In order to protect the privacy of participating patients in multicentric genetic research projects and to improve the working conditions for researchers in such projects a data protection framework needs to be installed. In the first place, all genetic data processed in the project has to be pseudonymized. In addition to that, contracts have to be concluded between the project and each project partner to guarantee that genetic data are used only within the project and that each partner complies with data security standards. Furthermore, a central data protection authority has to be installed in the project to control the partners’ compliance with these contracts and to serve as a central contact point for participants. If these conditions are fulfilled, only (de facto) anonymous data are used in the project, so that data protection legislation is not directly applicable. Second, each participant has to sign a special consent form for ethical reasons and as a fallback solution if the pseudonymization of the genetic data fails. With this safety net it is possible to protect the participants’ privacy and to improve the working conditions for researchers.

**Keywords:** genetic data; data protection; de facto anonymization; pseudonymization; central data protection agency; genetic research projects

1. Introduction

Multicentric clinical trials working with patients’ genetic data are of great value for the fight against (as yet) terminal illnesses. High-performance networks such as those based on grid applications provide increasingly better initial conditions for successful medical research. Moreover, 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 masses of information such as the person’s descent, ethnic origin, about possible future diseases and much more. Beside this, each individual’s genetic data are unique. This makes genetic data highly sensitive and requires strict regulations. Identification of and compliance with legal rules is therefore essential.

* Author for correspondence (kruegel@iri.uni-hannover.de).

One contribution of 16 to a Theme Issue ‘Crossing boundaries: computational science, e-Science and global e-Infrastructure II. Selected papers from the UK e-Science All Hands Meeting 2008’.
The question that needs to be answered is how to deal with these conflicting interests. Obviously the solution is not to generally prohibit any genetic research nor to just ignore patients’ right of privacy. It is an increasingly vital task for those who safeguard peoples’ privacy to combine these interests and provide practicable solutions combining all technical, organizational and liability-based measures than to battle against such research.

As most genetic research networks in Europe are transnational, different national laws are applicable for the various participants. All researchers primarily have to comply with the national legislation of their country of residence.\(^1\) This leads to the question of which law should be taken into account when building up a data protection framework for the whole project. Taking all different national legislations into account will lead, for different reasons, to a complex and legally uncertain infrastructure. On one hand, such a framework would have to be modified whenever a partner from another member state joins the project; and on the other hand, it is organizationally impossible to keep the framework in line with several national laws, as they may change permanently. Therefore, the only solution to this task is to comply with European legislation as it has to be transposed into national law by all member states of the EU. Hence, this article deals only with European legislation.\(^2\)

If personal data are to be processed in such trans-European projects, either a statutory permission or the informed consent of the person the data relate to is needed according to Articles 7 and 8 of the Data Protection Directive 95/46/EC.

Many of the trans-European genetic data research projects work with consent as the legal basis for data processing. However, the concept of consent causes several problems when it comes to research using genetic data. In the context of genetic research it is almost impossible to inform the person that the data relate to all (future) data-processing operations that might take place, because during a project new research methods may be developed which may demand other data-processing operations than those the patient has consented to. If, therefore, the wording of the consent is very general in order to cover future research methods too, the consent may not be seen as valid because it is not precise enough. If, by contrast, the wording of the consent is very specific, new research methods might not be covered by the consent\(^3\) (O’Brien & Chantler 2003; O’Neill 2003; Reymond et al. 2003; Alston 2005).

---

1 See Art. 4 para. 1 of the Data Protection Directive 95/46/EC.
2 If genetic data shall be transferred to a data controller established in a territory of a non-EU member state, this state must, according to Art. 25 of the Data Protection Directive, ensure an adequate level of protection. States such as Canada or Switzerland provide such a level of protection (see a comprehensive list of the countries at: http://ec.europa.eu/justice_home/fsj/privacy/thirdcountries/index_en.htm). However, if the particular country the genetic data shall be transferred to does not ensure an adequate level of protection, the data subject has to consent to the transfer of his or her data to this third country or the recipient of the data in the non-EU member state has to guarantee the adequate level of protection according to Art. 26 of the Data Protection Directive. Whenever genetic data shall be transferred to a research entity established in the territory of a non-EU member state ensuring no adequate level of protection, such as Australia or Japan, the entity shall sign a contract prior to the transfer of the genetic data, in which it ensures an adequate level of protection (see the model contracts of the Commission available at: http://ec.europa.eu/justice_home/fsj/privacy/modelcontracts/index_en.htm).
3 See references for different consent models (specific/general) and their legal validity.
Medical Research Council 2005; Hansson et al. 2006). Therefore, a data protection framework for genetic research projects relying only on consent contains legal uncertainty.

Apart from this, the best solution to protect the data subject’s privacy and to promote scientific research would be not to use any personal data at all. According to Article 2 lit. a of the Directive, personal data means 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 or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. If data subjects cannot be identified any more, their privacy is protected best. This would also be the best solution for researchers because where data protection legislation is not applicable they can carry out their research without taking it into consideration. In addition to that, the legal uncertainty regarding the validity of consent would be dispelled.

This article analyses the characteristics of European data protection legislation with regard to genetic data and the implementation of such legislation in a trans-European research project. It is motivated by the EU research project Advancing Clinico-Genomic Trials (ACGT on Cancer), which aims at the development of a trans-European grid-based network combining genetic data and cancer to promote better and more efficient curability (http://www.eu-acgt.org).

2. Legal background

The Data Protection Directive 95/46/EC introduces a fundamental distinction between personal and non-personal (=anonymous) data. According to Recital (26) of the European Data Protection Directive 95/46/EC ‘[...] the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable [...]’. If only non-personal data are used, the Data Protection Directive is not applicable, because in this case the privacy of the data subjects is not put at risk, so that protection by data protection legislation is not necessary. The best way to safeguard privacy is therefore to process only anonymous data. If this is the case, anonymous data can be collected, stored and published without restrictions, at least from a data protection point of view (Gola & Schomerus 2005).

However, as far as medical research is concerned, anonymized data are often not very useful. A patient must, in many cases, be identifiable in order to be able to follow the course of the patient’s disease, to grant the patient the benefits of newly explored therapy methods and to observe the patient’s reaction to the treatment. Therefore, the data subject’s name is often replaced beforehand with a label, in order to preclude identification of the data subject or to render such identification substantially difficult. The person can only be reidentified by using the appropriate key, which links the label to the real name. The data are ‘pseudonymized’.

4 Art. 3 (1) Dir. 95/46/EC: This Directive shall apply to the processing of personal data wholly or partly by automatic means, and to the processing otherwise than by automatic means of personal data which form part of a filing system or are intended to form part of a filing system.
The question at this stage is whether pseudonymous genetic data can in a legal sense also be qualified as anonymous data. As mentioned above, the European Data Protection Directive states in Recital (26) that data are anonymous if ‘[...] the data subject is no longer identifiable [...]’. Therefore, according to the wording of Recital (26), reidentification of the data subject has to be impossible for everybody. But can genetic data be rendered anonymous at all then? Owing to its uniqueness, genetic data cannot be rendered anonymous in such a way that nobody can ever reidentify the person concerned. If additional knowledge such as genetic information with identifying characteristics exists in another database, the identification of the data subject is always possible by a matching procedure. Therefore, a complete anonymization of genetic data is impossible. Does this mean that genetic data always have to be qualified as personal data because of their uniqueness? The consequence would be that, whenever genetic data are processed, data protection legislation would be applicable with all its obstacles for medical research.

Recital 26 of the Directive states: ‘Whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person’. That means account should not be taken of all means, but only of means likely reasonably to be used.

Therefore, the definition of ‘anonymization’ as stated in the European Data Protection Directive contains an element of risk regarding the possibility of deanonymization (Metschke & Wellbrock 2002a). According to that definition, the anonymization of genetic data might be possible. Recently, Article 29 of the Data Protection Working Party (http://ec.europa.eu/justice_home/fsj/privacy/workinggroup/index_en.htm) has even mentioned the anonymization of genetic data as a means to limit the dangers of genetic research. Even the European Commission is of the opinion that the term ‘anonymous data’ must not be interpreted too strictly: in the first report on the implementation of the Data Protection Directive 95/46/EC, the Commission states that the interpretation of the Directive must be sensible and flexible, and draws attention to an article of the European Privacy Officers Forum (EPOF), which emphasizes the practical orientation and exemplary function of the German definition of anonymization.

The German transposition of the Directive, the ‘Bundesdatenschutzgesetz’ (BDSG; English: Federal Data Protection Act), defines anonymization in Article 3 paragraph 6 BDSG as ‘the modification of personal data, so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual’. In conclusion, the BDSG accepts not only strictly anonymous data as anonymous data in a legal sense, but also de

facto anonymous data. De facto anonymous data can be transformed into personal data only with a disproportionate amount of time, expense and labour (Metschke & Wellbrock 2002b).

According to that definition, genetic data can be regarded as anonymous data if they can be deanonymized only with a disproportionate amount of time, expense and labour. In practice, Article 29 of the Data Protection Working Party stated, in the fields of clinical research, medical data could be seen as anonymous data if (i) in the specific framework, the reidentification is explicitly excluded and (ii) appropriate technical measures have been taken in this respect.

Another important legal question is: who is the person not able to identify the data subject with a proportionate amount of time, expense and labour so that the data can be qualified as non-personal? If the data controller can directly identify the data subject concerned, it is beyond dispute that these genetic data have to be qualified as personal for this data controller. Even if data controllers do not want to use knowledge or means they actually have and could use to identify the data subject concerned the genetic data have to be qualified as personal. Only if the data controller does not actually have any knowledge and/or means to identify the data subject concerned, the genetic data might be qualified as non-personal.

That leads us to the next question: what if the data controller cannot identify the data subject concerned directly with a proportionate amount of time, expense and labour, but a third party could? The data subject’s privacy is threatened also in this situation, as genetic data are unique. For example: researchers cannot identify a data subject with cancer themselves. If the genetic data relating to that data subject were to be qualified as non-personal data under these conditions, data controllers would be allowed (at least from a data protection point of view) to do whatever they want with this genetic data, for example they could publish the data on the Internet. A third person having a reference genetic dataset containing identifying characteristics (such as the clear name) could then download the data from the Internet, and run a matching procedure to find out whether one of the datasets stored at this third party is identical to one of the downloaded datasets. The third party could in this way identify the data subject that the downloaded genetic data relates to and find out (in this example) that this person has cancer. Especially life insurance companies, biobanks and law enforcement institutions run large databases containing genetic data with identifying characteristics.8 Data subjects’ privacy would be affected, although the (first) data controller cannot identify the data subjects itself with a proportionate amount of time, expense and labour. To eliminate this risk, genetic data also have to be qualified as personal data, if only a third party is able to identify the data subject concerned with knowledge and/or means it actually has or could use. This interpretation is also in line with European legislation. Recital 26 of the Data Protection Directive states that in order ‘to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person’. If this other person (the third party) gets access to the genetic data, for example via the Internet, he or she will use

8Huge databases containing genetic data are already in use, e.g. by law enforcement agencies, insurance companies and in the USA even in connection with employment contracts (see Weichert 2002).
means and knowledge he or she actually has or could use (for example, matching procedures) to identify the data subject concerned, at least if this person is interested in these data. As nobody can know for sure whether a third party might be interested in these data or not, the interest has to be presumed to guarantee the privacy of the data subject concerned (Bygrave 2003).

The problem arising for data controllers from this interpretation is that no data controller can know for sure for which genetic data a reference dataset exists, because they cannot know which genetic data are stored anywhere in the world. The consequence of this interpretation would be that the data controller would have to treat all genetic data as personal data as a reference dataset might exist somewhere in the world. Genetic research would be constricted very much, as data protection legislation would be applicable then in any case.

That is why this opinion must be applied restrictively. The aim of data protection legislation is to guarantee the privacy of the data subjects concerned. In this field the privacy of the data subjects is not threatened under two conditions:

(i) Data controllers cannot identify the data subject concerned with knowledge and/or means they actually have or could use.
(ii) The third party who could identify the data subject concerned does not have access to the original genetic data.

Whenever it can be guaranteed that these two conditions are fulfilled, the genetic data have to be qualified as non-personal as the data subject’s privacy is not threatened and the scope of data protection legislation would be extended too much otherwise. This interpretation is also in line with Recital 26 of the Data Protection Directive, as it states that in order ‘to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person’. Data controllers have no knowledge or means they could use to identify the data subject, and the other person has no access to the datasets of the data controller. That is why there are no means likely reasonably to be used to identify the said person, if these two conditions are fulfilled. The genetic data concerned have to be qualified as anonymous then, and data protection legislation is not applicable in this case. As a consequence for researchers in genetic research projects, one can state that, whenever they cannot identify the data subject concerned with knowledge and/or means they actually have or could use, genetic data are not personal, if no third party can access these data. Data protection legislation is not applicable if such genetic data are just stored or used for research by the data controller. But whenever a third party can access the data (e.g. following the transmission, disclosure or publication), the same data have to be qualified as personal data, as the second condition is not fulfilled any longer and the data subject’s privacy is threatened. For such processing operations, a statutory permission or the consent of the data subject concerned is needed, as data protection legislation is fully applicable in this case.

9 For the Austrian approach and the concept of indirectly personal data see especially Art. 2 Part 1 Sect. 4, No. 1, Art. 1 Part 2 Sect. 8 Para. 2 and Art. 1 Part 2 Sect. 9 Para. 1 No. 2 of the Austrian Federal Act concerning the Protection of Personal Data, available at: http://www.dsk.gv.at/site/6230/default.aspx. (See also Arning et al. 2006.)
This solution supports genetic research and promotes medical research. At the same time, this solution protects the privacy of the data subjects, as it promotes the use of anonymous genetic data. Furthermore, it is perfectly in line with European data protection legislation.

In a second step, these legal considerations have to be realized in practice.

3. Data protection framework

With respect to these legal requirements, we propose to set up a ‘data protection framework’ whenever it comes to trans-European genetic research projects, which consists of technical, organizational and legal measures equally.\(^\text{10}\)

(a) De facto anonymization

To ensure that the genetic data can be qualified as de facto anonymous, a legal and technical framework has to be set up, which builds on pseudonymization of genetic data, the establishment of a closed-user group (‘network of trust’), the establishment of a data protection authority (DPA) and on the integration of a trusted third party. We propose the following structure (figure 1).

(i) Data protection authority

From a practical point of view, the implementation of and compliance with a data protection framework is often one of the most crucial issues. Hence, to guarantee compliance of research projects with set-up policies and data protection legislation, it is essential to put the project consortium in the position to audit such compliance.

Therefore, building up a data protection framework for a trans-European research project requires, as one of the very first actions, the establishment of an authority within the project, which is both legally able to enter into binding contracts with the project participants as well as empowered to impose a penalty for infringement. To be able to conclude contracts, this DPA has to be a legal body, empowered by the project consortium, but independent in its decisions.

Once such an authority is established, policies integrated in binding contracts can be set up which implement measures such as the following.

(ii) Data flow

The data flow is organized as follows: genetic data of a patient, which are collected by the treating physician in a hospital, are analysed and stored at the hospital.

If a patient agrees to participate in the research project, the physician transmits the patient’s data to the research project database. This database is controlled by the DPA of that project, but can be physically situated in the particular hospital. With transmission of the data, the data are pseudonymized. All directly identifying characteristics are removed and replaced by a

\(^{10}\) The Data Protection Framework shall be subject to an audit prior to the start of the data exchange. Also throughout the operation audits shall be carried out on a regular basis.
pseudonym. The key, which forms the connection between the pseudonym and the particular patient and which allows the reidentification of the patient, is only stored at a trusted third party. To be trusted by everyone, this trusted third party has to be a separate legal entity, which is independent from the DPA, but bound to duties and responsibilities by an agreement with the DPA. The trusted third party is only allowed to use a key in order to deanonymize genetic data if a new medical treatment has been developed, which the patient concerned could benefit from, and if the particular patient wants to be informed. All partners within the research project do not have access to these keys. The researchers receive pseudonymous data only via the research project’s network.

But can this pseudonymous genetic data be qualified as de facto anonymous according to the definition presented in the theoretical part?

The very basic condition to qualify pseudonymous data used in such a research project as de facto anonymous is a closed-user group, meaning that, first, all project participants (for example the DPA, the researchers) are not able to identify the patient concerned with a proportionate amount of time, expense and labour and that, second, all data have to stay inside the network of the project. Within this closed-user group the genetic data would then be de facto anonymous. This contextual anonymity of genetic data within the closed-user group has to be created and guaranteed by both the technical design of the project’s data protection framework and a legal framework. The legal framework consists of contracts concluded between the DPA and each research organization that wants to take part in the research project (figure 2).

Access to the research project’s network is granted only if the research organization has signed such a contract, in particular one in which the following provisions are stipulated.
(iii) No matching

It is of vital importance that neither the DPA nor any of the researchers carry out any matching procedures between the pseudonymous data and the reference datasets (already existing at the research organization) containing identifying characteristics of the patient. With the help of these procedures, the identity of the patient concerned could be determined, especially as the pseudonymized datasets often come from hospitals that are simultaneously research organizations. That is why a contractual provision is needed that prohibits any matching procedure, so as to ensure that only de facto anonymous data are processed within the research project.

(iv) No publication and no transfer

Another crucial contractual provision is that no researcher is allowed to release any data received from the research project to any third party. As described above, de facto anonymous genetic data can be deanonymized by matching procedures. These matching procedures cannot only be carried out by a researcher participating in the project (which is forbidden), but also by a third party, if such a third party gains access to the pseudonymous genetic dataset.

If such a clause did not exist, the researcher could publish the pseudonymous data on the Internet, in a magazine or just transmit it to a third party. If a party receiving the dataset from the Internet, a magazine or directly from the researcher has a reference dataset including identifying characteristics, such as the patient’s name, this third party could also run a matching procedure and identify the patient concerned. An insurance company, for example, could use such genetic datasets published on the Internet to find out whether one of its clients has a certain disease, if the dataset is published in the context of a clinical trial on cancer for example. All pseudonymous datasets that escape out of the closed project environment could no longer be qualified as de facto anonymous.
data. On the contrary, they would have to be qualified as personal data, as described above, because a third party, having a reference database, could easily identify the patients concerned.

Since the DPA and the research organizations simply cannot know for which datasets a reference dataset including identifying characteristics exists, all published datasets would have to be qualified as personal data to avoid liability. In other words, if any pseudonymous genetic data escapes out of the closed project environment, the data concerned have to be qualified as personal data, as the contextual anonymity would be abrogated. In order to prevent this scenario, a contractual provision is needed, which prohibits any publication and transfer of pseudonymous data received from the research project to third parties not participating in the project.

With these provisions and the technical design of the data flow, a closed-user group and the contextual anonymity of data processed within the research project can be guaranteed.

(b) Informed consent for ethical reasons and as a fallback scenario

Although processing of pseudonymous data within the closed-user group of a research project does not need permission as only de facto anonymous data are processed, it is recommended to obtain informed consent from the participating patients anyway. First, it involves the patient in the whole procedure that leads to transparency and generates trust—an essential factor for the success of such research projects.

At the same time, informed consent can serve as a back-up solution if the genetic data of a patient are not pseudonymized accurately or pseudonymous data gets out of the closed project environment. In such cases, the pseudonymous data have to be qualified as personal data so that their processing requires permission pursuant to data protection legislation. Informed consent could fulfil this requirement. Therefore, compliance with national legislation concerning the validity of informed consent is (still) of crucial importance.

4. Conclusion

It is possible to keep the data flow for the most part outside of the scope of the Data Protection Directive 95/46/EC by de facto anonymizing genetic data. The very basic conditions for de facto anonymizing genetic data are pseudonymization of the data to prevent project partners running matching procedures to identify the data subject concerned and keeping the genetic data within the closed-user group of the project.

It is feasible to comply with these theoretical considerations via a data protection framework. The main parts of this framework are the establishment of a DPA within the project, a pseudonymization procedure, the introduction of a trusted third party, binding contracts between each project partner and the central DPA and finally informed consent of each patient for ethical reasons, on the one hand, and the unlikely case that the de facto anonymization fails, on the other hand. If this architecture is implemented in the research project, participating researchers could undertake their research without encountering big obstacles due to data protection reasons. They could concentrate on their
scientific research, so that this architecture would ensure and improve the efficiency of the research project. Researchers would comply with data protection legislation automatically just by using the proposed data protection framework.

The participants’ privacy is properly protected by this data protection framework as well, because only de facto anonymous data are used in the project. The participant can no longer be identified or only with a disproportionate amount of time, expense and labour.

By implementing this data protection framework, both the needs of the researchers and all other project partners as well the interests of the participating patients can be satisfied at the same time. This data protection framework will, therefore, play an important part in leading a genetic research project to success.

References

Weichert, T. 2002 Der Schutz genetischer Informationen. Datenschutz und Datensicherheit—DuD 26, 133.