Secondary Use of the EHR via Pseudonymisation

Klaus POMMERENING
Institut für Medizinische Biometrie, Epidemiologie und Informatik
Johannes-Gutenberg-Universität
D-55101 Mainz, Germany

Michael RENG
Klinik und Poliklinik für Innere Medizin I
Klinikum der Universität
D-93042 Regensburg, Germany

Abstract. The Electronic Health Record (EHR) has many secondary uses, such as health economy and health care research, or disease specific clinical or epidemiological research. For these uses in general the patient identity is not needed, therefore the data must be anonymised or pseudonymised. Whereas for one-time use of the data this procedure is straightforward, long-term data accumulation or the necessity of re-identification require a more sophisticated approach. This paper describes possible model architectures, developed for medical research networks, but useful in other contexts as well.

Introduction

The Electronic Health Record (EHR) is primarily used in the treatment context; here the identity data of the patient are needed and their processing is allowed. But the EHR also serves as a basis for secondary uses such as

- disease specific clinical or epidemiological research projects,
- health care research, assessment of treatment quality, health economy.

Typical aspects of these secondary uses are that

- the data leave the context of the physician where they are protected by professional discretion,
- the identity of the patient doesn't matter.

In such a context use of the data is allowed after anonymisation; therefore anonymisation should be performed whenever possible. But this is not always possible: In many cases of secondary use the correct association between a single patient's data from distinct sources or distinct points of time is essential. In some scenarios even a way back to the identity is required; 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; or a researcher might want to use a data pool to recruit suitable patients for a new clinical or epidemiological study.

Pseudonyms are the solution of these problems [1]. They represent the “golden mean” between perfect anonymity and exposing the identity data. Depending on the requirements one of two kinds of pseudonyms can be used:

- one-way pseudonyms, that cannot be reversed but allow record linkage,
- reversible pseudonyms, that allow the re-identification of the individual.

The use of reversible pseudonyms requires that the re-identification depends on a *secret key* and the pseudonymisation process is set up as a Trusted Third Party (TTP) service; see below. Moreover the use of a system that makes re-identification possible is allowed only after an *explicit informed consent* by the patient.

The technique of pseudonyms in information processing is not new, however rarely used as yet. Early examples are the untraceable electronic money (Chaum 1982 [2]) whose implementation by several banks was withdrawn, the electronic prescription (Struif ca 1990 [3]) and the pseudonymous settling of bills in health care (Pommerening, Bleumer, Schunter 1995 [4]) that were never implemented, and the Michaelis-Pommerening model [5] of cancer registry that is in actual use in several German states. Several recent German laws require pseudonymisation in appropriate contexts.

In her review of the first medical “Competence Networks” in Germany, the data protection commissioner of Nordrhein-Westfalen stated the following requirements (among others) [6]:

- Central data pools must only contain anonymous or at least pseudonymous data.
- A trusted third party (“Datentreuhänder”) that is protected by law (e. g. a notary) should carry out the pseudonymisation.
- The use of unique patient identifiers across distinct networks is not allowed.

The TMF – the Telematics Platform for the Medical Research Networks of the Federal Ministry of Education and Research – therefore started a project to develop and implement generic models for pseudonymisation that can be used in research networks, but in other health care scenarios as well.

1. Models of Pseudonymisation

We distinguish several scenarios where distinct procedures for anonymisation or pseudonymisation are appropriate. In particular for the long-term accumulation of patient data the TMF proposes two models, see sections 1.4 and 1.5; they differ with respect to the location of the data pool within the overall architecture.

There are technically more elegant procedures for pseudonymisation, based on blind signatures as proposed by Chaum [2]; they assume that the pseudonym owner is also the key owner and controls the use of the pseudonym. However these procedures don’t fit the needs and data flows of secondary uses of the EHR, neither for one-time uses nor for building a data pool, and are not used in the context of this paper. *Here a pseudonym is an encrypted patient identifier*.

1.1 Single Data Source, One-Time Secondary Use

This is the typical application case for anonymisation and is well-understood. As an example take a simple statistical evaluation of EHR data.

1.2 Overlapping Data Sources, One-Time Secondary Use

Here the data from diverse sources must be linked together. Think of a multi-centric study that uses data from EHRs, but also data or probes from a biomaterial bank, or follow-up data at a later point in time. This is the typical application case for one-way pseudonyms. An essential prerequisite is a unique patient identifier (PID) in the EHR and the other data sources. The pseudonymisation procedure then consists of a one-way encryption of the PID, and should be implemented as a TTP service. A typical feature of this service is the use of asymmetric encryption: The data source encrypts the medical data with the key of the secondary user and sends the PID (not the identity data) as well as the encrypted medical data to the pseudonymisation service that encrypts the PID and sends it to the secondary user, together with the encrypted medical data. Note that the TTP cannot read the medical data, only the secondary user can decrypt them. But he cannot decrypt the pseudonym. Figure 1 shows the data flow; MDAT stands for the medical data, IDAT for the identity data, and PSN for the pseudonym.

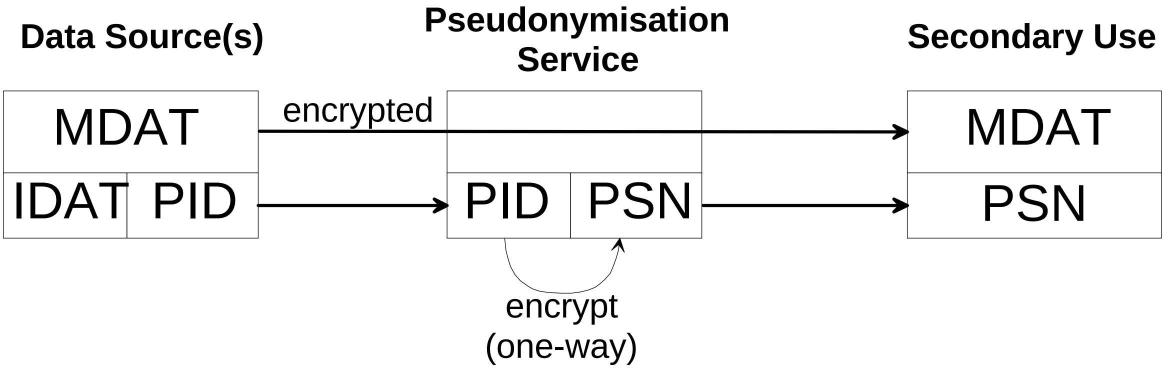

Figure 1. Data Flow for One-Time Secondary Use

Because the pseudonymisation service doesn't store the association between PIDs and pseudonyms, and cannot reverse the encryption, there is no need to treat the PID as secret, as long as the TTP implements an effective sender authentication and authorisation that prevents a "trial encryption" attack.

There was a TMF project that implemented this model in a health care research project [7], where it is routinely used since 2002.

1.3 One-Time Secondary Use with the Need of Re-identification

The conceptually simplest model of pseudonymisation with possible re-identification uses a reference list located at a trusted third party; in this model there is a big file containing patient identities and associated pseudonyms. This file is an attractive target for attacks and constitutes a single point of failure of the security concept. Moreover it stores patient identities outside of the proper treatment context and therefore violates the professional discretion of the participating physicians.

Therefore a refined reference list model is most suited, that extends the model in 1.2. It involves a two-step procedure for pseudonymisation and several keys and TTP services. First we need a PID that is not a "public" universal identifier (such as Patient Number, Insurance Number), but is project specific and is generated by a separate TTP service. This service stores the “patient list” – the association between identity data and PIDs; moreover it is responsible for the correct linkage between data from different sources. The PID is stored at the data source(s) but kept confidential. The pseudonymisation service works as a second TTP service and acts as in 1.2 but applies a reversible encryption procedure. Figure 2 shows the essential components of the data flow.

The pseudonymisation service doesn’t store the association between PID and pseudonym – not even the association between PID and data source – but can restore the PID from the pseudonym at any time with the help of its secret encryption key. For re-identification also the PID service is involved; it associates the PID with the identity data and notifies the data source.

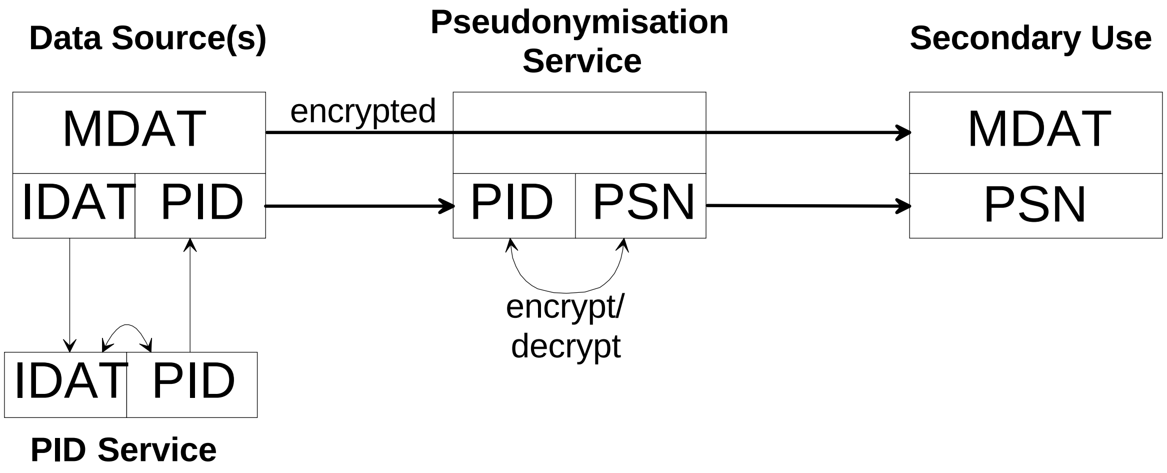

Figure 2. Data Flow with Possible Re-identification

1.4 Pseudonymous Research Data Pool

A new level of requirements appears with the need of long-term data accumulation, for example in building a disease specific registry. The “Model B” of the TMF uses the same procedure as in 1.3; the only additional feature is that on the “Secondary Use side” the data are collected in a data pool. This data pool is available for research projects. What projects may get access, depends on the situation, but as a rule the projects must be associated to the specific health care or research network by contracts; the data pool must not be a self-service database for arbitrary projects.

The data flow is basically the same as in 1.3, except that the “Secondary Use” is replaced by the “Data Pool” that permanently stores pseudonym and medical data and offers them for (possibly many) secondary uses. Because after pseudonymisation the quality assurance of the data would be much more difficult, careful *quality management should precede pseudonymisation*. This is the task of yet another TTP service. Note that – depending on the data protection policy – some of the TTP services might be offered by the same trusted third party.

1.5 Central Clinical Data Base, Many Secondary Uses

The alternative “Model A” of the TMF uses a somewhat different approach that better fits the needs of research networks with a “clinical focus”. It supports the long-term observation of patients with chronic diseases, and facilitates the individual feedback of research results to the patient or to the responsible physician. This model introduces a central clinical database as a TTP service with online access for the treating clinician who is also responsible for the quality of the data. The clinical database contains no identity data, but only the PID instead; the reference – in the case of authorised access – is established via the patient list. Additional data sources, for example biomaterial banks, use distinct references (LabID). If a research project needs data from this pool, the appropriate data set is exported in anonymised form or pseudonymised by a TTP with a project specific key; that means, different projects get different pseudonyms. Figure 3 shows the essential components of the data flow.

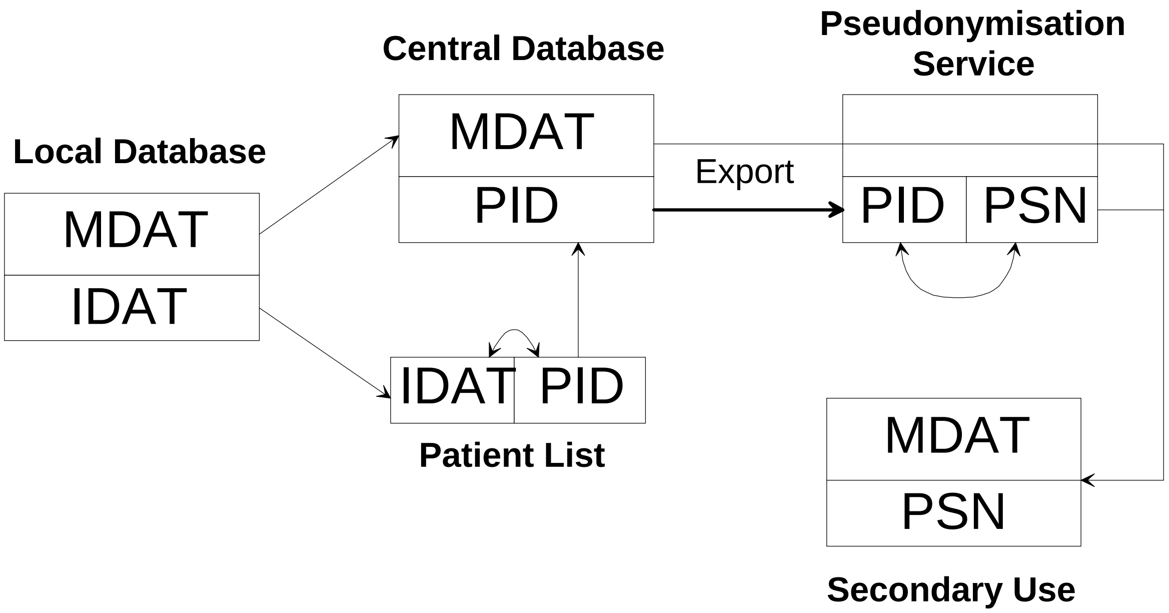

Figure 3. Data Flow for Model A

Note that Model A requires the implementation of sophisticated communication procedures. For example to access the data in the central data base, the participating physician gets temporary tokens that enable her to view or update her own data.

2. Results

The TMF Working Group on Data Protection developed the models A and B in close collaboration with the German Data Protection Commissioners (“Arbeitskreis Wissenschaft der Datenschutzbeauftragten des Bundes und der Länder”). The final version was consented by the Data Protection Commissioners (“Arbeitskreis Wissenschaft” and “Arbeitskreis Gesundheit”).

The TMF WG supports medical research networks with advice on adapting the “generic” models to their specific needs. Some networks already implemented one of the models, some other networks are in the process of implementation.

To support the implementation the TMF developed appropriate software tools for the specified communication paths and the involved TTP services. Moreover it provides sample forms for the patient’s consent, as well as policies and sample contracts for the participating members of the networks or projects.

3. Discussion

The TMF model architecture with its two variants provides ways for building medical research networks and central data pools that conform to the German and European data protection rules, respect the patients’ rights, and cover a wide range of situations. The transfer to other scenarios in health care is possible and recommended.

The generic TMF architecture is not a static structure. There are practical experiences and feedback from implementations, but also changing requirements in health care research and medical networks, for example with respect to genetic research. Therefore the TMF must continually keep its models up to date to meet new challenges.

Acknowledgement

This work was supported by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) as a project of the TMF and of the Competence Networks “Pediatric Oncology and Haematology”, “Chronic Inflammatory Bowel Disease”, “Systemic Inflammatory Rheumatic Diseases”, and “Acute and Chronic Leukaemia”.

References

[1] K. Pommerening: Pseudonyme - ein Kompromiß zwischen Anonymisierung und Personenbezug. In: H. J. Trampisch, S. Lange (Hrsg.), *Medizinische Forschung - Ärztliches Handeln,* 40. Jahrestagung der GMDS, Bochum, September 1995, MMV Medizin-Verlag, München 1995, 329–333.

[2] D. Chaum: Security without identification: Transaction systems to make Big Brother obsolete. Communications of the ACM 28 (1985), 1030–1044.

[3] B. Struif: Datenschutz bei elektronischen Rezepten und elektronischem Notfallausweis. In: *Vertrauenswürdige Informationstechnik für Medizin und Gesundheitsverwaltung*. Erfurt: TeleTrusT Deutschland e. V., 1994: 15/1–6.

[4] G. Bleumer, M. Schunter: Datenschutzorientierte Abrechnung medizinischer Leistungen. Datenschutz und Datensicherheit 2 (1997), 88–97.

[5] K. Pommerening, M. Miller, I. Schmidtmann, J. Michaelis: Pseudonyms for cancer registry. Methods of Information in Medicine 35 (1996), 112–121.

[6] Die Landesbeauftragte für den Datenschutz in Nordrhein-Westfalen: *15. Datenschutzbericht* 1999/2000.

[7] P. Ihle: Implementierung eines Pseudonymisierungsdienstes für versichertenbezogene Daten der gesetzlichen Krankenversicherung. Informatik, Biometrie und Epidemiologie 33 (2002), 350–351.

[8] M. Reng, P. Debold, K. Adelhard, K. Pommerening: *Generische Lösungen der TMF zum Datenschutz für die Forschungsnetze der Medizin*. Shaker-Verlag, Aachen 2004.