eHealth2014 – Health Informatics Meets eHealth
A. Hörbst et al. (Eds.)

doi:10.3233/978-1-61499-397-1-211

Patient Identity Management for Secondary Use of Biomedical Research Data in a Distributed Computing Environment

Michael NITZLNADER[a,1] and Günter SCHREIER[a]
[a]*AIT Austrian Institute of Technology, Austria*

Abstract. Dealing with data from different source domains is of increasing importance in today's large scale biomedical research endeavours. Within the European Network for Cancer research in Children and Adolescents (ENCCA) a solution to share such data for secondary use will be established. In this paper the solution arising from the aims of the ENCCA project and regulatory requirements concerning data protection and privacy is presented. Since the details of secondary biomedical dataset utilisation are often not known in advance, data protection regulations are met with an identity management concept that facilitates context-specific pseudonymisation and a way of data aggregation using a hidden reference table later on. Phonetic hashing is proposed to prevent duplicated patient registration and re-identification of patients is possible via a trusted third party only. Finally, the solution architecture allows for implementation in a distributed computing environment, including cloud-based elements.

Keywords. data protection, pseudonymisation, biomedical research, secondary use, cloud computing.

1. Introduction

Today, biomedical research relies more and more on linking research data from different source domains such as clinical trials, bio banks etc. This allows conducting larger scale investigations and research projects aiming at hypotheses that span multiple disease aspects. Aggregating data from the basic science level (e.g. genomics), the clinical level (e.g. medical images) and from the healthcare level (e.g. information on long-term survival) is crucial.

In the context of the ENCCA project (**E**uropean **N**etwork for **C**ancer Research in **C**hildren and **A**dolescents) [1] a so called **Virtual Institute (VI)** shall be established to facilitate the exchange of data between source (*Source Domain*) and consumer (*Consumer Domain*) systems in a distributed computing environment. Figure 1 shows the schema of the planned VI. The VI is responsible for storing the data, facilitating data aggregation as needed on the consumer side and for providing the aggregated data to the consumers. Accordingly, the VI itself is called *Integration Domain*. Besides numerous challenges concerning technical and organisational issues it is a central non-functional requirement of the VI to consider legal and data protection issues. This

[1] Corresponding Author: michael.nitzlnader@ait.ac.at

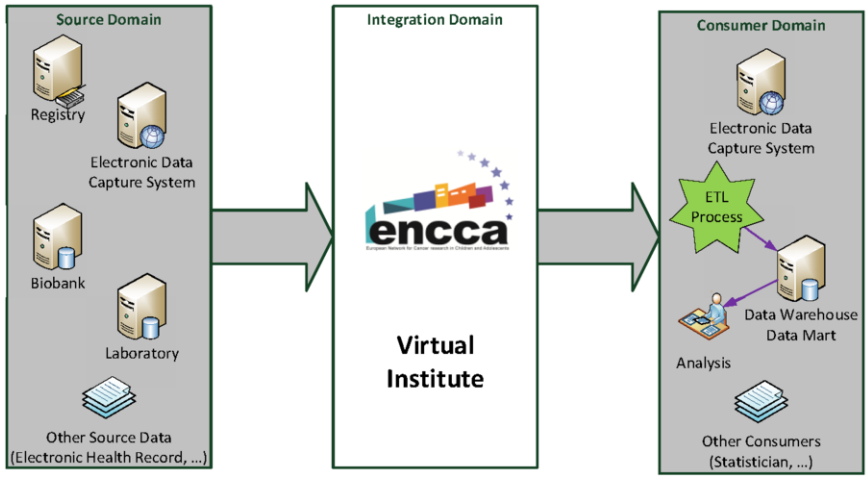

Figure 1. Schema of the ENCCA IT infrastructure for secondary use of biomedical research data.

paper deals with the demands and necessary steps to fulfil these requirements and proposes an appropriate concept for the ENCCA Patient Identity Management (PIM).

2. Methods

As a first step the requirements for PIM in the context of the ENCCA project were collected. These requirements arise from two different sources:

a) from the explicit aims of the ENCCA project itself and
b) from the implicit regulatory framework in which the ENCCA is embedded.

Both perspectives were considered and dealt with in the requirements analysis phase. Special emphasis was put to those requirements related to PIM.

3. Results

3.1. Requirements for patient identity management

3.1.1. Requirements arising from the aims of the ENCCA project

As shown in Figure 1 data from different data sources - in some cases related to the same patient - should be made available to the consumer domain in an integrated manner for joint analysis. Assuming the existence of an informed patient consent, these data will be registered for the purpose of secondary use. The registration of data will be done for a patient defined in the virtual institute and, subsequently, will allow queries on that data and to provide the resulting datasets, typically from more than one source system.

The creation of a new patient in the virtual institute is defined as the process of registering the patient under a new identifier (pseudonym). In addition, it has to be ensured that one and the same patient is not registered twice. Furthermore, upon their

request, it should be possible to inform patients about relevant research results. Therefore, the possibility of re-identification is needed [2]. Last but not least – according to the overall concept of the VI – the PIM concept has to be feasible in a distributed computing environment. For instance, it should be possible to operate some components in the cloud so as to take advantage of cloud computing features like scalability, accessibility, flexibility, reduced costs etc. and to integrate with other cloud-based services that support biomedical research.

3.1.2. Requirements arising from the regulatory perspective

Beside procedures required to comply with general IT security practices (e.g. access control, audit, encrypted transmission, etc.) especially the protection of sensitive data related to the health of patients (so-called medical data and hereinafter referred to as "MDAT") needs to be ensured. According to [3, 4] it makes a difference whether MDAT are processed inside the source context (intra-institutional) or whether they are transferred to a different context (cross-institutional). If MDAT are to be transferred to another system, several measures need to be taken to protect these data. The re-usage of an existing identifier from source systems should be avoided so as not to create universal identifiers across different systems. If the context of patient data processing changes, a new identifier / pseudonym should be used [2].

3.1.3. Overview on the requirements for the patient identity management concept

Table 1 summarises the requirements for the PIM as derived from both aspects above.

3.2. Patient Identity Management Concept

Based on the core requirements (Table 1) a concept for PIM as illustrated in Figure 2 was developed according to related work already done by other authors [5, 6, 7]. The proposed concepts were enhanced by a context-specific pseudonymisation to fulfil the requirements R3 and R4 and by phonetic hashing to fulfil requirement R1. Requirement R2 was achieved by incorporating asymmetric encryption and a Trusted Third Party (TTP) similar to [5]. Requirement R5 was fulfilled by the general architecture of the PIM concept as shown in Figure 2.

Table 1. Requirements for the patient identity management concept

R1	PIM should prevent a duplicate registration of one and the same patient.
R2	Preserve the possibility to re-identify subjects by a trusted third party in special cases, e.g. to inform a patient about relevant research outcomes on-demand.
R3	Different pseudonyms should be used also for one and the same patient for different contexts, e.g. different data sources. However, the PIM should provide a method to link the different pseudonyms in the background to allow secondary use of the data stored within the VI. In other words the PIM concept should provide the ability for the VI to link datasets that belong to the same patient but are stored under different pseudonyms of that patient. At the same time, an identification of patients should be rendered (almost) impossible for another context, e.g. for researchers from institutions that do not know the patient from a treatment context.
R4	Avoid creating a transparent universal patient ID that would impose re-identification threats by the potential availability of an increasing number of linked datasets.
R5	The PIM concept has to be feasible in a distributed computing environment.

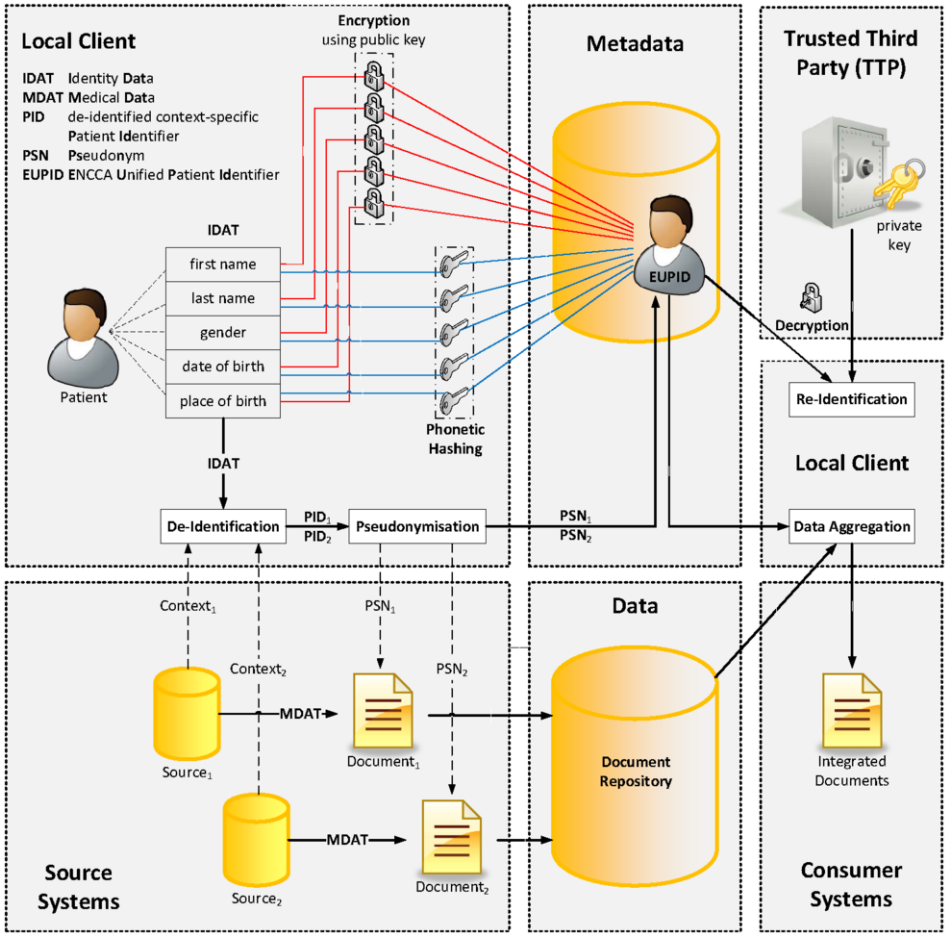

Figure 2. Overview on the patient identity management concept for secondary use of biomedical research data in a distributed computing environment.

Figure 2 illustrates the whole PIM concept. In a first step an internal context-specific patient ID (PID) is generated from the patient identification properties (IDAT) and a defined context (**De-Identification**). This PID is then transformed into a context-specific pseudonym (PSN) by using a one-way hashing algorithm (**Pseudonymisation**) [8, 9, 10].

Each context-specific PSN belonging to the same patient is assigned to a virtual and inaccessible PID called **ENCCA Unified Patient Identifier (EUPID)**. To prevent the generation of multiple EUPIDs for one and the same patient, e.g. in case of slightly different IDAT representations caused by spelling errors, the IDAT are transformed to unique values by using a phonetic hashing algorithm [11] field by field (**Phonetic Hashing**) and consequently assigned to the corresponding EUPID. This allows a reliable assessment whether a virtual patient had already been created previously. In order to enable a trusted third party (TTP) to perform on-demand re-identification of

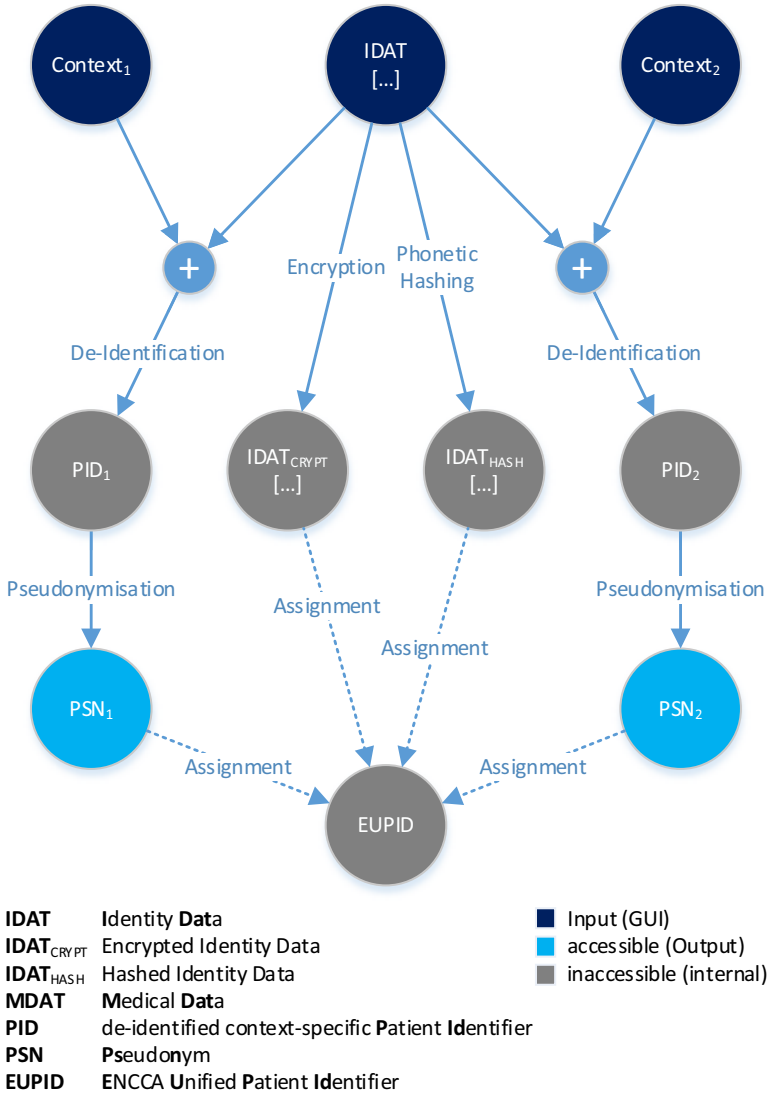

Figure 3. Relationships and pathways of the different identification items.

pseudonymised patients, the IDAT are additionally encrypted using an asymmetric cryptographic technique [12, 13] (**Encryption using public key**). This may be necessary in case a research finding comes up that the patient would like to be informed of. Only the TTP holds the private key necessary to perform the decryption of the IDAT (**Re-Identification**).

Subsequently to the patient registration the provided MDAT are assigned to the created PSNs and then stored in a database that only holds the MDAT (**Document Repository**). The different PSNs, $IDAT_{CRYPT}$ and $IDAT_{HASH}$ values for a EUPID are

stored in a dedicated metadata database (**Metadata**) that is completely separated from all other components. De-identification, pseudonymisation, encryption of the IDAT, hashing of the IDAT, re-identification as well as the aggregation of the data for secondary use are all done via a client application by authenticated and authorised users in their local environment (**Local Client**). An aggregation of MDAT from the document repository is initiated by a query carried out by the user (**Data Aggregation**). The desired data is linked and collated internally by the VI utilising the cross reference list for pseudonyms stored in the metadata database. Figure 3 illustrates the relationships and pathways of the different identification items.

3.3. Validation of the concept

Table 2 illustrates that all requirements can be fulfilled by the developed concept.

4. Discussion

The dataset that will be registered to the VI will be provided for secondary use. Hence, in many cases there will be no predefined purpose up-front on how these data will eventually be aggregated and used. Thus, a concept for PIM including a context-based pseudonymisation of the data had to be designed, which allows the aggregation of the data according to later needs. The approach described in this paper fulfils this requirement by storing the data using pseudonyms (PSNs), which also include a pseudonymisation context, and by providing a mechanism to make the desired and integrated data available.

It should be mentioned that a pseudonymisation is not defined by the Data Protection Directive 95/46/EC of the European Union[1] [14], which is the general basis for processing personal data in Europe. Although the utilisation of a pseudonymisation slightly conflicts with this directive, it had to be preferred instead of an anonymisation in accordance with [15] to allow the claimed re-identification of a registered patient.

An important property of a pseudonymisation context is that it defines a common regulatory environment. In that sense, different research endeavours require the definition of different contexts. For epidemiological purposes each participating institution could be considered a distinct pseudonymisation context. A more common case, however, would be that a group of institutions collaborate in a joint research project with a common patient consent. Here, the joint research project presents the context and all involved institutions would share the same PSN for a given subject. An application for a similar setting could be a situation, where two or more institutions are involved in the treatment of patients, each of which collecting additional research data that - at some point in the future - should be made available for a joint analysis.

It is important that the MDAT registered in the VI are stored in a different location than the metadata describing the correlation between the different PSNs. This reduces the hazards of unintended re-identification and increases data security.

[1] 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.

Table 2. Fulfillment of the requirements for the patient identity management concept

	Fulfilled by …
R1	… phonetic hashing of the IDAT.
R2	… asymmetric encryption and a TTP.
R3	… context-specific pseudonymisation.
R4	
R5	… the general architecture of the PIM concept.

The introduced concept supports a distributed computing environment and is suitable for approaches that would like to put some of the components to the cloud and, therefore, allows to take advantage of the benefits of cloud computing.

Utilising phonetic hashing allows minimising cases of duplicate patient registration, knowing that such cases cannot be avoided completely. The asymmetric encryption of the IDAT permits decryption only through a TTP. Thus, through a clear separation of the pseudonymisation and the re-identification processes [5] a secure possibility to contact a patient exists, which fully complies with data protection regulations. Since this concept increases complexity and will be associated with additional operational costs, one could think of utilising the context also to automatically decide whether or not to really add this encrypted information to the EUPID of given patients.

To further prevent any data aggregation that consequently would allow re-identification of patients based on aggregated datasets, exported data will be pseudonymised once more by creating new identifiers during data aggregation. In highly sensitive cases, further measures like k-anonymity and l-diversity concepts could be implemented in the data export process [16].

Recently, the present concept had been presented to the ENCCA board and cleared for implementation. Practical experience will be obtained during a demonstration phase and will reveal whether the patient identity management for secondary use of biomedical research data also fulfils the non-functional requirement of practicability.

Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7 2007-2013) under the project ENCCA grant agreement no. 261743.

References

[1] ENCCA - European Network for Cancer Research in Children and Adolescents, http://www.encca.eu, last access: 2013-11-05.

[2] Orientierungshilfe: Pseudonymisierung in der medizinischen Forschung: Der Bayerische Landesbeauftragte für den Datenschutz, http://www.datenschutz-bayern.de/technik/orient/ohilfe_psn_03.html, last access: 2013-11-10.

[3] W. Gall, W. Grossmann, G. Duftschmid, T. Wrba, W. Dorda. Analysis of EHRs for research, quality management and health politics. *Ehealth Beyond the Horizon - Get It There*. 2008;**136**:425-30. PubMed PMID: WOS:000274308700070.

[4] Metadaten für medizinische Dokumente: HL7 Deutschland; 2012, http://wiki.hl7.de/index.php/Metadaten_f%C3%BCr_medizinische_Dokumente, last access: 2013-12-15.
[5] H. Aamot, C. Kohl, D. Richter, P. Knaup-Gregori. Pseudonymization of patient identifiers for translational research. *BMC Medical Informatics and Decision Making*. 2013;**13**(1):75. PubMed PMID: doi:10.1186/1472-6947-13-75.
[6] R. Becker, P. Ihle, K. Pommerening, U. Harnischmacher. Ein generisches Datenschutzkonzept für Biomaterialbanken. Telematikplattform für Medizinische Forschungsnetze e.V., April 2006. Report No.
[7] K. Pommerening, M. Reng, P. Debold, S. Semler. Pseudonymization in medical research - the generic data protection concept of the TMF. *GMS Medizinische Informatik, Biometrie und Epidemiologie*. 2005;**1**:17.
[8] B.S. Elger, J. Iavindrasana, L. Lo Iacono, H. Müller, N. Roduit, P. Summers, et al. Strategies for health data exchange for secondary, cross-institutional clinical research. *Computer Methods and Programs in Biomedicine*. 2010 9//;**99**(3):230-51.
[9] Technische Arbeitsgruppe TR-03116-1. Technische Richtlinie BSI TR-03116-1, eCard-Projekte der Bundesregierung, Teil 1: Telematikinfrastruktur. 3.17 ed: Bundesamt für Sicherheit in der Informationstechnik; 2013. p. 39.
[10] National Institute of Standards and Technology. The Keyed-Hash Message Authentication Code (HMAC). 2008. p. 13.
[11] A. Karakasidis, V.S. Verykios. Privacy Preserving Record Linkage Using Phonetic Codes. *Proceedings of the 2009 Fourth Balkan Conference in Informatics*. 2009:101-6. PubMed PMID: WOS:000290888000017.
[12] A. Beutelspacher, J. Schwenk, K.D. Wolfenstetter. *Moderne Verfahren der Kryptographie: Von RSA zu Zero-Knowledge*: Springer; 2010.
[13] R.L. Rivest, A. Shamir, L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM*. 1978;**21**(2):120-6.
[14] 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, (1995).
[15] K. Holzer, W. Gall. Utilizing IHE-based Electronic Health Record Systems for Secondary Use. *Method Inform Med*. 2011;**50**(4):319-25.
[16] K. Stark, J. Eder, K. Zatloukal. Priority-based k-anonymity accomplished by weighted generalisation structures. *Data Warehousing and Knowledge Discovery, Proceedings*. 2006;**4081**:394-404. PubMed PMID: WOS:000241158200038.