Collaborative Patient Centred eHealth
E. De Clercq et al. (Eds.)
IOS Press, 2008
© 2008 The authors and IOS Press. All rights reserved.
doi:10.3233/978-1-58603-922-6-111

Privacy Protection through Pseudonymisation in eHealth

F. DE MEYER [a], G. DE MOOR [b], L. REED-FOURQUET [c]
[a] *Dept. Of Medical Informatics & Statistics, University Hospital Ghent, Belgium*
[b] *Dept. Of Medical Informatics & Statistics, Ghent University, Belgium*
[c] *e-HealthSign LLC, Wallingford, USA*

Abstract. The ISO TC215 WG4 pseudonymisation task group has produced in 2008 a first version of a technical specification for the application of pseudonymisation in Healthcare Informatics 0. This paper investigates the principles set out in the technical specification as well as its implications in eHealth. The technical specification starts out with a conceptual model and evolves from a theoretical model to a real life model by adding assumptions on the observability of personal data.

Keywords. Privacy, Privacy Enhancing Technology, Pseudonymisation

1. Introduction

In 2008, the current version of the ISO TS 25237 technical specification was released by the International Standardisation Organisation Health Informatics Technical Committee – security working group (ISO TC 251 WG4). Work started in 2005. The scope of this working group is "defining standards for technical measures to protect and enhance the confidentiality, availability and integrity of health information, and also accountability for users, as well as guidelines for security management in healthcare".

The TS 25237 document has been issued as a Technical Specification (TS) and not as a standard. ISO rules state that "When the subject in question is still under development or where for any other reason there is the future but not immediate possibility of an agreement to publish an International Standard (IS), the technical committee may decide that the publication of a technical specification would be appropriate".

Technical specifications shall be reviewed at least every three years to decide either to confirm the technical specification for a further three years, revise the technical specification, process it further to become an International Standard or withdraw the technical specification. After six years, a technical specification shall be either converted into an International Standard or withdrawn.

ISO member bodies may adopt technical specifications and publish them as documents having the same level of authority as the ISO/TS.

The document is a first in its kind on de-identification through pseudonymisation and constitutes the foundation for potential future standards.

2. The rationale for pseudonymisation

Pseudonymisation provides a means to link information together originating from the same entity across multiple data records or information systems without revealing the identity of the entity. The primary risk mitigated by pseudonymisation is privacy violation.

The term “pseudonymisation” may cause confusion. For readers unfamiliar with privacy protection, the word “pseudo” may invoke a negative bias. Moreover, “pseudonymisation” is only a specific instance of privacy enhancing technology 0. It is a de-identification concept. Nevertheless, for the sake of consistency with the technical specification, this document will continue to use the term “pseudonymisation”. A pseudo-identity is therefore a means to link data together and by definition does not automatically lead to identification of a data subject.

The advantage of using pseudonymised data instead of unrelated records from which all identifying data has been removed is the possibility to group data collected at different moments or coming from different sources. This is particularly useful in research applications that are interested in the aggregation of data or when cases are to be grouped without revealing identities.

Ethical and legal regulations prefer or may even require that research data be collected with the identifying data elements removed. In many applications this is only possible after the various data elements have been aggregated. The aggregation process itself requires a way to link the collected source data elements together. If privacy enhancing technology is not available, the linking is based on identifiable data, which in itself is against privacy protection. Pseudonymisation, when properly designed and implemented, allows the grouping of de-identified data, without revealing identities. It reconciles privacy requirements with flexibility in data linking.

Research is not the only beneficiary of pseudonymisation. Pseudonymisation can also be implemented in those branches of an application chain where there is no requirement to reveal identities. An example is the de-identified testing of body samples in a laboratory and the insertion of the results in the patient record. An example scenario of this is given in the informative annex of the technical specification.

3. History of pseudonymisation

Though the possibilities of de-identification through pseudonymisation are still not fully exploited throughout all application domains in eHealth, the concept is not completely new.

From 1993 onwards various papers have been published on the use of pseudonymisation and privacy enhancing techniques for eHealth applications.

In 1995, Germany regulated the setup of cancer registries. The setup was based upon a trusted pre-processing of identifying data by a separated trusted entity. The technology was based upon the management of encrypted control numbers 0. This has been implemented in the Cancer Registry of Lower Saxony (CARLOS).

In 1996, KITH, the Norwegian centre of medical informatics published a paper called “Socio-technical aspects of the use of health related personal information for management and research” 0. There, a description is given of the technical aspects of secure information management that includes a pseudonymisation model based on double layered third party pseudonymisation. The following application fields were thereby mentioned: epidemiological research, public health research, clinical research and evaluation as well as management, administration and finance. All these applications centred on the collection of epidemiological and research data by governmental institutions.

In those days, however, most security projects in eHealth were more targeted on the introduction and use of electronic signatures and on access control.

It was not until the end of the nineties that dedicated privacy enhancing technology provision became available. The EU has co-financed two projects that enhanced the take-up of privacy enhancing techniques: the PRIDEH (2001-2003) 0 and PRIDEH-GEN (2001-2004) 0 projects. The latter also includes privacy issues of genomic data.

The Healthgrid community also became interested in privacy enhancing technology 0.

Currently, privacy protection services are available on the eHealth market either as dedicated trust service providers or as part of data management applications.

The so called Article 29 Data Protection Working Party has issued a number of documents in which pseudonymisation and other de-identifcation practices are recommended (e.g. opinion 4/2007 on the concept of personal data and the working document on the processing of personal data relating to health in electronic health records).

4. Structure and scope of technical specification ISO TS 25237

4.1. Definitions, terminology and conceptual model

Terminology and definitions as defined in legislation, and more in particular EU directives, are the starting point for the conceptual model. These are key to understand what is meant by identifiability, anonymity, pseudonymisation etc.

Based on the relevant terminology, a conceptual model is derived that further clarifies the relationship between the entities mentioned in the definitions and clauses of the data protection regulations. The concept of identifiability in the specification is wider than normally used in legal documents. The bottom line is that a single data subject can be singled out from amongst its peers, even based on characteristics that in various circumstances can be seen as non-identifying.

The conceptual model includes paths from identifiable data to de-identified data, explained in generic terms, independent of technologies that can be used to achieve this.

4.2. Real world modelling

A compelling reason for drafting a technical specification is to provide guidance for decreasing the gap between theoretical conceptual models and the real world. The assessment of identifiability of data is strongly influenced by perception. From a legal point of view, the delineation between identifiable and anonymous is often seen as sharp because a theoretical definition is dualistic without a grey zone in between.

Yet, this grey zone exists. Neither theory nor practice can completely eliminate this grey zone. Instead of being stuck in an endless “yes-no” debate, the technical specification proposes a way to shift the “boundary of ambiguity”. This reconciliation is achieved by translating the legal clauses found in recital 26 of the Data Protection Directive (DPD) into what is called a “real life model”. Recital 26 states: “to determine whether a person is identifiable, account should be taken of all the means likely reasonably 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" 0.

Statements such as "all the means likely reasonable" and "by any other person" are rather vague. Since the definition of "identifiable" and "anonymous" depends upon the undefined behaviour ("all the means likely reasonable") of undefined actors ("by any other person") the conceptual model should include "reasonable" assumptions about "all the means" likely deployed by "any other person" to link observational data to data subjects. It should try to define as many elements as possible in a policy.

In the specification this translates into "levels of assurance of privacy protection". The classification provided in the specification is a first attempt. It consists of three levels. It is desirable to keep the number of levels to a minimum, while linking a particular level to a set of criteria that can easily be applied to each level.

This approach is an improvement with respect to current common practices that are more or less in line with level 1 assurance. This level consists of clearly identifying data or easily obtainable indirectly identifying data. Examples of level 1 assurance are the rules of thumb as they can for instance be found in the HIPAA rules 0. These include for instance that names, addresses, phone numbers should be deleted from the data.

Assurance level 2 consists of making assumptions about what kind of data that can be gathered by an attacker. This observational data is included into the model. Assurance level 2 requires a risk analysis that includes assumptions about types of attacks and attackers. An example of this is the assumption that an attacker can obtain discharge records from hospitals stating identities and discharge dates. In itself not highly revealing, but in combination with other observations, this can lead to privacy infringements. As a result it is possible to define more clearly what is stated in recital 26.

Levels 1 and 2 risk analysis are "static" analyses that are not modified by the amount of data or actual instances of the data. Level 3 assurance and the associated risk analysis takes into account live repositories. One of the issues encountered in live databases is the presence of outliers or rare data that could lead, in combination with observational data, to identification of data subjects.

4.3. Categories of data subjects

Though the majority of privacy issues are focused on patient privacy, other entities may require protection as well. Patient privacy is the focus because of compulsory privacy regulations. In practice, regulation is not the only motivation for protecting identities. An important group of entities that may require identity protection are health professionals or health care enterprises. This may be a requirement imposed by statistical blinding, but may also be required in e.g. peer ranking research of which the outcome can be considered sensitive for the participating organisations.

Not only persons or legal entities may have their identification protected. This goes as well for systems or even molecules (e.g. in drug discovery projects where pharmaceutical research institutes have a need to share basic research information, without exposing intellectual property details)

4.4. Re-identification

Pseudonymisation can be a one way process, but technically it is possible to reverse the de-identification in controlled circumstances and in a way foreseen during the design of the system. The technical specification describes this in a more detailed way. In fact, re-identification and de-identification can be considered building block of identity management systems.

Privacy protection should address more than just identifiability of an entity. Often privacy is already breached if a sensitive characteristic can be associated with an entity. An example is the assessment whether a data subject belongs to an HIV positive or negative group, without necessarily being able to identify the data subject in the group. An observer would not know which record is yours in the record set, but he would know that your record is in a particular group and therefore is able to tell if you have a particular disease or not.

4.5. The Pseudonymisation process

A considerable part of the content of the technical specification has been devoted to the pseudonymisation process. It does not claim exclusivity nor exhaustiveness, but aims at a realistic and trustworthy way to design, implement and use pseudonymisation in a practical way.

The specification contains an overview of the build-up of entities in the communication model from the data source up to the data target where the pseudonymised data is further handled.

A key concept in the model is that de-identification service and re-identification service provision are to be delivered by a trusted third party (TTP). A trusted third party is not to be confused with a TTP used for issuing cryptographic keys for digital signatures and encryption for confidentiality applications, though both share the concept of a trusted operation based upon policies and delivered by independent specialised providers.

The specification contains an example of a possible workflow and how the data can be prepared. Interoperability issues are also briefly mentioned.

4.6. A policy framework for pseudonymisation services

Pseudonymisation services have in common with other data security services that the operation of the components has to be driven by a policy in order to achieve the desired effects. The privacy policy document is not only important as a reference for the operation of a trust service, but should also explain to all parties involved what can be expected of the operation and what residual issues are left unsolved that should be countered by for instance access control.

Data security measures consist of a set of complimentary technical and organisational measures. Policy documents serve as references for the overall operation of all entities involved in the de-identification process and relevant parts of it should be reflected in the policy of each of these entities.

4.7. Pseudonymisation scenarios

In an informative annex, the specification lists a number of common healthcare pseudonymisation scenarios. These take into account the kind of identification that is being protected, the sensitivity of the data, single or multiple data sources and their relationships, primary and secondary use of data, the type of context and a number of other parameters. Eight scenarios are listed. The scenarios include examples where irreversible de-identification of the data is required for research purposes, and where de-identification is only a temporary episode when the users of the data during that episode have no need to know the identities of the data subjects but where after that episode, controlled reversible de-identification is required. This is for instance the case when biosamples are sent to a testing lab without the lab personnel being allowed to see the associated names of the data subjects. The results are re-identified and automatically entered into the EHR at the end of that episode.

5. Further work to be done

5.1. Re-identification risk analysis

The technical specification rests upon the interpretation of identifiability of data. It is clear that further research is required into the aspects of re-identification. The technical specification refers to re-identification risk analysis, but limits itself to presenting a model that extends a rather reduced view as encountered into the data protection directive towards a more realistic real life model that allows to formally take into account assumptions about data that can be obtained by an attacker. As a result, it shifts the border of ambiguity between identified, identifiable and anonymous data. It is more realistic to talk of "anonymised" data instead of "anonymous data", the latter being a rather theoretical concept, while "anonymised" reflects that all precautions reasonably possible and commensurate to the threats have been made to prevent identification. Further study of risk analysis models can significantly contribute to model for privacy protection.

5.2. Legal uncertainty

The inclusion in the technical specification of an easy to understand real life model that contains the elements to be taken into account to assess the level of assurance of the anonymity of the data is intended to bridge the gap between legal and ICT or data management experts.

Legal experts admit that a degree of legal uncertainty for various aspects of eHealth remains. Privacy is one group of these issues. In order to reduce the legal uncertainty, the European Commission has carried out a project called "Legally eHealth" that included a study on legal and regulatory aspects of eHealth 0. This is without any doubt an initiative that will contribute to the deployment of eHealth. This document does nevertheless contain a statement that is contradictory to the opinion of other legal experts in the field 0. The Legally eHealth document states that though anonymous data is not subject to data protection requirements, the processing carried out to render data anonymous is considered to be a processing of personal data. As a consequence the process of data anonymisation should be covered by data protection requirements as any other type of processing of personal data.

However, stating by default that the process of anonymisation of data is in itself considered a fully fledged form of processing of personal data is contradictory to the rationale for introducing de-identification services.

The technical specification agrees that privacy policies can reflect that even certain types of de-identified data may require additional data security protection (e.g. because of data that could easily be obtained by an attacker if brought outside the context for intended use).

As a minimum, the authors believe that a distinction should be made depending on the role of the de-identification trusted third party in a specific contractual setting with the controller(s) of the data on which behalf it is acting. The role of the trusted party may be limited to de-identifying the data that is sent to it under responsibility of the controller of the data, or the role may be more elaborate and consist of joining data from various controllers. In the latter case, the trusted party should take more elaborate legal precautions. In the first case however, the trusted party is only performing technical services on behalf of the controller of the data and thus should be exempted from the full legal procedure required for collecting and processing of personal data. This issue should be clearly resolved once and for all.

5.3. Interoperability

The objective of the technical specification is to lower the threshold for the use of de-identification and pseudonymisation in eHealth. Various methods and standards however co-exist in eHealth on how to store, process and communicate data. Some of these differences are touched upon in the informative technical annex of the technical specification, but these should be further elaborated as well.

It may be beneficial that subdomains in eHealth (e.g. clinical research) reflect how privacy protection through de-identification services can contribute to relevant business cases, thereby using the technical specification as a starting point and guide.

The domain of research in eHealth, based on patient data is especially interesting. It can unlock new applications in eHealth such as translational medicine and contribute to the cost efficiency of clinical research in general.

The technical specification has made a good start of that but in the coming years, more concrete applications of pseudonymisation should be worked out which will lead to more complete guidelines for the application of pseudonymisation.

6. The future of pseudonymisation specifications

The availability of the technical specification is an important stimulus to the take up of privacy enhancing technologies. Feedback to standardisation organisations will allow evolving the technical specification to a full standard.

Various parts of the specification have been included in the HITSP/C25 0, HITSP/T24 0 and HITSP/TP22 0 documents in the U.S. that give guidance to pseudonymisation and identity management.

The section on further requirements lists a number of issues that can be elaborated in the coming years in order to achieve a transparent and easy to integrate privacy protection. Especially areas that rely on the secondary use of patient data can greatly benefit from further development of de-identification solutions and supportive functions such as re-identification risk analysis.

By presenting models and policies that can be understood and agreed on by both the legal experts and the ICT experts, effective privacy protection can be further enhanced, but continuing efforts are needed to reach a common understanding of how to apply the basic principles to the various sub-domains in eHealth.

References

[1] ISO/IEC TC215/SC /WG4 ISO TS 25237 Health Informatics – Pseudonymisation

[2] DE MOOR GJE, CLAERHOUT B. Privacy Enhancing Techniques in eHealth: an Overview. Stud Health Technol Inform. 2004; vol 106;75-81.

[3] W. THOBEN, H.-J. Appelrath, Verschlusselung personenbezogener und Abgleich anonymisierter Daten durch Kontrollnummern, Verlässliche IT-Systeme, Rostock, 1995, p 193-206.

[4] KENNETH R. IVERSEN, Tor Olav Grotan, Socio-technical aspects of the use of health related personal information for management and research, Proceedings of IMIA Working Group 4 Working Conference, October 1996, Pages 83-91.

[5] DE MEYER F, CLAERHOUT B, DE MOOR G. The PRIDEH project: taking up Privacy Protection Services in e-Health. Proceedings MIC 2002 "Health Continuum and Data Echange". IOS Press, 2002:171-177.

[6] DE MOOR GJE, CLAERHOUT B, DE MEYER F. Privacy Enhancing Techniques: the Key to Secure Communication and Management of Clinical and Genomic Data, Methods Inf Med. 2003; 42 (2):148-153.

[7] DE MOOR GJE, CLAERHOUT B. Privacy Protection for Healthgrid Applications. Methods Inf Med. 2005; 44 (2):140-3.

[8] 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.

[9] HIPAA Privacy Rule and Public Health Guidance from CDC and the U.S. Department of Health and Human Services.

[10] IAN WALDEN. Anonymising Personal Data, Int J Law Info Tech 2002 10: 224-237.

[11] EUROPEAN HEALTH MANAGEMENT ASSOCIATION. Legally eHealth. Study on Legal and Regulatory Aspects of eHealth. Deliverable 2, Processing Medical Data: Data Protection, Confidentiality and Security. 2006; European Commission Contract 30-CE-0041734/00-55.

[12] Population Health Technical Committee. HITSP/C25 Anonymize Component. 2007.

[13] Population Health Technical Committee. HITSP/T24 Pseudonymize Transaction. 2007.

[14] Population Health Technical Committee. HITSP/TP22 Patient ID Cross-Referencing Transaction Package. 2007