INTERNATIONAL STANDARD

ISO 25237

First edition
2017-01

Health informatics — Pseudonymization

Informatique de santé — Pseudonymisation

Reference number
ISO 25237:2017(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword v

Introduction vi

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Abbreviated terms 6

5 Requirements for privacy protection of identities in healthcare 7
5.1 Objectives of privacy protection 7
5.2 General 7
5.3 De-identification as a process to reduce risk 8
5.3.1 General 8
5.3.2 Pseudonymization 8
5.3.3 Anonymization 9
5.3.4 Direct and indirect identifiers 9
5.4 Privacy protection of entities 9
5.4.1 Personal data versus de-identified data 9
5.4.2 Concept of pseudonymization 11
5.5 Real world pseudonymization 13
5.5.1 Rationale 13
5.5.2 Levels of assurance of privacy protection 14
5.6 Categories of data subject 16
5.6.1 General 16
5.6.2 Subject of care 16
5.6.3 Health professionals and organizations 16
5.6.4 Device data 16
5.7 Classification data 17
5.7.1 Payload data 17
5.7.2 Observational data 17
5.7.3 Pseudonymized data 17
5.7.4 Anonymized data 17
5.8 Research data 17
5.8.1 General 17
5.8.2 Generation of research data 18
5.8.3 Secondary use of personal health information 18
5.9 Identifying data 18
5.9.1 General 18
5.9.2 Healthcare identifiers 18
5.10 Data of victims of violence and publicly known persons 19
5.10.1 General 19
5.10.2 Genetic information 19
5.10.3 Trusted service 19
5.10.4 Need for re-identification of pseudonymized data 19
5.10.5 Pseudonymization service characteristics 20

6 Protecting privacy through pseudonymization 20
6.1 Conceptual model of the problem areas 20
6.2 Direct and indirect identifiability of personal information 21
6.2.1 General 21
6.2.2 Person identifying variables 21
6.2.3 Aggregation variables 21
6.2.4 Outlier variables 22
6.2.5 Structured data variables 22
6.2.6 Non-structured data variables 23

6.2.7 Inference risk assessment ... 23
6.2.8 Privacy and security ... 24

7 Re-identification process ... 24
7.1 General ... 24
7.2 Part of normal procedures ... 24
7.3 Exception ... 24
7.4 Technical feasibility ... 25

Annex A (informative) **Healthcare pseudonymization scenarios ... 26**

Annex B (informative) **Requirements for privacy risk analysis ... 39**

Annex C (informative) **Pseudonymization process (methods and implementation) ... 49**

Annex D (informative) **Specification of methods and implementation ... 55**

Annex E (informative) **Policy framework for operation of pseudonymization services (methods and implementation) ... 56**

Annex F (informative) **Genetic information ... 60**

Bibliography ... 61

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

The committee responsible for this document is ISO/TC 215, *Health informatics*.

Introduction

Pseudonymization is recognized as an important method for privacy protection of personal health information. Such services may be used nationally, as well as for trans-border communication.

Application areas include, but are not limited to:

- indirect use of clinical data (e.g. research);
- clinical trials and post-marketing surveillance;
- pseudonymous care;
- patient identification systems;
- public health monitoring and assessment;
- confidential patient-safety reporting (e.g. adverse drug effects);
- comparative quality indicator reporting;
- peer review;
- consumer groups;
- field service.

This document provides a conceptual model of the problem areas, requirements for trustworthy practices, and specifications to support the planning and implementation of pseudonymization services.

The specification of a general workflow, together with a policy for trustworthy operations, serve both as a general guide for implementers but also for quality assurance purposes, assisting users of the pseudonymization services to determine their trust in the services provided. This guide will serve to educate organizations so they can perform pseudonymization services themselves with sufficient proficiency to achieve the desired degree of quality and risk reduction.

Health informatics — Pseudonymization

1 Scope

This document contains principles and requirements for privacy protection using pseudonymization services for the protection of personal health information. This document is applicable to organizations who wish to undertake pseudonymization processes for themselves or to organizations who make a claim of trustworthiness for operations engaged in pseudonymization services.

This document

- defines one basic concept for pseudonymization (see Clause 5),
- defines one basic methodology for pseudonymization services including organizational, as well as technical aspects (see Clause 6),
- specifies a policy framework and minimal requirements for controlled re-identification (see Clause 7),
- gives an overview of different use cases for pseudonymization that can be both reversible and irreversible (see Annex A),
- gives a guide to risk assessment for re-identification (see Annex B),
- provides an example of a system that uses de-identification (see Annex C),
- provides informative requirements to an interoperability to pseudonymization services (see Annex D), and
- specifies a policy framework and minimal requirements for trustworthy practices for the operations of a pseudonymization service (see Annex E).

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 27799, *Health informatics — Information security management in health using ISO/IEC 27002*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at http://www.electropedia.org/
- ISO Online browsing platform: available at http://www.iso.org/obp

3.1
access control
means of ensuring that the resources of a data processing system can be accessed only by authorized entities in authorized ways

[SOURCE: ISO/IEC 2382:2015, 2126294]

3.2
anonymization
process by which *personal data* (3.37) is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party

Note 1 to entry: The concept is absolute, and in practice, it may be difficult to obtain.

[SOURCE: ISO/IEC 29100:2011, 2.2, modified.]

3.3
anonymized data
data (3.14) that has been produced as the output of an *anonymization* (3.2) process

[SOURCE: ISO/IEC 29100:2011, 2.3, modified.]

3.4
anonymous identifier
identifier (3.27) of a person which does not allow the *identification* (3.26) of the *natural person* (3.34)

3.5
authentication
assurance of the claimed identity

3.6
attacker
person deliberately exploiting vulnerabilities in technical and non-technical security controls in order to steal or compromise information systems and networks, or to compromise availability to legitimate users of information system and network resources

[SOURCE: ISO/IEC 27033-1:2015, 3.3]

3.7
ciphertext
data (3.14) produced through the use of encryption, the semantic content of which is not available without the use of cryptographic techniques

[SOURCE: ISO/IEC 2382:2015, 2126285]

3.8
confidentiality
property that *information* (3.29) is not made available or disclosed to unauthorized individuals, entities or processes

[SOURCE: ISO 7498-2:1989, 3.3.16]

3.9
content-encryption key
cryptographic key used to encrypt the content of a communication

3.10
controller
natural or legal person, public authority, agency or any other body which, alone or jointly with others, determines the purposes and means of the *processing of personal data* (3.40)

3.11
cryptography
discipline which embodies principles, means and methods for the transformation of *data* (3.14) in order to hide its information content, prevent its undetected modification and/or prevent its unauthorized use

[SOURCE: ISO 7498-2:1989, 3.3.20]

3.12
cryptographic algorithm
<cipher> method for the transformation of *data* (3.14) in order to hide its information content, prevent its undetected modification and/or prevent its unauthorized use

3.13
cryptographic key management
key management
generation, storage, distribution, deletion, archiving and application of *keys* (3.31) in accordance with a *security policy* (3.46)

[SOURCE: ISO 7498-2:1989, 3.3.33]

3.14
data
reinterpretable representation of *information* (3.29) in a formalized manner suitable for communication, interpretation or processing

Note 1 to entry: Data can be processed by humans or by automatic means.

[SOURCE: ISO/IEC 2382:2015, 2121272]

3.15
data integrity
property that *data* (3.14) has not been altered or destroyed in an unauthorized manner

[SOURCE: ISO 7498-2:1989, 3.3.21]

3.16
data linking
matching and combining *data* (3.14) from multiple databases

3.17
data protection
technical and social regimen for negotiating, managing and ensuring informational *privacy* (3.39), and security

3.18
data subject
person to whom *data* (3.14) refer

3.19
decryption
process of converting encrypted *data* (3.14) back into its original form so it can be understood

3.20
de-identification
general term for any process of reducing the association between a set of identifying *data* (3.14) and the *data subject* (3.18)

3.21
directly identifying data
data (3.14) that directly identifies a single individual

Note 1 to entry: Direct identifiers are those data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain.

3.22
disclosure
divulging of, or provision of access to, *data* (3.14)

Note 1 to entry: Whether the recipient actually looks at the data, takes them into knowledge or retains them, is irrelevant to whether disclosure has occurred.

3.23
encryption
process of converting *information* (3.29) or *data* (3.14) into a cipher or code

3.24
healthcare identifier
subject of care identifier
identifier (3.27) of a person for primary use by a healthcare system

3.25
identifiable person
one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity

[SOURCE: Directive 95/46/EC]

3.26
identification
process of using claimed or observed attributes of an entity to single out the entity among other entities in a set of identities

Note 1 to entry: The identification of an entity within a certain context enables another entity to distinguish between the entities with which it interacts.

3.27
identifier
information (3.29) used to claim an identity, before a potential corroboration by a corresponding authenticator

[SOURCE: ENV 13608-1:2000, 3.44]

3.28
indirectly identifying data
data (3.14) that can identify a single person only when used together with other indirectly identifying data

Note 1 to entry: Indirect identifiers can reduce the population to which the person belongs, possibly down to one if used in combination.

EXAMPLE Postcode, sex, age, date of birth.

3.29
information
knowledge concerning objects that within a certain context has a particular meaning

[SOURCE: ISO/IEC 2382:2015, 2121271, modified.]

3.30
irreversibility
situation when, for any passage from identifiable to pseudonymous, it is computationally unfeasible to trace back to the original *identifier* (3.27) from the *pseudonym* (3.43)

3.31
key
sequence of symbols which controls the operations of *encryption* (3.23) and *decryption* (3.19)

[SOURCE: ISO 7498-2:1989, 3.3.32]

3.32
linkage of information objects
process allowing a logical association to be established between different information objects

3.33
longitudinal or lifetime personal health record
permanent, coordinated record of significant information, in chronological sequence

Note 1 to entry: It may include all historical data collected or be retrieved as a user designated synopsis of significant demographic, genetic, clinical and environmental facts and events maintained within an automated system.

[SOURCE: ISO/TR 21089:2004, 3.61, modified]

3.34
natural person
real human being as opposed to a legal person which may be a private or public organization

3.35
person identification
process for establishing an association between an information object and a physical person

3.36
personal identifier
information with the purpose of uniquely identifying a person within a given context

3.37
personal data
information relating to an identified or identifiable *natural person* (3.34) ("data subject")

[SOURCE: Directive 95/46/EC]

3.38
primary use of personal data
uses and *disclosures* (3.22) that are intended for the *data* (3.14) collected

3.39
privacy
freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of *data* (3.14) about that individual

[SOURCE: ISO/IEC 2382:2015, 2126263]

3.40
processing of personal data
operation or set of operations that is performed upon *personal data* (3.37), whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction

[SOURCE: Directive 95/46/EC]

3.41
processor
natural or legal person, public authority, agency or any other body that processes *personal data* (3.37) on behalf of the *controller* (3.10)

Note 1 to entry: See Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

3.42
pseudonymization
particular type of *de-identification* (3.20) that both removes the association with a *data subject* (3.18) and adds an association between a particular set of characteristics relating to the data subject and one or more *pseudonyms* (3.43)

3.43
pseudonym
personal identifier (3.36) that is different from the normally used personal identifier and is used with pseudonymized data to provide dataset coherence linking all the information about a subject, without disclosing the real world person identity.

Note 1 to entry: This may be either derived from the normally used personal identifier in a reversible or irreversible way or be totally unrelated.

Note 2 to entry: Pseudonym is usually restricted to mean an identifier that does not allow the direct derivation of the normal personal identifier. Such pseudonymous information is thus functionally anonymous. A trusted third party may be able to obtain the normal personal identifier from the pseudonym.

3.44
recipient
natural or legal person, public authority, agency or any other body to whom *data* (3.14) are disclosed

3.45
secondary use of personal data
uses and *disclosures* (3.22) that are different than the initial intended use for the *data* (3.14) collected

3.46
security policy
plan or course of action adopted for providing computer security

[SOURCE: ISO/IEC 2382:2015, 2126246]

3.47
trusted third party
security authority, or its agent, trusted by other entities with respect to security-related activities

[SOURCE: ISO/IEC 18014-1:2008, 3.20]

4 Abbreviated terms

DICOM	Digital Imaging and Communication in Medicine
HIPAA	Health Insurance Portability and Accountability Act
HIS	Health Information System
HIV	Human Immunodeficiency Virus
IP	Internet Protocol
VoV	Victim of Violence use

5 Requirements for privacy protection of identities in healthcare

5.1 Objectives of privacy protection

The objective of privacy protection as part of the confidentiality objective of security is to prevent the unauthorized or unwanted disclosure of information about a person which may further influence legal, organizational and financial risk factors. Privacy protection is a subdomain of generic privacy protection that, by definition, includes other privacy sensitive entities such as organizations. As privacy is the best regulated and pervasive one, this conceptual model focuses on privacy. Protective solutions designed for privacy can also be transposed for the privacy protection of other entities. This may be useful in countries where the privacy of entities or organizations is regulated by law.

There are two objectives in the protection of personal data; one that is the protection of personal data in interaction with on-line applications (e.g. web browsing) and at the other is the protection of collected personal data in databases. This document will restrict itself to the latter objective.

Data can be extracted from databases. The objective is to reduce the risk that the identities of the data subjects are disclosed. Researchers work with “cases”, longitudinal histories of patients collected in time and/or from different sources. For the aggregation of various data elements into the cases, it is, however, necessary to use a technique that enables aggregations without endangering the privacy of the data subjects whose data are being aggregated. This can be achieved by pseudonymization of the data.

De-identification is used to reduce privacy risks in a wide variety of situations.

Extreme de-identification is used for educational materials that will be made widely public, yet should convey enough detail to be useful for medical education purposes (there is an IHE profile for automation assistance for performing this kind of de-identification. Much of the process is customized to the individual patient and educational purpose).

Public health uses de-identified databases to track and understand diseases.

Clinical trials use de-identification both to protect privacy and to avoid subconscious bias by removing other information such as whether the patient received a placebo or an experimental drug.

Slight de-identification is used in many clinical reviews, where the reviewers are kept ignorant of the treating physician, hospital, patient, etc. both to reduce privacy risks and to remove subconscious biases. This kind of de-identification only prevents incidental disclosure to reviewers. An intentional effort will easily discover the patient identity, etc.

When undertaking production of workload statistics or workload analysis within hospitals or of treatments provided against contracts with commissioners or purchasers of health care services, it is necessary to be able to separate individual patients without the need to know who the individual patients are. This is an example of the use of de-identification within a business setting.

The process of risk stratification (of re-hospitalization, for example) can be undertaken by using records from primary and secondary care services for patients. The records are de-identified for the analysis, but where the patients that are indicated as being of high risk, these patients can be re-identified by an appropriate clinician to enable follow-up interventions. For details on the healthcare pseudonymizaton, see Annex A.

5.2 General

De-identification is the general term for any process of reducing the association between a set of identifying data and the data subject with one or more intended use of the resulting data-set. Pseudonymization is a subcategory of de-identification. The pseudonym is the means by which pseudonymized data are linked to the same person or information systems without revealing the identity of the person. De-identification inherently can limit the utility of the resulting data. Pseudonymization can be performed with or without the possibility of re-identifying the subject of the data (reversible or irreversible pseudonymization). There are several use case scenarios in healthcare for pseudonymization with particular applicability in increasing electronic processing of patient data, together with increasing patient expectations for privacy protection. Several examples of these are provided in Annex A.

It is important to note that as long as there are any pseudonymized data, there is some risk of unauthorized re-identification. This is not unlike encryption, in that brute force can crack encryption, but the objective is to make it so difficult that the cost is prohibitive. There is less experience with de-identification than encryption so the risks are not as well understood.

5.3 De-identification as a process to reduce risk

5.3.1 General

The de-identification process should consider the security and privacy controls that will manage the resulting data-set. It is rare to lower the risk so much that the data-set needs no ongoing security controls.

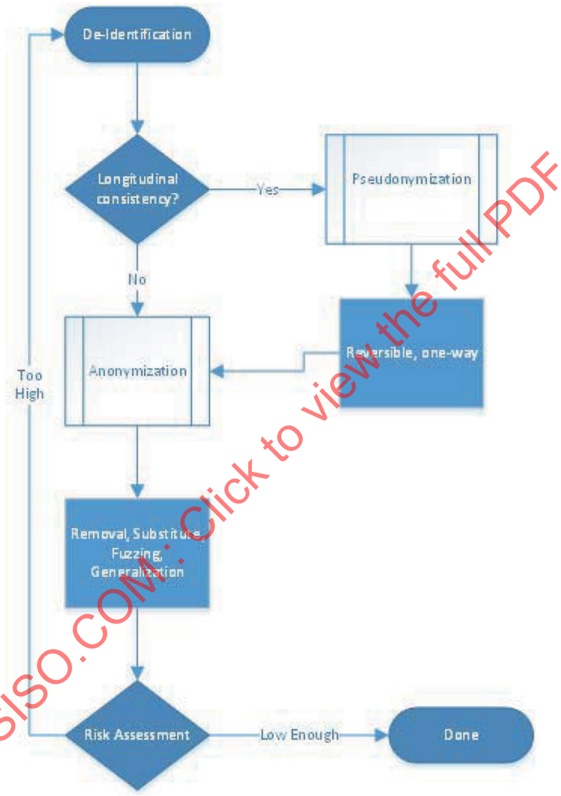


Figure 1 — Visualization of the de-identification process

Figure 1 is an informative diagram of a visualization of this de-identification process. This shows that the topmost concept is de-identification, as a process. This process utilizes sub-processes: pseudonymization and/or anonymization. These sub-processes use various tools that are specific to the type of data element they operate on, and the method of risk reduction.

The starting state is that zero data are allowed to pass through the system. Each element should be justified by the intended use of the resulting data-set. This intended use of the data-set greatly affects the de-identification process.

5.3.2 Pseudonymization

De-identification might leverage pseudonymization where longitudinal consistency is needed. This might be to keep a bunch of records together that should be associated with each other, where without this longitudinal consistency, they might get disassociated. This is useful to keep all of the records for a patient together, under a pseudonym. This also can be used to assure that each time data are extracted into a de-identified set that new entries are also associated with the same pseudonym. In pseudonymization, the algorithm used might be intentionally reversible or intentionally not-reversible. A reversible scheme might be a secret lookup-table that where authorized can be used to discover the original identity. In a non-reversible scheme, a temporary table might be used during the process, but is destroyed when the process completes.

5.3.3 Anonymization

Anonymization is the process and set of tools used where no longitudinal consistency is needed. The anonymization process is also used where pseudonymization has been used to address the remaining data attributes. Anonymization utilizes tools like redaction, removal, blanking, substitution, randomization, shifting, skewing, truncation, grouping, etc. Anonymization can lead to a reduced possibility of linkage.

Each element allowed to pass should be justified. Each element should present the minimal risk, given the intended use of the resulting data-set. Thus, where the intended use of the resulting data-set does not require fine-grain codes, a grouping of codes might be used.

5.3.4 Direct and indirect identifiers

De-identification process addresses three kinds of data: direct identifiers, which by themselves identify the patient; indirect identifiers, which provide correlation when used with other indirect or external knowledge; and non-identifying data, the rest of the data.

Usually, a de-identification process is applied to a data-set, made up of entries that have many attributes. For example, a spreadsheet made up of rows of data organized by column.

The de-identification process, including pseudonymization and anonymization, are applied to all the data. Pseudonymization generally are used against direct identifiers, but might be used against indirect identifiers, as appropriate to reduce risk while maintaining the longitudinal needs of the intended use of the resulting data-set. Anonymization tools are used against all forms of data, as appropriate to reduce risk.

5.4 Privacy protection of entities

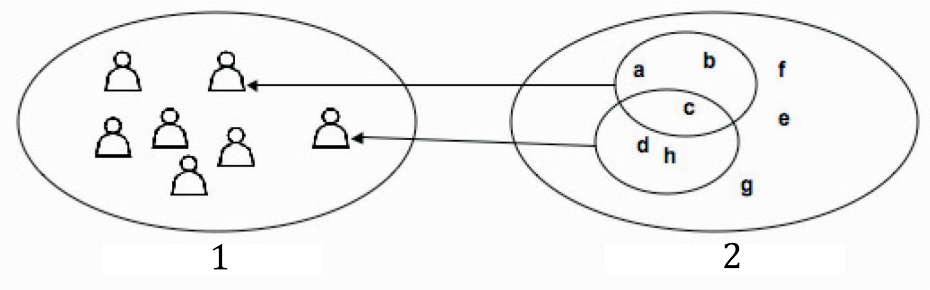
5.4.1 Personal data versus de-identified data

5.4.1.1 Definition of personal data

According to Reference [18], “personal data” shall mean any information relating to an identified or identifiable natural person (“data subject”); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

This concept is addressed in other national legislation with consideration for the same principles found in this definition (e.g. HIPAA).

5.4.1.2 Idealized concept of identification and de-identification



Key

1 set of data subjects

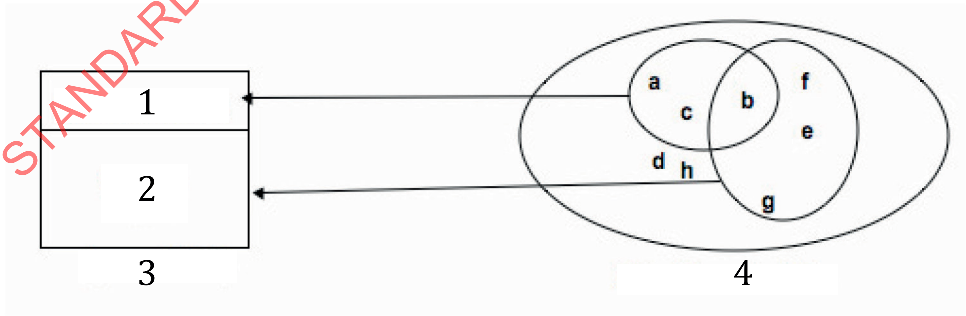
2 set of characteristics

Figure 2 — Identification of data subjects

This subclause describes an idealized concept of identification and de-identification. It is assumed that there are no data outside the model as shown in Figure 2, for example, that may be linked with data inside the model to achieve (indirect) identification of data subjects.

In 5.4.1, potential information sources outside the data model will be taken into account. This is necessary in order to discuss re-identification risks. Information and communication technology projects never picture data that are not used within the model when covering functional design aspects. However, when focusing on identifiability, critics bring in information that could be obtained by an attacker in order to identify data subjects or to gain more information on them (e.g. membership of a group).

As depicted in Figure 1, a data subject has a number of characteristics (e.g. name, date of birth, medical data) that are stored in a medical database and that are personal data of the data subject. A data subject is identified within a set of data subjects if they can be singled out. That means that a set of characteristics associated with the data subject can be found that uniquely identifies this data subject. In some cases, only one single characteristic is sufficient to identify the data subject (e.g. if the number is a unique national registration number). In other cases, more than one characteristic is needed to single out a data subject, such as when the address is used of a family member living at the same address. Some associations between characteristics and data subjects are more persistent in time (e.g. a date of birth, location of birth) than others (e.g. an e-mail address).



Key

1 identifying data

2 payload data

3 personal data

4 set of characteristics

Figure 3 — Separation of personal data from payload data

From a conceptual point of view, personal data can be split up into two parts according to identifiability criteria (see Figure 3):

- payload data: the data part, containing characteristics that do not allow unique identification of the data subject; conceptually, the payload contains anonymous data (e.g. clinical measurements, machine measurements);
- identifying data: the identifying part that contains a set of characteristics that allow unique identification of the data subject (e.g. demographic data).

Note that the conceptual distinction between "identifying data" and "payload data" can lead to contradictions. This is the case when directly identifying data are considered "payload data". Any pseudonymization method should strive to reduce the level of directly identifying data, for example, by aggregating these data into groups. In particular cases (e.g. date of birth of infants), where this is not possible, the risk should be pointed out in the policy document. A following section of this document deals with the splitting of the data into the payload part and the identifying part from a practical point of view, rather than from a conceptual point of view. From a conceptual point of view, it is sufficient that it is possible to obtain this division. It is important to note that the distinction between identifying characteristics and payload are not absolute. Some data that is also identifying might be needed for the research, e.g. year and month of birth. These distinctions are covered further on.

5.4.2 Concept of pseudonymization

The practice and advancement of medicine require that elements of private medical records be released for teaching, research, quality control and other purposes. For both scientific and privacy reasons, these record elements need to be modified to conceal the identities of the subjects.

There is no single de-identification procedure that will meet the diverse needs of all the medical uses while providing identity concealment. Every record release process shall be subject to risk analysis to evaluate the following:

a) the purpose for the data release (e.g. analysis);

b) the minimum information that shall be released to meet that purpose;

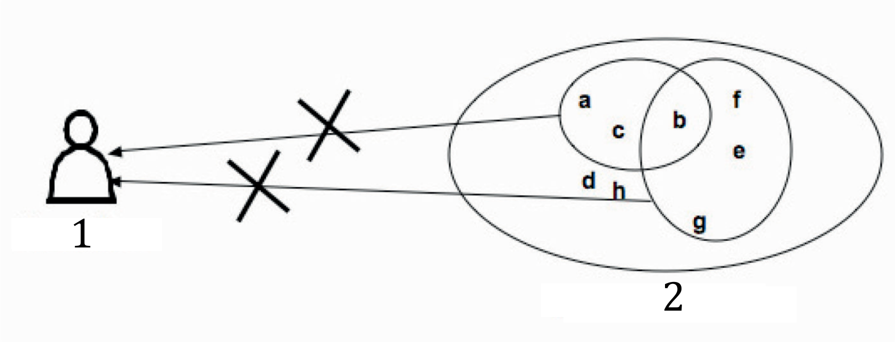
c) what the disclosure risks will be (including re-identification);

d) the information classification (e.g. tagging or labelling);

e) what release strategies are available.

From this, the details of the release process and the risk analysis, a strategy of identification concealment shall be determined. This determination shall be performed for each new release process, although many different release processes may select a common release strategy and details. Most teaching files will have common characteristics of purpose and minimum information content. Many clinical drug trials will have a common strategy with varying details. De-identification meets more needs than just confidentiality protection. There are often issues such as single-blinded and double-blinded experimental procedures that also require de-identification to provide the blinding. This will affect the decision on release procedures.

This subclause provides the terminology used for describing the concealment of identifying information.



Key

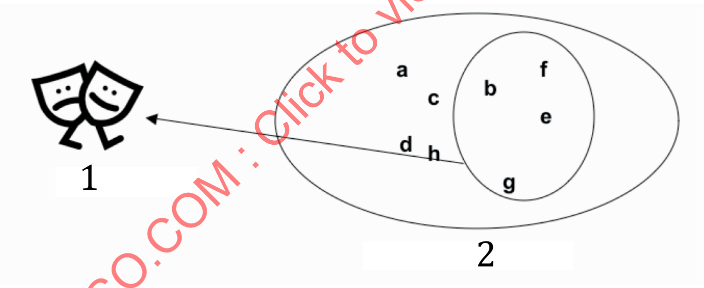
1 data subject

2 set of characteristics

Figure 4 — Anonymization

Anonymization (see Figure 4) is the process that removes the association between the identifying data set and the data subject. This can be done in two different ways:

- by removing or transforming characteristics in the associated characteristics-data-set so that the association is not unique anymore and relates to more than one data subject and no direct relation to an individual remains;
- by increasing the population in the data subjects set so that the association between the data set and the data subject is not unique anymore and no direct relation to an individual.



Key

1 pseudonym(s)

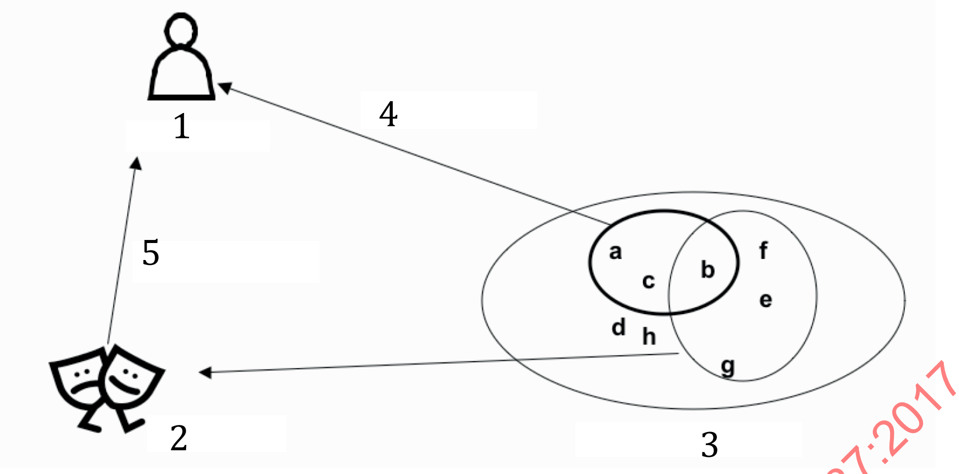
2 set of characteristics

Figure 5 — Pseudonymization

Pseudonymization (see Figure 5) removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.

From a functional point of view, pseudonymous data sets can be associated as the pseudonyms allow associations between sets of characteristics, while disallowing association with the data subject. As a result, it becomes possible, for example, to carry out longitudinal studies to build cases from real patient data while protecting their identity.

In irreversible pseudonymization, the conceptual model does not contain a method to derive the association between the data-subject and the set of characteristics from the pseudonym.



Key

1	data subject
2	pseudonyms
3	set of characteristics
4	a) derived from
5	b) derived from

Figure 6 — Reversible pseudonymization

In reversible pseudonymization (see Figure 6), the conceptual model includes a way of re-associating the data-set with the data subject.

There are two methods to achieve this goal:

a) derivation from the payload; this could be achieved by, for instance, encrypting identifiable information along with the payload;

b) derivation from the pseudonym or via a lookup-table.

Reversible pseudonymization can be established in several ways whereby it is understood that the reversal of the pseudonymization should only be done by an authorized entity in controlled circumstances. The policy framework regarding re-identification is described in Clause 7. Reversible pseudonymization compared to irreversible pseudonymization typically requires increased protection of the entity performing the pseudonymization.

Anonymized data differ from pseudonymized data as pseudonymized data contain a method to group data together based on criteria that are derived from the personal data from which they were derived.

5.5 Real world pseudonymization

5.5.1 Rationale

5.4 depicts the conceptual approach to pseudonymize where concepts such as “associated”, “identifiable”, “pseudonymous”, etc. are considered absolute. In practice, the risk for re-identification of data sets is often difficult to assess. This subclause refines the concepts of pseudonymization and unwanted/unintended identifiability. As a starting point, the European data privacy protection directive is here referred to.

There are many regulations in many jurisdictions that require creation of de-identified data for various purposes. There are also regulations that require protection of private information without specifying the mechanisms to be used. These regulations generally use effort and difficulty related phrases, which is appropriate given the rapidly changing degree of difficulty associated with de-identification technologies.

Statements such as “all the means likely reasonable” and “by any other person” are still too vague. Since the definition of “identifiable” and “pseudonymous” depend upon the undefined behaviour (“all the means likely reasonable”) of undefined actors (“by any other person”), the conceptual model in this document should include “reasonable” assumptions about “all the means” likely deployed by “any other person” to associate characteristics with data subjects.

The conceptual model will be refined to reflect differences in identifiability and the conceptual model will take into account “observational databases” and “attackers”.

5.5.2 Levels of assurance of privacy protection

5.5.2.1 General

Current definitions lack precision in the description of terms such as “pseudonymous” or “identifiable”. It is unrealistic to assume that all imprecision in the terminology can be removed, because pseudonymization is always a matter of statistics. But the level of the risk for unauthorized re-identification can be estimated. The scheme for the classification of this risk should take into account the likelihood of identifying the capability of data, as well as by a clear understanding of the entities in the model and their relationship to each other. The risk model may, in some cases, be limited to minimizing the risk of accidental exposure or to eliminate bias in situations of double-blinded studies, or the risks may be extended to the potential for malicious attacks. The objective of this estimation shall be that privacy policies, for instance, can shift the “boundaries of imprecision” and define within a concrete context what is understood by “identifiability” and as a result, liabilities will be easier to assess.

A classification is provided below, but further refinement is required, especially since quantification of re-identification risks requires the establishment of mathematical models. Running one record through one algorithm no matter how good the algorithm still carries risks of being re-identifiable. A critical step in the risk assessment process is the analysis of the resulting de-identified data set for any static groups that may be used for re-identification. This is particularly important in cases where some identifiers are needed for the intended use. This document does not specify such mathematical models; however, informative references are provided in the Bibliography.

Instead of an idealized conceptual model that does not take into account data sources (known or unknown) outside the data model, assumptions shall be made in the re-identification risk assessment method on what data are available outside the model.

A real-life model should take into account, both directly and indirectly, identifying data. Each use case shall be analysed to determine the information requirements for identifiers and to determine which identifiers can be simply blanked, which can be blurred, which are needed with full integrity, and which will need to be pseudonymized.

Three levels of the pseudonymization procedure, ensuring a certain level of privacy protection, are specified. These assurance levels consider risks of re-identification based upon consideration of both directly and indirectly identifying data. The assurance levels consider the following:

- level 1: the risks associated with the person identifying data elements;
- level 2: the risks associated with aggregating data variables;
- level 3: the risks associated with outliers in the populated database.

The re-identification risk assessment at all levels shall be established as a re-iterative process with regular re-assessments (as defined in the privacy policies). As experience is gained and the risk model is better understood, privacy protection and risk assessment levels should be reviewed.

Apart from regular re-assessments, reviews can also be triggered by events, such as a change in the captured data or introduction of new observational data into the model.

When referring to the assurance levels, the basic denomination of the levels as 1, 2 and 3 could be complemented by the number of revisions (e.g. level 2+ for a level 2 that has been revised; the latest revision data should be mentioned and a history of incidents and revisions kept up-to-date). The requested assurance level dictates what kind of technical and organizational safeguards need to be implemented to protect the privacy of the subject of data. A low level of pseudonymization will require more organizational measures to protect the privacy of data than will a high level of pseudonymization.

5.5.2.2 Assurance level 1 privacy protection: removal of clearly identifying data or easily obtainable indirectly identifying data.

A first, intuitive level of anonymity can be achieved by applying rules of thumb. This method is usually implicitly understood when pseudonymized data are discussed. In many contexts, especially when only attackers with poor capabilities have to be considered, this first level of anonymity may provide a sufficient guarantee. Identifiable data denotes that the information contained in the data itself is sufficient in a given context to pinpoint an entity. Names of persons are a typical example. 6.2.1 provides specification of data elements that should be considered for removal or aggregation to assert an anonymized data set.

5.5.2.3 Assurance level 2 privacy protection: considering attackers using external data.

The second level of privacy protection can be achieved when taking into account the global data model and the data flows inside the model. When defining the procedures to achieve this level, a static risk analysis that checks for re-identification vulnerabilities by different actors should be performed. Additionally, the presence of attackers who combine external data with the pseudonymized data to identify specific data sets should be considered. The available external data may depend on the legal situation in different countries and on the specific knowledge of the attacker. As an example, the required procedures may include the removal of absolute time references. A reference time marker “T” is defined as, for example, the admission of a patient for an episode of care and other events. Discharge is expressed with reference to this time marker. An *attacker* is an entity that gathers data (authorized or unauthorized) with the aim of attempting to attribute to data subjects, the gathered data in an unauthorized way and thus obtain information to which he is not entitled. From a risk analysis point of view, data gathered and used by an attacker are called “observational data”.

Note that the disallowed or undesired activity by the attacker is not necessarily the gathering of the data, rather the attempt to attribute the data to a data subject and consequently gain information about a data subject in an unauthorized way.

A risk analysis model may include assumptions about attacks and attackers. For example, in some countries, it may be possible to legally obtain discharge data by entities that are not implicitly involved in the care or associated administration of patients. The risk analysis model may take into account the likeliness of the availability of specific data sets.

From a conceptual point of view, an attacker brings data elements into the model that in the ideal world would not exist.

A policy document should contain an assessment of the possibility of attacks in the given context.

5.5.2.4 Assurance level 3 privacy protection: considering outliers of data.

The re-identification risk can be seriously influenced by the data itself, for example, by the presence of outliers or rare data. Outliers or rare data can indirectly lead to identification of a data subject. Outliers do not necessarily consist of medical data. For instance, if, on a specific day, only one patient with a specific pathology has visited a clinic, then observational data on who has visited the clinic that day can indirectly lead to identification.

When assessing a pseudonymization procedure, just a static model-based risk analysis cannot quantify the vulnerability due to the content of databases; therefore, running regular risk analyses on populated models is required to provide a higher level of anonymity.

In practice, proof of level 3 privacy protection will be difficult to achieve.

5.6 Categories of data subject

5.6.1 General

This document focuses on the pseudonymization of data pertaining to patients/health consumers. These principles can also be applied to other categories of data subjects such as health professionals and organizations.

5.6.2 to 5.6.3 enumerate specific categories of data subjects and list a number of issues related to these categories.

5.6.2 Subject of care

Decisions to protect the identity of the subject of care may be associated with the following:

- legal requirements for privacy protection;
- trust relationships between the health professional and the subject of care associated with medical secrecy principles;
- responsible handling of sensitive disease registries and other public health information resources;
- provision of minimum necessary disclosures of identifiers in the provision of care (e.g. laboratory testing);
- privacy protection to enable indirect use of clinical data for research purposes. Be aware that in some jurisdictions (e.g. in Germany), the indirect use of subject of care data require informed consent when the data are only pseudonymized and not fully anonymized.

Continuity of care requires uniform identification of patients and the ability to link information across different domains. Where data are pseudonymized in the context of clinical care, there is a risk to misidentification or missed linkages of the subject of care across multiple domains. In cases where pseudonymization is applied in a direct care environment, consideration shall be given to patient consent for those cases where the patient does not want pseudonymization for safety purposes.

5.6.3 Health professionals and organizations

Pseudonymization may also be used to protect the identity of health professionals for a number of purposes including the following:

- peer review;
- reporting of medical mishaps or adverse drug events;
- care process analysis;
- business analysis;
- physician profiling.

Such protections are subject to local jurisdiction legal requirements, which may be distinct from protection requirements of organization identities.

5.6.4 Device data

In healthcare, the security of devices, in support of the confidentiality of patient data is required for privacy protection. For patients, a consideration involves the consideration of implanted medical devices. Identifiable data on the device can be directly associable to the patient as can other medical and personal devices (e.g. respiratory assistive devices). As such, device identity or device data may be used to identify a person. Healthcare devices assigned to a healthcare professional or employee shall also be considered in identification risk assessment as it can identify the provider or organization, and hence, the patient.

5.7 Classification data

5.7.1 Payload data

According to the paradigm followed in this document, it should be possible to split data into data that can lead to identification and data that carry the medical information of interest. This assessment is fully dependent on the level of privacy protection that is targeted.

5.7.2 Observational data

In healthcare, the security of devices, in support of the confidentiality of patient data, is required for privacy protection. For patients, a consideration involves the consideration of implanted medical devices. Identifiable data on the device can be directly associable to the patient as can other medical and personal devices (e.g. respiratory assistive devices). As such, a device and its data may be able to identify a person. Healthcare devices assigned to a healthcare professional or employee shall also be considered in identification risk assessment as it can identify the provider or organization, and hence, the patient. Observational data which are gathered and used by an attacker reflects various properties of data-subjects recorded with the aim of describing the data subjects as completely as possible with the intent of re-identifying or identifying membership in certain classifications at a later stage.

5.7.3 Pseudonymized data

Two types of pseudonymized data are possible.

a) In irreversible pseudonymization, the pseudonymized data do not contain information that allows the re-establishment of the link between the pseudonymized data and the data subject.

b) In reversible pseudonymization, the pseudonymized data can be linked with the data subject by applying procedures restricted to duly authorized users.

NOTE Reversibility is a property that can be achieved by applying various methods such as: a) encrypt identifiable data along with the pseudonymized data; b) maintain a protected escrow list that links pseudonyms with identifiers.

5.7.4 Anonymized data

Anonymized data are data that do not contain information that can be used to link it with the data subject with whom the data are associated. Such linkage could, for instance, be obtained through names, date of birth, registration numbers or other identifying information.

5.8 Research data

5.8.1 General

Using health data for research is usually a secondary use of health data after/beside the primary use that is for patient treatment. In many jurisdictions, this may require the informed consent of the patient. It is a fundamental principle of data protection that identifiable personal data should only be processed as far as is necessary for the purpose at hand. There is a clear interest for organizations performing research to pseudonymize or even anonymize data, where possible. Concerns for privacy of individuals, particularly in the area of health information, triggered the development of new regulatory requirements to assure privacy rights. Researchers will need to comply with these rulings and in many cases, modify traditional methods for sharing individually identifiable health information.

Medical privacy and patient autonomy are crucial, but many traditional approaches to protection are not easily scalable to the increasing complexity of data, information flows and opportunities for enhanced value merged information sets. Classic informed consent for each data use may be difficult or impossible to obtain. For anonymized data, however, research may proceed without the data subject being affected or involved but not with pseudonymized data.

Trends and opportunities to accumulate, merge and reuse health information collected and gathered for secondary use (e.g. research) will continue to expand. Privacy enhancing technologies are well-suited to address the security and confidentiality implications surrounding this growth. Many important data applications do not require direct processing of identifiable personal information. Valuable analysis can be carried out on data without ever needing to know the identity of the actual individuals concerned.

5.8.2 Generation of research data

Pseudonymization may be used in the generation of research data. In this case, there is optimal opportunity to assess risks to privacy inherent in the research study and to mitigate these risks through anonymization techniques described in this document. Uses for research also more clearly facilitate consent and definition of rules surrounding circumstances and reasons for intentional re-identification needs.

5.8.3 Secondary use of personal health information

Where permitted by jurisdiction, pseudonymization may be used to protect the privacy of individuals whose personal health information is to be used for secondary use. Secondary uses are those that are different than the initial intended use for the data collected. Each secondary use shall undergo a privacy threat assessment and define mitigations to the identified risks. Assumptions shall not be made as to the sufficiency of an existing risk assessment and risk mitigation to extend the data resource to additional secondary use.

5.9 Identifying data

5.9.1 General

Data that contains information that allow unique identification of the data subject (e.g. demographic data).

5.9.2 Healthcare identifiers

In healthcare, conflicting identity requirements should be reconciled.

- When authorized, several medical data sources relating to a named data subject may be linked across different domains. Depending upon the use requirements for the linked data, linking may need to be:
  - correct (no linking of data sources relating to different patients);
  - complete (no missing links because of failure to correctly identify a data subject).
- When access to the data subject’s identifiable data is restricted, the data may, under controlled circumstances, be linked to the data subjects by authorized authorities, with the help of a trust service provider.

In some jurisdictions, linking between different domains may be restricted. This issue shall also be assessed. When a data subject has visited different healthcare providers, these providers often use their own internal numbering. Administrative and medical information is often handed over to other authorities with these locally issued numbers. Consequently, authorities that require aggregate data do not have assurance that the aggregated data are complete.

This can be avoided by the use of a structured approach to identity management. There are several approaches to identity management and therefore, a detailed discussion of identity management is outside the scope of this document. However, at the core of some identity, management solutions will be a pseudonymization solution.

5.10 Data of victims of violence and publicly known persons

5.10.1 General

Victims of violence, who are diagnosed or treated, often require extra shielding by hospital personnel as long as their identification poses specific threats. Caregivers in direct contact with the patient can identify the person but back-office personnel cannot.

Similar issues often arise when publicly well-known persons or persons otherwise known to the healthcare community, often wrongly denoted as "VIPs", are admitted (e.g. politicians, captains of industry, etc.).

5.10.2 Genetic information

There is no general consensus regarding genetic information and there are a variety of requirements based on the legal jurisdiction. See Annex F for further considerations.

5.10.3 Trusted service

In the case where the pseudonymization service is required to synchronize pseudonyms across multiple entities or enterprises, a trusted service provider may be employed. Trusted services may be implemented through numerous options, including commercial entities, membership organizations or government entities. Providers of trusted services may be governed through legislation or certification requirements in various jurisdictions.

5.10.4 Need for re-identification of pseudonymized data

Pseudonymization separates out personally identifying data from payload data by assigning a coded value to the sensitive data before splitting the data out. The reversible approach maintains a connection between payload data and personal identifiers, but can allow for re-identification under prescribed circumstances and protections. The irreversible approach does not maintain any connection between payload data and personal identifiers and consequently no re-identification is applicable.

This approach serves researchers well in that it provides a means of cleansing research data while retaining the ability to reference source identifiers for the many (controlled) circumstances under which such information may be needed. Such circumstances include the following coded values. This document defines a vocabulary. The vocabulary identification is: ISO (1) standard (0) pseudonymization (25237) re-identification purpose (1). The codes in this vocabulary are as follows:

a) data integrity verification/validation;

b) data duplicate record verification/validation;

c) request for additional data;

d) link to supplemental information variables;

e) compliance audit;

f) communicate significant findings;

g) follow-up research.

These values should be leveraged for audit purposes when facilitating authorized re-identification. Such re-identification methods shall be well-secured, and can be done through the use of a trusted service for the generation and management of the decoding keys. The criteria for re-identification can be defined, automated and securely managed using the trusted services.

5.10.5 Pseudonymization service characteristics

There are two primary scenarios for pseudonymization services:

a) pseudonyms maintained within or for an individual organization or single purpose: in this situation, typically, the service addressed identities assigned or known to the organization;

b) pseudonyms provided through pseudonymization services: in this situation, typically the service is providing pseudo identities across unaffiliated organizations enabling linking of patient health information while protecting the identity of those patients.

In both cases, the provision of the service shall be accomplished so as to minimize the risk of unauthorized re-identification of the subjects of the pseudonymization service.

The service entrusted to protect the patient identities shall conform to minimum trustworthy practices requirements.

- There is a need to assure the health consumer's confidence in the ability of the health system to manage the confidentiality of their information.
- There is a need for the service to provide physical security protection.
- There is a need for the service to provide operational security protection.
- Re-identification keys, transformation tables and protection need to be subject to multi-person controls and/or multi-organization controls consistent with the assurances claimed by the service.
- The service shall be under the control of (e.g. contractually or operationally) the custodian of the source identifiers.
- Legal and environmental constraints surrounding release of re-identification keys and protections need to be disclosed in support of the privacy protection levels claimed by the service.
- Quality and availability of service needs to be specified and provided in accordance with the information provision and access needs.
- Some identifiers may simply be blanked as they are unnecessary for the use.
- Some identifiers may be blurred in a way consistent with the intended use.

6 Protecting privacy through pseudonymization

6.1 Conceptual model of the problem areas

This document concentrates on information that is collected or stored and not so much on interactive use of systems by patients. Information entered or edited by the patient during interactive use can be considered stored information.

There are multiple reasons for protecting privacy by concealing identities. In all cases, the privacy policy shall set targets for the protection of privacy through pseudonymization in terms of what is considered identifying information and what is considered as non-identifying information.

From a functional point of view, it is important to specify if reversibility is required and what the finalities of the reversibility are, in order to procedurally and technically facilitate authorized application of reversibility while preventing others.

In identity management frameworks, complex pseudonymization functions that include pseudonym translations between identity domains may be required, depending on the identity management scheme.

Two important elements in the concept of pseudonymization are as follows:

- the domain where a pseudonym will be used;

— protection of the pseudonymization key or seed.

6.2 Direct and indirect identifiability of personal information

6.2.1 General

Personal data may be directly identifiable or indirectly identifiable. The data are considered directly identifiable in those cases where an individual can be identified through a data attribute or through linkage by that attribute to a publicly accessible resource or resource restricted access under an alternative policy that contains the identity. This would include cross reference with well-known identifiers (e.g. telephone number, address) or numeric identifiers (e.g. order numbers, study numbers, document OIDs, laboratory result numbers). An indirect identifier is an attribute that may be used in combination with indirectly identifying attributes to uniquely identify the individual (e.g. postal code, gender, date of birth). This would also include protected indirect identifiers (e.g. procedure date, image date) which may have more restricted access, but can be used to identify the patient.

6.2.2 Person identifying variables

Person identifying variables include the following:

- person's name (including preferred name, legal name, other names by which the person is known). Name includes all name data elements as specified in ISO/TS 22220;
- person identifiers (including, e.g. issuing authorities, types and designations such as patient account number, medical record number, certificate/license numbers, social security number, health plan beneficiary numbers, vehicle identifiers and serial numbers, including license plate numbers);
- biometrics (voice prints, finger prints, photographs, etc.);
- digital certificates that identify an individual;
- mother's maiden name and other similar relationship-based concept (e.g. family links);
- residential address;
- electronic communications (telephone, mobile telephone, fax, pager, e-mail, URL, IP addresses, device identifiers and device serial numbers);
- subject of care linkages (mother, father, sibling, child);
- descriptions of tattoos and identifying marks.

Depending on the data format standard used, there may be associated standard specifications available that should be followed (e.g. DICOM PS3.15:2016, Annex E).

6.2.3 Aggregation variables

For statistical purposes, absolute data references should be avoided.

a) Dates of birth, for example, are highly identifying. Ages are less identifying but can still pose a threat for linking observational data; therefore, it is better to use age groups or age categories. In order to determine safe ranges, re-identification risk analysis should be run, which is outside the scope of this document.

b) Admission, discharge dates, etc. can also be aggregated into categories of periods, but events could be expressed relatively to a milestone (e.g. *x* months after treatment).

c) Location data, if regional codes are too specific, should be aggregated. Where location codes are structured in a hierarchical way, the finer levels can be stripped, e.g. where postal codes or dialling codes contain 20 000 or fewer people, the code may be changed to 000 (HIPAA section 164.514).

Demographic data can be both direct and indirect identifiers and should be removed where possible, or aggregated at a threshold specified by the domain or jurisdiction. Where these data need to be retained, risk assessment of unauthorized re-identification and appropriate mitigations to identified risks of the resulting data resource shall be conducted. These demographic data include the following:

- language spoken at home;
- person's communication language;
- religion;
- ethnicity;
- person gender;
- country of birth;
- occupation;
- criminal history;
- person legal orders;
- other addresses (e.g. business address, temporary addresses, mailing addresses);
- birth plurality (second or later delivery from a multiple gestation).

A policy document shall be generated containing an assessment of the possibility of attacks in the given context as a risk assessment against level 2 privacy protection. The identified risks shall be coupled with a risk mitigation strategy.

6.2.4 Outlier variables

Outlier variables should be removed based upon risk assessment.

Outlier variables include the following:

- rare diagnoses;
- uncommon procedures;
- some occupations (e.g. tennis professional);
- certain recessive traits uncharacteristic of the population in the information resource;
- distinct deformities.

A policy document shall be generated containing an assessment of the possibility of attacks in the given context as a risk assessment against level 3 privacy protection. The identified risks shall be coupled with a risk mitigation strategy.

Persistent data resources claiming pseudonymity shall be subject to routine risk analysis for potentially identifying outlier variables. This risk analysis shall be conducted at least annually. The identified risks shall be coupled with a risk mitigation strategy.

6.2.5 Structured data variables

Structured data give some indication of what information can be expected and where it can be expected. It is then up to re-identification risk analysis to make assumptions about what can lead to (unacceptable) identification risks, ranging from simple rules of thumb up to analysis of populated databases and inference deductions. In "free text", as opposed to "structured", automated analysis for privacy purposes with guaranteed outcome is not possible.

6.2.6 Non-structured data variables

6.2.6.1 General

In the case of non-structured data variables, the pseudonymization decision of data separation into identifying and payload data remains the central issue. Freeform text shall be considered suspect and thus should be considered for removal. Non-structured data variables shall be subject to the following:

- single out what according to the privacy policy (and desired level of privacy protection) is identifiable information;
- delete data that is not needed;
- policies should state that the free text part shall not contain directly identifiable information.

Keep together as payload what is considered to be non-identifiable according to the policy.

6.2.6.2 Freeform text

Freeform text cannot be assured anonymity with current pseudonymization approaches. All freeform text shall be subject to risk analysis and a mitigation strategy for identified risks. Re-identification risks of retained freeform text may be mitigated through the following:

- implementation of policy surrounding freeform text content requiring that the freeform text data shall not contain directly identifiable information (e.g. patient numbers, names);
- verification that freeform content is unlikely to contain identifying data (e.g. where freeform text is generated from structured text);
- revising, rewriting or otherwise converting the data into coded form.

As parsing and natural language processing “data scrubbing” and pseudonymization algorithms progress, re-identification risks associated with freeform text may merit relaxation of this assertion.

Freeform text should be revised, rewritten or otherwise converted into coded form.

6.2.6.3 Text/voice data with non-parseable content

As with freeform text, non-parseable data, such as voice fields, should be removed.

6.2.6.4 Image data

Some medical data contain identifiable information within the data (e.g. a radiology image with patient identifiers on image). Mitigations of such identifiable data in the structured and coded DICOM header should be in accordance with DICOM PS3.15:2016, Annex E. DICOM (ISO 12052) has defined recommended de-identification processes for DICOM SOP Instances (documents) for some common situations. It defines a list of different de-identification algorithms that might be applied. Then it identifies some common use situations and characteristics, e.g. “need to retain device identification”. For each standard DICOM attribute (data element), it then recommends the algorithm that is most likely appropriate for that attribute in that situation.

These assignments are expected to be adjusted when appropriate, but providing a starting point for typical situations greatly reduces the work involved in defining a de-identification process. Additional risk assessment shall be considered for identifiable characteristics of the image or notations that are part of the image.

6.2.7 Inference risk assessment

It should be recognized that pseudonymization cannot fully protect data as it does not fully address inference attacks. Pseudonymization and anonymization services shall supplement practices with risk assessment, risk mitigation strategies and consent policies or other data analysis/pre-processing/post-processing. The custodian of pseudonymized repositories shall be responsible for reviewing data repositories for inference risk and to protect against disclosure of single record results. The information source shall be responsible for pre-viewing/pre-processing the source data disclosed to protect the disclosed data from inference based upon outliers, embedded identifiable data or other such unintentional disclosures. For more details on how to conduct an inference risk assessment, see Annex B.

6.2.8 Privacy and security

There is always the risk that pseudonymized data can be linked to the data subject. In light of this risk, the gathered data should be considered “personal data” and should be used only for the purposes for which it was collected. In many countries, legislation requires protection of pseudonymized data in the same manner as identifying data.

7 Re-identification process

7.1 General

Two distinct contexts of re-identification of pseudonymized information shall be considered:

- re-identification as part of the normal processing;
- re-identification as an exceptional event.

7.2 Part of normal procedures

If re-identification is part of the normal processing, conditions and procedures for re-identification should be part of the overall design of the processes. An example is, for instance, where pseudonymized requests are sent from a medical record application to a clinical pathology laboratory in a de-identified manner. The results are received in pseudonymous format, re-identified and automatically inserted into the medical record by the application.

Re-identification in normal procedures is characterized by the fact that re-identification is usually done in an automated, transparent way and that no authorization on a per-case basis should be required.

In cases where re-identification is part of a normal procedure, care will be taken as to the integrity of the data (completeness, not changes). In most of these cases, the processing requires and guarantees the same level of integrity as with personal data. This is not necessarily the case with research data, which falls for that reason under the category of the “exceptional procedure”.

7.3 Exception

When re-identification is an exception to the standard way of data processing, the re-identification process shall require

a) specific authentication procedures, and

b) exceptional interventions by the pseudonymization service provider.

When re-identification of de-identified data is considered the exception to the rule, the security policy shall describe the circumstances that can lead to re-identification.

The data processing security policy document should define the cases that can be foreseen and should cover the following.

- Each case should be described and one or more scenarios for re-identification per case should be described.

- Identification of the individual that initiates a request for re-identification.
- Verification of the requestor against the authorization rules that allow the re-identification. All entities involved in such cases shall be informed of the re-identification event. Re-identification described should only be started after proper authorization (electronic or otherwise) and should follow the scenario described in the policy.
- Exceptional re-identification should only be performed by a trust service provider (assuming that the pseudonymization service provider is required and capable of processing the re-identification).
- In all circumstances, care shall be taken that, apart from a trusted service provider, no one else shall have the technical capability of compiling lists that connect identifiers and pseudonyms. After processing, the trust service provider shall destroy these linking lists.
- The controller of the re-identified data shall carry out extensive testing of the integrity (correctness, completeness of the data). This is especially true in the case where the finality of the data changes. For example, pseudonymous research data are turned into data for diagnosis or treatment.
- The policy shall make clear who will be the controller of the personal data resulting from the re-identification process and what the finality of the data is. The recovered data should indicate its origin to the extent needed (de-identified data might not be as complete or reliable as the original personal data from which it was derived).

In exceptional cases that cannot be foreseen, the rules for cases that can be foreseen shall also apply. Unlike cases that can be foreseen, there is no *a priori* scenario for re-identification. The severity of the need for re-identification will have to be assessed. The controller of the data is responsible.

An exception to this rule may be re-identification for law enforcement. This is not treated in this document but it is assumed that the law-enforcement actors who take responsibility to re-identify also take care of proper privacy protection of the personal data that follows.

7.4 Technical feasibility

In cases where re-identification is part of the normal procedure or expected for a number of described scenarios, it will be technically feasible to re-identify.

There are several methods to enable re-identification.

Directly or indirectly identifying data (e.g. a list of local identifiers) can be encrypted and kept along with the pseudonymized data. Only a designated trust service provider can decrypt the data and re-associate the indirectly identifying data with the data subject.

A trust service provider (the pseudonymization service provider or an escrow service provider) can keep a linking list between pseudonyms and identifiers (directly or indirectly identifying).

Annex A
(informative)

Healthcare pseudonymization scenarios

A.1 General

This annex presents a series of high-level healthcare cases or “scenarios” representing core business and technical requirements for pseudonymization services that will support a broad cross-section of the healthcare industry.

General requirements are presented first, speaking of basic privacy and security principles and fundamental needs of the healthcare industry. The document then details each scenario as follows:

a) a description of the scenario or healthcare situation requiring healthcare pseudonymization services;

b) resulting business and technical requirements that a pseudonymization service shall provide.

A.2 Scenario explanation

The scenarios described in A.3.1 to A.3.5 show how pseudonymization services can be used in healthcare. Each scenario is intended to describe potential and probable uses of a healthcare pseudonymization service.

The following headers are used in the scenario description.

— **Kind of ID**

Denotes if the ID is a patient ID or if the ID is, e.g. a provider ID.

— **Uniqueness**

In anticipation of the use that will be made of a pseudonymized database, it is important to know if the input value uniquely identifies an individual in a given context. This is particularly important if data collected in time and over organizational boundaries is to be uniquely linked. It is also important to assess if data coming from the same entity will be linkable or if there is a risk that synonyms will exist in the target database(s).

— **Sensitivity of the data**

It is helpful to have an indication of the sensitivity of the data for the design of the solution. Sensitivity is to be interpreted against the background of legislation or against the importance in the business/application case. For example, collecting HIV-related information from physical persons will have a much higher degree of sensitivity from a legal point of view and will require a risk analysis that is commensurate. Collection of success rates for a particular treatment of a disease from participating institutions is non-sensitive from a legal point of view, but may be on the critical path of a business solution.

— **Data sources: single or multiple data sources and their relationships**

The number and context of data sources will strongly determine if the use of an intermediary organization delivering trust services is required or not.

- **Primary or secondary use of personal data**

  This is a characteristic that strongly influences the legal constraints that could result in different designs of the pseudonymization solution. It is also important to know if the data was collected directly from the data subject.

- **Context/finality: commercial, medical research, patient treatment**

  This gives a brief description of the context.

- **Searchability/linkability**

  Searchability is a very important element in the overall design of a pseudonymization solution. The granularity of the searchability shall be defined; searchability referring to a selection of pseudonymized data based on non-pseudonymized elements (e.g. per geographic region). The search function will require the use of a pseudonymization service and may be restricted.

- **Reversibility/re-identification**

  This is the consideration of whether re-identification is desirable, prohibited, desirable in controlled circumstances or whether it should be built in for yet unknown but future desirable circumstances. Consideration should be given to what amount of re-identification is acceptable.

- **Linkage in time**

  Re-identification risk is influenced by the amount of pseudonymized information that can be gathered. By limiting linkage in time, the amount of pseudonymized information can be limited. This, of course, may clash with the requirement of long-term longitudinal research.

- **Linkage across domains**

  The use of a particular key or method for pseudonymization should be limited to as narrow a domain as possible. Therefore, in scenarios, it is important to describe the domain in which a pseudonym will be used and for how long and what linking with other domains is required. This, in turn, will determine the need of an intermediary organization. This aspect could also take into account the cooperation of different intermediary organizations.

A.3 Healthcare scenarios

Table A.1 — Scenario characteristic

Scenario		Data subject			Data sources		Functional/performance requirements			
		Kind of ID	Uniqueness	Sensitivity (pers. data legisl.)	Data sources	Primary/ secondary	Context/ finality	Searchability	Re-identification	Linkage in time
1	Pseudon. Care	PAT ID	Unique in the initiating system (HIS)	High	Single data source	Primary	Care	N/A	Yes	Yes
2	Clin-trial	PAT ID	No guaranteed uniqueness	High	Multi-centre	Primary	Research	Yes	No (exc. policy)	Yes/No
3	Clin-res.	PAT ID/ Provider ID	No guaranteed uniqueness	High	Multi-centre	Secondary	Research	Yes	No (exc. policy)	Yes
4	Pub health monitor	PAT ID/ Provider	No guaranteed uniqueness	High	Multi-centre	Primary/ secondary	Public health management	Yes	Yes, under very controlled circumstances	Yes
5	Patient safety reporting	Pat ID/ provider	Unique	High	Multi-centre	Primary	Research	Yes	Yes	Yes
6	Non-HC research	Pat ID, other domain IDs	Very heterogeneous, no uniqueness	High for the medical data part	Multi-centre	Secondary	Non-medical research	Yes	No	Yes

Scenarios

1) Pseudonymous care (Pseudon. Care)

2) Clinical trials and post-marketing surveillance (Clin-trial)

3) Secondary use of clinical data, e.g. research (Clin-res)

4) Public health monitoring and assessment (Pub health monitor)

5) Confidential patient safety reporting (Patient safety reporting, includes adverse drug effects)

6) Non-healthcare research (Non-HC Research, previously consumer groups)

7) Healthcare market research (HC Market Research, includes comparative quality indicator reporting, peer review, utilization, clinical qualification/soundness of physician bills, financial billing)

8) Teaching files (educational material, student study material, physician special cases)

9) Field service (should preserve all machine details and machine measured data, but can usually remove all patient, physician, financial data.)

Table A.1 *(continued)*

Scenario		Data subject			Data sources		Functional/performance requirements			
		Kind of ID	Uniqueness	Sensitivity (pers. data legisl.)	Data sources	Primary/ secondary	Context/ finality	Searchability	Re-identification	Linkage in time
7	HC Market Research	Physician ID/ Pat ID	Unique	Low (physician) High (diagnosis)	Multi-centre	Secondary	Non-medical research	Yes	No	Yes
8	Teaching Files	Pat ID, other domain IDs	Very heterogeneous, no uniqueness	High for the medical data part	Multi-centre	Primary	Education	No	No	Yes
9	Field Service	Pat ID, other domain IDs	Very heterogeneous, no uniqueness	High for the medical data part	Multi-centre	Primary/ Secondary	Commercial operations	No	No	Limited (should preserve time, but not linkages)

Scenarios

1) Pseudonymous care (Pseudon. Care)

2) Clinical trials and post-marketing surveillance (Clin-trial)

3) Secondary use of clinical data, e.g. research (Clin-res)

4) Public health monitoring and assessment (Pub health monitor)

5) Confidential patient safety reporting (Patient safety reporting, includes adverse drug effects)

6) Non-healthcare research (Non-HC Research, previously consumer groups)

7) Healthcare market research (HC Market Research, includes comparative quality indicator reporting, peer review, utilization, clinical qualification/soundness of physician bills, financial billing)

8) Teaching files (educational material, student study material, physician special cases)

9) Field service (should preserve all machine details and machine measured data, but can usually remove all patient, physician, financial data.)

A.3.1 Clinical pathology order (pseudonymous care)

Scenarios taken as example

This scenario (see Table A.1) used the pseudonymization service for protecting patient identities and for the consistent tracking of patients across disparate systems.

A clinical care provider needs to send a sample for laboratory testing. The policy requires that the patient identifying information not be transmitted along with the order. It is, however, important to both match the order request with the order result, and for the laboratory service to be able to provide a comparative result over time for the same patient. A pseudonym is generated through a trusted pseudonymization service prior to sending the request to the laboratory, and the result set is returned with the pseudonym. The pseudonym is re-identified so as to post the result into the appropriate patient record.

Actors: placer of the order (e.g. care provider in hospital context), filler of the order (e.g. clinical pathology laboratory), pseudonymization service, HIS.

Pre-conditions: the placer of the order chooses a set of tests he wants the filler of the order to complete: the order set is related to the data subject by means of a hospital unique ID number.

Post-conditions: the placer of the order has received results from the filler of the order and has incorporated them in the HCR of the data subject using the data subject hospital unique ID number used for the order.

Workflow/events/actions

a) Submit order to health information system (HIS):

   1) the placer of the order authenticates towards the HIS;

   2) the placer of the order submits the order with the hospital unique ID number of the data subject to the HIS;

   3) the placer of the order checks order against policies (e.g. recipient not allowed to receive identifiable data, VIP, …) and decides on privacy protection measures;

b) Pseudonymize:

   1) the hospital information system invokes the pseudonymization service with, as input, the hospital unique ID number;

   2) the PS processes the hosp ids;

   3) the PS returns the pseudonym to the HIS;

c) The HIS sends the order with the pseudonym to the filler:

   1) establish communication;

   2) message sent;

   3) acknowledgement received;

d) The order is processed by the filler of the order using the pseudonym:

   1) (possible comparative analysis performed by specialist);

e) The filler of the order submits the result to the HIS with the pseudonym:

   1) establish;

   2) message sent;

   3) acknowledgement received;

f) Re-identify result:

   1) the HIS submits the pseudonym to the pseudonymization services;

   2) authenticated user (HIS) is verified against reverse ID policy;

   3) the PS processes the pseudonym;

   4) the PS sends the real ID to the HIS;

g) The HIS inserts the result with the hospital ID into the HCR.

Other examples/remarks

Online counselling services over the web (care provided to an individual) same individual time after time.

A person well-known to the public presents them self to a healthcare provider for clinical care. Wanting to assure that the episode of care and follow-up treatment remain confidential, the patient requests pseudonymized identifiers be used across the encounters.

A.3.2 Clinical trial

A.3.2.1 General

The clinical trials encompass a very wide range of situations. The clinical trials of drugs to gather data for submission to the FDA are subject to many procedural regulations. There are also trials of new equipment, e.g. ROC studies and trials of new procedures. The pseudonymization requirements are driven by more than just privacy regulations. For scientific reasons, there can be a need for pseudonymization of purely internal data in order to provide a suitable double blind analysis environment.

Figure A.1 indicates the various locations where the data might be modified to add clinical trial identification attributes (CTI) and/or remove attributes for pseudonymization.

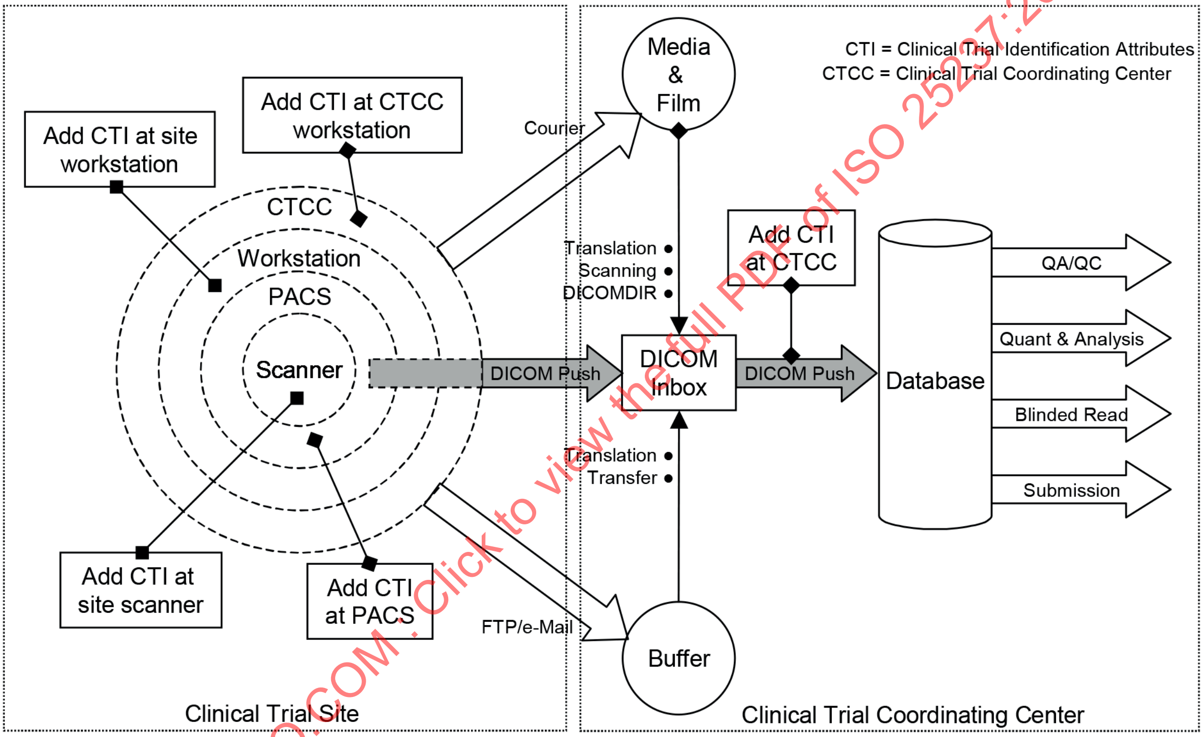


Figure A.1 — Clinical trial data modifications

Unlike the teaching files, there are usually multiple parties involved in the clinical trial process.

a) The **clinical trial sponsor**, who establishes the scientific requirements for the trial. This usually establishes the kinds of data that should be preserved, data that should be blinded and data that should be removed for scientific analysis purposes.

b) The **clinical trial coordinating centre**, which coordinates, gathers, and prepares the data. This centre may also provide the pseudonymization of data, depending upon the procedures chosen and the agreements made with the clinical trial sites.

c) Multiple **clinical trial sites**, where the actual clinical activity takes place. They pseudonymize the data in accordance with both their privacy policies and the needs of the clinical trial sponsor and in cooperation with the clinical trial coordinating centre.

d) Other **reviewers**, e.g. the FDA, who review the results of the clinical trial.

The trials may need reversibility so that actual patients can be notified of findings that are important to the patient's treatment. This can be implemented in various ways. The reviewer who makes the finds needs to be able to report to someone (e.g. the clinical trial agent) that "patient X in clinical trial Y should be notified of the finding ..."

A.3.2.2 Where pseudonymization is used

It is very difficult to make any specific statement in advance about what must be blinded or how. The range of topics that might be under investigation is very wide, and information about those topics often cannot be blinded. Each clinical trial needs to establish its own blinding and pseudonymization rules, although the work involved in doing this may be reduced by starting with the rules for similar previous trials.

A.3.2.3 Pseudonymization requirements

There are some unique regulatory concerns with data gathering for some clinical trials. These require complete audit trails and documentation of all data modifications. This includes modifications made for de-identification purposes. These regulatory requirements are a significant factor in the selection of de-identification techniques. Figure A.2 shows the use case diagram (ud) clinical trial flow.

Scenarios taken as example (in this group)

Submit data to clinical trials. This scenario describes the single source data collection for clinical care and clinical research study data resources.

Actors: System user (e.g. investigator, member of the care team), investigator health information system, clinical study information resource, care provider health information system (HIS).

Pre-conditions: Patient is in a clinical trial, investigator has data collection system that meets the needs of both the clinical trial and the external information resource, local information system available, patient consent obtained, a step of the clinical trial is concluded.

Post-conditions: Clinical study information resource has all relevant data from any patient encounter for a patient participating in the study. External information resource (e.g. HIS, EHR) has all relevant data from any patient encounter.

Workflow/events/actions

The process flow is as follows.

- Member of the care team authenticates to the HC system.
- Health information system initiates audit trail using consistent time.
- Clinician enters data into data collection system.
- System anonymizes or pseudonymizes data.
- System transmits relevant data to clinical study information resource.
- Clinical study information resource receives data.
- Data collection HIS posts clinical care information to external information resource and clinical trial investigator reviews and verifies (via eSignature or some verification mechanism) that these data accurately reflect the source data required for the trial.

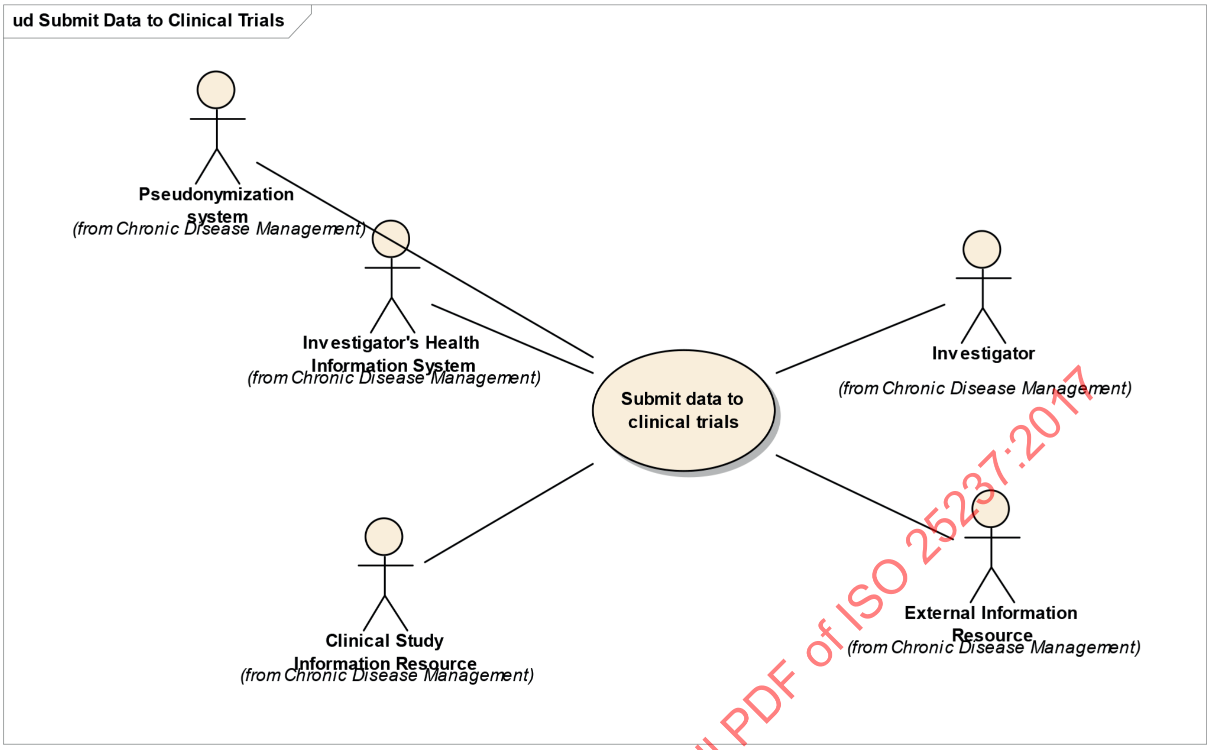

ud Submit Data to Clinical Trials
Pseudonymization system
(from Chronic Disease Management)
Investigator's Health Information System
(from Chronic Disease Management)
Submit data to clinical trials
Investigator
(from Chronic Disease Management)
Clinical Study Information Resource
(from Chronic Disease Management)
External Information Resource
(from Chronic Disease Management)

Figure A.2 — Clinical trial flows

A.3.3 Clinical research

Scenarios taken as example (in this group)

Secondary use of clinical data for research purposes.

Medical data have been collected in hospitals and treatment centres by the department of nephrology from diabetic patients in the context of medical treatment of their disease. Various medication schemes were used. Treatment data have been stored in several places, using the patient's national social security number.

At a later date, Ph.D. students decide to compare the success of the treatment. This constitutes secondary use of the data, as it is not intended for treatment. They have to collect data from various databases and combine the data per patient. However, the healthcare organizations that are holding the data will not release personal identifiable data. Therefore, all data are pseudonymized by sending it through a pseudonymization service that removes direct identifying data and replaces it on a one-to-one basis with a pseudonym. The researchers do not know the identity of the patients, but they are able to group information by patient.

In considering the success of the treatment through the research analysis, it is determined that correlation of the data with information not provided within the data set would provide valuable follow-up research. The appropriate research review board approves re-identification to request permission and interest to participate in a follow-up research study from the individuals that would make up the study cohort.

Multiple programmes on a similar set of data are as follows.

Actors: System user (e.g. investigator, member of the care team), investigator health information system, research information resource, care provider health information system (HIS).

Pre-conditions: Clinical record is determined to be of interest to researcher (may be all encounter data or only data with research topic cohort population criteria); investigator has data collection system that meets the needs of both the research and the local health information system; local information system available; patient consent obtained as required by local jurisdiction; a step of patient encounter is concluded.

Post-conditions: Research information resource has all relevant pseudonymized and privacy protected data from any patient encounter from patients within the cohort population. External information resource (e.g. HIS, EHR) has all relevant data from any patient encounter.

Workflow/events/actions

The process flow is as follows.

- Member of the care team authenticates to the HC system.
- Health information system initiates audit trail using consistent time.
- Clinician enters data into data collection system.
- System generates aggregate variables for privacy protection.
- System checks for uniquely identifiable characteristics in the data (e.g. rare diagnoses) or combined data variables.
- System anonymizes or pseudonymizes data.
- System transmits relevant data to research information resource.
- Research information resource receives data.
- Data collection HIS posts clinical care information to local HIS.

Other examples/remarks

Generation of teaching data:

Comparative quality indicator reporting: Encounter and discharge data are submitted by healthcare providers to a research database. Patient identifiers are pseudonymized through a pseudonymization service, as are identifiable grouping and risk adjustment data. Appropriate aggregations such as length of stay information are applied to further protect the research database from inference attacks. Provider identities are pseudonymized to protect the identity of practitioners and healthcare organizations.

Peer review: A new surgery technique is developed. Physicians use a pseudonymization service to submit case reports and adverse events to a common registry. This peer review registry is used to assess trends and compare experiences across multiple case mixes and co-morbidities. The confidentiality of the patients and practitioners are protected through the pseudonymization services provided by a pseudonymization service. This enables the patient data to be tracked across these providers to assess the full episode of care.

In assessing the cases in the study, it is found that a patient, having sought treatment from multiple providers, is at risk for a complication of the surgery. A case is made for re-identification to be able to contact the patient for follow-up assessment and treatment.

A.3.4 Public health monitoring

Scenarios taken as example (in this group)

The ability to detect events rapidly, manage the events and appropriately mobilize resources in response can save lives. Information from hospitals, other providers and ancillary facilities can be electronically reported to public health agencies and monitored without identifying patients and serve to provide a near real-time view of the health of our communities and inform decision-support processes in responding to the public's health threat event. These data can be shared with and among local, state and federal public authorities and the healthcare community to support coordinated response.

Actors: System user (e.g. public health official, member of the care team), public health information system, clinical information resource public health information resource.

Pre-conditions: Filter mechanisms criteria for data exchange have been established, event detection algorithms have been defined, patient is correctly identified, provider/information source is correctly identified, pseudonymization, de-identification and re-identification services are available.

Post-conditions: Data are submitted from multiple clinical information resources to the public health information system, data are received by the public health information system, the public health information system supports functions relevant to the public health event detection, i.e. the public health information system monitors, analyses, detects, investigates, notifies, alerts, reports and communicates data related to a public's health threat.

Workflow/events/actions

The process flow is as follows.

- Populate public health information system.
  - Clinical information resource supports entry of patient visit data into EMR.
  - Clinical information resource's EMR supports the public health information system data needs.
  - Clinical information resource initiates audit trail using consistent time.
  - Clinical information resource reviews and verifies (via eSignature or other verification mechanism) that these data accurately reflect the source data.
  - Clinical information resource selects information to submit (transmit) to the public health information system based upon filter criteria.
  - Clinical information resource invokes service to pseudonymize data.
  - Clinical information resource provides (transmits) relevant data to the public health information system through secured messaging and transmission.
  - Clinical information resource receives acknowledgement of receipt from the public health information system.
- Support detection of a public health threat event.
  - Provider receives notification from the public health information system of a suspected pattern through secure electronic means and via telephone.
  - Clinical information resource provides additional data to the public health information system as needed.
  - Provider receives health alert regarding the detected event through secure electronic means and via telephone.
  - Clinical information resource receives case-specific alert notifications from the public health information system for any pertinent patient follow-up.
- Support on-going monitoring of the event.
  - Clinical information resource captures and provides additional outbreak management data to the public health information system, particularly new and early diagnosed cases or suspect cases.
  - Authenticated clinical information resource invokes re-identification of patient identifiers through pseudonymization service to notify and provide follow-up treatment to patient and to request further screening of patient family/contacts as determined by outbreak management protocols.

  - Clinical information resource transmits daily data on utilization of resources to the public health information system.
  - Clinical information resource receives updates regarding the outbreak from the public health information system through secure electronic means.
- Support rapid response management of the event.
  - Clinical information resource receives recommendations/orders to conduct response-related activities in accordance with outbreak management protocols from the public health information system through secure electronic means.
  - Clinical information resource sends acknowledgement of receipt of recommendations/orders to conduct response-related activities in accordance with outbreak management protocols from the biosurveillance information system through secure electronic means.

Other examples/remarks

Once a week, general physician systems send influenza and allergy data to a central national repository. Before it reaches the repository, patient and physician identities are pseudonymized through a pseudonymization service, and the location information of the patient is aggregated into a larger area. The central repository is used for influenza and allergy alerts and has no need for identifiable data.

A.3.5 Patient safety reporting (adverse drug event)

Scenario description: monitor therapy safety. This scenario describes activities involved in monitoring therapy safety. This applies to both post-marketing surveillance and adverse event reporting. Figure A.3 shows the use case diagram (ud) patient safety flows.

Actors: System user (e.g. member of the care team), anonymization/pseudonymization system, health information system, event capture information resource.

Pre-conditions: patient receives care, patient is exposed (medication, device, environmental exposure such as poison ivy), member of care team has information system that meets the needs of adverse event reporting, local information system available, patient consent, and pseudonymization system is available to the local HIS.

Post-conditions: event reporting information resource has all relevant data and follow-up investigations supported where applicable.

Workflow/events/actions

The process flow is as follows.

- Member of the care team authenticates to the HC system.
- Health information system initiates audit trail using consistent time.
- Member of care team uniquely identifies patient.
- Member of care team documents patient symptoms, signs, diagnoses, and whether it was a causal or temporal relationship to exposure.
- Member of care team decides to put on a special report (e.g. AE report, allergy list).
- Member of care team posts information to patient record and anonymization/pseudonymization service is invoked.
- Member of care team posts information/reports for organization (patient safety, Comm, FDA, CDC, Public health).
- Follow-up with patient is determined to be required.

— Re-identification is conducted by the authenticated HC Provider.

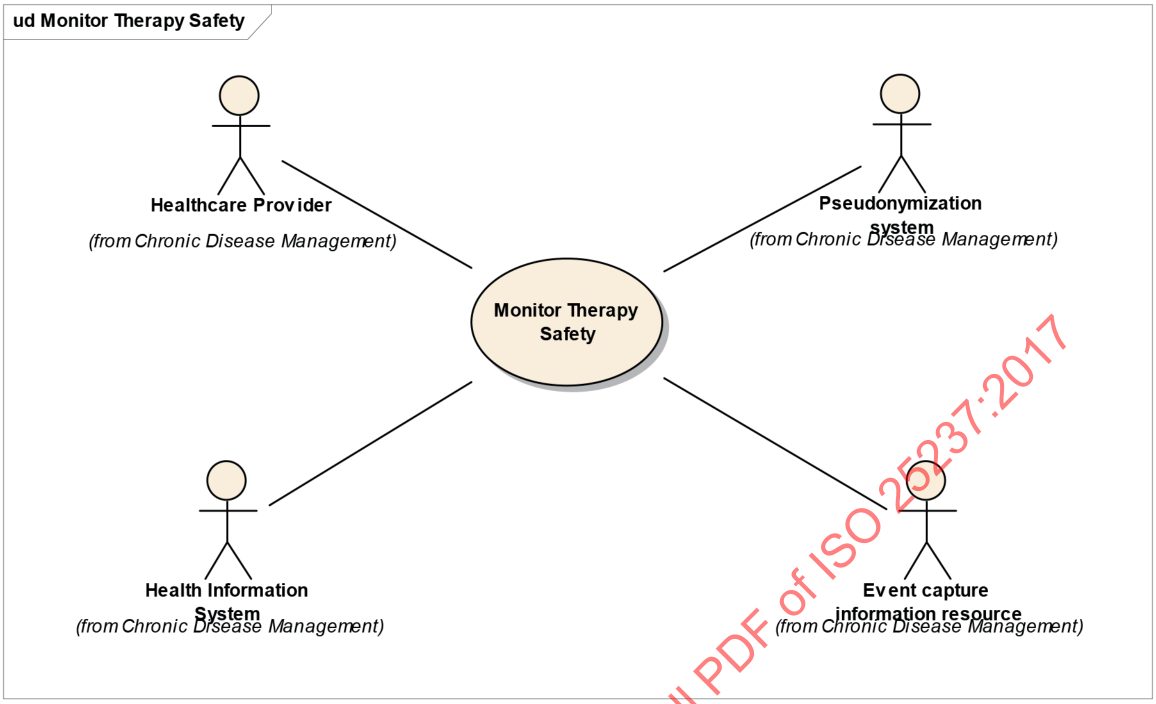

ud Monitor Therapy Safety
Healthcare Provider
(from Chronic Disease Management)
Pseudonymization system
(from Chronic Disease Management)
Monitor Therapy Safety
Health Information System
(from Chronic Disease Management)
Event capture information resource
(from Chronic Disease Management)

Figure A.3 — Patient safety flows

Other examples/remarks

A voluntary reporting system is used to generate a database in support of patient safety. Pseudonymization is used to protect the identity of both the patient and the provider submitting the data through the use of a pseudonymization service. Follow-up communications and requests for additional details on the submitted events are facilitated through the pseudonymization service without risk of identification of the patient or the provider.

A.3.6 Non-healthcare research using personal medial data

Scenario description

Regulatory policy requires evaluation of the long-term financial impact of seatbelt utilization. A study is approved to merge source data from crash reports, emergency medical response reports, motor vehicle license data, hospital records, rehabilitation records and community healthcare records. Identity and relevant risk adjustment data from these data sources is pseudonymized through a pseudonymization service and collected into a research database to be used by the study.

A.3.7 Market research

Scenario description

A group of healthcare providers agree to share information regarding the service utilization and characteristics. This includes market capture data, and as such, the organization identities are protected through pseudonymization techniques.

A.3.8 Classroom teaching files

A.3.8.1 General

Classroom teaching files are acquired by selecting interesting cases of real patients and then modifying the records to remove identifying and extraneous information. These can be generated using an anonymization process, but for living patients, there may be a need to update the records with new information at a later date. The data should be pseudonymized in order to preserve the relationships between the real patient and the teaching file so that these updates can be added to the teaching file.

The teaching files may be made available only to students at the generating facility, or they may be published for use by students around the world. In the former case, the rules for pseudonymization may permit greater detail, but for the latter use, the published medical records should be pruned to only the essentials for the tutorial purpose.

A more restricted but very common need is the creation of personal teaching files by students to capture cases that they found personally interesting. Privacy regulations prohibit the students from taking copies of those records.

A.3.8.2 Where pseudoymizatinon is needed

The key to pseudonymization is an assignment of name and patient ID for tutorial purposes. The typical phrasing in a tutorial report is something like "Mr. Smith is a 50-year old male with a history of ...." The pseudonymization should go through all of the medical records changing the name to Smith, assigning a new birth date (consistent with the relevant age range), and removing all other identifying information that is irrelevant to the tutorial purposes. The resulting new medical records can then be published as a teaching file.

A.3.8.3 Pseudonymization requirements

The generation of teaching files often requires creation of a secure database to maintain the relationship between the actual patient identity and the pseudonymous identity. The medical records are often on multiple systems, so the database needs to be either accessible or movable between the multiple systems. It also needs to be in a format that is understood by a variety of systems.

The generation of pseudonymous data will have both local rules, for such things as generation of pseudonymous IDs, and generic rules for such things as generation of blinded dates.

A clinician will need to establish the rules for what data attributes should be preserved, pseudonymized or removed. These rules need to be consistently applied by multiple systems at multiple times in the generation of the pseudonymous records.

A.3.9 Field service

The most common use of data blinding for field service is the anonymization of individual data records. For example, if a machine malfunction resulted in a flaw in the patient data, the service staff may need to take a copy of that data for analysis of the malfunction. This can usually be a single data record that is anonymized and substantially reduced in content. Only the machine parameters and anomalous data are needed for service analysis. Patient identification, history, etc. can be removed.

Sometimes a consistent set of records should be captured, rather than just a single record, but again, the data can be substantially reduced. In this situation, the data should be pseudonymized to preserve the relationships between records, but the pseudonymization can be irreversible.

Annex B
(informative)

Requirements for privacy risk analysis

B.1 General

The development of a method for privacy impact assessment is outside the scope of this document. As a result of a privacy impact assessment, you often have a confidentiality risk analysis. However, the following subclauses are intended to increase the awareness of those who will have to engage in privacy impact assessment. From a generic presentation of the issues, it should be possible to derive a number of requirements for privacy impact assessment design.

This document contains a model that takes into consideration three assurance levels. The levels have been chosen as a function of the complexity of re-identification of the data.

This document can, however, formulate a number of requirements that the re-identification risk assessment method should take into account for its design.

The common criteria (see ISO/IEC 15408-2) contain an informative annex on privacy (FPR). This can be used as a starting point but is more focused on the usage of resources.

The target of evaluation (TOE) security functions contain specifications that may be usable but do not incorporate the notion of levels of anonymity.

The remainder of this annex gives an overview of the risk assessment factors. The following text is adapted from Deliverable D2.1 with permission from the authors[37].

A key element in privacy risk assessment is to assess the effect of observational data that can be obtained by an attacker. Observational data can consist of events recorded by the attacker, but can also consist of information that can be legally obtained by the attacker. It could be that the attacker is a generic user of the system who has, either by accident or unauthorized effort, obtained extra data with which he should not have come into contact in the normal line of his duty.

It is important to note that this information is usually outside the scope of the data model of an application. Assumptions about observational data should however be made in order to assess the privacy of data contained in the system.

In order to create a methodology for privacy risk assessment, a formalized way of describing privacy threat and the risk of re-identification is needed.

A generic model of re-identification attacks, shown in its highest level of abstraction in Figure B.1, consists of three major entities.

Although it might look simple and straightforward in that form, the model is refined in this clause to such a level of complexity that it encompasses all real-life aspects of re-identification and privacy protection.

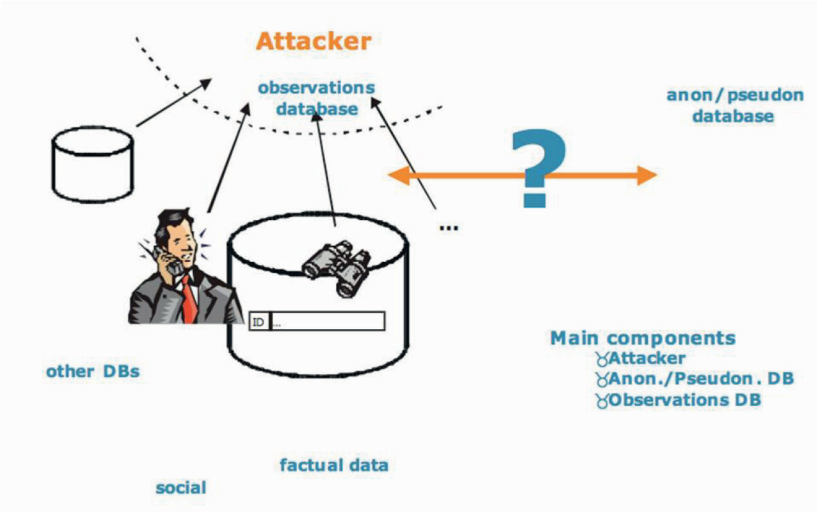


Figure B.1 — Re-identification attacks

There are three main entities in this model.

a) The anonymous database. This is the database containing anonymous records. It lists data on unknown subjects. It is the source of information which contains possible sensitive information which should not be disclosed.

NOTE Anonymous refers here to the fact that no direct identifying information (such as a subject's name) is included in the database, therefore, anonymous database should be interpreted in its broadest sense. It is trivial that privacy risk measurement only deals with such databases.

b) The attacker. The attacker is the individual who aims to abuse the information listed in the anonymous database. In order to do so, he needs to link the anonymous information to real-world persons, thus re-identify the subjects listed in that database.

c) The observations database. A database composed by the attacker, containing identified information.

Assumptions about the attackers will be important for risk analysis. Attackers may be opportunistic individuals looking for fun and visibility, or attackers may be highly trained and part of a team, focused on particular targets, backed up with substantial financial means and technical resources.

The attacker composes this latter database out of "observations" relevant for an attack. These observations may derive from different sources, such as

— information gathered from existing databases with identifiable information,

— social engineering, which is gathering information one is normally not entitled to, by exploiting social contacts, and

— factual data, e.g. data gathered by observation in the strict sense.

"Relevant" (for an attack) in this context refers to the fact that the gathered information shall be related to the content of the anonymous database. This means that this information is either listed in the anonymous database or closely related to it.

B.2 Threat model, goals and means of the attacker

One of the advantages of the concept of an observations database is that it allows definition of the extent of the methods that a real-life attacker would use. Indeed, security and safety analyses are usually (implicitly) based on an estimate of the threat to which a system will be exposed. The security model for modern public key cryptography, for example, is based on the fact that the attacker only has a limited amount of money (computing power) at his disposal.

Basically, there are two characteristics that define the threat level of an attacker as shown in Figure B.2. There is the goal of the attack (what information is an attacker after?) and there are the means at his disposal. The latter is linked with the "value" that the information that could be recovered from the anonymous database has for the attacker. It is clear that if the sensitive information enclosed in the anonymous data can lead to large gain for an attacker (e.g. medical records for an insurance company), he will be prepared to invest more into the re-identification process.

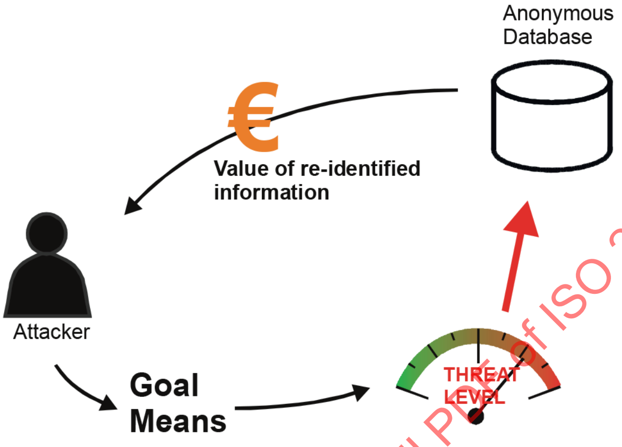


Figure B.2 — Goal and means of an attacker

Next to the level of determination of an attacker, it is important to include his "goals" into the threat model. Privacy protection is about protecting personal information and not simply about protecting the identity linked to a specific database record. This subtle difference is reflected in the three different attacker goals that are specified in the model. They are the following:

a) re-identification (full):

   1) identify to whom a specific anonymous record belongs;

   2) identify which anonymous record belongs to a certain person;

b) information recovery (or partial re-identification);

c) database membership:

   1) Is someone listed in the database?

   2) Is someone not listed in the database?

Full re-identification as a concept is well known. It is trying to (partially) convert an anonymous database to its identified equivalent. In the most general case, an attacker will try to re-identify a complete database. In practice, however, this is rarely the case, and an attacker will either want to find out to whom a specific, interesting anonymous record belongs (e.g. find out to whom a record, listing high income, belongs), or will want to retrieve all information about a specific person (e.g. an insurance broker trying to figure out if someone has a heart condition).

The two other goals (partial re-identification and database membership) are not often discussed, because analysis theory is quite complex. An attacker does not necessarily need to re-identify complete records to obtain the needed information. Sometimes it is sufficient for an attacker to recover only a single characteristic of a person from the anonymous database without ever knowing which records belong to that person.

Finally, in some situations, the target information is not listed within the database itself, but within the mere membership of a database. Being a member of a database can lead to private (sensitive) information, for example, HIV patient databases. Therefore, the goal of an attacker could be to merely determine if the people on his list are also in the anonymous database or not; he does not need to re-identify each anonymous record for that.

It is clear that the methods used by an attacker will depend on the goals he is trying to achieve. These attack strategies are closely related and although they fit within the same model, they differ strong enough to be elaborated separately at some points in the text.

Full and partial re-identification as defined in this document is obviously closely related. Partial re-identification is an intermediate stage between recovery of all information (on a particular subject) within the anonymous database and no recovery at all (see Figure B.3). In other words, it is the situation in which full re-identification fails, but in which the processes (algorithms) of re-identification used still succeed in recovering some information from the anonymous database.

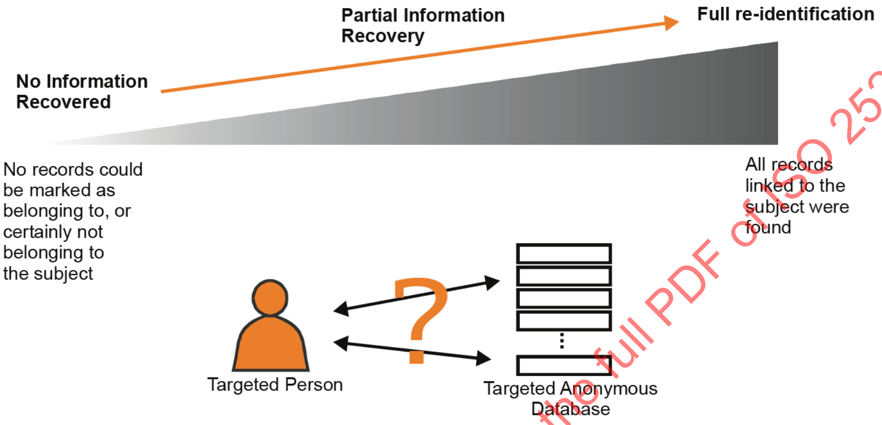

Partial Information Recovery
Full re-identification
No Information Recovered
No records could be marked as belonging to, or certainly not belonging to the subject
All records linked to the subject were found
?
Targeted Person
Targeted Anonymous Database

Figure B.3 — Full re-identification and partial information recovery

For all forms of re-identification, the attacker will mainly follow the same procedure. Based on his observations and the content of the anonymous database, he will list for each identifier (nom-ID) in the observations database, the anonymous identifiers (anon-ID) that could correspond with it.

The link between observations data and anonymous data can be made in numerous ways and will be situation-specific (see Figure B.4). It is, however, important to add some level of classification in the overall model, in order to understand the underlying mechanisms better.

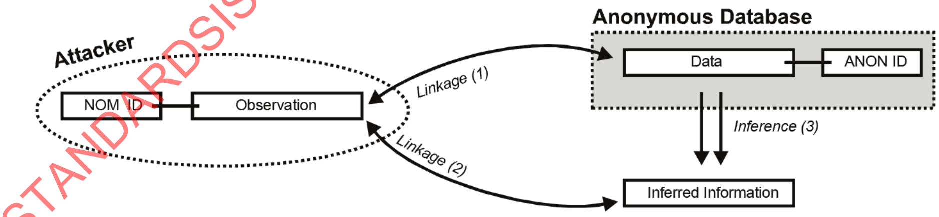


Figure B.4 — Linkage mechanisms

Figure B.4 shows that the link between nom-ID and anon-ID can be made directly through the variables listed in respective databases [linkage (1)], or through an intermediate step [linkage (2)]. In the first case, the data listed in the anonymous database corresponds directly with the observed data. This means that some of the variables in the anonymous database are observable by the attacker as a result of the database linkage. Through these shared variables, the attacker can determine if an anonymous record could correspond with an identifiable observations record.

In the second case, an intermediate step has to be taken in order to be able to link the two information sources. The observations are not listed literally in the anonymous database, but can be inferred from the variables present in the database. Note that for the reasoning in this document, this situation is equivalent with inferring from the observed database, to link with anonymous data.

The implementation of the linkage and inference algorithms themselves is usually data- and application-specific. However, several algorithms are able to deal with general data-types. At a higher abstraction level however, apart from the actual implementation, there is the important aspect of the “certainty” attached to a constructed link.

Both the linkage and inference algorithm(s) are not necessarily based on pure facts. A link made by an attacker based on observations is not necessarily a correct one. Depending on the assumptions that an attacker has made, the completeness and certainty of his observations, the complexity and fuzziness of the anonymous data, certain links will be more likely than others. Therefore, the attacker could need to associate a probability to the identifier link (off course, certainty gets a probability of 1).

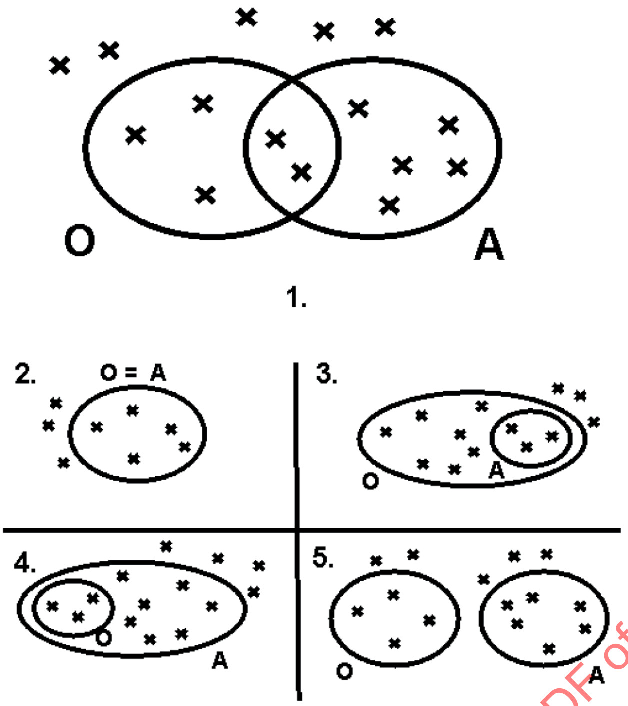
Examples can also be easily understood. Imagine that the anonymous database only lists salaries and that an attacker cannot get any direct salary information. He could, however, try to infer salary from observable variables such as job function, size of house, type of car. Clearly an attacker can never be sure, his assumptions are only true with a certain probability.

B.3 Re-identification, full or partial?

If at the end of this linking procedure, the attacker is able to match a single identifier with an anonymous identifier, then there is possible re-identification of the corresponding anonymous data records. The certainty of this re-identification depends on the probability of the used linkage rules and inference, and on a third factor, the relationship between the subjects listed in both the observations and the anonymous database.

Figure B.5 illustrates the different possible relations between an observations database (denoted as “O” on the figure) and an anonymous database (denoted as “A” on the figure). The crosses represent the data subjects themselves, not database records (a “cross” element of a set means that the corresponding subject is listed in that database).

As long as O is a subset of A, or vice-versa, re-identification is certain when a unique link between nom-ID and anon-ID is found. If this is however not the case, then a discovered unique link does not assure that there is a true correspondence between the identifiable and anonymous identifier.



Key

O	observational database
A	anonymous database
x	data subjects

Figure B.5 — Relations between observations and anonymous database

When there is no unique link found between a direct identifier and an anonymous identifier, this does not mean that there is no information disclosed at all. When a set of direct identifiers can be linked with multiple identifiers from the anonymous database, the common information enclosed among the anonymous identifiers can be associated with these direct identifiers (that is, if one is sure that all observed subjects are listed in the anonymous database; see Figure B.5). It is important to understand the full extent of this information leakage. If the attacker is only interested in the value of these common variables, then he has achieved his goal, and his attack has been successful.

When applying the described re-identification or information recovery method, the table with possible links, denoting which anonymous identifiers could correspond with which direct identifiers, should be updated and re-evaluated every time partial information is recovered, or a subject is fully re-identified.

B.4 Re-identification example

A simple example can be used to clarify the concepts explained above. In Figure B.6, the content of both an observations database and corresponding anonymous database is shown. As can be seen, the anonymous database contains three records with four static variables, which can have the values A or B (where a question mark indicates missing information). The attacker is able to observe only two of these variables directly and correctly, and knows all people who are listed in the anonymous database. The linkage rules for this situation are thus extremely simple, either a value is the same or it is not.

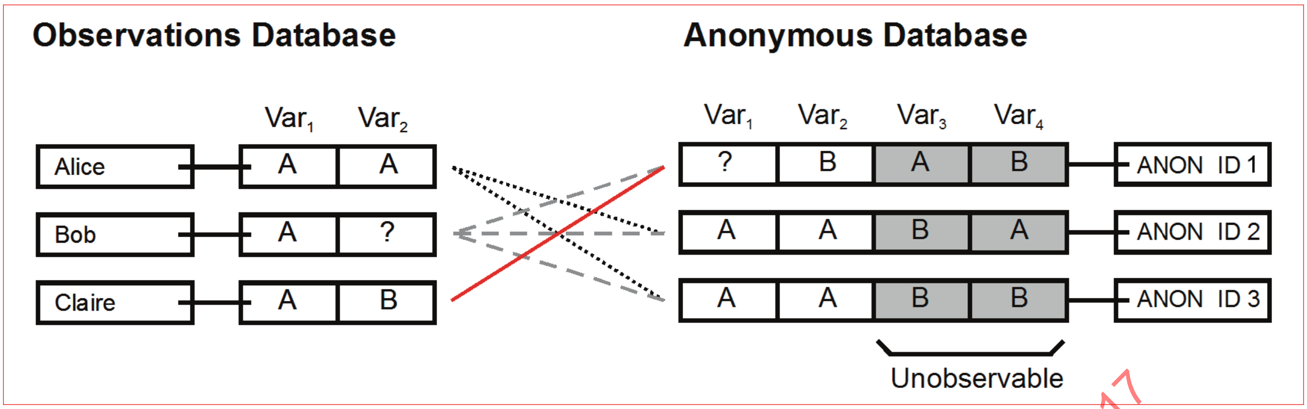


Figure B.6 — Re-identification example

When applying the simple linkage algorithm which links records by evaluating the variables listed in the anonymous and observations database directly [Figure B.6, linkage (1)], the following tables of correspondence can be composed (illustrated by the lines between the two record sets in Figure B.6).

Table B.1 — Direct identifier correspondence

Direct identifier	Can correspond with	Anonymous identifier	Can correspond with
Alice	A2, A3	A1	Bob, Claire
Bob	A1, A2, A3	A2	Alice, Bob
Claire	A1	A3	Alice, Bob

Table B.1 and Table B.2 represent how a direct identifier could correspond with an anonymous identifier and vice-versa under this particular linkage rule. If more than one algorithm is used, they should be evaluated together, which means that construction of the corresponding tables can become quite complex.

It can immediately be seen that record A1 can only belong to Claire, which means that the anonymous subject A1 is completely re-identified, and the attacker now knows that Claire has values (A, B) for the two unobservable variables. Taking this into account, the attacker could update the linkage tables as explained earlier, resulting in the following.

Table B.2 — Re-identification

Direct identifier	Can correspond with	Anonymous identifier	Can correspond with
Alice	A2, A3	A2	Alice, Bob
Bob	A2, A3	A3	Alice, Bob

From the remaining un-identified records, an attacker cannot clearly identify any more persons. He can however say that there is a 50 % chance that record A2 belongs to Alice or to Bob. In this case, this gives little extra knowledge on the data subjects in a realistic (large) database, however, such a situation can reveal useful information for the attacker.

Although no full re-identification is possible, there is still some information leakage, because anonymous record A2 and A3 have the same value for variable 3. From that, the attacker can conclude that both Alice and Bob have value B for variable 3. For the remaining variable 4, there is no information retrieved on Alice or Bob.

Finally, if the information that an attacker wanted to recover about Alice, Bob and Claire was listed in variable 3 only, then the attacker would have fully succeeded in his information recovery attempt. If variable 4 had contained important information, he would have only succeeded partially. Finally, note the fact that the value of variable 2 is now also known for Bob, although it could not be observed, thus again, more information was recovered.

The illustrated model and procedures do not only apply to such simple data structures as the one used in the example. Database records containing continuous variables, time-dependent information or a mixture of several types of data also fit within the presented model. However, the implementation of the corresponding linkage rules is much more complex.

B.5 Obtaining new information

The goal of re-identification is to obtain private information about someone. Although this is fairly straightforward and may seem unnecessary to mention, it is important to bear this fact in mind. When all variables of the anonymous or pseudonymous database records are observable by the attacker, the attacker's observations database is nothing but a complete identifiable version of the protected database. There can be no more information extracted from the anonymous database than is already available to the attacker; the latter cannot gain any knowledge from the anonymous data, hence, there cannot be a privacy risk associated with the data recorded in the anonymous database. All information is available elsewhere. Figure B.7 shows how two different data sets may be used to link information that is thought to be concealed.

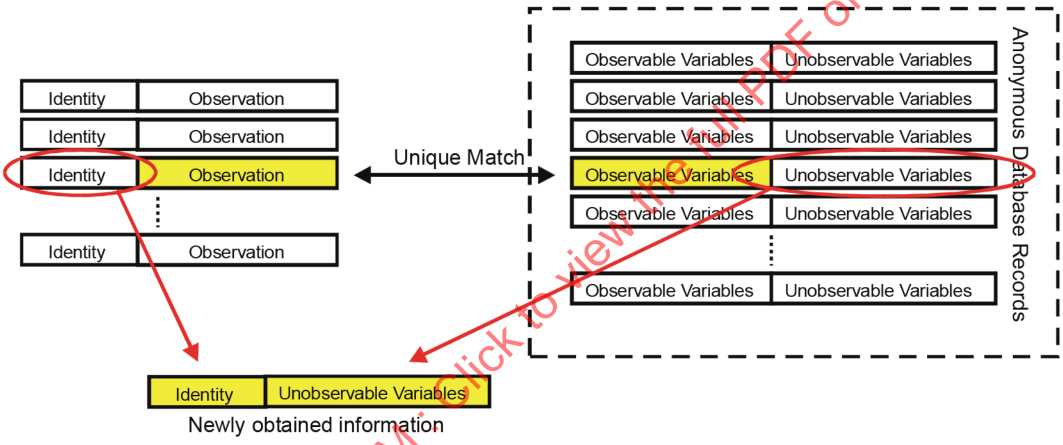


Figure B.7 — Extracting new information out of an anonymous or pseudonymous database through re-identification

Of course, there are some borderline cases. For instance, the observations could have a high uncertainty. The anonymous database then serves as a verification mechanism.

B.6 Database membership

Next to re-identification, the goal of an attacker can be to solely determine if a subject is listed in a database or not. The "membership" of a database itself can be the information the attacker needs. Real life examples are easily found, for instance, in disease management databases (e.g. a database with HIV infected persons).

Determining non-membership is relatively easy. When the linkage probability between the observations on the examined subject and every anonymous record is zero, there is no doubt that the subject is not listed in the database, that is, if the observations and anonymous databases contain no errors.

Unfortunately, this logic cannot simply be reversed. If an attacker cannot prove that a subject is not listed in a database, this does not necessarily imply that the subject is listed in the anonymous records. In terms of Figure B.8, if records are matching between the anonymous and the observations database, the records could belong to one single subject member of the A∩O subset or belong to two different subjects (one member of the A\O subset, another member of the O\A subset). A probability will have to be associated with the link, a probability which will then be translated into a probability of membership.

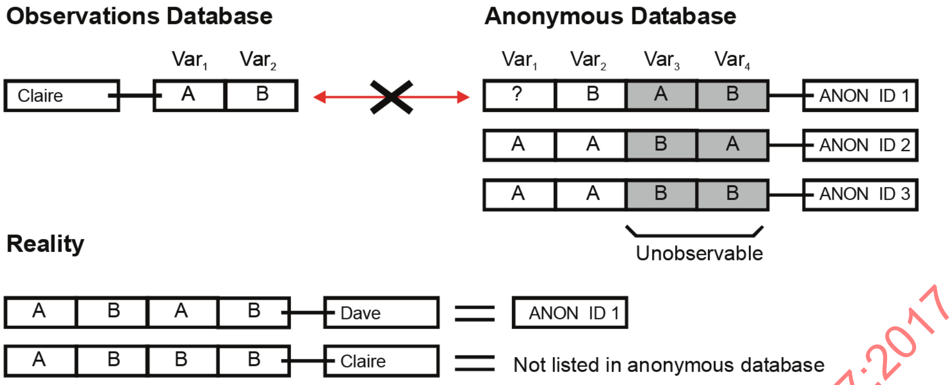


Figure B.8 — Example

The example in Figure B.8 illustrates this. Consider an anonymous database with three records. An attacker wants to know if Claire is a member of that database and using the observations and a simple linkage rule, there is a unique match between “Claire” and “anon-ID 1”. However, this does not mean that the anonymous record belongs to “Claire”, as demonstrated on the figure, it was actually “Dave” who was listed in the database. If the attacker has attributed a large probability to the unique link based on his linkage rules (here the comparison of 2 variables), then he might draw the incorrect conclusion that “anonymous identifier 1” is “Claire”.

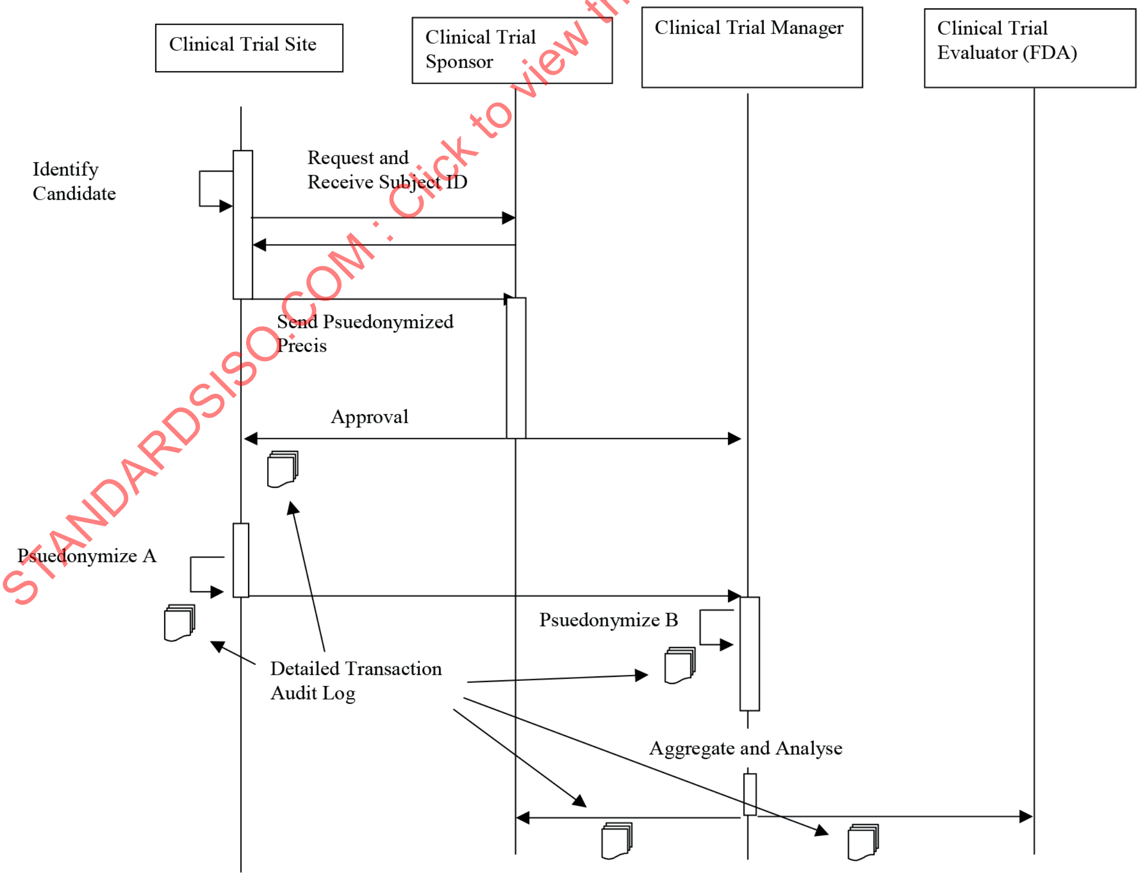


Figure B.9 — Pseudonymization for trial subjects

Figure B.9 represents a process flow for pseudonymization of trial subjects. The step are as follows.

a) The clinical trial sponsor has a clinical trial ID. The details of the trial are not normally revealed to candidates or healthcare providers. That would defeat the double blind purpose. Only enough information to assess the candidate and assess the appropriateness of the trial and the risks are disclosed.

b) The clinical trial site uses that to identify a candidate for the trial. This candidate is assigned a number by the clinical trial sponsor (or by the clinical trial manager). This is often just a sequence number of requests. The trial site has not disclosed anything more than an interest at this point.

c) The clinical trial site prepares a brief precis of the candidate medical record, using the clinical trial number and subject number instead of identifiers. The precis includes only that information needed for candidate evaluation.

d) The clinical trial sponsor (or manager) determines that the candidate is suitable and approves the trial to proceed.

e) At various times, the clinical trial site sends data to the clinical trial manager. This information is pseudonymized by removing unnecessary identifiers, removing some data and using the clinical trial ID and subject ID from the original request. All data modification, substitution and deletion are tracked in a detailed audit log, as is the data submission. This log is kept at the trial site.

f) These data are securely sent to the clinical trial manager. The receipt is logged and audited.

g) These data are further pseudonymized by further data removal. (The clinical trial site is aware of more details regarding the trial, and can eliminate more data.) Also, data that could identify the trial site are removed. This is audited with full data recovery.

h) This further pseudonymized data is analysed and aggregated to evaluate the trial results.

i) The aggregate data and supporting pseudonymized data are securely provided to the clinical trial sponsor, drug reviewers, etc. This transmission is fully audited.

NOTE 1 There is no pseudonymization service. It is performed internally by the systems involved.

NOTE 2 There is no use of crypto techniques for generating pseudonyms. The use of arbitrary sequence numbers assigned by the clinical trial sponsor is more robust. (Crypto-derived pseudonyms are highly vulnerable to dictionary attack unless very carefully implemented. The list of names is rather short.)

NOTE 3 Data recovery requires multiple steps. The audit trails and various site records are used, rather than embedding original data as an encrypted side payload. To track back to the original person, it is necessary to first visit the clinical trial manager, then examine audit logs, then visit the clinical trial site to finally identify the person.

Annex C
(informative)

Pseudonymization process (methods and implementation)

C.1 Design criteria

When data are being pseudonymized, identifying and payload data shall be separated.

The separation of identifying and payload data according to assurance levels and risk assessment as described, is a core step in the pseudonymization of data. Further processing steps will take the identifying part as input and leave the payload unchanged. The pseudonymization process translates the given identifiers into a pseudonym. For an observer, the resulting pseudonyms contain no identifying information (which is the basis of cryptographic transformations).

This transformation can be implemented differently according to the project requirements. Pseudonymization can:

- always map a given identifier with the same pseudonym. Because the combination of both preservation of linkage between records belonging to the same identity and the protection of privacy of the data subjects is the main reason for using pseudonymization, this variant is used most often;
- map a given identifier with a different pseudonym:
  - context dependent (context spanning aspect of a pseudonym);
  - time dependent (e.g. always varying or changing over specified time-intervals);
  - location dependent (e.g. changing when the data comes from different places).

C.2 Entities in the model

The pseudonymization model contains four entities that are denoted as (see Figure 6)

- data source,
- pseudonymization service provider,
- person identification service provider, and
- data target.

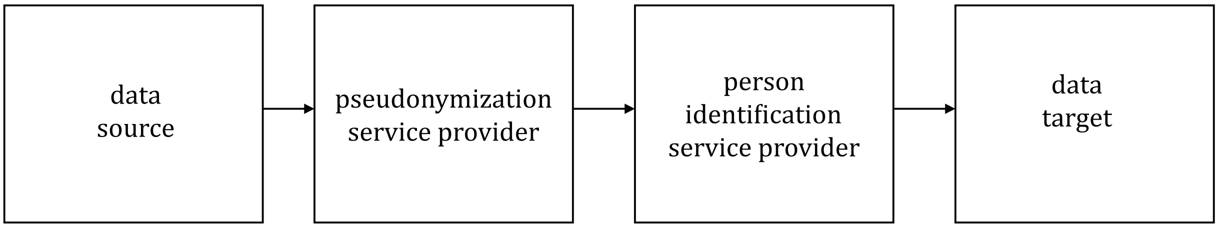

data source
pseudonymization service provider
person identification service provider
data target

Figure C.1 — Entities in the model

These entities can be complemented by, for example, authentication services, key escrow services or other services required by the process model (see Figure C.1).

A source is an entity that performs the following functions.

a) Preparing and structuring the data for submission to the person identification and the pseudonymization. The pseudonymization service has to know what it is expected to do with a data element. This can be done by either tagging the data elements or by positioning the data elements in a defined location, which will each be processed in a pre-defined way.

b) Submitting the data to the person identification service and then to the pseudonymization service. This can be done by calling an identifier service client and then a pseudonymization.

c) Reading and following-up the result code from calling the pseudonymization service. This can consist of simply logging the result in case of success or of retrying or sending warnings in the case of failure and depending on the return information.

A target is an entity that receives pseudonymized data from the pseudonymization service and that takes care of the further processing of the data. Depending on the local legislation and on the assurance level, even pseudonymized data may fall under the applicability of data privacy protection laws:

— decryption of the data received from the pseudonymization service;

— insertion of the received data into the target repositories according to the rules of the system (checking for doubles, updates, etc.);

— statistical analysis of the resulting data set.

The patient identification and pseudonymization services are the entities that perform the patient identification reconciliation and pseudonymization processes. All information needed on which to base its policy decision during a session shall be present in the session data. In cases where the pseudonymization is desired across unaffiliated entities or to decrease the risk of unauthorized re-identification, such services should be provided by a pseudonymization service.

The patient identification service manages identities communicated to the pseudonymization service. This patient identification service is associated with the data source either directly, or through a defined relationship. The source identity and that provided through the patient identification service may be different depending upon the architectural relationship.