A national example - Public Use File. An example of an open source Public Use File (PUF) for claims data from Germany. A ZIP file for each data model is provided, which contains all the variables of this data model.

- The data represented in the PUF underwent several anonymisation steps, which are described here. In this repository you can also find the code that was used to create the PUF, including breaking down the correlations, *k*-anonymity, ID replacement, or picking a simple random 1% sample from the original data.
- The PUFs published represent three different data models (i.e., DM1, DM2, DM3) as over time the amount and structure of the claims data adapted. A detailed description of the data models, entity relationship diagrams (Figure 7) and the content of the individual variables can be found in the publicly available dataset description. Please also consult this website from the health data lab. It is recommended to use the PUFs to develop scripts before applying to get data access, as the data that will be made accessible after a successful data access application to the health data lab, will be in the same format as the PUFs. This includes the variable names, for example. For more information, please consult the website on data usage. Please note that most of the information on the websites in in German.

Figure 7. Shows an excerpt of the entity relationship diagram for DM1 and DM2.



7.5.5 Tooling

The SPE should provide built-in and controllable tooling to support HDABs in fulfilling their regulatory obligations under the EHDS regulation. Such support may include anonymisation of data and data processing results, generating synthetic data[40], assessing anonymisation and privacy risks, validating synthetic data generation, and ensuring the traceability and auditability of all processing actions. These tools should assist HDABs in fulfilling their responsibilities under the EHDS regulation. Where appropriate, the tools should also be made available to data users and data holders. The HDAB is entitled to determine which tools are made available and to define the conditions for their use. All tools used in the SPE for anonymisation, or synthetic data generation must be either provided or explicitly approved by the HDAB to ensure regulatory compliance and auditability.

Data user-provided tools. The HDAB may provide mechanisms for approving, installing, or uploading additional tools requested by data users, where these tools support anonymisation, synthetic data generation, or privacy risk assessment.

Documentation generation. Where appropriate the tools should create anonymisation or synthetic data generation documentation (metadata) automatically or semi-automatically.

Privacy metrics and risk assessment. The functionalities below are considered useful in tools supporting privacy and privacy risk assessment. In particular, the functionalities will support the HDAB in carrying out its duties related to disclosure control[41]. Their applicability may depend on the specific context and type of data:

- Identifier and risk factor detection
    - Detect direct identifiers of data subjects, care personnel, and care organisations.
    - Detect quasi-identifiers or outliers that may lead to re-identification.
    - Classify variables and groups of data subjects according to their sensitivity levels enabling tailored protection strategies.
- Privacy metrics

[40] Synthetic data generation methods in healthcare: A review on open-source tools and methods - ScienceDirect.
[41] EHDS Regulation Article 73(2).

    - Perform similarity analysis between real and anonymised/synthetic datasets using standard distance metrics to quantify differences.
- Re-identification and inference risk assessment
    - Evaluate the overall re-identification risk of datasets and data processing results.
    - Support assessment of re-identification risks associated to membership inference attacks and attribute inference attacks.
    - Identify risks where known individuals may have their presence in the data revealed (membership inference).
    - Detect cases where unknown attributes of known individuals could be inferred from released data (attribute inference).
- Reporting and visualisation
    - Generate structured reports or visual summaries of privacy risk assessments and privacy metric outputs to support decision-making.
    - Parse anonymisation or synthetic data generation metadata to retrieve and visualise parameters used and metrics computed in the earlier processing steps.

Fidelity metrics. The following functionalities are considered useful in tools supporting anonymised and synthetic data fidelity assessment. Their applicability may depend on the specific context and type of data:

- Comparison of dataset variables using standard univariate and multivariate statistical analyses to quantify differences between original and anonymised datasets.
- Analysis of multivariate statistical relationships, such as comparing feature correlation properties between original and anonymised datasets.
- Visual comparison of original and anonymised datasets or related statistical properties such as heatmaps for comparing correlation matrices.
- ML based methods such as data labelling analysis.

Utility of anonymised or synthetic data. Utility assessment tools should support comparison of machine learning models trained on real versus anonymised or synthetic data[42]. This includes evaluating both the similarity of model outputs (e.g., predicted labels, probability distributions) and the model performance metrics (e.g., accuracy, precision, recall, F1-score).

Utility of ML model. Utility assessment[43] tools should enable evaluation of how the use of anonymised data or differential privacy during training affects model utility. This is typically done by comparing the resulting models to reference models trained on the original, non-anonymised dataset.

[42] Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare - ScienceDirect.
[43] Frontiers | Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees.

Examples of applicable tools. The following list is an exemplary, non-exhaustive list of open-source libraries.

For synthetic data generation:

- Synthcity (Python library that makes many existing synthesis methods for different data types easily accessible. Also implements utility and privacy metrics.)
- Synthpop (R package for fast synthesis for single tables. Showed high utility in several benchmark papers.)
- Synthetic Data Vault (Python library that includes several methods for synthetic data generation. Supports single tables, relational datasets including multiple tables and sequential or time series data. Also offers a quality report including several quality metrics. Please note that there is no real open-source license.)
- REaLTabFormer (Python library, GPT-2-based transformer model that can synthesize single tables and relational datasets including multiple tables.)

For synthetic data privacy evaluation:

- Anonymeter (Python library that evaluates different types of privacy risks (singling out, linkability, inference risks) in synthetic tabular data.)
- Shadow model attacks (Python library that can evaluate the privacy-utility trade-off of synthetic data publishing.)
- TAPAS (Python library that can evaluate the privacy of synthetic data within various attack scenarios.)
- SDmetrics (Python library that compares the original data with the synthetic data. Itis model agnostic and provides diagnostic, quality and privacy metrics.)
- Synthcity (Python library that makes many existing synthesis methods for different data types easily accessible. Also implements utility and privacy metrics.)

For synthetic data quality evaluation:

- SDmetrics (Python library that compares the original data with the synthetic data. It is model agnostic and provides diagnostic, quality and privacy metrics.)
- Synthcity (Python library that makes many existing synthesis methods for different data types easily accessible. Also implements utility and privacy metrics.)
- STDG evaluation metrics (Python library to evaluate resemblance, utility and privacy for synthetic tabular data.)

For anonymisation and residual privacy risk assessment

- ARX Data Anonymization Tool (Java software for anonymizing sensitive personal data with various privacy models—including k-anonymity, l-diversity, and differential privacy—and providing tools for risk analysis, data transformation, and utility evaluation).
- sdcTable (R package for applying statistical disclosure control (SDC) to tabular data).

6 Open questions and recommendations

Data minimisation

Open issues related to data minimisation:

Despite the centrality of data minimisation under Articles 5(1)(c) GDPR and 66(1) EHDS, several implementation challenges remain to be addressed:

- Estimating resources and ensuring process efficiency, particularly for large-scale datasets and time-constrained processing in SPEs;
- Determining the appropriate level of data granularity to meet the research purpose without exceeding necessity—especially in complex settings such as cohort selection or decision-tree-based models;
- Adapting variable selection strategies depending on the type of analysis — e.g., ensuring lawful minimisation while supporting feature-rich datasets in machine learning as opposed to traditional statistical models.

Privacy criteria for anonymised and synthetic data
Definition of quantitative privacy criteria for anonymised and synthesized datasets would be highly important for HDAB's but remains an open issue. The relevance of privacy assessment methods and related criteria varies on a case-by-case basis.

It is unlikely that single fixed values for privacy parameters—such as those used in *k*-anonymity, *l*-diversity, or *t*-closeness, or ε for privacy budget in differential privacy—can be defined to cover all use cases. For example, according to responses to the HDAB questionnaire conducted under WP7.2 (see Annex 2 – Methodology), selected k-values ranged between 3 and 100, highlighting the diversity of acceptable thresholds across different use cases. Instead of fixed quantitative values, it may be more practical to specify acceptable parameter ranges and examples depending on the data type, intended use, and context. Such parameter ranges should be available at least for the following parameters:

- re-identification risk
- inference risk
- similarity/distance metrics
- differential privacy parameters

Anonymisation and synthetic data documentation (metadata) structure
Anonymised and synthetic data should be documented in a harmonised format to support the HDAB's disclosure control process and any later usage of a dataset exported from an SPE. Thus, a common, machine-readable documentation format (metadata structure) should be defined. Such format could follow the example of the dataset description template defined for data linkage (see TEHDAS2 M7.5). As outlined in Section 7, this format should include:

- Metadata of the original dataset;
- Information about the anonymised or synthesized dataset size and statistics;
- Information about the process, methodology and tools used;
- Specific measures applied to ensure privacy protection;

- Quality metrics and privacy risk assessment results;
- Compliance statements and any remarks from the disclosure decision process.

Such a format should be standardised as part of the EHDS implementation and continuously updated to reflect evolving techniques and regulatory expectations.

To support harmonised and risk-based implementation across Member States, future EU-level work could focus on the development of:

- Quantitative privacy criteria and recommended parameter ranges adapted to various data types and use cases (e.g., anonymised and synthetic data., and
- A standardised, machine-readable metadata format for documenting anonymised and synthetic datasets (building on Milestone 7.5).

These initiatives would support Health Data Access Bodies (HDABs) in fulfilling their responsibilities under Articles 73, 78 and 79 of the EHDS Regulation, in particular regarding disclosure control, auditability, and quality labelling.

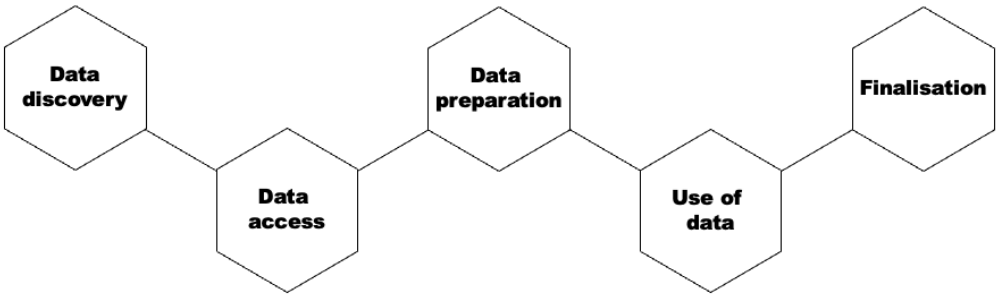
Annexes

Annex number	Annex title
1	EHDS user journey
2	Methodology
3	Glossary

Annex 1 – EHDS user journey

When a data applicant[44] applies for electronic health data for secondary use purposes, such as research and innovation activities, education, and policy-making, within the EHDS, the user journey consists of several phases (see Figure A1-1). 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 A1-1: 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 or data request form to a HDAB[45]. The user must complete the information

[44] Data applicant = a person applying to use electronic health data for a secondary use purpose
[45] 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

required in the form, upload necessary documents, and provide justifications as needed. The data minimisation principle (as per the GDPR) must be respected to ensure privacy.

Data access application 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.

Data preparation

During this phase, the data holder(s)[46] 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 (SPE)[47]. 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 SPE or of receiving the requested health data. The results should be provided in an anonymous format. The data user must inform the HDAB 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.

[46] 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)).

[47] Secure processing environment = an environment with strong technical and security safeguards in which the data user can process personal level electronic health data

Annex 2 – Methodology

The first input to this guideline was based on a survey we developed over the summer of 2024, which served as scoping aid and starting point. During weekly task meetings the results of the survey were discussed and additional external sources incorporated (such as the EDPB Guideline 01/2025 or scientific literature, for example).

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 surveys:

Data minimisation: A total of 122 responses were recorded. Incomplete responses and multiple entries by one Institution were cleaned, which resulted in 29 (full: 25, partial: 4) contributions. The top three personal fields of expertise were: Data management (12), Project management (11) and Data science (9). The most frequent represented countries were Italy (4), Spain (3), Germany (3), Finland (2) and Sweden (2). The top three roles that were represented were: Project coordination (6), Project lead (6) and Director (6). Among the respondents, there were data holders (15), HDABs (12), data users (6), a trusted third party (1) and others (5). The data these respondents are (planning) to make accessible are tables (23), relational databases (19), unstructured data (11), imaging data (9), genomic data (7), bio-sample data (4), and other (3).

Pseudonymisation: A total of 65 responses were recorded. Incomplete responses and multiple entries by one Institution were cleaned, which resulted in 31 (full: 23, partial: 8) contributions. The top three personal fields of expertise were: Data science (14), Project management (13) and Data management (11). The most frequent represented countries were Italy (4), Belgium (3), Germany (3), Finland (2), Sweden (2), Cyprus (2), Luxembourg (2) and Spain (2). The top three roles that were represented were: Project lead (7), Team lead (6), Project coordination (6) and other (6). Among the respondents, there were data holders (15), HDABs (11), data users (7), trusted third parties (3) and others (5). The data these respondents are (planning) to make accessible are tables (22), relational databases (20), unstructured data (15), imaging data (11), genomic data (7), bio-sample data (6), and other (4).

Anonymisation: A total of 43 responses were recorded. Incomplete responses and multiple entries by one Institution were cleaned, which resulted in 27 (full: 23, partial: 4) contributions. The top three personal fields of expertise were: Data science (11), Data management (11) and Project management (10). The most frequent represented countries were Finland (4), Italy (3), Belgium (2), Germany (2), Sweden (2), and Spain (2). The top three roles that were represented were: Project lead (9), Project coordination (7) and other (5). Among the respondents, there were data holders (15), HDABs (8), data users (5), trusted third parties (2) and others (4). The data these respondents are (planning) to make accessible are tables

(21), relational databases (17), unstructured data (11), imaging data (11), genomic data (5), bio-sample data (3), and other (3).

Synthetic data: A total of 54 responses were recorded. Incomplete responses and multiple entries by one Institution were cleaned, which resulted in 23 (full: 21, partial: 2) contributions. The top three personal fields of expertise were: Data science (12), Data management (12) and Project management (11). The most frequent represented countries were Italy (4), Germany (3), Finland (2), Belgium (2) Sweden (2), and Spain (2). The top three roles that were represented were: Project lead (7), Project coordination (6) and other (5). Among the respondents, there were data holders (10), HDABs (8), data users (6), trusted third parties (2) and others (5). The data these respondents are (planning) to make accessible are tables (19), relational databases (16), unstructured data (9), imaging data (7), genomic data (6), bio-sample data (4), and other (2).

Annex 3 – Glossary

Please note that we inserted the definitions below also in the master glossary.

Term	Description
Additional information (related to pseudonymisation)	Additional information is information whose use enables the attribution of ***pseudonymised data*** to identified or identifiable persons (EDPB Guideline 01/2025, Glossary). This term is specific to ***pseudonymisation*** and part of the "additional information" referred to in Regulation (EU) 2016/679 Article 4(5) (GDPR).
Anonymisation	Anonymisation is a process that renders personal data into data that does not relate to an identified or identifiable person, by any means "reasonably likely to be used", taking into account objective factors such as cost, time, and "available technologies at the time" and which therefore, no longer constitutes personal data, in accordance with Recital 26, Regulation (EU) 2016/679 (GDPR). It is recognised that anonymisation processes could be reverted in the future (e.g., by combining several anonymised datasets) and that anonymisation always involves a residual ***re-identification risk***. Robust anonymisation is assessed against singling out, linkability, and inference risks in their release context. (Regulation (EU) 2025/327 (EHDS) Recital 92; Findata, Producing anonymous results (accessed 10.04.2025); p.11-12 WP216; EDPS, 10 Misunderstandings related to anonymisation)
Anonymisation metadata	Refers to a structured set of detailed information describing (a) the methods and parameters used to anonymise a dataset, and (b) the resulting ***quality metrics*** used to anonymise a dataset or data processing result, or to assess their anonymisation. It includes details e.g., on applied techniques and transformation logs. This metadata helps assess data protection, track modifications, and ensure compliance with anonymisation criteria.
Anonymisation result	Refers to the output of anonymisation, which can be an anonymised dataset or a data processing result including ***anonymisation metadata***.
Anonymised statistical format	Refers to aggregated data that does not include information on individual data subjects or entities, also labelled as non-personal aggregated data.
Attribution of pseudonymised data to data subjects	Process that establishes that ***pseudonymised data*** relate to an already identified person, or links the data to other information with reference to

Term	Description
	which the data subjects could be identified. (EDPB Guideline 01/2025, Glossary)
Consistent pseudonymisation	Two sets of data are considered to be pseudonymised consistently if data contained in those sets and relating to the same person can be linked on the basis of the ***pseudonyms*** they contain (EDPB Guideline 01/2025, Glossary). Consistency is context-specific and may be limited to a ***pseudonymisation domain***.
Data aggregation	Process by which information is collected, manipulated and expressed in summary form (ISO/TR 12300:2014(en), 2.1.4)
Data anonymisation framework	Refers to the set of processes and practices designed to ensure data privacy through anonymisation and ***privacy risk assessment***.
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 linkage	The process of combining ***datasets*** "from several sources on one topic or data subject" (ISO 5127:2017, 3.1.11.12). This can be done using unique identifiers, probabilistic methods, or a combination of techniques.
Data minimisation	The data minimisation principle is expressed in Article 5(1)(c) Regulation (EU) 2016/679 (GDPR) and Article 4(1)(c) of Regulation (EU) 2018/1725. In Article 66(1) of the Regulation (EU) 2025/327 (EHDS) it is stated that access is only provided to electronic health data that are "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." Data minimisation applies to all phases of the data lifecycle.
Data permit	"'Data permit' means an administrative decision issued by a health data access body to process certain electronic health data specified in the data permit 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))
Data processing result	Refers to outputs from data processing activities carried out by the health data user. It may be generated from statistical analysis or machine

Term	Description
	learning algorithms, including descriptive statistics, model coefficients, performance indicators, visualisations (e.g., graphs).
Data protection	The “implementation of appropriate administrative, technical or physical means to guard against unauthorized intentional or accidental disclosure, modification, or destruction of data (ISO/IEC 20944-1:2013(en), 3.6.5.1).
Dataset	“’Dataset’ means a structured collection of electronic health data” Regulation (EU) 2025/327 (EHDS), Article 2(2)(w)).
Dataset provenance	Data provenance means a **description of the source** of the data, including context, purpose, method and technology of data generation, documenting agents involved in the provenance of data, data validation routines, source data verification, traceability of changes, and quality control of data (QUANTUM, D1.1).
Direct identifier	A direct identifier is a data element (or set thereof) that has been assigned or is being used to distinguish the data subject it refers to from all others in the given context without requiring the use of ***additional information***. Examples are passport or social security numbers, or the set consisting of first and last name as well as date of birth (EDPB Guideline 01/2025, Glossary).
Fidelity	Fidelity (or resemblance**)** refers to the extent to which processed data—such as anonymised data—retains the statistical properties, relationships, and structural characteristics of the **original data**. High fidelity means that distributions, correlations, and key patterns remain intact.
Health data access body (HDAB)	(…) ”the health data access body should assess the information provided by the health data applicant, based on which it should be able to issue a data permit for the processing of personal electronic health data pursuant to this Regulation that should fulfil the requirements and conditions set out in Chapter IV of this Regulation.” (…) (Regulation (EU) 2025/327 (EHDS), Recital 52). There can be several HDABs per member state and if this is true, there shall be one HDAB acting as coordinator (Regulation (EU) 2025/327 (EHDS), Article 55(1).
Health data holder	“‘Health data holder’ means any natural or legal person, public authority, agency or other body in

Term	Description
	the healthcare or the care sectors, including reimbursement services where necessary, as well as any natural or legal person developing products or services intended for the health, healthcare or care sectors, developing or manufacturing wellness applications, performing research in relation to the healthcare or care sectors or acting as a mortality registry, as well as any Union institution, body, office or agency, that has either: (i) the right or obligation, in accordance with applicable Union or national law and in its capacity as a controller or joint controller, to process personal electronic health data for the provision of healthcare or care or for the purposes of public health, reimbursement, research, innovation, policy making, official statistics or patient safety or for regulatory purposes; or (ii) the ability to make available non-personal electronic health data through the control of the technical design of a product and related services, including by registering, providing, restricting access to or exchanging such data;" (Regulation (EU) 2025/327 (EHDS), Article 2(2)(t))
Health data user	"'Health data user' means 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)).
Irreversible pseudonymisation	A pseudonymisation method where the ***pseudonymising transformation*** cannot be reversed. The information necessary to re-establish the link between the ***pseudonym*** and the ***original data*** has been permanently destroyed or is otherwise unavailable.
Original data	Original data refers to individual-level health data prior to any application of ***pseudonymisation***, ***anonymisation***, or ***synthetic data generation***. It consists of raw data that directly represent real-world individuals. While synthetic data may be informed by knowledge derived from original data

Term	Description
	(e.g., statistical models), the original data itself is not directly transformed into synthetic data and should not be conflated with it.
Privacy (of synthetic or anonymised data)	Privacy measures the extent to which anonymised or synthetic data protects individuals from re-identification, membership inference, or sensitive information leakage. High privacy ensures that no single individual can be traced back to the real dataset, nor can their participation in the ***dataset*** be inferred.
Privacy risk assessment	"Overall process of identifying, analysing, evaluating, consulting, communicating and planning the treatment of potential ***privacy*** impacts with regard to the processing of personally identifiable information (3.7), framed within an organisation's broader risk management framework" (ISO/IEC 29100:2024(en), 3.18). ***Re-identification risk assessment*** falls under privacy risk assessment, together with attribute inference and group membership, for example.
Pseudonym	Identifier that is added to data in the course of the ***pseudonymising transformation*** and set in such a way that it can be attributed to data subjects only using ***additional information***. (EDPB Guideline 01/2025, Glossary)
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))
Pseudonymisation domain	Environment in which the controller or processor wishes to preclude attribution of data to specific data subjects. May incorporate persons acting under the authority of the controller or processor, respectively, other natural or legal persons, public authorities, agencies or other bodies, and their respective technological and informational resources. Does not include persons authorised to process additional data allowing the attribution of the ***pseudonymised data*** to data subjects. (EDPB Guideline 01/2025, Glossary)
Pseudonymisation entity	The entity responsible of processing identifiers into pseudonyms using the pseudonymisation function. It can be a data controller, a data processor (performing pseudonymisation on

Term	Description
	behalf of a controller), a ***trusted third party*** or a data subject, depending on the pseudonymisation scenario. It should be stressed that, following this definition, the role of the pseudonymisation entity is strictly relevant to the practical implementation of pseudonymisation under a specific scenario. (ENISA, Pseudonymisation techniques and best practices, p. 10 - *modified*)
Pseudonymisation secrets	Data that is used in the application of the ***pseudonymising transformation*** or is created during that process. Usually *(…)* cryptographic keys or salts, for example. Allows the computation of pseudonyms from certain identifying attributes. Part of ***additional information***. (EDPB Guideline 01/2025, Glossary, *modified*)
Pseudonymised data	Result of applying the ***pseudonymising transformation*** to some personal data. Cannot be attributed to a specific data subject without ***additional information***. (EDPB Guideline 01/2025, Glossary)
Pseudonymising controller or processor	Controller or processor that uses pseudonymisation as a safeguard and modifies ***original data*** according to Regulation (EU) 2016/679 (GDPR) Article 4(5). (EDPB Guideline 01/2025, Glossary)
Pseudonymising transformation	Procedure that modifies ***original data*** in a way that the result cannot be attributed to a specific data subject without ***additional information***. (EDPB Guideline 01/2025, Glossary)
Public use file	A ***dataset*** made available to the public, typically containing anonymised, synthetic or aggregated data to protect individual privacy. These files can be released to data users for information and testing purposes before they apply for a data permit. It is based on **original data**.
Quality metrics	Refer to qualitative and quantitative indicators used to assess the fitness for purpose of a dataset. In the context of synthetic and anonymised data, quality metrics are particularly relevant to evaluate how transformations affect the data's ***utility***, ***fidelity***, and ***privacy***. Quality metrics may also be used to assess pseudonymised or original datasets, particularly when serving as a benchmark or when evaluating fitness for specific secondary use purposes. (Adapted from ISO and EHDS principles, see Regulation (EU) 2015/327 (EHDS) Article 66 and Recital 58).

Term	Description
Quality metrics evaluation	Refers to the calculation or derivation of the ***quality metrics***.
Quality metrics tool	Quality metrics tool (or "metrics tool") refers to a software, an algorithm, a processing pipeline, a documented manual process, or a combination of these, designed to perform ***quality metrics evaluation***.
Quasi-identifier	A ***dataset*** attribute that, when considered in conjunction with other attributes are sufficient to attribute at least part of the pseudonymised data to data subjects (EDPB Guideline 01/2025, §101).
Re-identification	The "process of associating data in a de-identified ***dataset*** with the original data principal" (i.e., data subject) (ISO/IEC 20889:2018(en), 3.31).
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).
Re-pseudonymisation	The processing of ***pseudonymised data***, where project ***pseudonyms*** are generated using a pseudonymisation algorithm, replacing previously generated pseudonyms. Re-pseudonymisation should not be confused with attempts to reverse the pseudonymisation, which is not meant here.
Reversible pseudonymisation	The ***pseudonymisation entity*** uses a ***pseudonymising transformation*** process that allows the pseudonymisation entity to reverse the ***pseudonym***, if necessary. For example, by using separately kept matching tables of pseudonyms and identifying data, or computable secrets allowing for calculating back to the original input.
Secure Processing Environment (SPE)	"'Secure processing environment' means the physical or virtual environment and organisational means to ensure compliance with Union law, such as Regulation (EU) 2016/679, in particular with regard to data subjects' rights, intellectual property rights, and commercial and statistical confidentiality, integrity and accessibility, as well as with applicable national law, and to allow the entity providing the secure processing environment to determine and supervise all data processing actions, including the display, storage, download and export of data and the calculation of derivative data through computational algorithms;" (Regulation (EU) 2022/868 (DGA), Article 2(20); and reference to Regulation (EU) 2025/327 (EHDS), Article 2(1)(c)).

Term	Description
Sensitive data	Data with potentially harmful effects in the event of disclosure (i.e., providing access to data to a third party) or misuse (ISO 5127:2017(en), 3.1.10.16)).
Statistical disclosure control	Statistical disclosure control can be defined as the set of "methods to reduce the risk of disclosing information on the statistical units (natural persons, households, economic operators and other undertakings, referred to by the data), usually based on restricting the amount of, or modifying, the data released" (Eurostat CROS).
Synthetic data	"The concept of synthetic data generation is to take an original data source (dataset) and create new, artificial data, with similar statistical properties from it. Keeping the statistical properties means that anyone analysing the synthetic data, a data analyst for example, should be able to draw *similar* statistical conclusions from the analysis of a given dataset of synthetic data as he/she would if given the real ***(original) data***." (...) (Phiri glossary, *modified*, see also synthetic data vs mockup data).
Synthetic data documentation	Documentation of a synthetic dataset generated automatically or semi-automatically by the ***synthetic data generator***. The documentation shall be anonymised so that it can be accompanied by the synthetic dataset when released for the data user or for public use.
Synthetic data generator	A synthetic data generator is a software application, model or algorithm designed to generate ***synthetic data***. It uses real-world data as input and generates a synthetic dataset. It is also possible to use parameters derived from the ***original data*** as input and/or modify additional parameters entered by the user.
Trusted health data holder (TDH)	A member state designated health data holder for whom a simplified procedure can be followed for the issuance of data permits. Trusted data holders leverage their expertise on the data they hold to assist the HDAB by providing assessments of data requests or access applications. Once data permits are authorised, these trusted data holders provide the data within an SPE that they manage (Regulation (EU) 2025/327 (EHDS), Article 72).
Trusted third party (TTP)	A ***pseudonymisation entity*** which is independent from the data user and data holder that processes identifiers into pseudonyms. (ENISA, Pseudonymisation techniques and best practices, p. 10, *modified*). The TTP needs only to

Term	Description
	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).
Utility	Utility refers to how well the data supports its intended use, such as syntactical testing, analytical tasks, decision-making, or machine learning model performance. In the context of anonymised and synthetic data high utility means that insights, predictions, or outcomes derived from the data closely match those obtained using the ***original data***.