BS EN ISO 25237:2017



BSI Standards Publication

Health informatics — Pseudonymization (ISO 25237:2017)

Documents
Document Sharing Hub

Full document details: BS EN ISO 25237:2017

National foreword

This British Standard is the UK implementation of EN ISO 25237:2017.

The UK participation in its preparation was entrusted to Technical Committee IST/35, Health informatics.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2017.
Published by BSI Standards Limited 2017

ISBN 978 0 580 83466 0

ICS 35.240.80

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 31 January 2017.

Amendments/Corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS EN ISO 25237:2017

EUROPEAN STANDARD

NORME EUROPÉENNE

EUROPÄISCHE NORM

EN ISO 25237

January 2017

ICS 35.240.80

English Version

Health informatics - Pseudonymization (ISO 25237:2017)

Informatique de santé - Pseudonymisation (ISO 25237:2017)

Medizinische Informatik - Pseudonymisierung (ISO 25237:2017)

This European Standard was approved by CEN on 14 December 2016.

CEN members are bound to comply with the CEN/CENELEC Internal Regulations which stipulate the conditions for giving this European Standard the status of a national standard without any alteration. Up-to-date lists and bibliographical references concerning such national standards may be obtained on application to the CEN-CENELEC Management Centre or to any CEN member.

This European Standard exists in three official versions (English, French, German). A version in any other language made by translation under the responsibility of a CEN member into its own language and notified to the CEN-CENELEC Management Centre has the same status as the official versions.

CEN members are the national standards bodies of Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Former Yugoslav Republic of Macedonia, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and United Kingdom.



EUROPEAN COMMITTEE FOR STANDARDIZATION
COMITÉ EUROPÉEN DE NORMALISATION
EUROPÄISCHES KOMITEE FÜR NORMUNG

CEN-CENELEC Management Centre: Avenue Marnix 17, B-1000 Brussels

© 2017 CEN All rights of exploitation in any form and by any means reserved worldwide for CEN national Members.

Ref. No. EN ISO 25237:2017 E

Documents
Document Sharing Hub

Full document details: BS EN ISO 25237:2017

European foreword

This document (EN ISO 25237:2017) has been prepared by Technical Committee ISO/TC 215 “Health informatics” in collaboration with Technical Committee CEN/TC 251 “Health informatics” the secretariat of which is held by NEN.

This European Standard shall be given the status of a national standard, either by publication of an identical text or by endorsement, at the latest by July 2017, and conflicting national standards shall be withdrawn at the latest by July 2017.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. CEN [and/or CENELEC] shall not be held responsible for identifying any or all such patent rights.

According to the CEN-CENELEC Internal Regulations, the national standards organizations of the following countries are bound to implement this European Standard: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Former Yugoslav Republic of Macedonia, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and the United Kingdom.

Endorsement notice

The text of ISO 25237:2017 has been approved by CEN as EN ISO 25237:2017 without any modification.

Type de document : Norme européenne
Sous-type de document :
Stade du document : Publication / Adoption
Langue du document : E

Y:\STD_MGT\STDDEL\PRODUCTION\Standards\00251\316\64_e_stf.doc STD Version 2.5a

Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

Contents

Page

Foreword ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Abbreviated terms **6**
5 Requirements for privacy protection of identities in healthcare **7**
5.1 Objectives of privacy protection 7
5.2 General 7
5.3 De-identification as a process to reduce risk 8
5.3.1 General 8
5.3.2 Pseudonymization 8
5.3.3 Anonymization 9
5.3.4 Direct and indirect identifiers 9
5.4 Privacy protection of entities 9
5.4.1 Personal data versus de-identified data 9
5.4.2 Concept of pseudonymization 11
5.5 Real world pseudonymization 13
5.5.1 Rationale 13
5.5.2 Levels of assurance of privacy protection 14
5.6 Categories of data subject 16
5.6.1 General 16
5.6.2 Subject of care 16
5.6.3 Health professionals and organizations 16
5.6.4 Device data 16
5.7 Classification data 17
5.7.1 Payload data 17
5.7.2 Observational data 17
5.7.3 Pseudonymized data 17
5.7.4 Anonymized data 17
5.8 Research data 17
5.8.1 General 17
5.8.2 Generation of research data 18
5.8.3 Secondary use of personal health information 18
5.9 Identifying data 18
5.9.1 General 18
5.9.2 Healthcare identifiers 18
5.10 Data of victims of violence and publicly known persons 19
5.10.1 General 19
5.10.2 Genetic information 19
5.10.3 Trusted service 19
5.10.4 Need for re-identification of pseudonymized data 19
5.10.5 Pseudonymization service characteristics 20
6 Protecting privacy through pseudonymization **20**
6.1 Conceptual model of the problem areas 20
6.2 Direct and indirect identifiability of personal information 21
6.2.1 General 21
6.2.2 Person identifying variables 21
6.2.3 Aggregation variables 21
6.2.4 Outlier variables 22
6.2.5 Structured data variables 22
6.2.6 Non-structured data variables 23

© ISO 2017 – All rights reserved

AVA Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

6.2.7 Inference risk assessment 23
6.2.8 Privacy and security 24

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

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

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

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

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

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

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

Bibliography 61

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

Foreword

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

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

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

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

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

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

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS EN ISO 25237:2017

Introduction

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

Application areas include, but are not limited to:

— indirect use of clinical data (e.g. research);

— clinical trials and post-marketing surveillance;

— pseudonymous care;

— patient identification systems;

— public health monitoring and assessment;

— confidential patient-safety reporting (e.g. adverse drug effects);

— comparative quality indicator reporting;

— peer review;

— consumer groups;

— field service.

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

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

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

Health informatics — Pseudonymization

1 Scope

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

This document

— defines one basic concept for pseudonymization (see Clause 5),

— defines one basic methodology for pseudonymization services including organizational, as well as technical aspects (see Clause 6),

— specifies a policy framework and minimal requirements for controlled re-identification (see Clause 7),

— gives an overview of different use cases for pseudonymization that can be both reversible and irreversible (see Annex A),

— gives a guide to risk assessment for re-identification (see Annex B),

— provides an example of a system that uses de-identification (see Annex C),

— provides informative requirements to an interoperability to pseudonymization services (see Annex D), and

— specifies a policy framework and minimal requirements for trustworthy practices for the operations of a pseudonymization service (see Annex E).

2 Normative references

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

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

3 Terms and definitions

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

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

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

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

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

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS EN ISO 25237:2017

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

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

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

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

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

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

3.5
authentication
assurance of the claimed identity

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

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

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

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

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

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

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

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

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

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

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

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

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

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

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

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

[SOURCE: Directive 95/46/EC]

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

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

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

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

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

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

EXAMPLE Postcode, sex, age, date of birth.

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

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

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

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

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

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

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

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

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

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

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

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

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

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

[SOURCE: Directive 95/46/EC]

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

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

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

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

[SOURCE: Directive 95/46/EC]

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

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

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

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

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

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

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

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

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

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

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

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

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

4 Abbreviated terms

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

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

5 Requirements for privacy protection of identities in healthcare

5.1 Objectives of privacy protection

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

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

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

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

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

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

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

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

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

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

5.2 General

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

© ISO 2017 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS EN ISO 25237:2017

together with increasing patient expectations for privacy protection. Several examples of these are provided in Annex A.

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

5.3 De-identification as a process to reduce risk

5.3.1 General

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

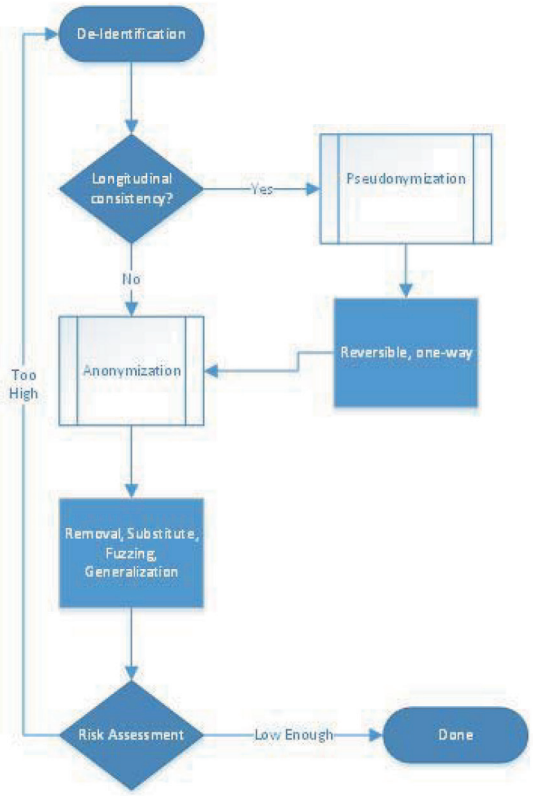


Figure 1 — Visualization of the de-identification process

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

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

5.3.2 Pseudonymization

De-identification might leverage pseudonymization where longitudinal consistency is needed. This might be to keep a bunch of records together that should be associated with each other, where without this longitudinal consistency, they might get disassociated. This is useful to keep all of the records

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

for a patient together, under a pseudonym. This also can be used to assure that each time data are extracted into a de-identified set that new entries are also associated with the same pseudonym. In pseudonymization, the algorithm used might be intentionally reversible or intentionally not-reversible. A reversible scheme might be a secret lookup-table that where authorized can be used to discover the original identity. In a non-reversible scheme, a temporary table might be used during the process, but is destroyed when the process completes.

5.3.3 Anonymization

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

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

5.3.4 Direct and indirect identifiers

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

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

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

5.4 Privacy protection of entities

5.4.1 Personal data versus de-identified data

5.4.1.1 Definition of personal data

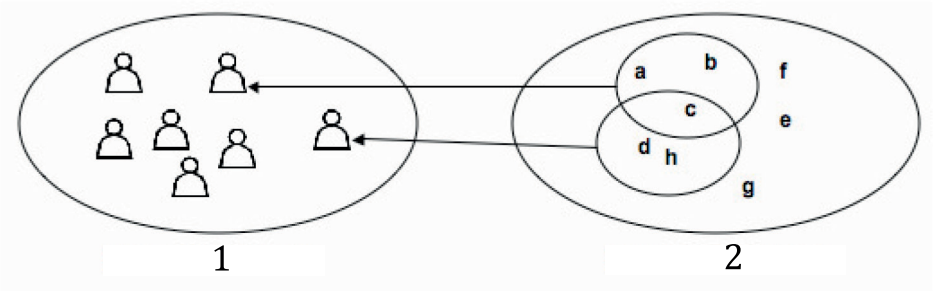
According to Reference [18], “personal data” shall mean any information relating to an identified or identifiable natural person (“data subject”); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors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

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

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

5.4.1.2 Idealized concept of identification and de-identification



Key

1 set of data subjects

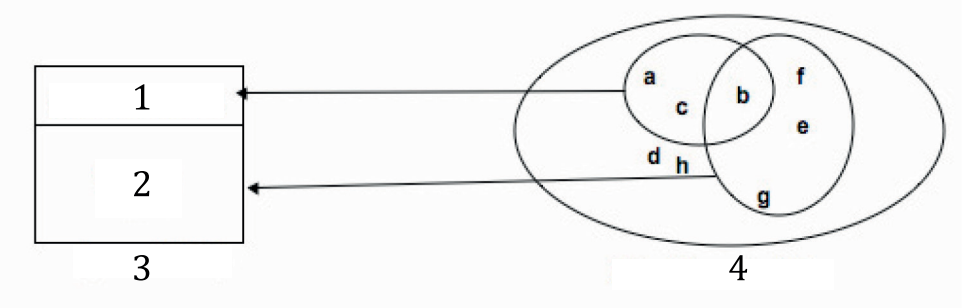
2 set of characteristics

Figure 2 — Identification of data subjects

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

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

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



Key

1 identifying data

2 payload data

3 personal data

4 set of characteristics

Figure 3 — Separation of personal data from payload data

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

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

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

— identifying data: the identifying part that contains a set of characteristics that allow unique identification of the data subject (e.g. demographic data).

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

5.4.2 Concept of pseudonymization

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

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

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

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

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

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

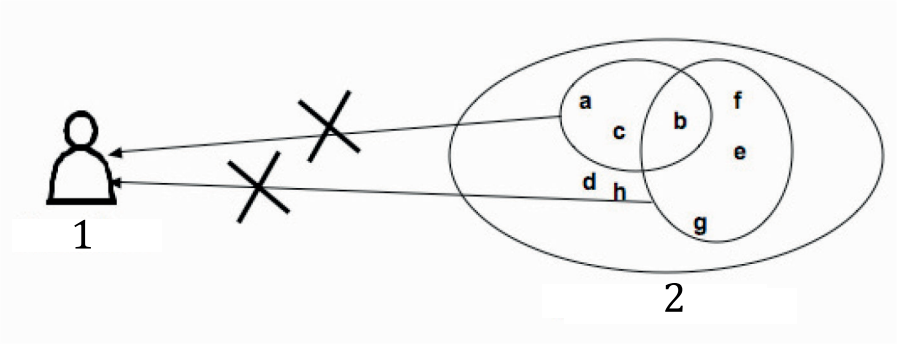
e) what release strategies are available.

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

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

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017



Key

1 data subject

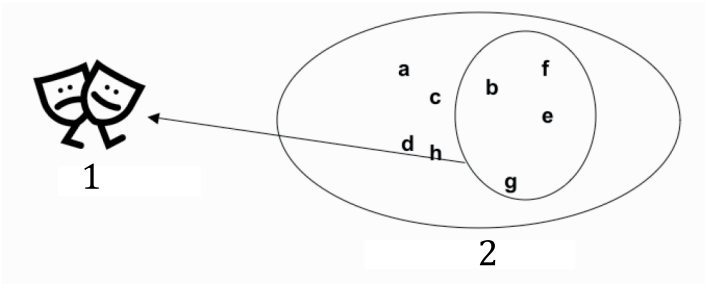
2 set of characteristics

Figure 4 — Anonymization

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

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

— by increasing the population in the data subjects set so that the association between the data set and the data subject is not unique anymore and no direct relation to an individual.



Key

1 pseudonym(s)

2 set of characteristics

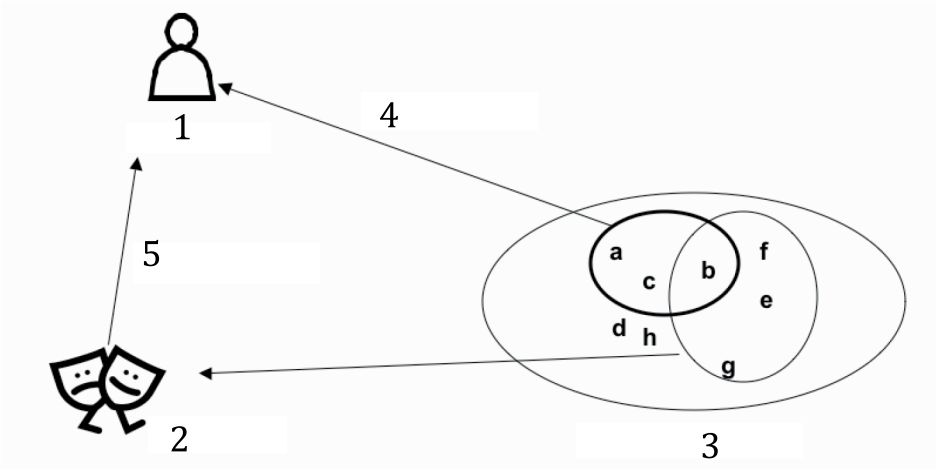
Figure 5 — Pseudonymization

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

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

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

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017



Key

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

Figure 6 — Reversible pseudonymization

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

There are two methods to achieve this goal:

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

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

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

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

5.5 Real world pseudonymization

5.5.1 Rationale

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

There are many regulations in many jurisdictions that require creation of de-identified data for various purposes. There are also regulations that require protection of private information without specifying the mechanisms to be used. These regulations generally use effort and difficulty related phrases,

© ISO 2017 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS EN ISO 25237:2017

which is appropriate given the rapidly changing degree of difficulty associated with de-identification technologies.

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

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

5.5.2 Levels of assurance of privacy protection

5.5.2.1 General

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

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

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

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

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

— level 1: the risks associated with the person identifying data elements;

— level 2: the risks associated with aggregating data variables;

— level 3: the risks associated with outliers in the populated database.

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

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

© ISO 2017 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017

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

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

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

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

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

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

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

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

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

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

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

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

© ISO 2017 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS EN ISO 25237:2017