TECHNICAL SPECIFICATION

# ISO/TS 25237

First edition
2008-12-01

# Health informatics — Pseudonymization

*Informatique de santé — Pseudonymisation*



Reference number
ISO/TS 25237:2008(E)



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order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn.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again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.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

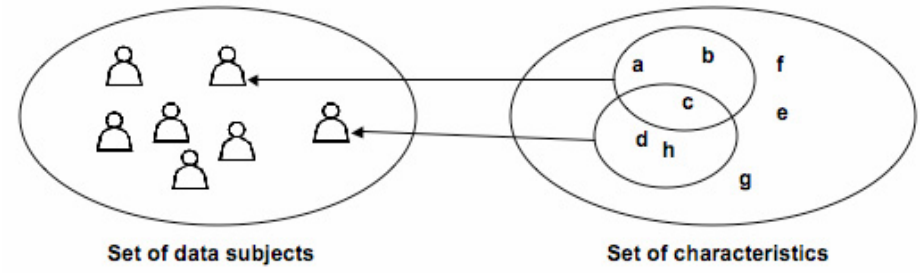


**Figure 1 — Identification of data subjects**

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).

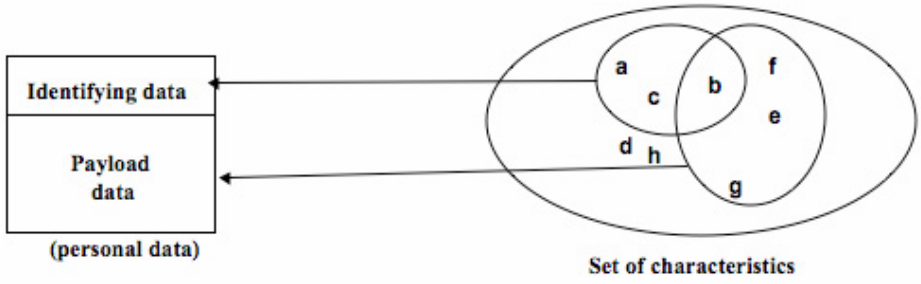


**Figure 2 — Separation of personal data from payload data**

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.

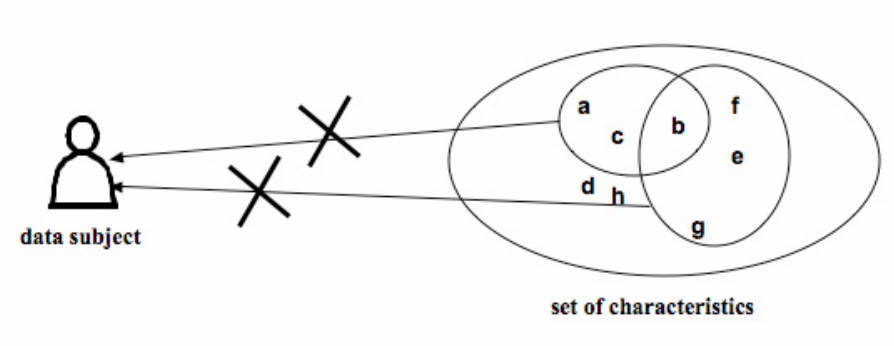


**Figure 3 — Anonymization**



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.

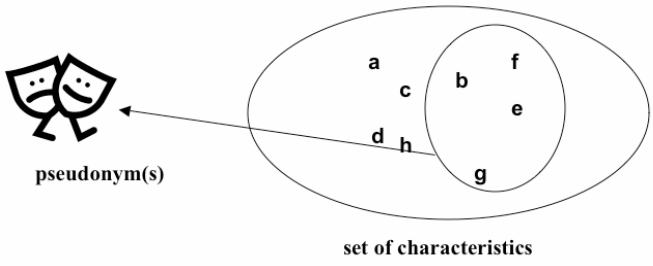


**Figure 4 — Pseudonymization**

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.