Collaborative Patient Centred eHealth
E. De Clercq et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-922-6-67

A Data Protection Framework for Transeuropean genetic research projects

Brecht CLAERHOUT [a,1], Nikolaus FORGÓ [b,2], Tina KRÜGEL [b], Marian ARNING [b]
Georges DE MOOR [a]

[a] *Custodix, Belgium*
[b] *Institute for legal Informatics, Leibniz University of Hanover, Germany*

Abstract. The paper proposes a data protection framework for trans-European medical research projects, which is based on a technical security infrastructure as well as on organizational measures and contractual obligations. It mainly relies on pseudonymization, an internal Data Protection Authority and on a Trusted Third Party. The outcome is an environment that combines both good research conditions and an extensive protection of patients´ privacy.

Keywords. Data protection**,** pseudonymization, anonymization, genetic research

1. Introduction

This paper is motivated by the EU research project ACGT (Advancing Clinico-Genomic Trials on Cancer[3]), which aims at the development of a trans-European cancer/gene grid network to promote better and more efficient curability. Within ACGT the authors are responsible for the technical security structure on the one hand and legal, especially data protection, issues on the other hand.

Trans-European genetic research projects such as ACGT are of great value for the fight against diseases such as cancer. At the same time it is of high importance to safeguard patients´ rights, in particular their right of privacy concerning medical data. The tension between human-genetic research and the legal aspects of data protection is obvious: A person's genetic data provides a massive amount of information such as the person's descent, ethnical origin, information on possible future medical conditions (with a certain probability), and much more. Each individual's genetic data is unique and can be of importance even for unborn blood relatives. This makes genetic data highly sensitive and its processing has to be carried out under strict regulations, combining all technical, organizational and liability based measures.

[1] Brecht Claerhout, Custodix NV, Verlorenbroodstraat 120 bus 14, 9820 Merelbeke, Belgium; brecht.claerhout@custodix.com.
[2] Prof. Dr. Nikolaus Forgó, Institute for legal Informatics, Leibniz Univerity of Hanover, Königsworther Platz 1, 30167 Hanover, Germany; nikolaus.forgo@iri.uni-hannover.de.
[3] http://www.eu-acgt.org/.

2. Data protection issues in trans-European genetic research projects

According to Art. 8 para 1 of Directive 95/46/EC the processing of genetic data is in general prohibited, as it has to be qualified as personal data. Personal data shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number […].

The most relevant exemption to the prohibition of data processing is the data subject's consent. However, the wording of the consent causes several problems when it comes to research in genetic data: It must be considered that to consent in advance to each data processing is almost impossible as normally, in the course of a project, new research methods are developed which may demand other operations to be performed upon the data than those the patient has consented to. But a vague consent, that covers all these unclarities, may not be seen as valid. If, in contrast, the wording of the consent is very specific, new research methods are not covered and even after years the patients would have to give new consents. The expenditure in organization and the practical problems which arise (is the patient still able to give consent?) are obvious. In the end medical progress would be jeopardized.

Therefore it would be best if the researcher used non-personal data for the research, because anonymized data are out of the scope of the Directive 95/46/EC. The Directive is applicable in cases of the processing of personal data only. If data is rendered anonymous, the data subject requires no further protection, because re-identification is impossible due to the lack of reference to the said person. Therefore the processing of anonymous data offers the best protection for the said person. Consequently, when genetic data has to be processed, it must be considered carefully, whether it is possible to process it anonymously. But the question arising is whether genetic data can be rendered anonymous at all or in contrast always has to be qualified as personal data because of its uniqueness.

The crucial point is how to define the term "anonymous". The Directive itself does not contain any explicit definition of this term. Only Recital (26) of the Directive contains an explanation: "(26) […] whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable […]." According to this wording, data can only be classified as anonymous if re-identification of the data subject is impossible for everybody. But the unique quality of genetic data causes the problem that despite comprehensive anonymization the re-identification of the said person still is possible if relevant additional knowledge such as genetic information exists in another database. In this case the identification of the data subject would always be possible by a matching procedure. Therefore a complete anonymization of genetic data is impossible [1].

Apart from that, as far as medical research is concerned, anonymized data is often not helpful anyway. In order to be able to follow the course of a patient's disease and to observe the patient's reaction to the treatment, the patient must be identifiable. At the same time researchers often replace the data subject's name etc. with a label, in order to preclude identification of the data subject or to render such identification substantially difficult. The person can only be re-identified by using the appropriate key. The data is "pseudonymized". The question arising here is whether such pseudonymous data still has to be considered to relate to an "identifiable" person or if this data could be seen as (de facto) anonymous data for the researcher not having the appropriate key. According to a recent opinion of the Article 29 Data Protection Working Party in the fields of

clinical research medical data could be seen as anonymous data if a) in the specific framework the re-identification is explicitly excluded and b) appropriate technical measures have been taken in this respect [2]. In those cases such key coded data are not subject to the rules of data protection legislation.

3. Data protection framework

With respect to these legal requirements the following data protection framework was designed for ACGT:

3.1. Data Protection Authority

From a practical point of view in research projects compliance often is a crucial issue. Hence to guarantee compliance of the project with data protection legislation it is in a first step essential to put the project consortium in the position to audit such compliance. Otherwise all investment in project policies, technical infrastructure or organizational measures is not worth the effort. Therefore it is appropriate to establish an authority that is both legally able to enter into binding contracts with the project participants and empowered to inflict a penalty for infringement. To be able to conclude contracts, this Data Protection Authority has to be a legal body, empowered by the project consortium, but independent in its decisions. Once this authority is established, policies integrated in binding contracts can be set up, which implement and/or legally confirm measures such as the following:

3.2. Pseudonymization

As shown above pseudonymous genetic data in the context of clinical research may not be subject to the rules of data protection legislation, if appropriate policies as well as organizational and technical measures are set up. To ensure that the processing of genetic data within ACGT is de facto anonymous a legal/technical framework is set up that builds on a state of the art pseudonymization and the integration of a Trusted Third Party. The primary legal conditions of this framework are:

- All technical and organizational measures as well as obligations, such as the irrevocable prohibition of matching procedure in order to re-identify a patient, are codified in binding contracts signed by all participants of the project.
- All data transmitted to and processed within the project must be pseudonymized before entering the network by a unique state of the art pseudonymization.
- To monitor and audit compliance with data protection policies as well as to give patients one central contact for any questions or complaints concerning the processing of their data, it has to be guaranteed that the internal Data Protection Authority is the central data controller within the project, whereas all users of the network process data only on behalf of this controller.
- To build up a network with only one central data controller makes a strict organizational and technical separation necessary between data stored and analyzed in the hospitals for medical treatment and the data stored and analyzed on behalf of the research project. Separate databases, adequate

access control and contractual obligations have to be implemented to ensure this separation.

- Re-identification where it is needed for therapy reasons only must solely be possible involving the Trusted Third Party that provides the software tool for the pseudonymization and holds the cryptographic authorizing the re-identification.

3.3. First fall back scenario: informed consents

Nevertheless the data protection framework in projects like ACGT should be structured like a safety net, as for all genetic data that can NOT be qualified as "de facto anonymous" a different solution is needed. If the researcher can establish the link between data and data subject there is need for a legal permission, as in those cases the genetic data is personal data in the sense of the Directive 95/46/EC, and the Directive therefore is applicable.

Thus, in a second step, the data processing is legitimated in the traditional way: by informed consents of the patients. To obtain informed consents in every case has several advantages: First of all, it involves the patient in the whole procedure. This leads to transparency, generates trust and is required for ethical reasons anyway. At the same time it gives legal certainty in those cases where researchers can establish the link between data and patient, even though they might not even know they can.

Hence the data protection safety net builds on the de facto anonymization of genetic data as a basic principle, with a legitimation via informed consents as a fallback scenario. It therefore combines both, the protection of the patients' privacy and the legal certainty for the researchers involved.

3.4. Second fallback scenario: Exceptions for genetic research in national legislation

For the unlikely event that for a specific patient both the de facto anonymization fails and the informed consent does not exist, does not cover the specific use or is invalid, as a second fallback scenario the particular national legislation has to be analyzed with regard to an exemption according to Art. 8 para. 4 Dir. 95/46/EC. Member States may, for reasons of substantial public interest, lay down further exemptions from the general prohibition on processing sensitive data, e.g. scientific research, see Recital (34).

The problem with this exemption is that Member States are free to implement it. Whether the Member State whose law is applicable for the data processing operation in question, has introduced such an exemption in its national law, has to be analyzed individually. However, this analysis ought to be made by the Data Protection Authority for each individual case as those national provisions differ and can change at any time.

4. Technical implementation of the data protection framework

The ACGT Service Oriented Architecture (SOA) has a layered structure. The lower layers of the platform that provide basic functionality such as resource allocation, job management, etc. are based on the Globus Toolkit [3] and GRIDGE [4]. On top of those, the ACGT Business Processes Services reside providing a "biomedical grid layer" to ACGT [5] (i.e. semantic mediation, a master ontology service, knowledge

discovery tools, etc.). Common security functionality such as secure communication, user authentication, virtual organization (VO) management, etc. are based upon the functionality provided by these layers (e.g. the Gridge Authorization Service GAS responsible for VO management).

However, such standard components fall short of offering a means to meet the demands for treating sensitive biomedical data explained in the previous sections. The two major implications of the data protection framework on the ACGT platform are the data de-identification and pseudonymization requirement and the need for controlling the context in which data is used (so that this data can be treated as de facto anonymous for the data controller).

4.1. De-identification and Pseudonymisation Toolkit

In order to meet the de-identification demands laid out in the legal analysis, a tool [6] was created for exporting pseudonymous data from the (internal) hospital data stores to their anonymous ACGT counterparts (i.e. the ACGT accessible data sources, also physically residing in the hospitals). The tool is innovative in a sense that it offers a generic solution regardless of the type of data to be treated or of de-identification requirements. It consists of a "workbench" and a "wizard". The "workbench" serves at defining the mechanics (data protection profile) through which data is exported for sharing, the "wizard" allows to easily apply those profiles on various data sources. The workbench allows domain experts and privacy professionals to:

- create a mapping from a specific data format such as flat files (e.g. CSV), imaging data (e.g. DICOM), microarray data, structured data (e.g. XML, databases) to a generic data model
- define the set of actions that should be performed on the generic data model in order to de-identify data (i.e. the data protection profile)

Once that a data mapping and a data protection profile is created in the "workbench", end users (i.e. physicians) can easily export several data sources at once by using the wizard (logically, this operation can also be automated).

Privacy processing actions such as creating a pseudonym (randomly assigned, through encryption of immutable identifiers, etc.), freetext de-identification, basic encryption, calculation of relative dates (to obfuscate absolute birthdates), etc. are defined towards the internal generic data model. The big advantage of this approach is that a single privacy protection profile can be applied to various data sources. It also provides a base for extending the tool with privacy risk analysis functionality and more complex information content reductions algorithms (e.g. local suppression and generalization routines).

The privacy transformations are provided as a library (also usable by developers through an API), hence the functionality of the tool can be easily extended to suit new requirements (e.g. a new freetext de-identification routine or a new input data format).

4.2. Protecting the Context of Anonymity

Sensitive data processed within ACGT can only be treated as "de facto" anonymous if the context in which it is used can be controlled by the internal Data Protection Authority (data controller): i.e. sensitive data should only be accessed by people and organizations legally bound to the ACGT policies. ACGT relies on the (legally bound)

service and data providers on the platform to enforce this "contextual anonymity". In this task they are (technically) supported by a separate central authorization service managed by the Data Protection Authority, which specifically deals with data protection policy decisions. Decisions made by this separate authorization service are only based on data protection aspects of the access request (i.e. the request could still be denied based upon evaluation of other security rules regardless of the data protection decision). These decisions are made though interpretation (rules engine) of data protection policies based on the existing legal contracts and data protection metadata associated with the datasets that need protection. This system of privacy related metadata relies on the "ACGT data wrappers" which are part of the basic ACGT framework and allow to associate generic metadata to data handled on the ACGT infrastructure. Note that this approach is quite similar to the concept of "sticky policies". This system of central data protection policy management allows the ACGT data providers and services to comply effortlessly with the rules laid down in the data protection framework, and relieves the ACGT Data Protection Authority liability in case that one of the data providers intently violates data protection legislation. In addition to policy management, this system provides a form of information flow management and audit trail for the ACGT Data Controller.

5. Conclusion

This paper proposes a data protection framework for trans-European genetic research projects. It recommends a graded safety net and the establishment of an internal Data Protection Authority as well as the involvement of a Trusted Third Party.

The ultimate ambition is to process only de facto anonymous genetic data. To gain de facto anonymous genetic data, we propose a pseudonymization which can only be revoked in cooperation with a Trusted Third Party. The Data Protection Authority is the central data controller that monitors and audits paticipants´compliance with the project's data protection policies. This has to be obtained by binding contracts which empower the Data Protecting Authority to inflict a penalty for infringement.

If achieved, the proposed framework is on the one hand in line with European regulations and on the other hand easily manageable for researchers.

References

[1] T. Weichert, Der Schutz genetischer Informationen, in: DuD 2002, p. 133 (134).
[2] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, p. 20.
[3] I. Foster, Globus Toolkit Version 4: Software for Service-Oriented Systems, IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp. 2-13, 2006.
[4] J. Pukacki, M. Kosiedowski, M. Kupczyk, M. Wolski, M. Adamski, P. Grabowski, M. Jankowski, C. Mazurek, N. Meyer, R. Mikolajczak, J. Nabrzyski, T. Piontek, M. Russell, M. Stroiński, M. Wolski, Programming Grid Applications with Gridge, Computational Methods in Science and Technology, 12(1), 47-68 (2006).
[5] M. Tsiknakis, M. Brochhausen, J. Nabrzyski, J. Pucacki, S. Sfakianakis, G. Potamias, C. Desmedt, D. Kafetzopoulos, A Semantic Grid Infrastructure Enabling Integrated Access and Analysis of Multilevel Biomedical Data in Support of Post-Genomic Clinical Trials on Cancer, Digital Object Identifier: 10.1109/TITB.2007.903519 (to appear), http://ieeexplore.ieee.org/xpl/tocpreprint.jsp?isnumber=26793&punumber=4233.
[6] CAT – Custodix Anonymisation Tool. Retreived August 17, 2008 from http://cat.custodix.com/.