The DeCODE Proposal for an Icelandic Health Database

Ross Anderson

October 20, 1998

1 Executive Summary

I have been invited by the Icelandic Medical Association to evaluate the privacy aspects of DeCODE's proposal for a central database of Icelanders' medical records, genealogy and genetic data. The primary use of the proposed database is research into hereditary diseases by or on behalf of drug companies; its secondary uses will include providing management information to the health service and supporting other research.

Of the three components of the database, the genealogies are essentially public domain, and the genetic data will be gathered from patients who have given their consent to its use in research. The medical records will, however, be collected from hospitals and health centres, de-identified only to the extent that obvious identifiers such as names and social security numbers will be replaced with a single pseudonym. Patients will have the right to opt out of the database, but will not be asked to give explicit consent.

This creates a serious conflict with medical ethics and with data protection principles, both of which demand that with few exceptions, patients' consent be sought for the use of their personal health information.

Many countries permit data which have been made anonymous to be used in certain circumstances without consent. For example, health service managers routinely gather statistics such as numbers of operations and consumption of drugs. These statistics are typically compiled from current records which give only a snapshot of healthcare activity at a certain time or over a short period; de-identifying such records is relatively easy.

Some countries maintain databases of de-identified medical records which link together all, or many, of the health care encounters in a patient's life. Such records are in practice impossible to de-identify completely, as the combination of data is frequently enough to identify the patient. They do not even meet the more usual test of requiring unreasonable effort by an attacker who wishes to identify a patient. It is therefore necessary to have quite extensive controls to prevent abuse.

For example, New Zealand maintains a database called the National Medical Data Set which contains most citizens' health records, identified by an encrypted social security number. In addition, the system limits access to a small group of health service statisticians, limits the type of enquiry that can be made, and rejects any enquiry which would be answered by reference to the records of less than six patients. Even in the presence of such controls, special administrative measures are also thought necessary; all the national databases of which I am aware are operated by government agencies, and in many cases special legislation, or data protection regulation, is thought necessary.

But, for a number of reasons, the database proposed in Iceland lies well beyond the limits of established precedent:

- Firstly, the proposed system will be intrinsically more difficult to protect than existing health databases, because it contains information on genealogy as well as health, and because it is proposed to allow access to a large and transient population of commercial subscribers, rather than a few carefully vetted statisticians;
- Secondly, the proposed database will also be available to the Ministry of Health for tasks such as cost control. This will be the first time that medical records on Icelanders are available centrally to the government rather than being kept locally in health centres and clinics. This raises a number of ethical and other issues, which appear to have escaped debate;
- Finally, despite this environment of greatly increased privacy risk compared with existing systems, the measures which are proposed to limit the scope of users' enquiries, and to provide technical protection in other ways, are not credible. There is not even enough information about the proposed use of the database to determine whether effective protection measures are feasible.

In conclusion, the proposed database falls outside the boundaries of what would be acceptable elsewhere in Europe. If established as proposed, it would likely cause serious conflict with the ethical principle that identifiable health information should only be made available with the consent of the patient.

I therefore recommend that the Icelandic Medical Association oppose the current bill. This need not rule out the possibility of supporting an amended proposal, in which the uses of the database are clarified and appropriate security measures included.

Finally, I wish to point out that the proposed database is also in conflict with established data protection principles. If data protection authorities overseas acquire the view that Iceland is a country in which normal data protection controls can be bypassed easily by powerful vested interests, then this could have extremely grave consequences for trade and development. I therefore caution Icelanders against considering the matter to be a simple choice between national development and medical privacy.

2 Introduction

DeCODE Genetics Inc has sponsored legislation, currently before the Icelandic parliament, which would enable it to construct a database of Icelanders' medical records, genealogy and genetic data [1]. The stated objective is to facilitate research into hereditary diseases and thus enable DeCODE's clients, who will be mainly drug companies, to develop and test new products [2]. A number of secondary uses are envisaged, such as providing management information to the health service and supporting other research.

Of the three components of the database, the genealogies are essentially public domain, although the genealogical database being developed for DeCODE may be much more complete than the online sources which are currently available. This component of the database appears to have few privacy implications, as the underlying paper records are publicly available.

The genetic data will be gathered from patients who have given their consent to its use in research (there was an implication that historical data might be used, such as pathology samples from post mortems [3], but this appears to have been dropped). Privacy protection is a requirement for this data, in order to prevent its use in applications for which the patients have not consented.

As for the medical records, it is proposed that they will be collected from hospitals and health centres, de-identified to the extent that obvious identifiers such as names and social security numbers will be replaced with a single pseudonym (an encrypted social security number), and provided to the database [4]. Patients will have the right to opt out of the database, but will not be asked to give explicit consent [1].

Non-consensual secondary uses of medical records raise very sharp ethical concerns, which can sometimes be dealt with by de-identifying the records. The usual test for this technology is whether it will take an unreasonable amount of time and effort to identify a patient in the information that is subsequently made available. However, de-identification is not a panacea and it is important to understand its limits.

3 When are de-identified data not anonymous?

Firstly, although it is not too difficult to de-identify data that provide only a time-limited snapshot of a population's health – such as the data which health services use to compile monthly management statistics of numbers of operations, consumption of drugs and the like – it is effectively impossible to de-identify longitudonal records, that is, records which link together all (or even many) of the health care encounters in a patient's life. Someone wishing to abuse the database to investigate a business or political rival, for example, is likely to know some facts about the target of investigation (that he broke his ankle playing football on the 14th October 1974, that he was absent from Iceland for 1978-1982 doing postgraduate work, and so on) and wish to know other facts (such as whether he has ever been treated for alcoholism or for psychiatric disorders). In many cases, the known facts will enable the target patient to be identified despite the use of a pseudonym in the database itself [5] [6].

For this reason, a database of longitudonal medical records must be considered to be personal health information; although some of the patients may be protected by the use of pseudonyms, many will not be. In a database which also contains genealogies, individuals will be even more easy to identify; one could search for people with four uncles, two aunts, eight great-uncles, seven great-aunts, etc, and if the data for several generations are available then most groups of siblings could be identified.

This point – that the database contains identifiable medical information – was readily conceded by DeCODE management on the 12th October during a briefing at the Medical Association [4], although a subsequent press release claimed that the concession had not been made [7].

So those countries whose health services maintain central databases of medical records, such as New Zealand, do not consider pseudonyms to be enough protection. There are also stringent use controls. The New Zealand system, as noted above, limits access to a small group of health service statisticians, limits the type of enquiry that can be made, and rejects any enquiry which would be answered by reference to the records of less than six patients [8].

There is a large literature on such mechanisms, or 'inference security' as the subject is called. The basic ideas were initially developed by the world's census bureaux to prevent statistical enquiries made of census databases being abused in ways that could leak information about identifiable individuals. It is of critical and growing interest to medical research organisations as well, and is being actively promoted by data protection authorities in Europe and elsewhere [9]. The standard introductory textbook is [10].

It is not sufficient to merely require that enquiries be based on a minimum size of query set; one must also ensure that combinations of queries cannot be used to identify individuals. For example, it might be possible to make one enquiry about the target plus ten other individuals, and a second about the ten others (see [10] for many more complex examples and powerful attack techniques). Common protection mechanisms include logging and analysing queries, adding noise to the underlying data, making each query depend on a pseudorandomly selected fixed size subset of the data, and suppressing records with particular data values (such as census records indicating very high incomes, or in the medical context, subscriptions for AZT). None of these techniques will prevent all possible inference attacks, and whether a system provides an adequate level of protection depends closely on the nature of the application.

Systems that use de-identified data fall into two broad categories. In the first, the data are processed once and for all to remove identifiers and then released for arbitrary processing by untrusted programs. An example of this is given by the US census, which has in the past distributed a tape containing the record of one household in every thousand, with the names and exact geographical locations removed. In the second, the data are held in a trusted system and only a restricted set of enquiries are permitted; an example of this is the New Zealand medical records system mentioned above.

In both kinds of system, effective de-identification depends on detailed knowledge of the application. For example, I recently evaluated on behalf of the BMA a proposed system for collecting de-identified data from pharmacies for resale to drug companies. In this case it was required to protect the privacy of doctors as well as patients. The original design had proposed grouping doctors into cells of about 20 doctors, within which they would be known as 'doctor 1', 'doctor 2', and so on. The system would provide total prescriptions of each drug per week. However, it was possible for an experienced drug salesman to look at the figures and say, for example, "Doctor 7 must be Susan Smith, because she went skiing during the second week of February, and look at the drop off in prescriptions there." So the system had to be redesigned to show percentage market share rather than absolute volumes (and with other controls as well).

The above prescription system is of the first kind (pre-process then release). The DeCODE proposal is of the second kind; the data held in the database are in many cases identifiable, and privacy depends on the mechanisms used to restrict queries. This makes it necessary to control the kind of programs which an enquirer can run on the database. For example, if the system merely compelled enquiry programs to read at least ten records, then an attacker who wished to find out about a target patient might write a program which read the target patient's record and those of nine others selected at random, and then returned the value '1' if the target were an alcoholic, '2' if he had received psychiatric treatment, '3' if both and '0' if neither. For this reason, arbitrary enquiries should not be permitted; the database user must not have access to a query language that is Turing powerful (this is a well known concept in computer science for describing a language that is as powerful as a general computer, in the sense that one may write arbitrary programs in it.)

4 Why the DeCODE Proposals are Inadequate

This leads to the reasons why I consider the security proposals made by DeCODE to be unsatisfactory, and the level of technical expertise shown by them so far to be inadequate.

The point that users must not have access to a Turing powerful query language is a point that DeCODE have failed to understand; at the 12th October briefing, it emerged that their technical expert did not even understand the phrase 'Turing powerful'. I am convinced that this is not simply a linguistic misunderstanding, as even after I had explained the requirement for user queries to be strictly limited, and the difficulty of doing so, during the morning on the 12th October, DeCODE continued to maintain at a further meeting during the afternoon that writing a filter to police user queries would be simple.

A security expert should have been aware that this is not the case. For example, much of the expenditure in banking computer security relates to extensive quality control procedures whereby all programs are examined and tested by multiple independent people, to reduce the risk that a programmer could credit a large sum of money to his own account. Another example comes from military computer security, where systems prevent information flows from a higher security level to a lower one independently of the application programs, in order to prevent an application programmer from writing code that could leak information. Yet another example is given by the popular 'Java' programming language, which is designed in order to let users download programs from the Internet and run them in their web browsers with relatively little risk that these programs could steal personal information, destroy data or otherwise misbehave. In short, the problem of which software one must trust, and to what extent, is the central issue in computer security.

The other security proposals by DeCODE, and in particular the claims made about encryption, also indicate a lack of expertise:

- it was claimed that one-way functions can be used to process social security numbers and thus turn them into pseudonyms. However the file of Iceland's 280,000 or so social security numbers is publicly available, and an attacker could simply pass them through the one-way function and build a look-up table to link numbers with pseudonyms. When this was pointed out, DeCODE claimed that the one-way function would involve a different key at each hospital or health centre, and that a trusted party such as the data protection commission would then translate these institution specific pseudonyms into nationally uniform pseudonyms for the database. But in that case, the appropriate mechanism would not be a one-way function, but a block cipher (the use of a one-way function would compel the trusted party to use the key to build a look-up table for decryption as described above);

- it was also claimed that the disease codes would be encrypted by a public key, so that they would be coded in the database. But then anyone could use the public key to encrypt the known ICD disease codes giving a look-up table to decrypt the database. When this was pointed out, DeCODE claimed that the public key encryption would include a random number to prevent this. But then how would the codes in the database be accessed by authorised users? We are told that the trusted party would have the private key and decrypt them. But in that case, again, the appropriate mechanism would not be a public key encryption function, but a symmetric block cipher (with under 100 healthcare providers in Iceland, the use of public key mechanisms is hard to justify);

- most of DeCODE's presentation slides on cryptography were not shown to me at the 12th October briefing, on the grounds that 'you know this stuff anyway'. The exception was a slide in which it is proposed to guard against the risk of a breakthrough in cryptanalysis by using three block ciphers (DES, IDEA and RC5) one after the other. This idea is suggested by outsiders from time to time, but has not appealed to professional cryptologists for many years (only if ciphers commute can one prove that their composition is no weaker than any of the components, and block ciphers should not commute);

- it is claimed that a separation of duty policy can be enforced in the database, in order to prevent system administrators having access to the full patient records, by encrypting different families of disease codes under different symmetric keys, and by encrypting the genealogic and genotypic databases with different keys. I am very sceptical of this claim; having experience of designing databases which use encryption for copy protection, I am aware of many difficulties that need to be overcome and of which DeCODE appear unaware. In any case, the principal issue with the database is not encryption but how one controls the programs that are run on it and the people who have access to the program output.

  For example, I cannot accept the claim that encrypting some of the records with different keys will prevent system administrators having access to the database. If the decryption is performed in software, the system administrators would have access to the keys; if it were performed in tamper resistant hardware, they would still have access to the plaintext whenever it was decrypted; and if all the processing were performed in a tamper-resistant computer, then the system administration of this computer would now become the issue. Automating system administration might be a solution eventually but is a long way off in practice.

For these reasons, I cannot accept DeCODE's claim to have adequate expertise in computer security, or their claim that they do have adequate security plans but that these have simply not been disclosed to me [7]. The lack of competence at computer security is quite evident in their proposal.

5 Should a Health Database be Built?

The question that now arises is whether, given access to security expertise, the problems could be fixed.

As the New Zealand example has shown, it is possible to construct and operate a national healthcare database in a way that satisfies both medical and privacy interests. The obvious question to ask is whether a database can be built which would deliver adequate value to DeCODE and its customers for the exercise to be worthwhile, and also provide adequate privacy protection.

As noted above, in order to design or evaluate a de-identified health record system, it is necessary to have a detailed understanding of the use which will be made of the data.

I have had significant difficulty in finding out precisely what the database will be used for. The DeCODE proposals are not only very vague, but different accounts have been given at different times to different people. Their 'non-confidential corporate summary' claims that the database will be marketed for two principal uses: to design disease management programs and to search for drug targets through genotypic/phenotypic correlation. Other claims are to 'assess interplay of genes encoding members of a pathway' and to 'identify biological pathways that are affected by a particular disease, into which a gene product fits, (or) that provide approaches to the search for drug targets'.

It is envisaged that subscribers to the database – which DeCODE said at the briefing on the 12th would be a large and changing population of users – would be able 'to perform in silico mapping of individual disease genes as well as to determine how constellations of genes influence pathogenesis, natural history, response to treatment and complications of diseases'. These users will include pharmaceutical companies, biotechnology companies and insurance firms.

This would appear to require that analysts would be able to make very complex enquiries of the database and would thus need a powerful query language. However, it is in stark contrast with the version we heard on the 12th, following ethical objections by the IMA and others. We are now told that the database will not be used to identify possible subjects for genetic investigation, and that queries will only be answered if they are based on the records of ten or more individuals. When I asked what sort of queries could be made under such restrictions, the example given was 'what is the likelihood that someone diagnosed with a disease such as asthma, and who has had a cancer case in the family, will also develop cancer?' This could indeed be done with simple, restricted queries, but one wonders whether it would justify the investment.

When I pressed for more details, the example I was given was that a disease might be traced to a certain marker on a certain chromosome by correlating available health records, genealogies and genotypic data. But as genotypic information is only available on patients who have given consent for their doctor to enter them in DeCODE's research programmes, such enquiries do not appear to require the records of patients who have not given consent and thus the proposed legislation is not required.

There thus remains the serious concern that if DeCODE were to construct a database which supported only very restrictive queries then they might find it uneconomic and would be forced to extend its functionality to that originally envisaged in [3].

6 Access by the Ministry of Health

The bill provides for the database to be made available to the Ministry of Health for administrative purposes. These will presumably include cost control, clinical audit and other tasks related to the performance analysis of health providers and perhaps individual health service staff. There may also be public health missions. While many of these tasks are unobjectionable, and may be performed using a mechanism along the New Zealand lines, there appears to have been no public discussion of the issues (e.g., what sort of institutional arrangements will be necessary to prevent 'function creep').

I understand that this will be the first time that medical records on Icelanders have been available centrally to administrators rather than being kept locally in health centres and clinics. It is prudent to see to it that there is an open and informed public debate on the issues; if Icelanders simply wake up one morning and realise that the Ministry of Health has a copy of the medical record which they believed to be kept safely in the local health centre, then the reaction could be disruptive and harmful.

In the UK, some health information systems were developed without consultation and then apparently adapted to unethical ends. For example, in 1996 the BMA became concerned that police access to prescription records, which had been granted in order to trace doctors and nurses who were stealing heroin, was being used to search for illegal immigrants. There were other problem systems too. The resulting public row led to the establishment of a commission to look into secondary uses of health records, and the development of health networking was held up for over a year as this commission deliberated.

I can assure all parties in Iceland that such an experience is to be avoided if at all possible. In order to maintain trust between doctors and patients, between administrators and public health professionals, and between politicians and the healthcare sector generally, it is much better to have the necessary debate before such systems are built rather than afterwards.

7 Recommendations to the Medical Association

In the initial opinion I gave to the Medical Association following the meetings on the 12th October, and in the interviews I gave to the media, I went out of my way to give DeCODE the benefit of the doubt. Rather than simply dismissing the proposals as unacceptable (which in their current form they are), I considered it better to give DeCODE the opportunity to step back and consider whether they can produce a system that would respect the ethical constraints and still be a viable business asset. Despite the abusive tone of their press release [7] I feel that this is still an appropriate response.

I understand that DeCODE decided to delay the detailed design of the database until after the bill was passed. In my view, this is unacceptable. It is unclear that a database can be built that is simultaneously ethical and useful for the purposes DeCODE claim to have in mind. If the bill is passed, and it turns out that an ethical useful database is impossible, then a likely outcome is an unethical but useful database. Even if an ethical useful database is possible, parliamentary endorsement of DeCODE's current plans might embolden its management to cut corners in order to save money.

I therefore recommend that the Icelandic Medical Association insist that DeCODE produce a functional specification of the database, and a security specification, which are sufficiently detailed for an independent evaluation to be carried out. This will mean, at the very least, specifying which data items will be be stored, what the restrictions on processing will be, and how they will be enforced.

I also recommend that the Icelandic Medical Association insist that the custodian of the health data should be a body which they consider to be trustworthy. This might mean vesting control in the Chief Medical Officer or the Data Protection Commissioner; or it might mean keeping health records distributed in the health centres and hospitals and having a mechanism allowing queries to be sent to them. The latter kind of system is used in the UK and it is most helpful in maintaining medical confidence: doctors can observe what sort of queries are being made and can always unplug the modem if they believe that the system is being abused.

In the absence of a functional specification, a security specification and a trustworthy custodian, my recommendation is that the Medical Association oppose the current bill and, should it be passed, advise members not to cooperate with the resulting data collection exercise. This need not rule out the possibility of supporting an amended proposal which satisfies an independent evaluation.

8 International and Economic Issues

There is also an economic point that the people of Iceland ought to consider. The future prosperity of Iceland, as of every country, is tied up to some extent with the information economy. For this reason, it would not be prudent to do anything that is seen as a grave breach of the letter (or even the spirit) of European data protection law. Even though governments can in theory grant exemptions to this law on the grounds of national interest, such exemptions are designed for police and national intelligence purposes. But transferring the medical records of non-consenting patients to a private company, which will sell access to them to clients who are outside the European Union, will be seen as outrageous. I quote a statement made about the DeCODE proposals by the Data Protection Commissioners of the EU and EES countries made in Santiago de Compostela in September 1998 at the 20th International Conference on Data Protection:

‘(The Commissioners) stress the importance of the following elements:

- the principle of free and informed consent of the person concerned to the storage and further processing of his or her data must be fully respected. The data subject must also be given the right to withdraw from the base once his or her data have been entered. Exemption from these principles would only be acceptable for exceptional reasons and with adequate safeguards for the correct use of the data.

- the definition of "personal data" must be explicitly clear and the method of securing anonymity must be effective. In a country with a relatively small population, information or genetics is likely to indicate biological lineage and to reveal identities of persons concerned. The use of a code to replace identifiers is in any case not sufficient to secure anonymity.
- the commercial interests of the user must not lead to expansion of the original purpose of the register.

They express their serious concern about the matter and recommend the Icelandic authorities to reconsider the project in the light of the fundamental principles laid down in the European Convention on Human Rights, the Council of Europe Convention 108 on Data Protection and Recommendation (97)5 on medical data, and the EC Directive 95/46 on the protection of personal data.'

9 Conclusion

In my opinion, the privacy protection which the DeCODE database appears likely to provide falls well short of the minimum standards demanded elsewhere in the developed world, and supplying information to it will thus raise severe ethical problems. The Icelandic Medical Association should therefore oppose it.

There is also a grave risk that this bill, if enacted, will undermine confidence in Iceland's ability to be trusted with the processing of personal information from other countries. This would isolate Iceland from the EU's information economy, and may well impose costs on the Icelandic economy which will greatly exceed any benefits.

References

[1] Draft – Bill on a Health Sector Database, available from `<http://brunnur.stjr.is/interpro/htr/htr.nsf/pages/gagnagr-ensk>`

[2] Home page of DeCODE Genetics Inc, `<http://www.decode.is>`

[3] "A non-confidential corporate summary", DeCODE Genetics Inc., 7th June 1998

[4] Presentation by DeCODE Genetics Inc., Icelandic Medical Association, 12th October 1998

[5] 'Security in Clinical Information Systems', RJ Anderson, British Medical Association, January 1996

[6] The Caldicott Report, Department of Health, UK, 1998

[7] Press release from DeCODE Genetics Inc., 13th October 1998

[8] "Managing Health Data Privacy and Security: A Case Study from New Zealand", R Neame, in Personal Medical Information – Security, Engineering and Ethics, Springer (1997), pp 225-232

[9] 'Privacy-Enhancing Technologies: The Path to Anonymity', Information/Privacy Commissioner, Ontario, Canada, and Registratiekammer, The Netherlands, August 1995; available from `<http://www.ipc.on.ca/web_site.ups/matters/sum_pap/papers/anon-e.htm>`

[10] 'Cryptography and Data Security', Dorothy Denning (Wiley, 1983)