The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

Secure Communication and Management of Clinical and Genomic Data: the Use of Pseudonymisation as Privacy Enhancing Technique

Brecht Claerhout[a], Georges J.E. De Moor[b], Filip De Meyer[c]

[a]*CUSTODIX nv, Belgium*
[b]*Department of Medical Informatics and Statistics, University Hospital Gent, Belgium.*
[c]*RAMIT vzw, Belgium*

Abstract

The growing need of managing both clinical and genetic data raises important legal and ethical challenges. This article introduces some of the privacy protection problems related to genomic medicine and highlights the relevance of Trusted Third Parties and of Privacy Enhancing Techniques (PETs) in the context of e.g. research. Practical approaches based on two pseudonymisation models, for both batch and interactive data collection and exchange, are presented.

Keywords:
Genomic medicine; confidentiality; stigmatisation; discrimination; Trusted Third Parties (TTPs); Privacy Enhancing Techniques (PETs); pseudonymisation

1 Background

Although genomic medicine is still in its infancy, it is already evident that Medicine, Genomics and Information and Communication Technologies (ICT) will continue to develop in some sort of symbiotic evolution [1,2].

Genomic medicine encompasses predictive and diagnostic genetic testing. It can also use the information that derives from this testing to select or to fashion the best drug and therapeutic regimen for a patient, i.e. one that maximizes efficacy and minimizes side effects: pharmacogenetics is one of the avenues which will lead toward individualized healthcare and health maintenance.

Both genetic testing and pharmacogenetics give rise to concerns about the proper collection, storage and use of individually identifiable genetic information [3]. As the practice of genomic medicine develops, researchers and healthcare providers may want to store genetic profiles to determine treatment modalities as the need arises. The existence of such genetic databases will even increase the risk that unauthorized persons will obtain access.

Clinicians and researchers will therefore need to safeguard the confidentiality of such sensitive patient information.

Institutional Review Boards (IRBs) already pay careful attention to the requirement of obtaining the informed consent from subjects [4]. Research ethics and security guidelines demand research units to divert more and more resources and time to privacy and identity protection, but burdensome requirements governing the transmission of medical and genetic information could unnecessarily discourage research. Protecting human rights (e.g. privacy) while maximizing research productivity is one of the coming challenges. Well-intentioned privacy laws should not clash with the legitimate use of information when clearly to the public's benefit.

This paper mainly focuses on the possible use of Privacy Enhancing Techniques in the context of research and statistics.

2 Threats and Problems

The differences between genetic information and other medical information can be summarized as follows:

- Genetic data not only concern individuals, but also their relatives, thus people who have not been tested directly;
- Medical data deal with past and current health statuses of persons, whereas genetic testing can also give indications about future health or disease conditions;
- Personal genetic profiles can directly be derived from tissue samples;
- An individual person's genotype is almost unique and stable.

A widely discussed problem is that, unlike other data from e.g. clinical health records, genetic information is rarely about one single individual. A person's consent to release his or her genetic information constitutes a de facto release of information about other individuals, i.e. his or her relatives. In the case of genomic medicine, there is a complex interaction between individual rights and collective requirements.

Any collection of blood samples linked to identifiable persons can have an enormous impact on privacy; any material containing DNA is a potentially attractive source that can be mined for improper purposes.

Considering the risk of stigmatisation of particular subpopulations, the predictive and diagnostic testing for susceptibilities to disorders also remains problematic. This is even being complicated by the fact that some patients suspected of having a genetic disorder (e.g. Alzheimer) may lack the capacity to give their informed consent for a genetic test [5].

Given the potentially long latency period before symptoms develop, discrimination is another threat (e.g. insurers might use the results of diagnostic and predictive testing to calculate health risks and set premiums).

The question will be whether the perceived short and long term benefits exceed the risks of "improper access and use" and what security measures can be taken to reduce such risks. Finding the right balance between privacy protection and clinical utility will therefore become an issue, given the fact that physicians who will prescribe drugs without genetic testing could even face the risk of malpractice liability.

A couple of basic approaches to safeguarding confidentiality have been identified in the past. The first approach focuses on the creators and maintainers of the information, prohibiting them from disclosing the information to inappropriate parties. An alternative approach focuses on the use by Trusted Third Parties (TTPs) of so called Privacy-Enhancing-Techniques (PETs) and other measures using cryptographic techniques. In contrast with horizontal types of data exchange (e.g. for direct care), vertical communication scenarios (e.g. in the context of disease management studies and other research) do not require identities as such: here pseudonymisation can help find solutions.

3 Pseudonymisation and Trusted Third Parties

Pseudonymisation refers to Privacy Enhancing Techniques and methods that are used to withdraw and replace the true identities of individuals or organizations. Contrary to simple anonymisation, it still enables the linkage of data associated to the pseudo-identities (pseudo-Ids).

Therefore, pseudonymisation is a powerful and secure solution to the problem of reconciling the two following conflicting requirements:

- the adequate protection of individuals and organizations with respect to their identity and privacy;
- the linkability of data associated with the pseudo-IDs irrespective of the collection time and place (this being important in e.g. longitudinal studies).

Simply put, pseudonymisation translates a given identifier into a pseudo-identifier (a.k.a. 'digital pseudonym'). This is preferably done by using secure, dynamic and irreversible cryptographic techniques (and not static translation tables). The choice between an irreversible versus a reversible approach depends on users' needs.

Generated pseudonyms are thus represented by (to an observer) complete random selections of characters (letters, numbers and/or other marks). It is a flexible technique which can be employed in different ways, e.g. the transformation method can be time-related (a given identifier can always map with the same pseudo-ID or every time with a different pseudo-ID, the transformation method can change at specified time-intervals, ...)

Trusted Third Parties or TTPs have in common that they provide as independent intermediaries "trust services" to other parties. When the security solution is based on pseudonymisation, the trustee is a pseudonymisation TTP. Proper and secure pseudonymisation can only be performed with the support of such a pseudonymisation trust service provider, whose main features are:

- its strict independence as an organization;
- its strict code of conducts, trust practice statement and secrecy agreement policy.
- the trustworthiness of its methods, implementations and infrastructure;
- its adherence to the principles of openness and transparency regarding its methods;
- the provision of professional expertise related to the domain of relevance;
- its project-specific privacy and security policies;
- its documentation, operating reporting and auditing systems;

4 Batch versus Interactive Data Collection Pseudonymisation Model

The models and techniques explained below have already been tested and implemented in several different contexts, e.g. in Phase 4 Clinical Trials and for processing drug prescriptions.

A first possible scenario is the use of pseudonymisation in batch data collection. Generally, there are three entities in such a pseudonymisation process:

1. data suppliers or 'sources' (e.g. electronic medical record systems);
2. the pseudonymisation server or 'TTP-server' (Trusted Third Party-server);
3. one or several 'data registers' where the pseudonymised data are stored.

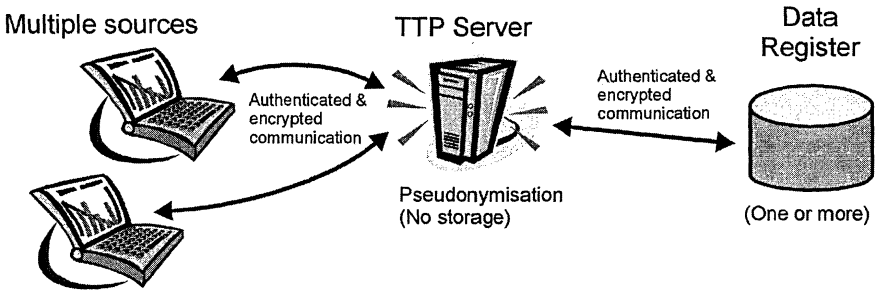

Figure 1: Communicating entities

In contrast to traditional data collection, the sources do not necessarily interact directly with the database, and vice versa. Communication is routed through the TTP server, where the pseudonymisation and the processing of relevant data take place, as required.

Data are being gathered and packed at the sources, typically in local databases for on-site use. The data is transmitted on a regular basis to the register through the TTP server where it is pseudonymised. At the source side, the pseudonymisation service should include:

- a basic pre-pseudonymisation functionality. To ensure maximum security, the pseudonymisation process is split into source-pseudonymisation (or pre-pseudonymisation) and pseudonymisation at the TTP level;
- Software that provides encryption and signing of the data, secure communication (authentication, authorisation, ...)

As sources are not allowed to supply batches of data directly to the register, they interact through the TTP. Before the transfer of data to the TTP, the data are split into two parts:

1. the identities (identity related data, e.g. social security number, name, internal reference number, etc);
2. the assessment data (also called 'payload data') related to those identities.

This split up is done following strict privacy protection policies. Assessment data should be screened for possible privacy threatening information. When defining the separation of identities and assessment data, one should not only filter direct identifying fields out of the assessment data (e.g. name, social security number), but also indirect identifying information (e.g. information that would make a person unique within a dataset).

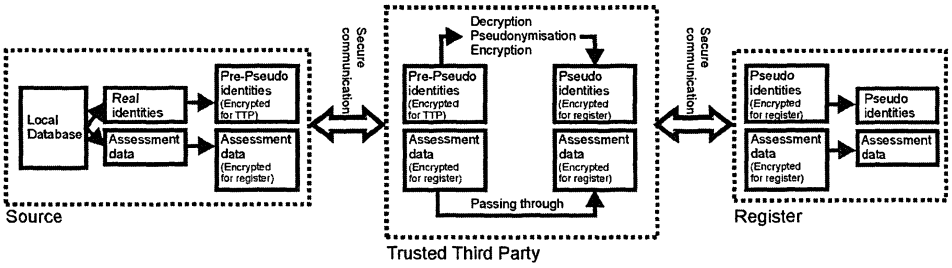

Figure 2: Data-flow (identity-data versus assessment-data)

The identities are pre-pseudonymised and pre-processed at the source. Thus no real identities leave the sources, and the TTP is never actually processing real identities. The pre-pseudonymised data are then encrypted using a public-key scheme for decryption by the TTP server exclusively. The payload data are public-key encrypted to the register, so that only the register can read the data. This means that although information passes through the TTP server, the latter can neither interpret nor modify the assessment data. Thus, full trustworthiness and integrity of the service is guaranteed not only by means of policy but also on a technical level. As an additional safety measure, all files transmitted by the sources are digitally signed.

Still, it must be understood that although the pre-pseudonymised information leaving the source no longer contains any real identities, this does not always guarantee absolute privacy. As the pre-pseudonymisation software is available at many sources/locations, an intruder might find a way to obtain real identities by entering known identities and creating

a translation table in order to map identities with their corresponding pseudo-identities (a 'dictionary attack').

By performing a second transformation in a centrally controlled location, i.e. in the TTP server, optimum security can be offered against such malicious attacks. In a second stage, the pre-pseudonymised ID is then transformed by using a stronger and even more secure algorithm.

Privacy protection is increased by extra measures on the pseudonymisation server, like e.g. monitoring of incoming identities against dictionary attacks. Furthermore, authorized sources and registers should also be bound by a code of conduct, as specified in a privacy and security policy agreement.

After the processing of the identifiers by the TTP - transforming the pre-pseudonymised identifiers into the final pseudo-IDs - both the payload data and the pseudo-IDs are transferred to the register via secure communication. At the register, the data can then be stored and processed without raising any privacy concerns.

Pseudonymisation services are not limited to privacy protection for batch data collection. Projects requiring immediate source interactions, typically remote database access, need a different model in which there is no need for local storage of data at the source, and therefore, no need for local database extraction and batch processing. All information is interactively obtained from, or delivered to the user at the source. Both privacy of the data subjects and the submitter are at risk.

Today, remote database access is often implemented with web browser technology because this requires only a minimum of extra software to be installed. In such a case, Privacy Enhanced Web Forms can protect all the privacy sensitive fields of the database (and the submitter's identity), while maintaining the flexibility of an interactive on-line database accessed by web browser technology.

Privacy Enhanced Web Forms involve placing a (transparent) intermediary entity between the users and the database web server. This third entity consists of a proxy/pseudonymisation server located at a TTP. When data is submitted at the source to the remote database, it passes the TTP proxy/pseudonymisation server before being forwarded to the database. At the TTP, identities are being transformed into pseudo-IDs, assessment data is left unchanged. The pseudonymised data is then being forwarded to the register. When a database request is submitted by the user, again the proxy/pseudonymisation server performs the necessary pseudonymisation on that request before forwarding it to the database. The database answer also passes through the proxy/pseudonymisation server, where it is filtered (e.g. pseudo-IDs can not be seen by the user) or translated (in the case of reversible pseudonymisation).

As with batch data collection, the integrity and confidentiality of all data streams should be ensured by use of appropriate techniques (authentication, encryption).

Such solution renders the interposition of a privacy protection service largely transparent to both the user and the database web server. Nothing is changed from the perspective of the user of the on-line database application. Traditional web browsing software is used to access the application, but instead of communicating with the client's web server, the user interacts with the proxy set-up.

5 Conclusions

Privacy includes the right of individuals and organisations to determine for themselves when, how and to what extent information about them is communicated to others. Advanced pseudonymisation techniques can provide optimal privacy protection of individuals while still allowing the grouping of data collected over different time periods (cf. longitudinal studies) and from different sites (cf. multi-center studies).

Protecting human rights in the realm of privacy, while optimising research potential and other statistical activities is a challenge that can easily be overcome with the assistance of a trust service provider offering advanced privacy enabling/enhancing solutions. As such, the use of pseudonymisation and other innovative Privacy Enhancing Techniques can unlock valuable data sources.

6 References

[1] Goodman KW. Ethics, Genomics, and Information Retrieval. Comput. Biol. Med. 1996; vol 26, no.3:223-229.
[2] Martin-Sanchez F. Integrating Genomics into Health Information Systems. In: Methods Inf Med 2002; 41:25-30.
[3] Fedder RS. To Know or Not to Know. Legal Perspectives on Genetic Privacy and Disclosure of an Individual's Genetic Profile. The Journal of Legal Medicine; 21:557-592.
[4] Mehlman MJ. The effect of Genomics on Health Services Management: Ethical and Legal Perspectives. Frontiers of Health Services Management; 17;37:17-26.
[5] McConnell LM, Koenig BA, Greely HT, Raffin TA. Genetic Testing and Alzheimer Disease: Recommendations of the Stanford Program in Genomics, Ethics, and Society. Genetic Testing 1993; vol. 3, nr 1:3-12.
[6] European Directive 95/46 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
[7] Schneier B. Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd edition; John Wiley & Sons, 1996
[8] Brands SA. Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy. Library of Congress Cataloging-in-Publication Data. ISBN 0-262-2491-8.

7 Address for correspondence

Prof. Dr. Georges De Moor
Department Medical Informatics and Statistics
University Hospital Gent
Building 3 – 5th floor
De Pintelaan 185
9000 Gent
BELGIUM
Tel: +32 9 240 34 36, Fax: +32 9 240 34 39
Email : georges.demoor@rug.ac.be