The use of probabilistic record linkage, public key cryptography and trusted third parties to improve the protection of personal privacy and confidentiality in disease registers and tissue banks

Tim Churches

Medical epidemiologist, Epidemiology and Surveillance Branch, New South Wales Health Department, Locked Mail Bag 961, North Sydney NSW 2059, Australia, email tchur@doh.health.nsw.gov.au, ph: +61 2 9391 9193, fax: +61 2 9391 9232

Abstract

Disease registers (DRs) aim to collect information about all instances of a disease or condition in a defined population. Traditionally DRs have required that notifications of cases of the target diseases be fully identified with items such as name and date of birth so that multiple notifications relating to the same case can be identified and merged. However, growing concern over the privacy and confidentiality aspects of DRs is beginning to hinder their operation, particularly in Europe. An alternative method of operation is proposed which involves splitting the personal identifiers from the medical details at the source of notification, and separately encrypting each part using asymmetrical public key cryptography. The identifying information is sent to a single population register (PR), and the medical details to the relevant DR. The shared PR does not need to capture identifying details of every person in the population, only those of people notified to a DR. The PR uses probabilistic record linkage to assign a unique personal identification (UPI) number for each person notified to it. This UPI is shared only with a single trusted third party whose only function is to translate between the UPI and separate series of personal identification numbers which are specific to each DR. The proposed scheme, which extends an algorithm described by Blobel et al. in 1995 for use in German cancer registries, would also allow linkage of records between DRs with minimal extra effort, under the supervision of a trusted third party. The scheme is directly extensible for use with tissue banks and other repositories of genetic material as well as disease registers and other health status or health event data collections. With the exception of a probabilistic record linkage engine, all of the components required by the proposed scheme are raedily available in the form of reliable and well-tested free, open source software. It should be possible to retrofit existing health information systems to interoperate with the proposed system without enormous effort or expense.

Introduction

It now appears that wealthier nations are on the threshold of a minor revolution in the delivery of health care – a revolution mediated by highly affordable computers, electronic health records and ubiquitous access to national and international networks such as the Internet. These developments promise substantial improvements in the quality of health care and the efficiency with which it can be delivered. They also present a number of concomitant challenges. One particular challenge – that of protecting individual privacy and maintaining confidentiality in an environment in which large volumes of health information can be copied and transmitted ad infinitum in just seconds – is attracting increasing attention from health care providers, regulators and consumers alike.

An example of these concerns can be found in the recent debate in Britain (and elsewhere) over the automatic transfer of personal health information to disease registers, either with or without explicit and informed consent [1] [2] [3] [4] [5].

This paper does not attempt to address the philosophical and societal issues underlying such debates. It does however describe a technological solution which would enhance the protection afforded to the large volumes of often highly confidential personal health information which disease registers necessarily accumulate.

Cancer registries are perhaps the best known type of population-based disease register. However, over the last few decades, a number of other types of disease register have been established in many countries. These include registers of birth defects, diabetes and chronic infectious diseases.

The core function of all such registers is to measure the incidence or prevalence of their respective target diseases or conditions in a defined population. Some registers have additional functions, such as providing population-based cases for case-control or cohort epidemiological studies, or collecting information which can be used to monitor the effectiveness of the treatment and clinical management of their target disease.

Traditionally, disease registers have required that health service providers notify them of each instance of the target disease or condition by forwarding details of the instance, together with identifying information for the person to whom it relates.

Notifications to most disease registers need to be identified to enable the register to assemble a single record for each unique case of the target disease from the multiple notifications which might be received about that case. For example, a patient might

receive a clinical diagnosis of a particular type of cancer from their general practitioner, who will send a notification to the relevant cancer registry. A fine needle biopsy of the tumour may be taken, and this will result in a histopathology laboratory sending another notification to the cancer registry. The patient may then be admitted to hospital for surgery, which results in yet another notification of the same case to the cancer registry. Such redundancy in the notification process is desirable because it minimises the likelihood of a case being overlooked by the disease register, but also means that the disease register must be able to determine that all these notifications relate to a single case of disease in an individual patient.

Disease registers have successfully used this method of operation for many decades. However recent advances in computing, cryptography and communication networks have made alternative methods of operation possible.

Related work

The idea of linking files by a trusted third party using only "identifiers" appears to have been first described by Boruch and Cecil [6] in 1979 in the context of linking social surveys to administrtaive data held by social services agencies. Subsequently, Pommerening, Miller, Schmidtmann and Michaelis [7] described a similar method, to be used for improving privacy and security in cancer registries, necessitated by changes to German privacy legislation. Their method involves dividing the cancer registry into two operationally distinct offices. The first office receives notifications and handles all communication with notifying health care providers. The personal identifiers on each notification are encrypted before passing the records to the second office which links the new data to its database using the encrypted identifiers. Blobel provides further details of this system [8]. A related method for protecting the confidentiality of clinical records through "secret splitting" was later described and patented in the USA by Ho [9] [10]. Kohane, Dong and Szolovits [11] have described a system, the Health Information Identification and De-identification Toolkit (HIDIT) which provides a framework of unique personal identifiers of varying scope.

The system proposed in this paper combines a method similar to that described by Blobel with public key encryption techniques widely used in "e-commerce" to secure electronic financial transactions. A trusted third party is used to act as a non-privileged intermediary between parties and to provide a series of limited-scope unique personal identifiers, similar to those used in the HIDIT system. For the sake of convenience the proposed system will be referred to henceforth by the acronym PKISS (Public Key Infrastructure and Secret Splitting).

Public key cryptography

PKISS relies on public key cryptography to ensure that any information exchanged between parties can be read only by the intended recipients. Public key cryptography uses properties of large prime numbers to encrypt data using a pair of complementary keys (equivalent to passwords) belonging to each party [12]. These are known as the public key and the private key. The public key is published and can used by anyone wishing to encrypt information in such a way that it can only be decrypted (read) by the holder of the matching private key, and by no-one else. In practice, due to efficiency considerations, public key encryption and decryption algorithms are used to pass "session keys" securely between parties and these session keys are used with conventional encryption algorithms to protect the actual data – however, the effect is the same as if the entire message is encrypted or decrypted using public or private keys. Each party's private key can also be used to digitally "sign" messages to guard against tampering or substitution of the contents of the message during transit and to prove to the recipient that the party sending the message is in fact whom they claim to be. Usually a trusted agency known as a certificate authority handles the distribution of public keys and vouches for the authenticity of the other parties involved. Together, these facilities are often referred to as public key infrastructure (PKI).

Elements of the system

For the purposes of describing the system, the following definitions will be used.

A Disease Register is an organisation which collects relevant information about all incident or prevalent cases of a particular disease or condition which occur in a defined population. Usually the information collected includes demographic details of the person in whom the disease occurs and medical details of the specific diagnosis, disease or condition. Information about the treatment, complications and outcomes (such as death) in each case may also be collected or obtained from other sources such as death registers.

Health Care Providers are organisations or individuals which provide some form of health care service to patients (persons) and which are therefore in a position to capture the information about health events which might be required by a Disease Register. Examples of Health Care Providers include hospitals, general practitioners and pathology laboratories.

A Notifiable Health Event is any event about which a Disease Register requires information. Examples of Notifiable Health Events include the diagnosis of a new case of cancer, an admission to hospital for a particular reason, or births and deaths (in which case the statutory body responsible for registering vital events is regarded as a type of health care provider).

The Population Register is a trusted agency which is organisationally and physically distinct from all other parties which participate in the system. The function of the Population Register is to maintain a database of personal identifying information, such as name, date of birth, sex, country of birth and residential address. The database is used to assign a unique Population Register identifier (ID), typically a number, to each person of whom the Population Register is notified (which is not necessarily every person in the wider population). However, unlike other widely used unique identifiers such as the NHS Tracking Number, the Population Register ID is divulged to only one

other party: the Identifier Translation Agency.

The Identifier Translation Agency is another trusted third party which is organisationally and physically distinct from all other parties (including the Population Register) which participate in the system. Its role is to translate the unique identifier assigned to each person by the Population Register into a separate unique identifier which is specific to both each person and to each of the Disease Registers which participate in the system. This person and Disease Register-specific identifier (again typically a number) is shared only with the Disease Register to which it relates and with no-one else. The Identifier Translation Agency also provides temporary storage and forwarding facilities for encrypted messages, although this role could also be handled by yet another trusted agency if required.

Flow of information

The following sequence of message exchanges and operations correspond to the numbers in Figure 1.

A Health Care Provider produces or captures information about a Notifiable Health Event, such as the diagnosis of a case of cancer.

The Health Care Provider's information system sends a Health Event Notification message to the Identifier Translation Agency. This message comprises two parts. The first part contains only the personal identifying details (such as name, address and date of birth) of the person to whom the Notifiable Health Event relates. These identifying details are encrypted by the Health Care Provider's information system using the public key of the Population Register, effectively rendering the information unreadable by any party other than the Population Register. The second part of the message contains only the medical or other details of the Notifiable Health Event in question, but not the personal identifiers of the person to whom it relates. This second part is also encrypted prior to dispatch, this time using the public key of the target Disease Register for this particular Notifiable Health Event. This renders the information unreadable by any party other than the target Disease Register.

Upon receipt of the Health Event Notification message, the Identifier Translation Agency "unpacks" the two parts and tags each with the same arbitrary, unique random number (a "nonce") for tracking purposes. The first part of the message, which contains the encrypted personal identifiers, is forwarded to the Population Register in the form of a request to "look up" the Population Register ID for that person. The purpose of interposing the Identifier Translation Agency between the Health Care Provider and the Population Register is to prevent the Population Register from discovering the source of the Health Event Notification message and thereby being able to infer information about the Notifiable Health Event which triggered it. The Identifier Translation Agency also temporarily stores the second part of the Health Event Notification message, which contains the encrypted medical details of the Notifiable Health Event in question.

Upon receipt of a look-up request message, the Population Register decrypts (using its private key) the personal identifying information which the message contains and attempts to match this information against its database of persons. Probabilistic record linkage or other "fuzzy", error-tolerant matching techniques would be used for this look-up operation, possibly assisted by human intervention where required. If a match can be made, then the previously assigned Population Register unique identifier (Population Register ID) for that person is retrieved, otherwise that set of identifying information is added to the database as a previously unencountered person to whom a new Population Register ID is assigned.

The Population Register ID which has been retrieved or assigned is encrypted using the public key of the Identifier Translation Agency and returned to it in the form of a response message.

The Identifier Translation Agency maintains a database which translates each Population Register ID to a series of unique alternative ID numbers which are specific to each combination of a person and a Disease Register. The Identifier Translation Agency decrypts (using its private key) the response message which it has received from the Population Register and extracts the Population Register ID contained in it, together with the nonce which identifies the message. The Identifier Translation Agency then uses the nonce to retrieve the temporarily stored second part of the Health Event Notification message which it received previously from the Health Service Provider. From this it determines to which Disease Register the information should be sent. It then uses the Population Register ID to retrieve from its translation table the corresponding person/Disease Register-specific ID, or if one does not exist, it assigns one.

The person/Disease Register-specific ID is encrypted using the public key of the target Disease Register and packaged with the retrieved medical details of the Notifiable Health Event (which are still encrypted with the private key of the target Disease Register). This package is sent as a message to the target Disease Register.

The target Disease Register decrypts both parts of the message using its private key and updates its database with the medical details of the person identified by the Population Register ID, without needing to know who that person is.

In practice, each element of the system would acknowledge the receipt of messages and periodically resend unacknowledged messages in order to guarantee delivery. These "handshaking" messages are not shown in Figure 1 in the interests of clarity. Normal Internet (SMTP) e-mail could be safely used to convey the messages between parties because of the end-to-end use of high-level encryption. Disease registers are rarely required to operate in "real time" or "near real time", so delays in transmission and processing of messages are unlikely to be a problem.

Discussion

The PKISS system differs from previously described systems in a number of ways. Firstly, all directly identifying informa-

tion is split from the medical details at the earliest opportunity – that is, at the source of the notification. Secondly, public key encryption is used throughout the system to ensure that the identifying information remains effectively split from the medical details. Thirdly, a single Population Register is shared by multiple Disease Registers. Fourthly, a trusted third party, the Identifier Translation Agency, is used as a proxy to obfuscate the source of information flowing into the Population Register and to limit the scope and use of the unique ID assigned by the Population Register. Fifthly, the trusted third party does not have access to any privileged information – it acts solely as a conduit for encrypted messages which it cannot itself decrypt, and as a translator of sets of arbitrary ID numbers which have no meaning other than to the Population Register and to particular Disease Registers. In this way the hazards of a unique identifying number which must be widely shared throughout the health system are avoided [13].

However, the most important feature of the PKISS system is that the improvement in the protection of privacy and maintenance of confidentiality stems from its underlying architecture, rather than from the need for perpetual and unfailing observance of administrative and procedural safeguards by the staff of disease registers. Although disease registers have an excellent track record on security and the maintenance of confidentiality, it must be recognised that as they and other electronic health data collections become more numerous and access extended to more people, there is an increased likelihood of accidental or deliberate breaches of confidentiality, possibly on a large scale. The architectural protection provided by PKISS flows from the fact that at no stage are personal identifiers such as name stored by the same organisation or in the same database as (potentially very sensitive) medical or other health details.

When assessing any health information system it is important to consider not only the risk of security breaches but also the hazard associated with them. Perhaps the worst case scenario for a disease register, and therefore the maximum hazard, would be the misappropriation of its entire database and complete publication of the information it contains on the Internet. Although such a scenario is unlikely, it is nevertheless possible and must therefore be contemplated.

In the case of a conventional disease register which holds fully identified information, such an event could be devastating for individuals whose details were recorded on the disease register, and would almost certainly curtail further operation of the register (and others like it) for a considerable time due to public outrage .

For the PKISS system, such an event would still be serious, but not necessarily disastrous. If the Population Register were compromised, then at worst a list of names, dates of birth and other demographic details of selected members of the population would be discovered, but nothing more sensitive. The publication of the unique identification number assigned by the Population Register in conjunction with names and other identifying information would not compromise the entire system because the number is used only by a trusted third party, the Identifier Translation Agency, which also does not hold any medical or health information. Similarly publication of the identification number translation tables held by the Identifier Translation Agency would also have only a limited impact since the information has meaning only to Disease Registers and the Population Register, which are unlikely to be compromised at the same time.

However, in the PKISS system, a breach in the security of a Disease Register could still have serious consequences. Although the information held by each Disease Register is "de-identified" in the sense that names, dates of birth and addresses are not available or stored, it is not necessarily anonymous. In the presence of additional health and demographic information about individuals, such as that which might be available to a private health insurer, it may be possible to re-identify the "de-identified" information held by a Disease Register with a reasonable degree of certainty. It is therefore important that Disease Registers are supplied with only as much medical and other health information as they need to fulfil their core functions. It is also clear that the system described here would not obviate the need for careful attention to security by Disease Registers, although it would substantially mitigate the impact of a security failure compared the same failure in a Disease Register which used conventional methods of operation.

The PKISS system is applicable not only to registers for chronic diseases such as cancer, renal failure or diabetes registers but also to many other data collections which operate on the same or a similar basis. Examples include communicable disease surveillance systems, longitudinal health and social surveys and even tissue banks used for genetic epidemiology, such as the Icelandic Health Sector Database [14] [15] or the proposed UK Population Biomedical Collection [16].

The PKISS system also permits information from different Disease Registers to be linked at the level of individuals without the need for any of the Registers to obtain identifying information such as names or dates of birth – the Identifier Translation Agency need only provide the corresponding Person/Disease Register-specific IDs to an independent organisation, such as a research institution. Each of the Disease Registers would then provide the relevant de-identified disease information. If the statutory authority which registers deaths were to participate in a PKISS system, this mechanism could be used to routinely provide survival information to a number of Disease Registers.

Given that the PKISS system potentially facilitates the combination the data collected by different Disease Registers, the issue of legal protection and governance must be carefully considered. Ideally, the Identifier Translation Agency would be established under legislative arrangements which provide it with independence from other agencies such as government departments and with protection from legal processes (such as subpoena). Additionally the Identifier Translation Agency should be governed by a set of rules which require appropriate ethical review of all proposals which involve the linkage of information held by individual Disease Registers. It would be desirable if similar legislative protection and governance arrangements were also available for the Population Register

and for each Disease Register, but this is not essential. Indeed, the PKISS system could even be implemented entirely within an organisation provided that adequate administrative and physical independence of the various elements of the system could be established and maintained.

There do not appear to be any major technical impediments to the implementation of the PKISS system. Public key cryptography and its associated infrastructure is now commonplace, as are frameworks such as HL7 [17] or CorbaMED [18] for the communication of standardised, structured health information. The functionality of the Population Register is available in a number of off-the-shelve software products which conform to the CORBAmed Person Identification Service specification [19].

Weaknesses of the PKISS system include its apparent complexity and the need for all participating parties to adopt and adhere to information standards and protocols. Despite the apparent complexity, it should be possible to promote the system to sponsors and to the general public in quite simple terms: names, addresses and other identifying information are split off from medical details at source and are separately transmitted and stored at all stages thereafter. Incorporation of the required data processing standards and protocols should not present difficulties for new information systems or existing systems undergoing major revision, but may be problematic for "legacy" systems.

Despite these potential barriers, those responsible for implementing disease registers and related information systems need to actively consider the potential afforded by the modern, networked computing environment for intrinsically more secure ways of assembling and storing health information at the population level.

An expanded version of this paper is now available at http://www.biomedcentral.com/1471-2288/3/1/.

References

1. G. Kelly, Patient data, confidentiality, and electronics, BMJ 316 (1998) 718–719.

2. J.P. Vandenbroucke, Maintaining privacy and the health of the public, BMJ 316 (1998) 1331–1332.

3. R. Al-Shahi and C. Warlow, Using patient-identifiable data for observational research and audit, BMJ 321 (2000) 1031–1032.

4. T. Helliwell, S. Hinde and V. Warren, Cancer registries fear collapse letter], BMJ 322 (2001) 730.

5. Z. Kmietowicz, Registries will have to apply for right to collect patients' data without consent, BMJ 322 (2001) 1199.

6. R. Boruch and J. Cecil, Assuring the Confidentiality of Social Research Data. University of Phiadelphia Press, Philadelphia, 1979.

7. K. Pommerening, M. Miller, I. Schidtmann and J. Michaelis, Pseudonyms for cancer registries, Methods of Information in Medicine, 35 (1996) 112–121.

8. B. Blobel, Clinical Record Systems in Oncology. Experiences and Developments on Cancer Registers in Eastern Germany, in: R. Anderson (Ed.), Personal Medical Information. Security, Engineering, and Ethics, pp 39–56. Springer, Berlin, 1997.

9. A.P. Ho, US Patent 6,148,342: Secure database management system for confidential records using separately encrypted identifier and access request, United States Patent Office, 2000.

10. A.P. Ho, A Secret Splitting Method for the Protection of Confidentiality in Computer Records, in Research Advances in Database and Information Systems Security (V. Atluri and J. Hale, eds.), Kluwer Academic Publishers, Boston, 2000.

11. AS Kohane, H Dong and P Szolovits. Health Information Identification and de-Indentification Toolkit. In: C Chute, ed. Proceedings, Annual Fall Symposium of the American Medical Informatics Association; Florida: Hanley and Belfus, Inc; 1998. P356–60.

12. Y. Etheridge, PKI (public key infrastructure) – how and why it works, Health Management Technology, 22 (2001) 20–21.

13. R.J. Anderson, Remarks on the Caldicott Report. Available at http://www.cl.cam.ac.uk/~rja14/caldicott/caldicott.html

14. J. Gulcher and K. Stefansson, The Icelandic Healthcare Database and Informed Consent. N Engl J Med 342 (2000) 1827–1830.

15. H. Rose, The Commodification of Bioinformation: The Icelandic Health Sector Database, The Wellcome Trust, London, 2001. Available at http://www.wellcome.ac.uk/en/images/hilaryrose1_3975.PDF

16. J. Kaye and P. Martin, Safeguards for research using large scale DNA collections, BMJ 321 (2000) 1146–1149.

17. Health Level Seven: An Application Protocol for Electronic Data Exchange In Healthcare Environments. Draft Version 2.3. Health Level Seven, Inc., Ann Arbor, 1996.

18. Object Management Group, CORBAmed: OMG's Healthcare domain task force, available at http://www.omg.org/corbamed

19. Object Management Group, Person Identification Service specification (document formal/99–03–05), available at http://cgi.omg.org/cgi-bin/doc?formal/99–03–05