European Journal of Human Genetics (2000) 8, 739–742

www.nature.com/ejhg

ARTICLE

Protection of privacy by third-party encryption in genetic research in Iceland

Jeffrey R Gulcher, Kristleifur Kristjánsson, Hákon Gudbjartsson and Kári Stefánsson

Decode Genetics, Inc., Lynghals 1, Reykjavik, Iceland

As the new human genetics continues its dramatic expansion into many laboratories and medical institutions, the concern for the protection of the personal privacy of individuals who participate increases. It seems that even the smallest of laboratories must confront the issue of how to protect the genetic and phenotypic information of participants in their research. Some have promoted the use of anonymity as a way out of this dilemma. But we are reminded by others that the future cannot be predicted, and that future benefits may be lost when the links to these benevolent volunteers are gone forever. More recently, some ethical bodies have suggested, without specific recommendations, that a reversible third-party encryption system may be a solution to this problem. However, they have not provided a route or even examples of how to proceed. We present here the Icelandic approach to this issue by developing a third-party encryption system in direct collaboration with the Data Protection Commission (DPC) of Iceland. We have incorporated the encryption system within our sample collection and storage software, which minimises inconvenience but enhances security. The strategy assures a barrier between the laboratory and the outside world that can only be crossed by the DPC. *European Journal of Human Genetics* (2000) **8**, 739–742.

Keywords: ethics; genetics; encryption; third-party; privacy; database; sample bank; biobank

Introduction

Society has paid considerable attention to the importance of protection from unauthorized or unethical use of personal information gathered in the process of delivering health care. Most countries have paid little attention to the importance of protecting personal information on health and disease gathered in the process of medical research. This may be because such information gathered in research has allegedly never or rarely been used to harm people.[1] However, even if true, this does not in our view reduce the need to protect the privacy of people who are the objects of medical research. In addition to upholding the basic rights of participants, there are at least two compelling reasons for developing methods to protect their privacy. One is that participation in medical research is almost universally voluntary; the sole exception is epidemiological research using information gathered in the process of delivering health care. Therefore just a few cases of abuse of information collected for research could deter people from participating in scientific studies. The second reason is that it is relatively easy to develop systems to protect the privacy of subjects of medical research because such research is always aimed at gaining knowledge about the nature of a group of people rather than information about individuals. Data gathering for medical research tends to be different in nature from that in health care, which aims to produce readily accessible information about individuals and is used to serve them.

Collection of information and blood samples and how they are used in research challenges the protection of privacy. This is especially true today when there are fears that governments, insurance companies, and employers may use or abuse genetic information, even when gained in research in a non-clinical laboratory. Almost all genetic research carried out today is without encryption of personal identifiers attached to the samples or medical data. Only in recent years has there been a formal call to consider a third party encryption system.[2,3] The American Society of Human Genetics ethics committee recommended researchers to 'consider a way of coding samples by a third independent

Correspondence: Jeffrey R Gulcher or Kári Stefánsson
Tel: + 354 570 1900; E-mail: jgulcher@decode.is; kstefans@decode.is
Received 17 February 2000; revised 26 May 2000; accepted 7 June 2000

party who would keep the codes inaccessible unless there are specific circumstances in which the code needs to be broken'.[4] A set of guidelines recently published in *Science* was less explicit: 'Current practices to protect confidentiality of experimental research data should be studied and best practices should be developed.'[5]

In Iceland deCODE genetics and the Data Protection Commission of Iceland (DPC) have developed a third-party encryption system supervised by the DPC. It has been used for all disease-based gene discovery projects at deCODE for over three years. Here we show that a sophisticated third party encryption is feasible with little inconvenience to the researcher if integrated with sample encryption and storage software. This describes how deCODE proceeds in its current disease-based projects and does not apply to the Icelandic Healthcare Database.[6–8]

Materials and methods

At deCODE genetics all genetic studies begin with collaboration with the physicians who attend patients with a particular disease (see Figure 1). The research consortium formed by the company and the physicians must then receive permission from two national committees – the National Bioethics Committee and the DPC – before it can proceed. The National Bioethics Committee and DPC are appointed by the Ministry of Health and the Ministry of Justice, respectively, and consist of lawyers, ethicists, physicians and other representatives of the community. The physicians create for the DPC a population-based list of the patients with a particular disease. The list is reversibly encrypted by the DPC using a 128-bit symmetric encryption algorithm – TwoFish – a candidate for the Advanced American Encryption Standard (AES).[9] This process converts the social security

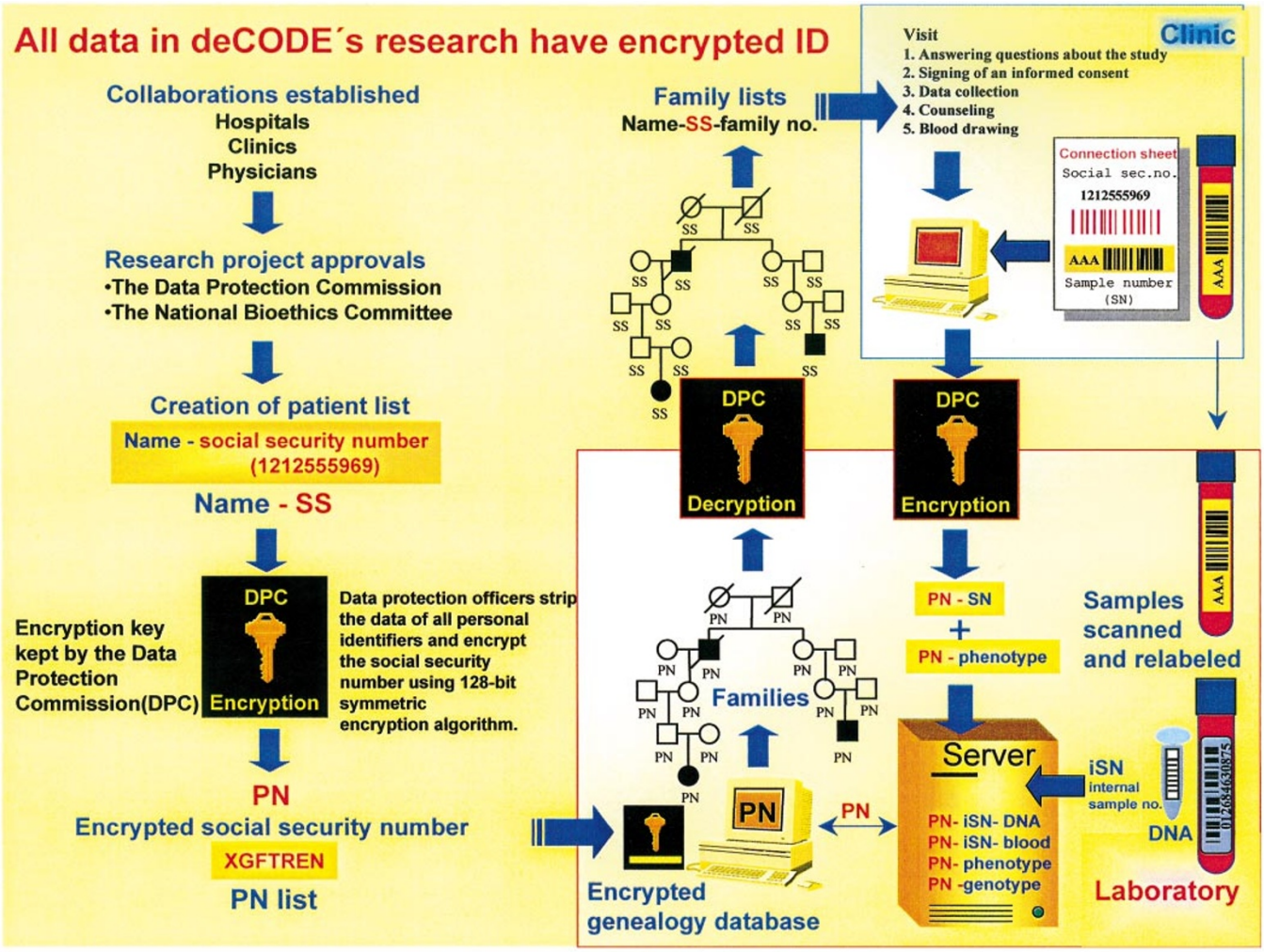

Figure 1 The figure shows the reversible third-party encryption system that is used for disease-based projects in Iceland. The DPC (data protection commission) is the only information avenue in to or out of the laboratory, since they hold the key for encryption and decryption of personal identifiers attached to phenotypic information and samples.

number (SS) to an alphabet-derived character string (PN). For example, SS number 2802343344 might become TZRXMRW. The DPC delivers the PN list of patients along with the phenotype classification to the laboratory. The laboratory then runs the PN list against a population-based computerised genealogy database that has been encrypted by the DPC in the same way (the uncoded computerised genealogy database is updated and maintained outside the laboratory). Those patients who are most closely connected within a certain number of meioses (dictated by the density of markers used) are selected to make up deCODE's 'wish list' for the initial stages of our genetic study. This list is handed over to the DPC which decodes the list off-site, generating a list with the social security numbers of the patients. The patients are contacted by the physicians by phone or letter, inviting them to participate. The patients who are willing to participate come to a clinic run on our behalf by the Data Protection Commission of Iceland and staffed by nurses who can answer questions related to the informed consent form and medical questionnaire. From those who sign an informed consent, blood is taken directly into vacutainers that are labelled only with barcode stickers representing a third number, the sample number (SN) which acts as a temporary coded identifier. The SN is distinct from the PN and is selected at random from a preprinted roll of sample labels. The SN is immediately scanned from the barcode on the tube into a computer and the patient's social security number keyed in the presence of the patient. This establishes the link of SN to SS. When the blood is taken at a facility where immediate scanning of the sample into a computer is not possible, another barcode sticker with the name SN is placed on a printed form (connection sheet) with the patient's social security number. The stack of connection sheets is subsequently scanned into a computer at a facility outside deCODE under the supervision of DPC. Before the blood is sent to the laboratory, the DPC officer encrypts the list of SS – SN to PN – SN. This establishes the link of SN with PN. The PN – SN list is sent on a sealed computer diskette along with the blood to the laboratory. The PN is not used outside the laboratory and direct personal identifiers, such as names or SS, never enter the laboratory.

Once in the laboratory, the tubes of blood are scanned into a sample storage program and the temporary SN number is replaced by relabelling with a second and more permanent in-house sample number (iSN), barcoded, with the iSN clearly visible. This iSN will remain with the DNA samples isolated from the blood and is directly linked by the sample storage program to the PN used to label individuals within the genealogy database. This re-labelling of the tubes of blood excludes any link between the PN and sample numbers seen and perhaps recorded on the clinical side which might have been improperly connected to a direct personal identifier. The temporary SN numbers are limited in number and are rotated every few weeks. Therefore, the only connection between samples or data in the laboratory and the patient or data on the clinical side through the PN and the DPC, the sole keeper of the encryption code.

Note that the original list of individuals with their diagnoses is considered to be epidemiological data. Therefore, informed consent from each individual is not required before we crossmatch it with the genealogy database. However, individual informed consent is indeed required for both the generation and cross-matching of any molecular genetic data. The informed consent form covers only one disease or a collection of closely related diseases. Individuals from families who have other diseases must be re-contacted by the DPC-run clinic and asked to sign another specific informed consent form before the laboratory is allowed to use their blood sample or genotypes previously acquired for another disease.

Results

This system ensures a solid wall between those who work in the laboratory and those in the outside world. No personal identifiers associated with genetic or medical data ever reach the laboratory; the encryption labels, PN and iSN and the genetic data on individuals never leave the laboratory. Just as for the blood samples, communication about phenotypes and genealogy is by encryption or decryption of the SS or PN, respectively. The DPC holds the key. Our system was built with scientists in mind as well. We should not unnecessarily burden them with genetic information linked to personal identifiers, especially in a small community like Iceland where a scientist may be related to or be an acquaintance of the patient. We have successfully used this system over the last three years to work with patient lists covering almost 30 common diseases averaging about 2500 patients per disease along with blood samples from over 35 000 Icelanders.

Discussion

The issues of possible access by insurance companies and government to research information raised in a recent article in *Science*,[5] can be addressed by a third party encryption system such as the one we described. In general, medical research can proceed without scientists knowing the identities of participants in the study. One approach has been to anonymise irreversibly the samples by stripping off all labels. However, this may not be the most ethical and practical approach for several reasons. First, the information derived may later benefit the patient, and the patient or the physician may later request access to the research data.[3,4,10] For example, studies anonymously testing the HIV status of cohorts have been criticised for not offering participants the results of their test if they so desire.[11] The system that we have described here allows for high security of research results, but there remains recourse to obtain the information later on request from a patient if ever the data were to become clinically relevant. Second is the issue of the individual's right to know and not to know the genetic

information. While the system we have in place incorporates research data that will need to be validated later to become clinically useful, our system could be applied to a clinical genetics program.[12] Third, anonymisation of data prevents the prospective addition of clinical data or samples from the participants in the future for research, which greatly reduces the long-term value of the research to the community. The third party could be a representative of the local institutional ethics committee such as an Institutional Review Board (IRB) or a national or regional ethics or data protection board. This system also decreases the chance of bias in the phenotyping of patients by providing a simple way to blind the physician to the genotypes of his own patients.

Informatics systems set up in this way can keep the intrusiveness of encryption systems more manageable, even for a small laboratory. Furthermore, the major cost of such a system once implemented is essentially the labour involving the actual encryption and decryption by a trusted third party or representative. We estimate that it takes about 3 hours for a list of 1000 individuals to be encrypted or decrypted, including checking and verifying the personal identifiers. A couple of computers (one for the laboratory and one for the third-party) with associated bar-code reader systems would be needed. However, a core encryption facility within an institution may increase efficiency. Such costs are likely to be nominal compared with the laboratory costs of genetic research. Sample handling and storage costs within the laboratory could be lower than with conventional systems if the encryption system is integrated with sample management software as in our case. Also, we have designed software that enables secure remote third-party encryption, thereby dispensing with an independent representative at each encryption session of personal data; thereby our encryption process would be even less intrusive. Description of this software is beyond the scope of this paper, but we intend to publish its design and use in due course.

This system protects the individual from invasion of privacy since personal identifiers are encrypted. The system protects ethnic groups better than the current method of little or no coding with personal identifiers, although it does not prevent indirect tagging of a group via alleles. But irreversible anonymisation would also fail to guarantee that protection. In our view legislation is needed in most countries, not to regulate creation of new knowledge, but its application.

In our quest for privacy for subjects of medical research it is important to keep in mind that when they volunteer to participate in a study and sign properly written and executed informed consent forms, they register faith in scientists and the science. They may not want anonymity and scientists may seek inspiration in personal contact. It is clear that a third-party encryption has a price. In this article we have described one that works and we have chosen to use. However, the benefits and disadvantages of a third-party encryption system must be weighed by each research group and organisation where the research is carried out.

References

1 Axelson O, Tondel M: Concerns about privacy in research may be exaggerated. *Brit Med J* 1999; **319**: 706.
2 HUGO-ELSI Committee: Statement on the principled conduct of genetic research, 1996. http://www.gene.ucl.ac.uk/hugo/conduct.htm
3 Knoppers BM, Hirtle M, Lormeau S, Laberge CM, Laflamme M: Control of DNA samples and information. *Genomics* 1998; **50**: 385–401.
4 American Society of Human Genetics: Response to the NBAC on ethical issues surrounding research using human biological samples, 1999, http://www.faseb/org/genetics/ashg/policy/pol-33.htm
5 Fuller BP *et al*: Privacy in genetic research. *Science* 1999; **285**: 1359–1361.
6 http://www.database.is
7 Gulcher J, Stefánsson K: Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin Chem Lab Med* 1998; **36**: 523–527.
8 Gulcher J, Stefánsson K: The Icelandic Healthcare Database: A tool to create knowledge, a social debate, a bioethical and privacy challenge. *Medscape* 1999. (http://molecularmedicine.medscape-.com/10835.rhtml)
9 Two-fish: a new block cipher, 1999. http://www.counterpane-.com/twofish.html
10 HUGO Ethics Committee: Statement on DNA sampling: control and access, 1998. http://www.gene.ucl.ac.uk/hugo/sampling.html
11 de Zulueta P: Reducing the vertical transmission of HIV. Anonymous testing is unethical. *Brit Med J* 1998; **316**: 1901.
12 Stone DH, Stewart S: Screening and the new genetics; a public health perspective on the ethical debate. *J Public Health Med* 1996; **18**: 3–5.