TECHNICAL SPECIFICATION

ISO/TS 25237

First edition
2008-12-01

Health informatics — Pseudonymization

Informatique de santé — Pseudonymisation

Reference number
ISO/TS 25237:2008(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimised for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In other circumstances, particularly when there is an urgent market requirement for such documents, a technical committee may decide to publish other types of document:

— an ISO Publicly Available Specification (ISO/PAS) represents an agreement between technical experts in an ISO working group and is accepted for publication if it is approved by more than 50 % of the members of the parent committee casting a vote;

— an ISO Technical Specification (ISO/TS) represents an agreement between the members of a technical committee and is accepted for publication if it is approved by 2/3 of the members of the committee casting a vote.

An ISO/PAS or ISO/TS is reviewed after three years in order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn. If the ISO/PAS or ISO/TS is confirmed, it is reviewed again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.

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.

ISO/TS 25237 was prepared by Technical Committee ISO/TC 215, *Healthcare informatics*.

Introduction

Pseudonymization is recognised 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:

- secondary 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;
- equipment maintenance.

This Technical Specification 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 Technical Specification also defines the interfaces to pseudonymization services to ensure interoperability between pseudonymization service systems, identity management systems, information providers and recipients of pseudonyms.

Health informatics — Pseudonymization

1 Scope

This Technical Specification contains principles and requirements for privacy protection using pseudonymization services for the protection of personal health information. This technical specification is applicable to organizations who make a claim of trustworthiness for operations engaged in pseudonymization services.

This Technical Specification:

- defines one basic concept for pseudonymization;
- gives an overview of different use cases for pseudonymization that can be both reversible and irreversible;
- defines one basic methodology for pseudonymization services including organizational as well as technical aspects;
- gives a guide to risk assessment for re-identification;
- specifies a policy framework and minimal requirements for trustworthy practices for the operations of a pseudonymization service;
- specifies a policy framework and minimal requirements for controlled re-identification;
- specifies interfaces for the interoperability of services interfaces.

2 Normative references

The following referenced documents are indispensable for the application 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.

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

[ISO/IEC 2382-8:1998, definition 08.04.01]

3.2
anonymization
process that removes the association between the identifying data set and the data subject

3.3
anonymized data
data from which the patient cannot be identified by the recipient of the information

[General Medical Council Confidentiality Guidance]

3.4
anonymous identifier
identifier of a person which does not allow the unambiguous identification of the natural person

3.5
authentication
assurance of the claimed identity

3.6
ciphertext
data produced through the use of encryption, the semantic content of which is not available without the use of cryptographic techniques

[ISO/IEC 2382-8:1998, definition 08-03-8]

3.7
confidentiality
property that information is not made available or disclosed to unauthorized individuals, entities or processes

[ISO 7498-2:1989, definition 3.3.16]

3.8
content-encryption key
cryptographic key used to encrypt the content of a communication

3.9
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

3.10
cryptography
discipline which embodies principles, means and methods for the transformation of data in order to hide its information content, prevent its undetected modification and/or prevent its unauthorized use

[ISO 7498-2:1989, definition 3.3.20]

3.11
cryptographic algorithm
⟨cipher⟩ method for the transformation of data in order to hide its information content, prevent its undetected modification and/or prevent its unauthorized use

3.12
key management
cryptographic key management
generation, storage, distribution, deletion, archiving and application of keys in accordance with a **security policy** (3.43)

[ISO 7498-2:1989, definition 3.3.33]

3.13
data integrity
property that data have not been altered or destroyed in an unauthorized manner

[ISO 7498-2:1989, definition 3.3.21]

3.14
data linking
matching and combining data from multiple databases

3.15
data protection
technical and social regimen for negotiating, managing and ensuring informational privacy, confidentiality and security

3.16
data-subjects
persons to whom data refer

3.17
decipherment
decryption
process of obtaining, from a ciphertext, the original corresponding data

[ISO/IEC 2382-8:1998, definition 08-03-04]

NOTE A ciphertext can be enciphered a second time, in which case a single decipherment does not produce the original plaintext.

3.18
de-identification
general term for any process of removing the association between a set of identifying data and the data subject

3.19
direct identifying data
data that directly identifies a single individual

NOTE 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.

3.20
disclosure
divulging of, or provision of access to, data

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

3.21
encipherment
encryption
cryptographic transformation of data to produce **ciphertext** (3.6)

[ISO 7498-2:1989, definition 3.3.27]

NOTE See **cryptography** (3.10).

3.22
subject of care identifier
healthcare identifier
identifier of a person for exclusive use by a healthcare system

3.23
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

[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]

3.24
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 The identification of an entity within a certain context enables another entity to distinguish between the entities with which it interacts.

3.25
identifier
information used to claim an identity, before a potential corroboration by a corresponding authenticator

[ENV 13608-1]

3.26
indirectly identifying data
data that can identify a single person only when used together with other indirectly identifying data

NOTE 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.

3.27
information
data set within a context of meaning

3.28
irreversibility
situation when, for any passage from identifiable to pseudonymous, it is computationally unfeasible to trace back to the original identifier from the pseudonym

3.29
key
sequence of symbols which controls the operations of **encipherment** (3.21) and **decipherment** (3.17)

[ISO 7498-2:1989, definition 3.3.32]

3.30
linkage of information objects
process allowing a logical association to be established between different information objects

3.31
other names
name(s) by which the patient has been known at some time [HL7]

3.32
person identification
process for establishing an association between an information object and a physical person

3.33
personal identifier
information with the purpose of uniquely identifying a person within a given context

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

[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]

3.35
primary use of personal data
use of personal data for delivering healthcare

3.36
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 about that individual

[ISO/IEC 2382-8:1998, definition 08-01-23]

3.37
processing of personal data
any operation or set of operations that is performed upon personal data, 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

[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]

3.38
processor
natural or legal person, public authority, agency or any other body that processes personal data on behalf of the controller

[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]

3.39
pseudonymization
particular type of anonymization that both 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

3.40
pseudonym
personal identifier that is different from the normally used personal identifier

NOTE 1 This may be either derived from the normally used personal identifier in a reversible or irreversible way, or alternatively be totally unrelated.

NOTE 2 Pseudonym is usually restricted to mean an identifier that does not allow the derivation of the normal personal identifier. Such pseudonymous information is thus functionally anonymous.

3.41
recipient
natural or legal person, public authority, agency or any other body to whom data are disclosed

3.42
secondary use of personal data
any use different from primary use

3.43
security policy
plan or course of action adopted for providing computer security

[ISO/IEC 2382-8:1998, definition 08-01-06]

4 Symbols (and abbreviated terms)

HIPAA	Health Insurance Portability and Accountability Act
HIS	Hospital Information System
HIV	Human Immunodeficiency Virus
IP	Internet Protocol
VoV	Victim of Violence

5 Requirements for privacy protection of identities in healthcare

5.1 A conceptual model for pseudonymization of personal data

5.1.1 General

De-identification is the general term for any process of removing the association between a set of identifying data and the data subject. Pseudonymization is a subcategoy of de-identification. The pseudonym is the means by which pseudonymized data are linked to the same person across multiple data records or information systems without revealing the identity of the person. 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.

NOTE Anonymization is another subcategory of de-identification. Unlike pseudonymization, it does not provide a means by which the information may be linked to the same person across multiple data records or information systems. Hence re-identification of anonymized data is not possible.

5.1.2 Objectives of privacy protection

The objective of privacy protection, e.g. by using pseudonymization, 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 strands in the protection of personal data, one that is oriented towards the protection of personal data in interaction with on-line applications (e.g. web browsing) and another strand that looks at the protection of collected personal data in databases. This Technical Specification will restrict itself to the latter context.

A pre-requisite of this conceptual model is that data can be extracted from, e.g. treatment or diagnostic databases. This conceptual model ensures that the identities of the data subjects are not 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.

5.1.3 Privacy protection of entities

The conceptual model uses the privacy of personal data as a starting point, but the term "data subject" is not limited to persons and can denote any other entity such as an organization, a device or an application. It is however useful to focus on physical persons as their privacy is covered in legislation and the focus of privacy protection is on them. Privacy legislation contains specifications on some of the concepts covered in this model. In the healthcare context, the privacy protection of persons is much more complicated than the privacy protection of, e.g., devices, because phenotype data can potentially help to identify the data subject.

5.1.4 Personal data *versus* de-identified data

5.1.4.1 Definition of personal data

According to the Data Privacy Protection Directive (Directive 95/46/EC) of the European Parliament and of the Council of 24th October 1995[7] (European Data Protection Directive), *“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 legislations with consideration for the same principals found in this definition (e.g. HIPAA).

5.1.4.2 The idealized concept of identification and de-identification

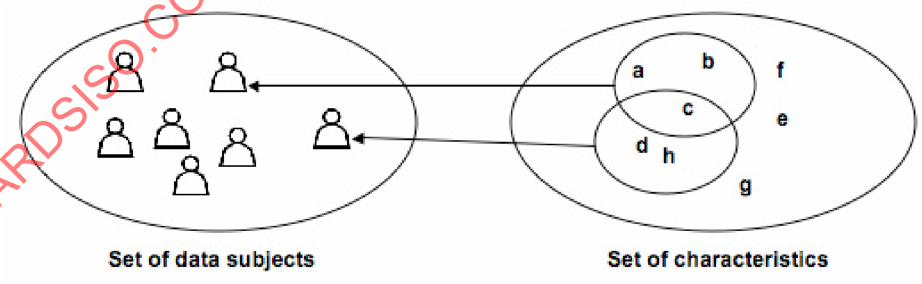


Figure 1 — Identification of data subjects

This subclause describes an idealized concept of identification and de-identification. It is assumed that there are no data outside the model, e.g. that may be linked with data inside the model to achieve (indirect) identification of data subjects.

In 5.1.5, 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).

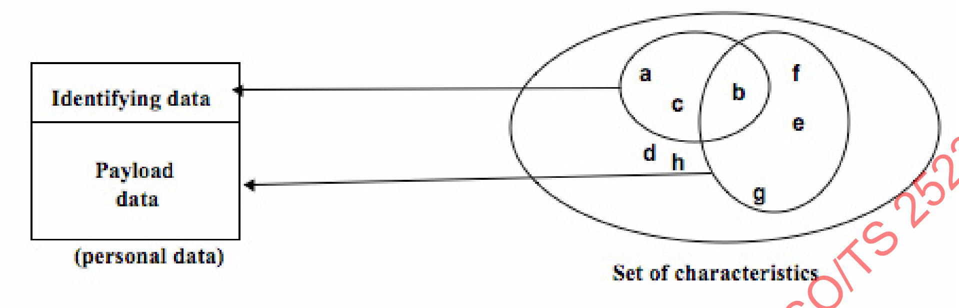


Figure 2 — Separation of personal data from payload data

From a conceptual point of view, personal data can be split up into two parts according to identifiability criteria (see. Figure 2):

- payload data: the data part, containing characteristics that do not allow unique identification of the data subject; conceptually, the payload contains anonymous data;
- 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, e.g. 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.

5.1.4.3 The 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 one 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:

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) 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 privacy 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.

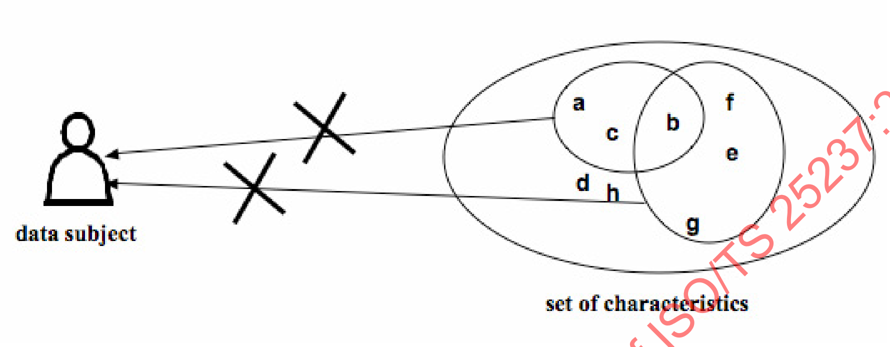


Figure 3 — Anonymization

Anonymization (see Figure 3) 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;
- 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.

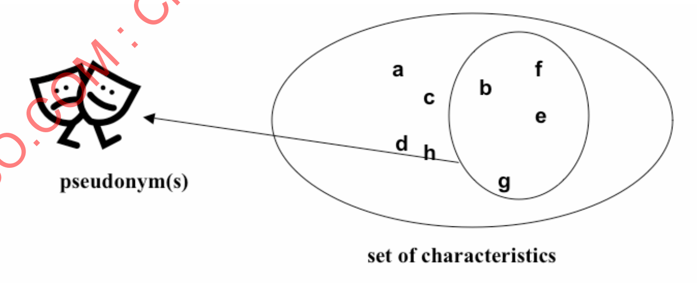


Figure 4 — Pseudonymization

Pseudonymization (see Figure 4) 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, e.g., 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.

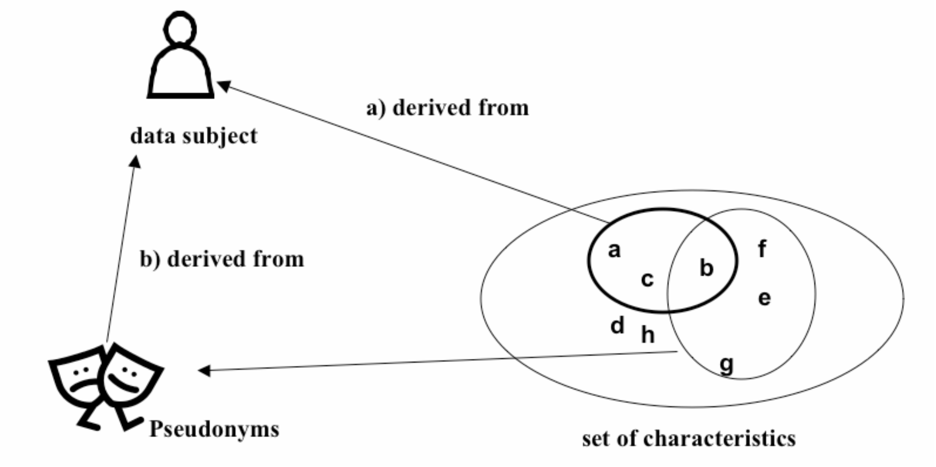


Figure 5 — Reversible pseudonymization

In reversible pseudonymization (see Figure 5), 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 9 and should take care of this. 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.

5.1.5 Real world pseudonymization

5.1.5.1 Rationale

Subclause 5.1.4 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.

Recital 26 of the European Data Privacy Protection Directive states that “*to determine whether a person is identifiable, account should be taken of all the means likely reasonable to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible*”.

The recital focuses, as the definition of personal data itself, on “identification”, i.e. the association between data and data subject.

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”.

5.1.5.2 Levels of assurance of privacy protection

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 that algorithm is, 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 pseudonimized.

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:

— 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.

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. Subclause 6.6.2.2 provides specification of data elements that should be considered for removal or aggregation to assert an anonymized data set.

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, e.g., the admission of a patient for an episode of care and other events, e.g. 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. E.g. 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.

Assurance level 3 privacy protection: considering outliers of data.

The re-identification risk can be seriously influenced by the data itself, e.g. 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.

5.2 Categories of data subject

5.2.1 General

This Technical Specification 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.

Subclauses 5.2.2 to 5.2.4 enumerate specific categories of data subjects and list a number of issues related to these categories.

5.2.2 Patient/healthcare consumer

Decisions to protect the identity of the patient may be associated with:

- legal requirements for privacy protection;
- trust relationships between the health professional and patient associated with medical secrecy principals;
- 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 secondary use of clinical data for research purposes; be aware that in some legislations (e.g. in Germany), the secondary use of patient 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 patient 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.

5.2.3 Health professionals and organizations

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

- 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.

5.2.4 Device data

Device considerations in healthcare are required for privacy protection. For patients, a primary device consideration involves implanted medical devices. Associated device identifiable data are directly associable to the patient. Other medical and personal devices may also be associated with a patient (e.g. respiratory assistive devices). As such, a device as a data subject may be identifiable as well as characteristic output from the device that may distinguish a patient. Healthcare devices assigned to a healthcare professional or employee shall also be considered in identification risk assessment as it pertains to the provider or organization.

5.3 Classification of data

5.3.1 Payload data

According to the paradigm followed in this Technical Specification, 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.

5.3.2 Observational data

Observational data reflect 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.

5.3.3 Pseudonymized data

Two types of pseudonymized data are possible.

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.

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

NOTE Reversiblity is a property that can be achieved by applying various methods such as: 1) encrypt identifiable data along with the pseudonymized data; 2) maintain a protected escrow list that links pseudonyms with identifiers.

5.3.4 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.

5.3.5 Research data

5.3.5.1 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 scaleable 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 re-use 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.

5.3.5.2 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 facilitates consent and definition of rules surrounding circumstances and reasons for intentional re-identification needs.

5.3.5.3 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 purposes. 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.

5.3.6 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 Technical Specification. However, at the core of some identity management solutions will be a pseudonymization solution.

5.3.7 Data of VoV and publicly known persons

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.).

5.3.8 Genetic information

There are two schools of thought regarding genetic information. One point of view is that the genetic information is different from other medical information. The second point of view considers the genetic data in the same way as any other medical information. The point of view that genetic data are different than other medical data, has been coined “genetic exceptionalism”.

In 2004, the European Directorate C (Science and Society), unit C3 (Ethics and Science) issued twenty-five recommendations on the ethical, legal and social implications of genetic testing. The recommendations are:

- “genetic exceptionalism” should be avoided, internationally, in the context of the EU and at the level of its member states. However, the public perception that genetic testing is different needs to be acknowledged and addressed;
- all medical data, including genetic data, should satisfy equally high standards of quality and confidentiality.

On confidentiality, privacy and autonomy, the recommendations are:

- genetic data of importance in a clinical and/or family context should receive the same level of protection as other comparably sensitive medical data;
- the relevance for other family members should be addressed;
- the importance of a patient's right to know or not to know should be recognised and mechanisms incorporated into professional practice that respect this;
- in the context of genetic testing, encompassing information provisions, counselling, informed consent procedures and communication of test results, practices should be established to meet this need.

There are of course other opinions that do not agree with this vision. One argument, for instance, is that the information content of genetic information collected in the context of population research is not exactly known and that the information content of data in databanks or human tissue repositories may be more sensitive than currently can be assessed.

Other groups do not follow this line of thought, such as the so-called "Montreux declaration" made at the 27th International Conference of Data Protection and Privacy Commissioners in September 2005 where, in the preamble, it is stated that: “Aware that the fast increase in knowledge in the field of genetics may make human DNA the most sensitive personal data of all; Aware also that this acceleration in knowledge raises the importance of adequate legal protection and privacy”.

Nevertheless, existing legislation, guidelines, and publications on related subject matter that often deals with the broader context of population genetic research and human tissue repositories shall be considered where genetic information is part of the dataset to be pseudonymized. In generic terms, data resources may be classified as either the identification of disease susceptibility genes or diagnostic biomarkers.

5.4 Trusted services

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.

5.5 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. This approach maintains a connection between payload data and personal identifiers, but can allow for re-identification under prescribed circumstances and protections.

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.

Vocabulary identification: iso (1) standard (0) pseudonymization (25237) re-identification purpose (1)

1) data integrity verification/validation;

2) data duplicate record verification/validation;

3) request for additional data;

4) link to supplemental information variables;

5) compliance audit;

6) communicate significant findings;

7) 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.

5.6 Pseudonymization service characteristics

There are two primary scenarios for pseudonymization services:

1) 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;

2) pseudonyms provided through pseudonymization services – in this situation, typically the service is providing pseudoidentities 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 Pseudonymization process (methods and implementation)

6.1 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 dependant (e.g. always varying or changing over specified time-intervals);
  - location dependant (e.g. changing when the data comes from different places).

6.2 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
- data target

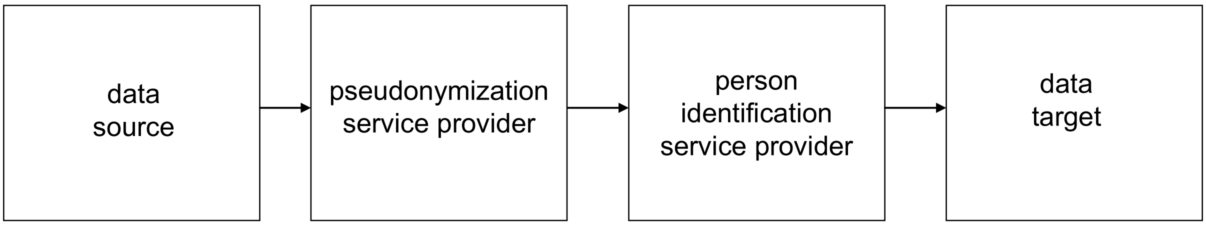

data source
pseudonymization service provider
person identification service provider
data target

Figure 6 — 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.

A source is an entity that performs the following functions.

- 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.

- 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.

- 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;

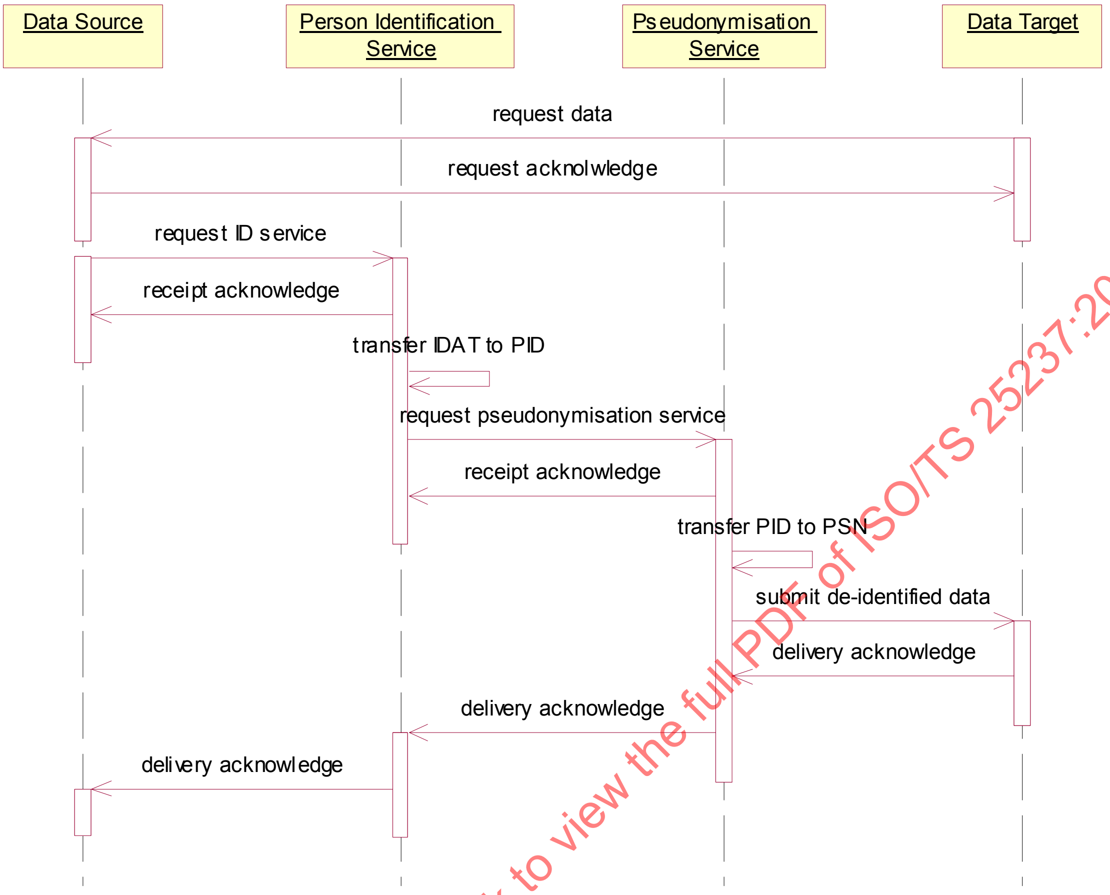
- insertion of the received data into the target repositories according to the rules of the system (checking for doubles, updates, etc.);

- 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.

6.3 Workflow in the model



Abbreviations:

ID service:	identification service	PID:	personal identifier
IDAT:	identifying data	PSN:	pseudonym

Figure 7 — Data flows

Figure 7 shows typical data flow between the entities in the workflow model. These identify also the message types that will be required for passing data or signalling the status of the operation through acknowledgement messages. The extraction of the data and integration of the data is outside the scope of this model and are handled by the source application and target application respectively.

The flow events include:

— request of data from the data target to be included in the identification and pseudonymization requests and associated acknowledgement;

— request of domain patient identifiers for communication with the pseudonymization service and associated receipt acknowledgement;

— transmission of the pseudonymization request;

— receipt acknowledgement by the pseudonymization services of the pseudonymization request;

- submit pseudonymized data to data target; when the pseudonymization process has been completed the data are sent to the target application(s).
- delivery acknowledgement: the data target acknowledges the receipt of the data (this process may also include the result of checks, e.g. on the validity of the data format);
- if round trip acknowledgement is required, the pseudonymization service transmits the acknowledgement to the person identification service that transmits an acknowledgemement to the data source.

In the case where exceptional events occur during the processing by the pseudonymization service, exception codes are returned (e.g. data malformed, failed authentication of data source) to the source.

6.4 Preparation of data

Before data can be submitted to the pseudonymization service, it has to be prepared at the source. The preparation is necessary in order to apply the privacy principles defined in the privacy policy.

The conceptual model for the use of pseudonymization services requires that the data be split up in a part containing identifying data and in another part containing nothing but anonymous data.

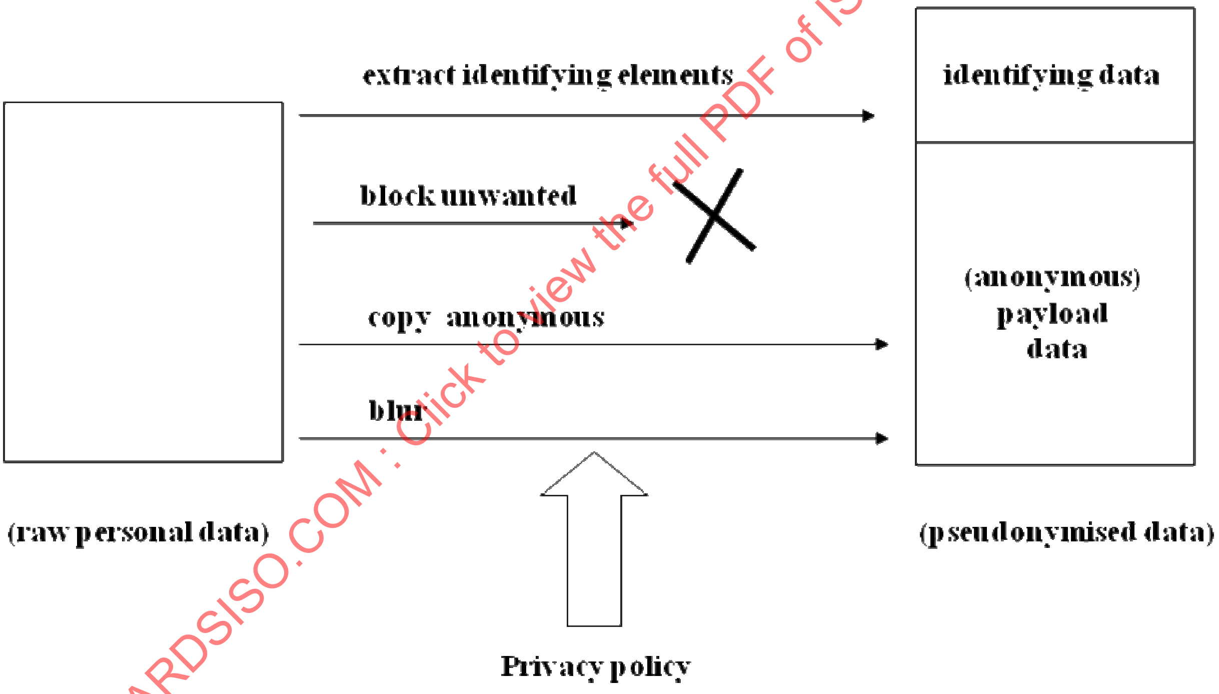


Figure 8 — Data preparation

The structuring can be done by tagging the data elements, by creating a table with vectors to the data elements or by putting the data elements in a pre-defined location.

The following preparations can be distinguished in accordance with the conceptual model on the levels of anonymity.

- Data elements that will be used for linking, grouping, anonymous searching, matching, etc. shall be indicated and marked in such a way that the pseudonymization service knows where to find them and how to handle them.

- Depending on the privacy policy, convert elements that need specific transformations, e.g. for changing absolute time references into relative time references, dates of birth into age groups, need similar marking.

- Identifying elements that, according to the privacy policy, are not needed in the further processing in the target applications, shall be discarded.

- The anonymous part of the raw personal data is put into the payload part of the personal data element.

6.5 Processing steps in the workflow

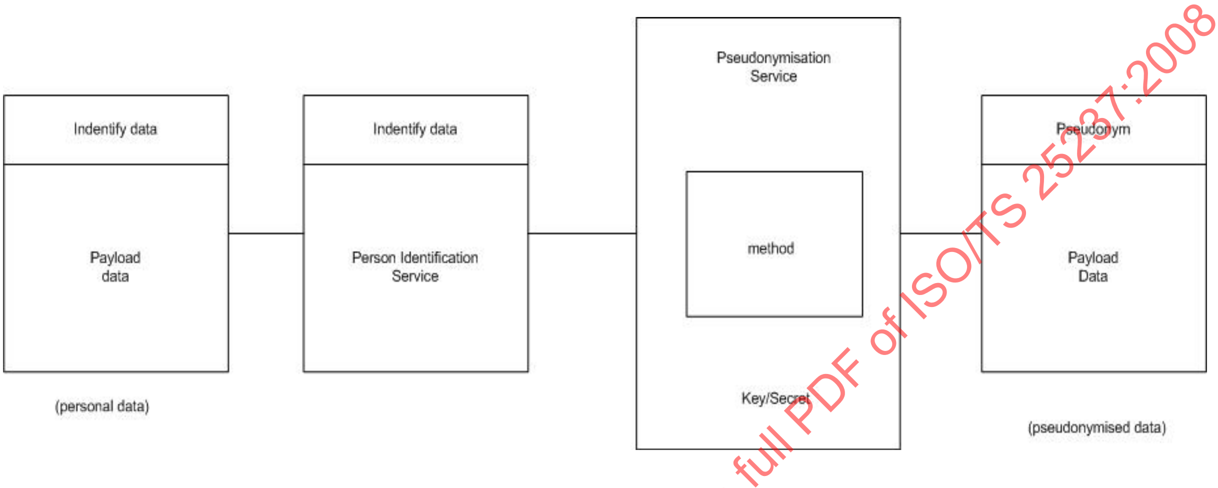


Figure 9 — Pseudonymization process

The basic steps of a pseudonymization process consist of the following.

1) At the source that is submitting personal data to the pseudonymization service, data are split in an identifying and in an anonymous payload data. What is considered identifying data and what is considered anonymous data depend upon the target level of anonymity in the security policy of the data collection project.

2) The pseudonymization service parses the header and performs the tasks laid down in its policy. This comprises the pseudonymization of data items, calculation of relative dates, removal of data items and possible encryption of specific data items where this is defined in the privacy protection policy. Consequently, the content of the payload is not visible to the pseudonymization service. This is the preferable way of operation for the pseudonymization service. The policy can, however, define otherwise, such that the content is parsed as well. Typically the parsing is done to check for unwanted content (e.g. identifiers in the data items). All checking is done in a “stateless” way. The pseudonymization service does not store data it has previously processed and therefore it cannot compare or check with that data. Only the current session can be taken into account.

   NOTE The processing of pseudonyms and anonymization of payload data can be conducted through separate entities and services.

3) Once the processing is done, the pseudonymization service sends the pseudonymized data to the data repository over a secure channel.

4) When the repository application receives the data, it applies its business rules to it in order to incorporate the data into the repository. This may include rules such as checking for double entries, checking for missing entries, required acknowledgement procedures, etc.

Indeed, a pseudonymization service or trusted third party performing a pseudonymizing transformation is necessary for trustworthy implementation of the pseudonymization technique across multiple entities. There are three main reasons for this.

- As one communicating party does not always trust the other, trust can be established indirectly because the two parties trust a third, independent party. Both parties are bound by a code of conduct, as specified in a privacy and security policy agreement they agree on with the pseudonymization service.
- Use of a pseudonymization service offers the only reliable protection against several types of attack on the pseudonymization process.
- Complementary privacy enhancements technology (PETs) and data processing features can easily be implemented.

EXAMPLES Controlled reversibility without endangering privacy, interactive databases.

6.6 Protecting privacy protection through pseudonymization

6.6.1 Conceptual model of the problem areas

This Technical Specification 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 may be required that include pseudonym translations between identity domains, depending on the identity management scheme.

Two important elements in the concept of pseudonymization are:

- the domain where a pseudonym will be used;
- protection of the pseudonymization key or seed.

6.6.2 Direct and indirect identifiability of personal information

6.6.2.1 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.

6.6.2.2 Person identifying variables

Person identifying variables include:

- person's name (including preferred name, legal name, other names by which the person is known); by name, we are referring to the name and 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 supplement 55).

6.6.2.3 Aggregation variables

For statistical purposes, absolute data references should be avoided.

- Dates of birth, e.g. 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 Technical Specification.
- 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).
- 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[1].

Demographic data are 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:

- language spoken at home;
- person's communication language;
- religion;
- ethnicity;
- person gender;
- country of birth;

[1] HIPAA section 164.514.

- 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.

6.6.2.4 Outlier variables

Outlier variables should be removed based upon risk assessment.

Outlier variables include:

- 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.

6.6.2.5 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.

6.6.2.6 Non-structured data variables

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.

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:

- 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.

Text/voice data with non-parseable content

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

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 Supplement 55, Attribute Level Confidentiality Supplement. Additional risk assessment shall be considered for identifiable characteristics of the image or notations that are part of the image.

6.6.2.7 Inference risk assessment

It must be recognised 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/preprocessing/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.

6.6.2.8 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 (methods and implementation)

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.

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 characterised 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 can guarantee 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”.

Exception

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

- specific authentication procedures;
- exceptional interventions by the pseudonymization service provider.

When re-identification of de-identified data are 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 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. e.g. 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-dentification process and what the finality of the data are. The recovered data should indicate their origin (as a caveat, as de-identified data might not be as complete or reliable as the original personal data from which it was derived in research databases or in clinical data warehouses).

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.

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).

8 Specification of interoperability of interfaces (methods and implementation)

Interoperability of pseudonymization services can be defined on several levels. A solution may, for instance, make use of an intermediary pseudonymization service to perform the pseudonymization, or may be an in-house module added to extraction software.

Pseudonymization services may be batch oriented or may be built on the principle of pseudonymized access of live databases.

This Technical Specification however, only concentrates on the core mechanisms in which a pseudonymization service is used.

Interoperability should cover the following elements.

- One or more mechanisms for exchanging the data between the entities in the model (source, pseudonymization service, target) and for controlling the operation. This is less of an issue and existing protocols can be used, such as html. Where needed it is possible to design converters between formats as part of pre- or post-processing.

- Choice of cryptographic algorithms. Pseudonymization based on cryptographic algorithms will consist of a chain of basis cryptographic and related algorithms: hashing, random number generators, key generation, encryption and basis logic bit-string functions.

- Key exchange issues.

Since there are many different contexts in which pseudonymization can be deployed, it is important to restrict the context in which interoperability will be defined.

The common service of most pseudonymization services consists of cryptographic transformations of indentifiable data.

For two independent pseudonymization service providers to be interoperable, it should be possible either

- to integrate each other's data: data from the same date subject processed by any of the service providers should be linkable to each other without direct re-identification of the data subject;
- to convert the pseudonymization results from one or more service providers in a controlled way without direct re-identification of the data subject.

9 Policy framework for operation of pseudonymization services (methods and implementation)

9.1 General

It is important to complement the technical measures with appropriate non-technical measures. Such non-technical measures are generally expressed through policies, agreements and codes of conduct.

9.2 Privacy policy

Each data processing or collecting project that uses pseudonymization should have a privacy policy dealing with the pseudonymization aspects, either as a self-contained policy or as part of an overall security policy.

This policy should include:

- description of the processing in which pseudonymization plays a role;
- identification of the controller of the personal data;
- identification of the controller of the pseudonymized data;
- description of the pseudonymization method;
- identification of the entity carrying out the pseudonymization;
- protection, storage and handling of the pseudonymization “secrets” (usually a cryptographic key or a linking table); description of what will happen if the organization is discontinued (or at least the pseudonymization part of its activities), description also for which domains and applications the secret will be used and or how long it is valid (in case of change of secret, describe possibilities and procedures to link with legacy data);
- detailed description if the pseudonymization is reversible and what authorization by whom is required;
- definition of the limitations of the receiver of pseudonymized data (e.g. information actions, onward forwarding, retention policies):
  - he cannot make it publicly available;
  - he shall protect it from unauthorized access and only use it internally to produce de-identified data that has been aggregated and that can be made public or sold to customers;
  - he shall destroy it when not used and when he chooses not to secure the data anymore.

9.3 Trustworthy practices for operations

Provision of pseudonymization services in healthcare shall meet the following objectives in order to be effective in securing the privacy protection of personal health information:

- the reliable and secure binding of unique pseudonyms to individuals or organizations that are the subject of pseudonymized personal health information;
- the protection of the pseudonyms from unauthorized re-identification;
- the provision of authorized re-identification of the subject's source identifier(s) in accordance with re-identification policy parameters as agreed between the service provider and the service subscriber.

The above objectives shall be accomplished in a manner that maintains the trust of all who rely upon the confidentiality of the personal health information that is protected through the pseudonymization service. As pseudonymization is particularly suited to primary and secondary research and analysis purposes, information resources that rely upon these services to protect personal health information may implicitly require the trust of the patients whose information is being examined, as well as that of the general public. It is unlikely that either healthcare providers or patients will cooperate in the contribution of personal health information for analytical purposes if such identity protection services are believed to be insecure.

In order to satisfy these requirements, a pseudonymization service:

- should be strictly independent of the organizations supplying source data;
- shall be able to guarantee security and trustworthiness of its methods by publishing to its subscribers its operating practices;
- shall be able to guarantee security and trustworthiness of its software modules:
  - shall provide assurances as to the source, processes and integrity of its software modules,
  - code integrity shall be asserted through code signing;
- shall be able to guarantee security and trustworthiness of its operating envirionment, platforms and infrastructure:
  - shall restrict network traffic to restrict all unnecessary traffic,
  - shall disable all unnecessary operating system services,
  - shall provide technical, physical, procedural, and personnel controls in accordance with ISO 27799;
- shall implement monitoring and quality assurance services and programmes:
  - to assure quality of service,
  - to monitor against network penetration and malicious attacks;
- cryptographic key management:
  - shall be under multi-person control,
  - identifiers shall be encrypted by two keys, one under control of the data source, and one under the control of the pseudonymization service;
- instantiation of the pseudonymization service:

  - shall be documented,
  - shall be recorded and audited to be able to demonstrate service integrity;
- business continuity of the pseudonymization service:
  - shall be assured through backup,
  - shall be assured through a disaster recovery plan;
- internal audit procedures:
  - shall be documented,
  - shall be executed on no less than a monthly basis;
- external audit procedures:
  - shall be used to establish to the satisfaction of subscribers and any relying party that it fully complies with published operating procedures,
  - the auditor shall be completely independent of the audited party by belonging to a separate organization from the pseudonymization service provider,
  - the auditor shall have no financial interest in the audited party,
  - the auditor shall be a qualified information systems auditor to the extent necessary for admission to the relevant professional body,
  - service subscribers shall immediately be notified of any pseudonymization services that are found by an auditor to be deficient;
- participants:
  - shall maintain integrity of the organization's key(s) associated with pseudonymization,
  - shall maintain physical, network, personnel, and technical controls of the associated systems in accordance with ISO 27799,
  - are responsible for the anonymization of payload data and privacy protection of any pseudonymized information resources maintained by the organization.
- risk assessment shall be conducted regarding access by the data source to the resulting pseudonyms and specification of such restrictions shall be expressed in operational policies.

9.4 Implementation of trustworthy practices for re-identification

The pseudonymization service should provide support for controlled re-identification. A pseudonymization service providing such support shall make available to subscribers and data subjects a re-identification policy that specifies the criteria required for an authorized re-identification event.

- Re-identification shall be subject to multi-person controls, and multi-organization control.
- Time-sensitive re-identification shall be accommodated within the defined policy and process communicated to those needing time-sensitive processes (e.g. public health authorities).

- Audits shall:
  - be provided for all re-identification events in accordance with RFC 3881;
  - minimally include:
    - the party to whom the identity was disclosed;
    - the time/date of the re-identification;
    - the reason for re-identification as defined in 5.5.
- Re-identification from the pseudonymization service shall re-identify only the local pseudonym from the source organization.
- The controller of the data is responsible for re-identification of the patient, and may, as permitted by local jurisdiction, validate further the re-identification request.

Annex A
(informative)

Healthcare pseudonymization scenarios

A.1 Introduction

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:

1) A description of the scenario, or healthcare situation requiring healthcare pseudonymization services.

2) 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. E.g. 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 must 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 co-operation 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-identifcation	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

Table A.1 — (*contiued*)

Scenario		Data subject			Data sources		Functional/performance requirements			
		Kind of ID	Uniqueness	Sensitivity (pers. data legisl.)	Data sources	Primary/ secondary	Context/ finality	Searchability	Re-identifcation	Linkeage in time
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
7	HC Market Research	Physician ID/Pat ID	Unique	Low (physician) High (diagnosis)	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) Patient/health consumer identification systems (consumer ID)

A.3.1 Clinical pathology order (pseudonymous care)

Scenarios taken as example (in this group):

This scenario 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

- Submit order to health information system (HIS)
  - the placer of the order authenticates towards the HIS;
  - the placer of the order submits the order with the hospital unique ID number of the data subject to the HIS;
  - 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.
- Pseudonymize
  - the hospital information system invokes the pseudonymization service with, as input, the hospital unique ID number;
  - the PS processes the hosp ids;
  - the PS returns the pseudonym to the HIS.
- The HIS sends the order with the pseudonym to the filler
  - establish comm;
  - message sent;
  - ack received.
- The order is processed by the filler of the order using the pseudonym
  - (possible comparative analysis performed by specialist).
- The filler of the order submits the result to the HIS with the pseudonym
  - establish;
  - message sent;
  - ack received.
- Re-identify results
  - the HIS submits the pseudonym to the pseudonymization services;
  - authenticated user (HIS) is verified against reverse ID policy;
  - the PS processes the pseudonym;
  - the PS sends the real ID to the HIS.
- 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 themself to a healthcare provider for clinical care. Wishing 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.

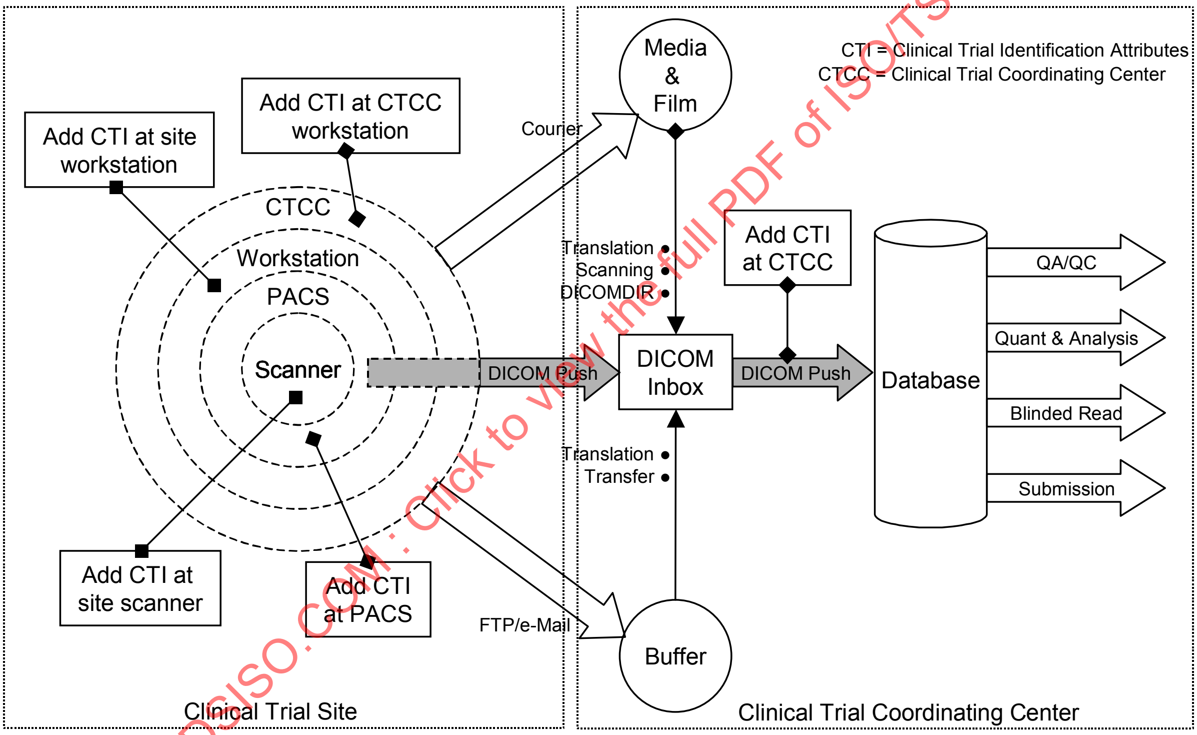


Figure A.1 — Clinical trial data modifications

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 must be preserved, data that must be blinded and data that must be removed for scientific analysis purposes.

b) The *clinical trial coordinating centre*, who 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.

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 resource (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

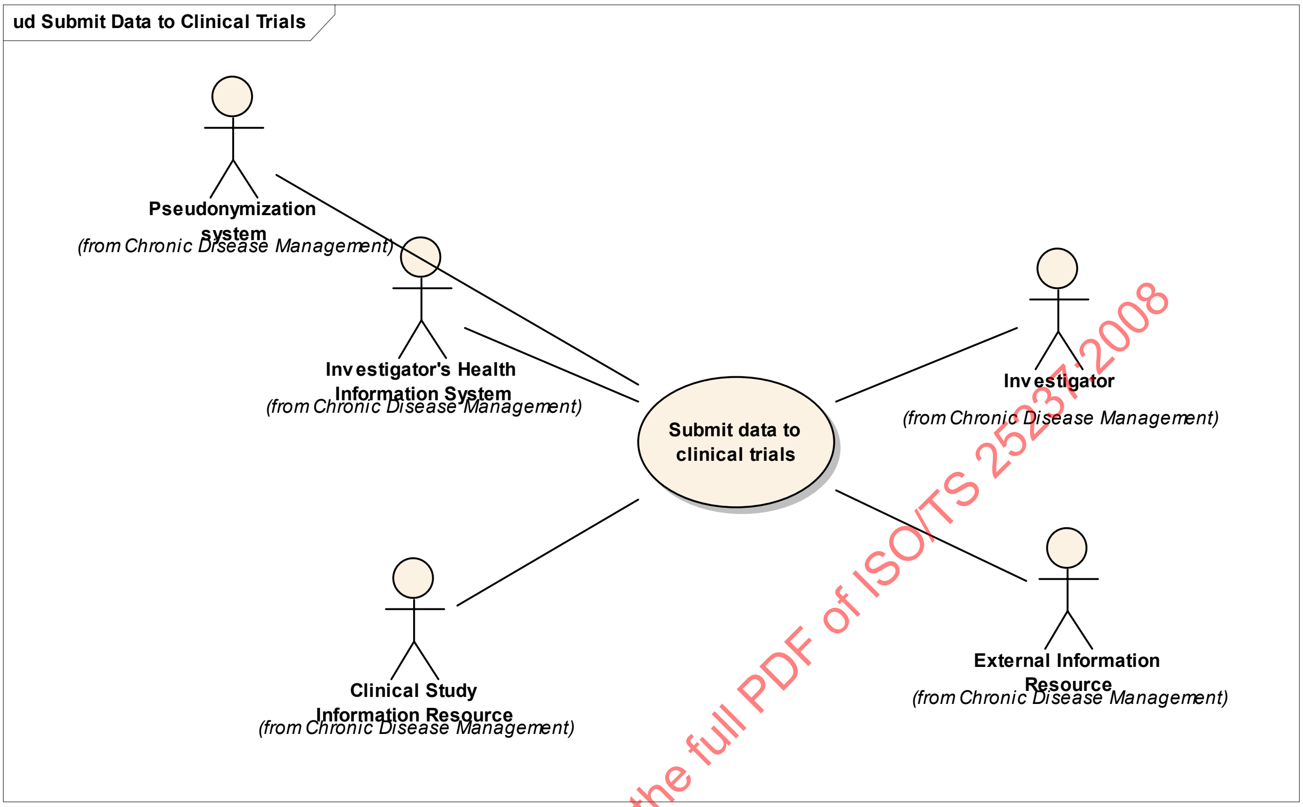

ud Submit Data to Clinical Trials
Pseudonymization system
(from Chronic Disease Management)
Investigator's Health Information System
(from Chronic Disease Management)
Submit data to clinical trials
Investigator
(from Chronic Disease Management)
Clinical Study Information Resource
(from Chronic Disease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

Figure A.2 — Clinical trial flows

Process flow:

- 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.

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 nefrology 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:

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

- 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

- 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.

Actors

- System user (e.g. member of the care team);
- Anonymization/pseudonymization system;

- Health information system;
- Event capture information resource.

Pre-conditions

- patient receives ordinary patient 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;
- pseudonymization system is available to the local HIS.

Post-conditions

- event reporting information resource has all relevant data;
- follow-up investigations supported where applicable.

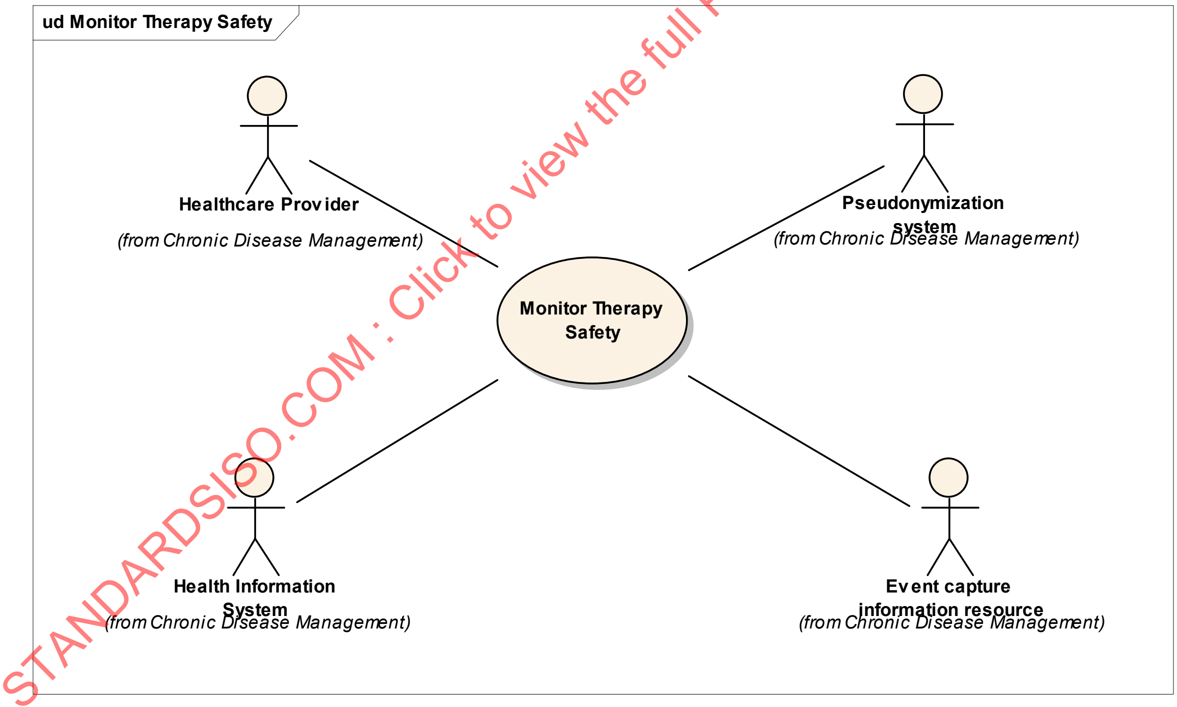

ud Monitor Therapy Safety
Healthcare Provider
(from Chronic Disease Management)
Pseudonymization system
(from Chronic Disease Management)
Monitor Therapy Safety
Health Information System
(from Chronic Disease Management)
Event capture information resource
(from Chronic Disease Management)

Figure A.3 — Patient safety flows

Process Flow

- 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 conducted by the authenticated HC Provider.

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 medical 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 must 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 must 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. With properly supervised generation of pseudonymous records the student may be permitted to keep the pseudonymous files for personal educational purposes.

A.3.8.2 Where pseudonymization is used

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-yr old male with a history of ....” The pseudonymization must 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 must be captured, rather than just a single record, but again the data can be substantially reduced. In this situation the data must be pseudonymized to preserve the relationships between records, but the pseudonymization can be irreversible.