Pseudonymization Service and Data Custodians in Medical Research Networks and Biobanks

Klaus Pommerening[1], Markus Schröder[2], Denis Petrov[2], Marc Schlösser-Faßbender[2], Sebastian C. Semler[3], Johannes Drepper[3]

[1]Institut für Medizinische Biometrie, Epidemiologie und Informatik
Johannes-Gutenberg-Universität
D-55101 Mainz
pommerening@imbei.uni-mainz.de

[2]Tembit Software GmbH
Am Borsigturm 42
13507 Berlin
schroeder@tembit.de
petrov@tembit.de
msf@tembit.de

[3]Telematikplattform für Medizinische Forschungsnetze e.V. (TMF)
Geschäftsstelle
Neustädtische Kirchstr. 6
10117 Berlin
sebastian.semler@tmf-ev.de
johannes.drepper@tmf-ev.de

Abstract: Medical research networks collect data consisting in large data pools, and collect samples in biobanks, and want to keep them for future research projects. Telematik-Plattform für medizinische Forschungsnetze (TMF) has developed data protection concepts for research networks and biobanks that use pseudonymization as an essential tool. Various networks subsequently adapted their system and thus gained practical experiences in doing so. TMF additionally offers tools for the data custodian services of "Identity Management" and "Pseudonymization"; both services that can be invoked as web services or can be integrated into a given communication software within a network.

1 Background

Medical research requires data and biomaterial. Therefore, medical research networks collect data in large data bases or registries as well as samples in biobanks. As building data bases and biobanks of high quality is laborious and expensive, it is essential to use data and material long-term and store these for future research projects that may not be foreseen at the time of acquisition.

Anonymization makes long-term storage possible; therefore, researchers should use anonymous data and material whenever possible. However, in many cases of research the correct association between a single patient's data from distinct sources or distinct points of time is crucial. Some scenarios even require a way back to the identity; it could be important for the patient, and be in his interest, to learn about results of a research project, for example a genetic disposition. Pseudonyms are the solution of these problems [Po96].

The situation with biomaterial is peculiar. Samples of biomaterial contain complete biochemical and molecular genetic information about their respective donors. Although we assume that material without accompanying personal data will continue to remain anonymous during the next few years, a long-term concept cannot rely on this assumption; the recommendation is to make use of pseudonymization.

Data and material is acquired in a treatment context or directly for a research project. Particularly in medical research networks, the patients often participate in clinical multicenter studies; their data is transferred to a central study data base (SDB). Here, they are stored until the end of the study or are transferred to a research database (RDB). Figure 1 illustrates the various data sources.

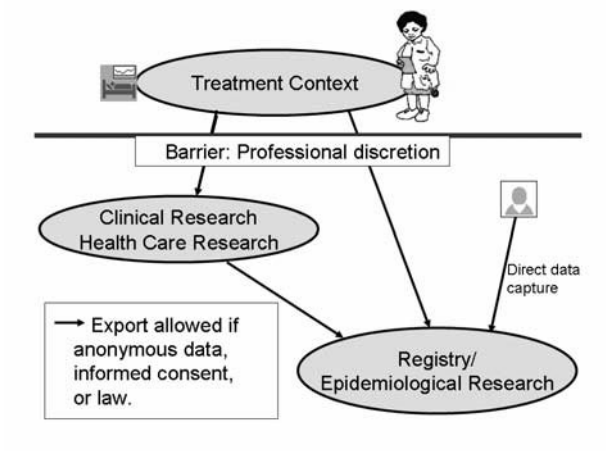

Figure 1: Data sources for medical research

Several legal conditions increase restrictions for the transfer of data and material for research purposes or for long-term storage: data protection regulations, professional discretion – if the sources of data or sample are within a treatment context, for biobanks, one has to pay attention to proprietary rights and to individuals' rights on parts of their body; while property may be given up, personality rights cannot.

The use of data and material for research presupposes the informed consent by the patients. But also with consent data and material must be transferred and stored only for defined purpose, restricted time frame and explicitly listed users, all mentioned in the patients' information [EU95].

Extending these restrictions is possible in certain circumstances, but only by applying additional safeguards and conditions in a rigid organizational framework, where the risk of re-identification is strictly observed.

2 Concepts

The TMF approach overcomes the barriers for keeping data indefinitely by additional safeguards; this is possible because research is privileged by the constitution. The additional safeguards are

- establishing the medical research network as a legal instance with clear accountability,
- offering state-of-the-art information and communication security, including Public Key Infrastructure (PKI) techniques and access control,
- dividing informational powers by designating information and procedures to several independent parties; in particular, the establishment of Trusted Third Parties (TTPs) and separate storage of data, medical images, biomaterial, and corresponding analysis results,
- using pseudonymization,

We distribute information and tasks over the network and establish independent TTPs serving the following purpose:

- legal and organizational: implement data custodians as persons or organizations with legal accountability,
- technical: implement data custodians as TTP *services* that are part of the network architecture.

Necessary tools for the processing of information in these networks are unique patient identifiers (PID), record linkage, quality management, and pseudonyms. Figure 2 shows the basic principle. We keep the identity management (with PID service and record linkage capabilities) separate from the pseudonymization service. Thus, we require two TTP services (and additional TTPs for other tasks, such as quality management):

- the Identity Management service that has a reference list, performs record linkage, and assigns unique PIDs,
- the Pseudonymization service, that encrypts a PID to a pseudonym PSN.

To perform this task, the PSN service uses a secret key. Different applications or information stores use different keys, namely different (un-linkable) pseudonyms.

The Identity Management service only can view the basic identity information (IDAT) and the PID and retains the link to the data source. The Pseudonymization service only can view the PID and the PSN – the medical data (MDAT) are encrypted with the public key of the research database.

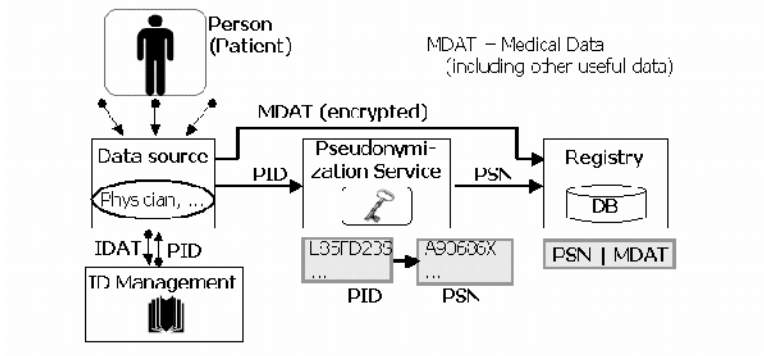

Figure 2: The basic pseudonymization setting with two different Trusted Third Parties, "Identity Management" and "Pseudonymization Service".

De-pseudonymization facilitates the feedback of a research result to the patient. De-pseudonymization presupposes the consent of a specifically authorized committee of the network. It involves the Pseudonymization service that reconstructs the PID from the PSN as well as the Identity Management service who possesses the corresponding identity information (IDAT) and knows the data source, namely the contact person for the patient.

All these procedures must be published in the statutes and standard operating procedures (SOPs) of the medical research network, allowing the patients' consent to be referred to and consequently can be kept perspicuous.

3 Software

TMF offers tools for the two TTP services of "Identity Management" and "Pseudonymization" that can be invoked as web services or can be integrated into the communication software of an existing network: the PID generator and the PSN service.

The PID generator was developed by the working group on Medical Informatics at the University of Mainz. It essentially has two parts: a record linkage module and a reference database. The PID generator is in production use for the Competence Network for Pediatric Oncology and Hematology since 2002 and – in the meantime – has a database comprising approximately 47000 patients. Other networks are using pilot installations and plan to go into production in the near future [GHP06].

TMF e.V. has authorized the reimplementation of the pseudonymization service in order to achieve better integration into existing heterogeneous IT-infrastructures, among other things. Both the initial implementation [Se04] and the recently authorized reimplementation represent the technical transformation of the concept. The reimplemented pseudonymization service now mainly consists of four components illustrated below as follows.

3.1 Services

PSNService

The PSNService transforms a pseudonymized patient number (PID) from a study data base (SDB) into a pseudonym (PSN) with the purpose of long-term data storage in a research database (RDB). This service is executed on an independent computer. Communication is provided by means of safe https-protocols based on mutual authentification. Pseudonymization itself is performed by means of a symmetric, cryptographic algorithm of high security (AES). The key is stored on a SmartCard made specifically for this purpose and is safeguarded against readout. On this SmartCard, the PID number is transformed into the PSN number. The same procedure, also for the refeeding of data, is applied – only that the PSN number is de-pseudonymized by using the SmartCard and transforming the latter into the respective PID number

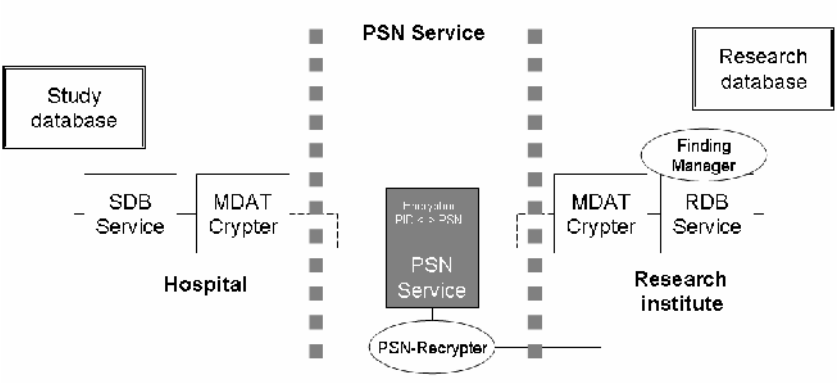

Figure 3: Data Encryption with PSN Service

SDBService and RDBService

The SDBService is file-based from the study database and/or is invoked based on web services that monitor those actions performed. The SDBService first monitors whether the received data are valid and complete, then whether the requested action is possible and useful as well as whether the user is indeed eligible to perform the requested action. Finally, SDBService monitors whether the PID supplied is valid before transmitting the received data to the PSNService. The RDBService is invoked by the PSNService and acts inversely to the SDBService.

Crypter

The Crypter reads out medical data (MDAT) without PID and encrypts these. The data are signed by the CrypterService before they are transmitted via the PSNService to the RDBService.

3.2 Components

The four services detailed above entirely represent the true purpose of the pseudonymization service. However, certain cases requiring additional modules should be taken into account and illustrated as such for the service of pseudonymization.

- **FindingManager:** The purpose is to supply the attending physician of a patient with data pertaining to important diagnostic findings that his or her patient should or would like to be informed of.
- **PSNRecrypter:** In the case of loss or theft of the SmartCard, all PSN numbers are transformed back into PID numbers by means of the PSNRecrypter and a copy of the original SmartCard; subsequently, they are encrypted with a new SmartCard.

3.3 Communication

Communication between the respective services is invoked through web services, and supported by XML RPC via https. Mutual authentification takes place when the https connection is made; both the SERVER as well as the CLIENT must show proper certification (CLIENT-CERT).

4 Results

The TMF working group on Data Protection published a "generic" data protection concept [Po05; Re06] with two variants – A and B – the description given in this paper mainly refers to variant B. Based on variant B, the TMF project group on Biobanks created a data protection concept for biobanks [Po06]. Both of these concepts received positive feedback by the German Data Protection Commissioners as a reliable base for individual concepts to be applied to single networks.

From a technical point of view, the TTP services are ready for use. From the legal and organizational point of view, there has been an ongoing debate regarding who should be the designated TTP, the data custodian. Proposals range from a notary – who is highly protected by law, even against confiscation by the public prosecutor – to the data center of a university hospital that, however, should at least be independent of other parties involved within the network.

5 Lessons Learned

Nearly 20 networks have already adapted the TMF data protection concept; the implementations are more or less advanced. In order to learn from the experiences in implementing the TMF data protection concept, TMF conducted a workshop where both achievements and problems were discussed. TMF will revise its concept in order to meet the following requirements:

- The concept architecture should be modular and scalable; in particular, criteria to be able to assess the appropriateness of measures and safeguards in different situations are called for. The concept for biobanks already fulfills this requirement.
- Health care and medical research is experiencing increasing integration. Medical research networks concepts should reflect this trend.
- The processes of quality management – particularly the highly formalized process for clinical studies – will be better adjusted to the data protection concept.
- Greater support for single networks by means of centrally organized TTP services will be given.

A revision of the generic TMF data protection concept according to the above-mentioned requirements is already underway and is expected to be completed by the end of 2006.

Acknowledgement

This work was supported by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) as a project of TMF (Telematikplattform für Medizinische Forschungsnetze e. V.). Prof. Klaus Pommerening participated as representative of the Kompetenznetz für die Pädiatrische Onkologie und Hämatologie (competence network for pediatric oncology and hematology).

References

[EU95] The European Parliament and the Council: Directive 95/46/EC. Online under http://www.cdt.org/privacy/eudirective/EU_Directive_.html

[GHP06] Glock, J.; Herold, R.; Pommerening, K.: Personal identifiers in medical research networks: Evaluation of the personal identifier generator in the Competence Network Paediatric Oncology and Hematology. GMD Med Inform Biom Epidemiol. 2/2 (2006)

[Po05] Pommerening K. et al.: Pseudonymization in Medical Research - The Generic Data Protection Concept of the TMF. GMS Med Inform BiomEpidemiol 2005; 1(3): Doc17.

[Po06] Pommerening, K. et al.: Datenschutz in Biomaterialbanken. In: Steyer G, Tolxdorff T, [Hrsg.]. TELEMED 2006: Gesundheitsversorgung im Netz. Tagungsband zur 11. Fortbildungsveranstaltung und Arbeitstagung - Nationales Forum zur Telematik für die Gesundheit. Berlin: Aka GmbH; 2006: 89-99.

[Po96] Pommerening, K. et al.: Pseudonyms for cancer registry. Methods of Information in Medicine 35 (1996), 112-121.

[Re06] Reng, C.M. et al.: Generische Lösungen der TMF zum Datenschutz für die Forschungsnetze der Medizin. Medizinisch Wissenschaftliche Verlagsgesellschaft, München 2006.

[Se04] Semler, S.C. et al.: Pseudonymisierung für Forschungsdatenbanken und Register - TMF Pseudonymisierungsdienst für Medizinische Forschungsnetze. In: Jäckel, A. [Hrsg.]. Telemedizinführer Deutschland - Ausgabe 2005. Ober-Mörlen: Medizin-Forum, 2004s: 209-214.

[TMF06] TMF e. V.: Ansprechpartner für Fragen der vernetzten medizinischen Forschung. Online under http://www.tmf-ev.de/.