electronic Journal of Health Informatics
http://www.ejhi.net
2008; Vol 3(1): e5

Privacy and the Use of Health Data - Reducing Disclosure Risk

Christine M O'Keefe
CSIRO Preventative Health National Research Flagship

Abstract
In this paper we provide a review of some technological approaches to the problem of enabling the use of health data for research and policy analysis while protecting privacy and confidentiality. We take a statistical viewpoint, and assume that such approaches are implemented in an appropriate governance structure and secure environment. We give five examples of current implementations of such approaches in Australia and one in Canada, and indicate one emerging technology. We note that none of the technologies discussed here is the only solution to the problem, since there are many different scenarios for the use of health data, each with a different set of requirements. It is clear that different technologies and approaches have different strengths and weaknesses, and so are suitable for different scenarios.

Keywords: Health data, privacy, statistics, disclosure, privacy-enhancing technology

1. Introduction

As the health care industry moves from paper-based to electronic records, electronic data archives are accumulating in health care facilities and administrative agencies. Analysis of these health system usage and clinical data can yield information vital to effective health policy development and evaluation, as well as to enhanced clinical care through evidence-based practice and safety and quality monitoring.

At the same time, the analysis of these health data archives must be conducted in such a way as not to compromise standards of privacy and confidentiality for individual health care consumers, health care providers, health care facilities and health data custodians. Compliance with privacy legislation and codes of practice is a minimum requirement and health data custodians' responsibilities to protect sensitive data must be supported.

Data analysis and data mining tools are constantly being developed to be more powerful and to extract more information from data. Even if an analyst does not have direct access to the data, just the results of analyses can be enough to reveal private information.

In this paper we provide a review of some technological approaches to the problem of enabling the use of health data for research and policy analysis while protecting privacy and confidentiality. We note that none of these technologies provides the full answer, for each must be implemented within an appropriate legislative and policy environment and governance structure, with appropriate management of the community of authorised users and with an appropriate level of IT security including user authentication, access control, system audit and follow-up.

In addition, none of the technologies discussed here is the only solution to the problem, since there are many different scenarios for the use of health data, each with a different set of requirements. It is clear that different technologies and approaches have different strengths and weaknesses, and so are suitable for different scenarios.

A high level discussion of the problem of enabling the use of health data while protecting privacy and confidentiality typically discusses two broad approaches. The first is *restricted access,* where access is only provided to approved individuals, for approved analyses, possibly at a restricted data centre, possibly at a cost and possibly with further conditions such as restrictions on the types of analyses which can be conducted and restrictions on the types of outputs which can be taken out of the room. The second is *restricted or altered data,* where something less than the full data set is published or the data are altered in some way before publication. Restricted data might involve removing attributes, aggregating geographic classifications or aggregating small groups of

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381

data. For altered data, some technique is applied to the data so that the released dataset does not reveal private or confidential information. Common examples here include the addition of noise, data swapping or the release of synthetic data. Often these broad approaches are used in combination.

In this paper we review three current technological approaches to the problem. These fall into the category *restricted or altered data* above, and all are used in combination with *restricted access*. Where possible, we provide examples of successful initiatives which are using each approach, focussing on current implementations in Australia.

The first approach we discuss is *de-identification,* where obvious identifying features are removed from the data set and it is released under strict controls. We discuss the example of the Western Australian Data Linkage Unit, which has been providing de-identified data sets to researchers since 1995, and the newly-established Centre for Health Record Linkage in New South Wales and the Australian Capital Territory. The British Columbia Linked Health Database is an initiative with similar aims.

We next discuss *statistical disclosure control* which seeks to reduce the risk of a breach of privacy or confidentiality while still providing useful data to an analyst. In this setting we discuss the example of the Australian Bureau of Statistics, and the release of Confidentialised Unit Record Files, or CURFs, to researchers on CD-ROM.

We also discuss *remote analysis servers,* which are designed to deliver useful results of user-specified statistical analyses with acceptably low risk of a breach of privacy and confidentiality. One example is the Australian Bureau of Statistics Remote Access Data Laboratory, which provides access to more detailed CURF data than is available on CD-ROM. The RADL is a secure online data query service that clients can access via the Australian Bureau of Statistics web site. Authorised users submit queries written in the SAS, Stata or SPSS language through a web interface. The queries are run against CURFs that are kept within the Australian Bureau of Statistics environment. The results of the queries are checked for confidentiality then made available to the users via their desktop computers.

In closing, we give an introduction to CSIRO's Privacy-Preserving Analytics®. This approach is that of a remote analysis server; however it differs from the RADL in that it is designed to enable analyses of the original unit record files, not a confidentialised version. An analyst has no direct access to any data at all but submits analysis requests to the server, and receives results that have been filtered to reduce the risk of releasing private information.

Each of the broad technologies is implemented within the context that the analyst is trusted to comply with legal and ethical undertakings made. However, the different approaches have been designed with different risks of disclosure of private information, and so rely more or less heavily on trust. De-identification requires the greatest trust in the researcher, while Privacy-Preserving Analytics® requires the least. Statistical disclosure control, whether used alone or in combination with a remote analysis server such as the RADL, is somewhere in between these two extremes. De-identification provides the most detailed information to the researcher, while Privacy-Preserving Analytics® provides the least. Again, Statistical Disclosure Control and the RADL are in between.

2. De-identification

One way that data custodians seek to solve the problem of enabling the use of sensitive data while protecting privacy and confidentiality is to release de-identified data to researchers under strict controls. We provide a brief discussion of de-identification and examples of successful implementations of this approach.

De-identification is a very complex issue surrounded by some lack of clarity and standard terminology. It is also very important as it underpins many health information privacy guidelines and legislation.

First, it is often not at all clear what is meant when the term "de-identified" is used to refer to data. Sometimes it appears to mean simply that nominated identifiers such as name, address, date of birth and Medicare number have been removed from the data. At other times its use appears to imply that individuals represented in a data set cannot be identified from the data – though in turn it can be unclear what this means. Of course simply removing nominated identifiers is often insufficient to ensure that individuals represented in a data set cannot be identified – it can be a straightforward matter to match some of the available data fields with the corresponding fields from external data sets, and thereby obtain enough information to determine individuals' names either uniquely or with a low uncertainty. In addition, sufficiently unusual records in a database without nominated identifiers can sometimes be recognised. This is particularly true of health information or of information which contains times and/or dates of events.

In Australia the National Statement on Ethical Conduct in Human Research [1] avoids the term 'de-identified data' as its meaning is unclear. Instead, it proposed that data may be collected, stored or disclosed in there mutually exclusive forms, as follow:

a) *individually identifiable data,* where the identity of a specific individual can reasonably be ascertained. Examples of identifiers include the individual's name, image, date of birth or address;

b) *re-identifiable data,* from which identifiers have been removed and replaced by a code, but it remains possible to re-identify a specific individual by, for example, using the code or linking different data sets;

c) *non-identifiable data,* which have never been labelled with individual identifiers or from

which identifiers have been permanently removed, and by means of which no specific individuals can be identified. A subset of non-identifiable data are those that can be linked with other data so it can be known that they are about the same data subject, although the person's identity remains unknown.

One problem is that is not difficult to imagine datasets which do not fit into any of these categories. For example, a dataset of detailed health information from which all identifiers have been permanently removed may still allow the identification of an individual by matching to an external database, so these data could not fit into any of these categories.

On the other hand, the US Health Insurance Portability and Accountability Act 1996 (US) (HIPAA) provides a useful legislative test for de-identification that provides certainty for the research community and for ethics committees. It allows for a small risk of de-identification through reverse engineering or multiple, complex queries. In this case de-identification can require some modification of the data as well as removal of identifying fields, as discussed in Section 3. The relevant HIPAA section includes:

Section 164.514 Other requirements relating to uses and disclosures of protected health information.

(a)Standard: de-identification of protected health information. Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

(b)Implementation specifications: requirements for de-identification of protected health information. A covered entity may determine that health information is not individually identifiable health information only if:

- A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
  - o Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and-Documents the methods and results of the analysis that justify such determination;

    or
- The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:
  - o Names; All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, [some further tests deleted here]; All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older; Telephone numbers; Fax numbers; Electronic mail addresses; Social security numbers; Medical record numbers; Health plan beneficiary numbers; Account numbers; Certificate/license numbers; Vehicle identifiers and serial numbers, including license plate numbers; Device identifiers and serial numbers; Web Universal Resource Locators (URLs); Internet Protocol (IP) address numbers; Biometric identifiers, including finger and voice prints; Full face photographic images and any comparable images; and Any other unique identifying number, characteristic, or code;

    and
- The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

One potential difficulty with the HIPAA requirement is the burden of compliance – if an organisation has many data sets then it would take a great deal of time for a person to perform the tasks outlined.

A common response to situations of increased risk, such as data regarding small regional communities, is to allow the release of only aggregated data. This option is covered in Section 3 below.

2.1. The Western Australian Data Linkage Unit

The Western Australian Data Linkage Unit (WADLU) enables the provision of de-identified linked health data sets to bona-fide researchers in Western Australia who meet the stringent requirements of ethics committee clearance. Data can be requested for ethically approved research, planning and evaluation projects which aim to improve the health of Western Australians, see [2].

Because linked data sets generally contain more information than single data sets, they are generally regarded as requiring a higher level of protection.

The WADLU was established in 1995 to develop and maintain a system which enables the linkage of data about health events from separate databases for individuals in Western Australia. Strict protocols are followed to ensure confidentiality and security of linked data and research is performed on de-identified data files, as far as possible. (In some cases access to identified data may be granted.)

More than 340 projects have made use of WA linked data since 1995. These projects have originated from University, Government and hospital-based settings and have often involved collaboration between diverse research groups.

A list of projects from 1995 to 2003, detailing the title, investigators, institution, research field, project status and research output production is available [3]. A register of all applications for data is maintained, along with outputs from the project. Researchers who obtain linked data are asked to submit a copy of any report, journal article, other publication, conference presentation or media interest generated from data supplied by the WADLU. To date, there have been 700+ research outputs of projects of the WADLU. A Summary Report of the Research Outputs Project, including a list of project outputs from 1995 to 2003 is available [4].

The core Data Linkage System consists of links within and between the State's seven core population health datasets including: midwives' notifications, birth registrations, morbidity, mortality, mental health registrations, cancer registrations, ambulance, emergency and electoral roll, spanning 35 years. This is augmented through links to an extensive collection of external research and clinical datasets and cross-jurisdictional links to residential aged care data and Medicare Benefits Scheme [5] and Pharmaceutical Benefits Scheme [6] claims data.

The protocols and procedures developed for the Western Australian Data Linkage Project are fully described in [7]. The basic idea is that the WADLU acts as highly trusted third party. Participating data custodians extract the personal identifiers from their databases and send them to the WADLU. The WADLU uses probabilistic data matching algorithms and extensive clerical review to determine the links between records in the databases. The WADLU generates appropriate internal linkage keys for the linked records. When a project is approved, the WADLU generates a new set of project-specific linkage keys which enable the data custodians to supply only de-identified health information to the researcher or user. The researcher uses the project-specific linkage keys to assemble the linked dataset, but cannot join together two datasets created for different projects.

The separation of personal identifiers from health or clinical information gives excellent privacy protection in this case where there is a trusted third party linkage unit, and where data sources are permitted to release identifying information to that linkage unit. In fact, a recent study has shown that data linkage conducted in accordance with this best practice protocol is an effective way to conserve patient privacy in a research rich environment, see [8].

2.2. The Centre for Health Record Linkage

The Centre for Health Record Linkage (CHRL) is a unit that was established in July 2006 to provide record linkage services in New South Wales (NSW) and the Australian Capital Territory (ACT). It is an eight partner joint-venture between: ACT Health, NSW Health, the Cancer Institute NSW, the Sax Institute, the Clinical Excellence Commission, the University of Sydney, the University of New South Wales and the University of Newcastle, see [9].

All record linkage projects carried out by the CHRL must have the approval of the owners of the databases and a Human Research Ethics Committee. The CHRL uses identifying information to create a master linkage key that points to the locations of records for the same person in different health databases. Information about people's health is not revealed to the CHRL, but stays in the original database. Once the linkage is complete, the master linkage key is used to create a project-specific key. The CHRL provides the owners of each database with a list of records to be provided to the researchers and the project key. The owners of each database can then each provide a project database to the researcher, who links the project databases using the project key.

2.3. The British Columbia Linked Health Database

The University of British Columbia Centre for Health Services and Policy Research (CHSPR) in Canada houses the British Columbia Linked Health Database (BCLHD), [10]. This database integrates health service records, population health data and census statistics at the individual level, to form a longitudinal, person-specific, anonymous health record on each of British Columbia's four million residents, from 1985 forward. It is one of the world's largest collections of health services utilization and population health data.

The data holdings include data on: services provided under the province's universal insurance program, prescription medications provided under the public drug insurance program, hospital separations, continuing care, cancer incidences from registers, workers' compensation claims, births, deaths and mental health.

CHSPR is the central access point for researchers wishing to obtain and use these data in de-identified format for research in the public interest. The CHSPR ensures data confidentiality and security in a number of ways:

1) *Limiting Access:* Data in the BCLHD are only used for research proposals that meet conditions relating to scientific merit, ethical acceptability, and public interest. Requests for data are carefully reviewed in a process coordinated by the BC Ministry of Health, and operate on a minimum rights model: only data that are absolutely essential for the conduct of the research project are extracted and stripped of any personally identifiable information.

2) *Physical Protection:* CHSPR maintains a tightly controlled physical workspace with multiple layers of protection, including locked and alarmed premises, and a separately locked computer server room.

3) *Technical Safeguards:* CHSPR has stringent technological safeguards in place to protect data against loss, theft, unauthorized access, disclosure, copying, use or modification. These safeguards are constantly evaluated and evolve to meet new challenges and threats.

4) *Organizational Structure:* All CHSPR staff sign an oath of confidentiality agreeing to adhere to strict privacy policies and procedures for ensuring the confidentiality and security of data held at the Centre. Only a small number of specially trained staff can access the BCLHD.

The CHSPR encourages and facilitates the use of the BCLHD for research in the public interest. The approval process is coordinated through the BC Ministry of Health and is governed by an access policy written specifically to conform to the BC Freedom of Information and Protection of Privacy Act. In addition, the Data Access Request/Research Agreement requires that researchers:

1) Become familiar with all material outlining access requirements, data holdings, cohort definitions, and privacy considerations

2) Ensure that the ethics and peer review requirements of BCLHD data stewards have been met

3) Submit a research protocol and completed Data Access Request/ Research Agreement to the BC Ministry of Health, which reviews them together with the designated data stewards.

4) If approval is granted, both the researchers and CHSPR are notified. CHSPR staff then provide a cost estimate for the data preparation, and once the researchers have signed an agreement to reimburse costs, prepare and deliver the data. The data sets resulting from a request are large (they fill a CD-ROM) and often comprise hundreds of thousands of records spanning multiple files.

5) The BCLHD has been used in more than 120 health care and health services research projects since 1996.

3. Statistical Disclosure Control

The goal of a statistical disclosure control technique is to provide researchers with useful statistical data at the same time as preserving privacy and confidentiality.

Fienberg [11] has pointed out that any release of data or statistical summaries increases the risk of identification of some individual in the relevant population, with the consequent risk of harm to that individual through inference of private information about them. On the other hand, attempts to limit such disclosures can adversely affect the outcomes or usefulness of statistical analyses conducted on the data. Statistical disclosure control theory attempts to find a balance between these opposing objectives. Good general references to statistical disclosure control include [12, 13, 14, 15, 16, and 17].

Statistical disclosure control techniques can be organised into categories in several different ways. First, there are different techniques for tabular data (where data are aggregated into cells) versus microdata (individual level data). Second, techniques can be perturbative or non-perturbative. Perturbative methods operate by modifying the data, whereas non-perturbative methods do not modify the data. Perhaps the most well-known perturbative method is the addition of random "noise" to a dataset, and perhaps the most well-known non-perturbative method is cell suppression. In fact, current non-perturbative methods operate by suppressing or reducing the amount of information released, and there is much ongoing debate on whether a good perturbative method gives more useful information than a non-perturbative method. On the other hand, it has been noted that perturbative techniques which involve adding noise provide weak protection and are vulnerable to repeated queries [18], essentially because the noise becomes error in models of the data. There is much activity directed at developing perturbative techniques that do not suffer from this problem.

Virtually every statistical disclosure control technique can be implemented with differing degrees of intensity, and hence depends on a parameter which is usually pre-specified.

In this section we describe the aims and techniques developed for tabular data and for microdata (individual level data), then discuss the example of the successful Australian Bureau of Statistics CURFs.

3.1. Statistical Disclosure Control for Tabular Data

Tabular data results when data are summarised and presented as a table where the axes of the table correspond to observed variables.

Table cells can contain counts, where each data record contributes 1 to its tabulation cells and 0 to all other cells, in which case the data is called tabular count data and the table is called a contingency table. The counts can also be scaled and presented as frequencies, proportions or percentages, giving rise to tabular frequency data, proportion data or percentage data. In tables of rates (conditional observed frequencies) a cell entry represents the proportion of individuals who share an attribute with respect to the marginal count.

Table cells can also contain aggregates of one contributed variable, for example the total or average value of that variable for individuals contributing to that cell, in which case the data is called magnitude data. Thus count data comprises a table of integer counts, while magnitude data comprises a table of values, usually together with the integer count of contributing individuals.

In traditional Statistical Disclosure Control, the first task is to determine whether any of the cells are sensitive, where a sensitive cell is one for which the release of the data in the cell could lead to a disclosure. The main methods are:

- Threshold rule - a cell is sensitive if less than n individuals contribute to its value
- (n,k) -rule - a cell is sensitive if less than n individuals contribute at least k% of its value [18]
- p-percent rule - ensures that any contributed value cannot be determined by a coalition of c intruders to within p%
- (p,q)-percent rule - similar to the p-percent rule, but in addition we assume that the sensitive value is known to within q% before the data release.

For a discussion of the shortcomings of these techniques, see [19, 20].

After the sensitive cells in a table have been identified, the second task is to take steps to address the disclosure risk. The main techniques are:

- Deletion of variables - remove sensitive variables and/or variables which lead to sensitive cells.
- Collapsing cells - merging pairs of cells until no sensitive cell remains.
- Recoding variables - adjusting the level of aggregation of variables to reduce the number of sensitive cells.
- Cell suppression - suppression of the entry in each sensitive cell, then suppression of entries in non-sensitive cells sufficient to prevent reconstruction of the sensitive value.
- Rounding - all cells are rounded to a multiple of a chosen positive integer, for example, 3 or 5.
- Other perturbation techniques - sensitive cell values are altered, and normally also non-sensitive cell values are altered too.

There is much discussion in the literature about the shortcomings of these approaches, which include: reduction of information available to the analyst, release of table structures which are difficult to interpret or analyse, uneven impact on data subjects and uncertainty that the application of the technique reduces disclosure risk sufficiently.

Perhaps the most widely known software package for statistical disclosure control on tabular data, *t*-ARGUS (see [21]) uses variable recoding and cell suppression as its main techniques for protecting sensitive information, see [22].

Recently there has been a great increase in academic research activity internationally, much of which is centered on researchers at the US National Institute of Statistical Science and the Digital Government projects, see [23]. This activity is directed largely at putting the area of statistical disclosure control onto a sounder theoretical footing. There is now an emphasis on developing quantitative measures for disclosure risk and data utility, which enable an informed tradeoff to be made.

Despite this significant increase in research activity, there would appear to be little direct evidence of take-up of the results by statistical agencies.

3.2. Statistical Disclosure Control for Microdata

We will use the term microdata to refer to data where each record is contributed by an individual in the population, so that the record typically comprises values of a number of variables for the corresponding individual. An individual may contribute more than one record, for example, if the data are time-stamped hospital event data.

In contrast to the situation with tabular data, the identification of sensitive records in microdata is more an art than a science.

The two main ways that a disclosure occurs from a microdata file are:

- Reading identifiers contained in the microdata file
- Spontaneous recognition, where a data user knows enough about an individual to recognise their record in the data file, and
- Linking to an external database, by variables which are common o the two databases.

Therefore a record is considered to be sensitive if it has an unusual combination of characteristics or if it is unique in the ample and in the population, see [24]. Geographic information is regarded as being particularly sensitive.

The main techniques used for microdata confidentialisation are:

- Sampling - a random sample drawn from the data is released.
- Data swapping - one or more attributes are interchange between records in a microdata file.
- Addition of noise - random values are added to the data.
- Microaggregation - records are clustered into small groups and the averages are released.
- Synthetic data methods - synthetic data with similar characteristics to the original data are generated and released.

The National Institute of Statistical Sciences package NISSWebSwap [25] is a Web Service that swaps one or more attributes between user-specified records in a microdata file, uploading the original data file from the user's computer and downloading the file containing the swapped records.

Again, much debate about the advantages and disadvantages of these approaches can be found in the literature. The disadvantages are mainly the reduction in the amount of information available to analysts and the corresponding reduction in the precision of estimates and bias can be introduced. If synthetic data is too similar to the original data then protection may not be achieved.

3.2.1. The Australian Bureau of Statistics CURFs

The Australian Bureau of Statistics (ABS) Confidentialised Unit Record Files (CURFs) contain data from ABS surveys in the form of unit records. CURFs contain the most detailed statistical information available from the ABS for researchers and analysts to run statistical queries on the data using SAS or SPSS software.

CURFs have been confidentialised by removing name and address information, by controlling the amount of detail and by changing a small number of values through the application of statistical disclosure control techniques. There are three levels of data detail available - Basic, Expanded and Specialist – corresponding to three different access modes - CD-ROM, Remote Access Data Laboratory and ABS Data Laboratory. The Remote Access Data Laboratory is a secure online data query service while the Data Laboratory is a secure on-site facility.

Basic level CURFs, which is those that are the least detailed, are available on CD-ROM. Each CURF is released for an individual's specified statistical purposes and for a stated period to the nominated Responsible Officer and Individual Authorised Users. Both the Responsible Officer and Individual Authorised Users are required to sign and agree to be bound by a legal Undertaking which if breached can result in a fine or imprisonment or both. The Responsible Officer and Individual Authorised Users consequently have an obligation to ensure that the CURF and any copies of the CD-ROM, or RADL CURF microdata output, remain secure. The release of a CURF is at the sole discretion of the Australian Statistician, who must approve each release under the specified conditions. Users can work unrestricted on them on their own computers, but are bound by the legal undertaking.

More detailed CURFs may be accessed via the Remote Access Data Laboratory (RADL) (see Section 4.1) and the most detailed, specialist level CURFS may be accessed through the ABS Data Laboratory (ABSDL) (see Section 4.2).

The majority of CURFs produced before 2003 are only available at the Basic level on CD-ROM or via the RADL, but from 2003 onward most CURFs are produced at both the Basic and Expanded level. Some CURFs will only be produced at the Basic level and be available on CD-ROM or via the RADL, while some CURFs will only be produced at the Expanded or Specialist level and will only be available via the RADL or ABSDL respectively.

4. Remote Analysis Servers

An alternative to de-identifying or confidentialising data before release to analysts is the technology of remote analysis servers. Such servers do not provide data to users, but rather allow statistical analysis to be carried out via a remote server. A user submits statistical queries by some means, analyses are carried out on the original data in a secure environment, and the user then receives the results of the analyses. In some cases the output is designed so that it does not reveal private information about the individuals in the database.

This approach was discussed in [26] and the characteristics of such a server are explored in [27] and [28]. Techniques for regression model diagnostics are provided in [29].

The approach has several advantages. First, no information is lost through confidentialisation, and there is no need for special analysis techniques to deal with perturbed data. In many cases it is found to be easier to confidentialise the output of an analysis, in comparison to trying to confidentialise a dataset when it is not known which analyses will be performed.

Karr et al [30] in particular note that analysis servers are not free from the risk of disclosure, especially in the face of multiple, interacting queries. They describe the risks and propose quantifiable measures of risk and data utility that can be used to specify which queries can be answered and with what output. The risk-utility framework is illustrated for regression models.

4.1. The Australian Bureau of Statistics Remote Access Data Laboratory

The Remote Access Data Laboratory (RADL) is a secure online data query service that clients can access via the Australian Bureau of Statistics web site. Authorised users submit queries in the SAS, Stata or SPSS language against CURFs that are kept within the Australian Bureau of Statistics environment via the RADL web interface. The results of the queries are checked for confidentiality then made available to the users via their desktop computers [31].

The RADL provides access to CURFs from any computer with an internet connection and batch processing with quick turnaround time. All data and output are stored in a user's individual secure workspace.

The List of Available CURFs can be obtained from the Australian Bureau of Statistics, and is regularly updated. Examples of health data currently available as CURFs include: Disability, Ageing and Carers (Basic), Mental Health and Wellbeing of Adults, Australia and Western Australia (Basic), National Aboriginal and Torres Strait Islander Health Survey (Expanded), National Health Survey (Basic and Expanded), National Health Survey, Indigenous (Expanded) and National Nutrition Survey (Basic). Most Basic CURFs that are currently available on CD-ROM can also be accessed via the RADL.

Users apply for access to CURFs through their organisation, and can only apply to use CURFs for which their organisation has approved access. The Australian Bureau of Statistics has developed a training manual on responsible access to ABS CURFs, and a user guide for the RADL.

If the application is approved, access to each CURF is granted for the specified statistical purpose and for the stated period. Both the user and the organisation's nominated Responsible Officer are required to sign and to be bound by legal undertakings which provide for a fine, imprisonment or both in the case of a breach.

A researcher can obtain a limited number of (confidentialised) unit records through the RADL. The undertaking requires that no such unit record may be disclosed. Tables or

other aggregated output (e.g. averages, model parameters) released as ``general output'' via the Remote Access Data Laboratory (RADL) may be disclosed or disseminated by the user.

The RADL user agreements require that any output from RADL which is labelled ``keep secure'' must not be disclosed to any other person. The output must be kept in a locked room or secured in a locked cabinet when the researcher is not present, and destroyed in a secure manner if no longer needed. Any unit level information should also be handled in this way.

Each year, the ABS publishes a detailed description of CURF Research Activities. For example, the report of 2005 Research Activities is available at [32].

The ABS publishes annually a detailed description of CURF Research Activities to enable CURF users to understand the range of research activities undertaken by individuals using CURFs. This information is created from annual renewal information provided to the ABS by CURF users.

4.2. The Australian Bureau of Statistics Data Laboratory

We include a brief discussion of this data access mode for completeness.

The ABS Data Laboratory (ABSDL) is an on-site facility offering a high level of data analysis of specialist level CURFs, with both SAS and SPSS software provided within the system. In addition, users may be able to integrate the CURFs with other datasets.

All unit records remain within the ABS IT environment, and restrictions apply to the nature of the queries which may be run, and to the nature and the size of the outputs which may be obtained. ABSDL users are less restricted than RADL users in terms of the data they can access and the analyses they can run.

The access mode provided is interactive and on-site consultancy support is available. ABSDL users are bound by a legal undertaking, and their activity is monitored.

4.3. CSIRO's Privacy-Preserving Analytics®.

CSIRO's Privacy-Preserving Analytics® is a remote analysis server designed to run analyses on original unconfidentialised microdata, in a secure computing environment. The software is hosted by a data custodian, and made available to analysts through a web interface.

An analyst has no direct access to the data, but submits queries through a menu-driven interface. More-or-less traditional statistical analyses can be run, and the results are presented with the aim that no individual unit record is disclosed through the output, or can be deduced or inferred from the output. This is achieved by ensuring that no directly identifying information is released, and no values can be deduced or inferred that can lead to spontaneous recognition or can be matched to an external database.

For more details about how the outputs are presented in order to avoid disclosures, see [33].

The software is currently demonstrator grade, and we are working with several agencies to run an evaluation and demonstration.

5. Summary

In this paper we considered the problem of enabling the use of health data for research on clinical practice and policy analysis while protecting the privacy and confidentiality of individuals, health care providers, health care facilities and data custodians. Because health policy and practice affect all of us, it is vital that information extracted from health databases is reliable and free from bias. Therefore, the objective of sufficiently high data utility needs to be balanced against the objective of sufficiently low disclosure risk.

In this paper we provided a review of three technological approaches to the problem of balancing disclosure risk with data utility, namely: the provision of de-identified data by a trusted third party, the release of confidentialised data and the use of remote servers. These examples were chosen because each has been implemented and used over a number of years in the Australian setting. In each of implementations, the analyst is trusted to comply with legal and ethical undertakings. However, the different approaches have been designed to entail different risks of disclosure of private information, and so rely more or less heavily on trust.

We note that none of these technologies provides the full answer, for each must be implemented within an appropriate legislative and policy environment and governance structure, with appropriate management of the community of authorised users and with an appropriate level of IT security including user authentication, access control, system audit and follow-up. In addition, none of the technologies discussed here is the only solution to the problem, since there are many different scenarios for the use of health data, each with a different set of requirements. Different technologies and approaches have different strengths and weaknesses, and so are suitable for different scenarios.

Acknowledgements

The author thanks Peter Lamb for useful discussions.

References

1. Australian Government National Health and Medical Research Council (2007), National Statement on Ethical Conduct in Human Research, http://www.nhmrc.gov.au/publications/synopses/e35syn.htm

2. Western Australian Data Linkage Unit, http://www.populationhealth.uwa.edu.au/welcome/research/dlu/linkage

3. Western Australian Data Linkage Unit (2003). Western Australian Data Linkage Unit Projects 1995 - 2003, http://www.populationhealth.uwa.edu.au/_data/page/63033/Projects_1995-2003.pdf

4. Western Australian Data Linkage Unit (2003). Summary of Research Outputs Project, WA Data Linkage Unit (1995 - 2003), http://www.population-health.uwa.edu.au_data/page/63033/ROP_SUMMARY3.pdf

5. Australian Government Department of Health and Ageing, Medicare Benefits Scheme (MBS), http://www.health.gov.au/internet/wcms/publishing.nsf/Content/health-medicarebenefits-healthpro

6. Australian Government Department of Health and Ageing, Pharmaceutical Benefits Scheme (PBS), http://www.health.gov.au/pbs

7. CW Kelman, AJ Bass and CDJ Holman (2002). Research use of linked health data - a best practice protocol, Australian and New Zealand Journal of Public Health, 26: 251-255.

8. B Trutwein, CD Holman and DL Rosman (2006). Health data linkage conserves privacy in a research-rich environment, Ann Epidemiol, 16:279-280.

9. Centre for Health Record Linkage, http://www.saxinstitute.com.au/public-docs/Sax_CHRL_Flyer.pdf

10. British Columbia Linked Health Database, http://www.chspr.ubc.ca/data

11. Stephen E Fienberg (2005). Confidentiality and Disclosure Limitation, Encyclopedia of Social Measurement, Elsevier, 1.

12. NR Adam and JC Wortmann (1989). Security-Control Methods for Statistical Databases: A Comparative Study. ACM Computing Surveys 21:515-556.

13. J Domingo-Ferrer (Ed) (2002). Inference Control in Statistical Databases: From Theory to Practice, State-of-the-Art Survey, LNCS 2316 Springer-Verlag Berlin Heidelberg 2002.

14. J Domingo-Ferrer and V Torra (Eds) (2004). Privacy in Statistical Databases, Lecture Notes in Computer Science, Vol 3050, Springer-Verlag Berlin Heidelberg.

15. P Doyle, JI Lane, JJM Theeuwes and L Zayatz (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier, Amsterdam.

16. Office of Information and Regulatory Affairs (1994). Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology, Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

17. L Willenborg and T de Waal (2001). Elements of Statistical Disclosure Control, Lecture Notes in Statistics, v155, Springer.

18. L Cox (1981). Linear Sensitivity Measures in Statistical Disclosure Control, Journal of Statistical Planning and Inference 5: 153-164.

19. J Domingo-Ferrer and V Torra (1993). A critique of the sensitivity rules usually employed for statistical table protection, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

20. DA Robertston and R Ethier (2002). Cell Suppression: Experience and Theory, In Domingo-Ferrer, Josep (Ed.) Inference Control in Statistical Databases: From Theory to Practice. State-of-the-Art Survey, LNCS 2316. Springer-Verlag Berlin Heidelberg.

21. http://neon.vb.cbs.nl/CASC/TAU.html

22. ES Nordholt (2003). Application of statistical disclosure control methods, Proceeding of Statistics Canada Symposium.

23. National Institute of Statistical Sciences, http://www.niss.org/

24. CJ Skinner and MJ Elliot (2002). A measure of Disclosure Risk for Microdata, Journal of the Royal Statisical Society, Series B 64:855-867.

25. National Institute of Statistical Science, NISSWebSwap, Version 1.1 http://www.niss.org/WebServices/dg/WebSwap.html

26. GT Duncan and RW Pearson (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future, Statistical Science 6:219-239.

27. S Keller-McNulty and EA Unger (1998). A database system prototype for remote access to information based on confidential data, Journal of Official Statistics 14:347--360.

28. B Schouten and M Cigrang (2003). Remote access systems for statistical analysis of microdata, Statistics and Computing 13:371--380.

29. JP Reiter (2003). Model diagnostics for remote access regression servers, Statistics and Computing, 13:71--380.

30. S Gomatam, AF Karr, JP Reiter and AP Sanil (2005). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers, Statistical Science} 20:163-177.

31. Australian Bureau of Statistics, Remote Access Data Laboratory, http://www.abs.gov.au/Websitedbs/D3110129.NSF/f578250c9c9b9ee1ca256de4002ca08b/36bd92d8f488355dca256f4a0010c110!OpenDocument

32. Australian Bureau of Statistics, 2005 CURF Research Activities, http://www.abs.gov.au/Websitedbs/D3110129.NSF/85255e31005a1918852558ac00697645/90ef40fecb7a7058ca25717900150b7e!OpenDocument

33. R Sparks, C Carter, JB Donnelly, CM O'Keefe, J Duncan, T Keighley and D McAullay (2008).Remote Access Methods for Exploratory Data Analysis and StatisticalModelling: Privacy-Preserving Analytics.Preprint.

Correspondence

Dr. Christine O'Keefe
Research and Business Leader, Health Data and Information
CSIRO Preventative Health National Research Flagship
GPO Box 664
Canberra ACT 2601, Australia

Phone: +61 (0)2 6216 7021
Fax: +61 (0)2 6216 7111
http://www.csiro.au

Christine.OKeefe@csiro.au