National Health Information Management Group

NHIMG

October 2002

Issues for the use of unique patient identifiers in statistical collections

information paper

Introduction

At its April 2000 meeting the National Health Information Management Group agreed to accept the following responsibilities in relation to the development and use of unique patient identifiers:

- Identify issues for national minimum data set management raised by proposals for the introduction of unique patient identifiers.
- Draft business rules for the use of unique patient identifiers for linkage for statistical purposes.
- Provide comment and advice on these matters to agencies developing unique patient identifiers.
- Provide comment and advice on these matters to agencies developing privacy legislation and guidelines.

This paper addresses the first of these four objectives by discussing some of the issues for health and statistical data set management raised by proposals for the introduction of unique patient identifiers (UPIs). The discussion covers UPIs with and without explicitly identifying details such as names and addresses.

Appendix A addresses the second objective by setting out, as a basis for more detailed discussion, some principles for the use of unique patient identifiers for linkage between statistical collections. These principles have a narrower focus restricted to UPIs that are anonymous in that they do not include explicitly identifying information such as names and addresses.

Table of contents

AIHW

AUSTRALIAN INSTITUTE OF HEALTH & WELFARE

Background

Developments in this area are occurring rapidly. For example, Territory Health Services has already introduced a UPI providing clients with access to health services in acute care and in urban and rural community care throughout the Northern Territory. The UPI is included in demographic and clinical data that are downloaded into a data warehouse where all data are managed for statistical reporting and analysis. At this level the data can be de-identified by encrypting the UPI and the encrypted UPI can be used for data linkage.

Similar projects are either in place or under development by most other State/Territory health authorities although in general not yet on a statewide scale.

The New South Wales Health Council recommended that New South Wales Health establish a UPI for every individual in that State so that health care providers could identify with certainty the particular patient they were dealing with irrespective of where the patient had entered the health system.

A strategy to achieve this aim was endorsed in November 2000. The strategy recommends a tiered approach from area health service level to State level, primarily to ensure that data quality issues related to patient identification and reconciliation are performed at an area health service level. The initial implementation will be limited to links between in-patient records and any outpatient department or community health centre that has the capability of linking to an Area UPI system. The target for UPI implementation in all area health services and a central State location is November 2002.

In Western Australia, a system of linkages within and between hospital discharge records, death registrations, and cancer and mental health records has been established using all available identifying information. UPIs are assigned during the linkage process and the system of linkages is maintained separately from clinical information. Access to the links in the system is dependent upon institutional ethics committee clearance and approval by each of the data custodians. A summary of the protocol developed for a cross-jurisdictional linkage project on diabetes is included in Appendix B as an example of 'best practice' in record linkage.

At the national level, where full identifying information may not be available, the date of birth and some of the characters of the client's name may be used for linkage. These statistical linkage keys (SLKs) are not classified as UPIs and are used to create links between records for statistical purposes only. A report entitled *Statistical Data Linkage in Community Services Data Collections* was submitted to the National Community Services Information Management Group (NCSIMG) in October 2001. The focus of the NCSIMG report is the management of data linkage using SLKs rather than the management of the original records.

The Health*Connect* project endorsed by the Australian Health Ministers' Conference is investigating a national health information network. It will complete its research and development stage by mid-2003 and if a decision is made to roll out Health*Connect* nationally, a system of reliable patient identification will be required. Subject to robust consent and privacy arrangements, information collected under Health*Connect* has the potential to be used for statistical data linkage.

It is also worth noting that although several systems of coded patient identifiers already exist in administrative systems, few are truly unique in that more than one identifier may be assigned to an individual client and more than one individual client may be assigned the same identifier. This can occur in both manual and electronic systems and is very difficult to avoid completely. However, a draft standard for a framework for the positive identification of clients within health care organisations has been developed by Standards Australia (Subcommittee IT-14-9-3) and, if adopted, should reduce the risk of mistaken identity.

Privacy issues

A primary requirement in any health data collection is to protect the privacy of the individual. In statistical collections this is usually achieved by the use of de-identified records and by adopting a rigorous protocol to minimise the risk of re-identification. The unit records held in de-identified statistical collections are said to be 'anonymous'; in that transparently identifying information (such as names and addresses) is either not collected or is removed before the unit records are made available for statistical or research purposes.

The inclusion of unique patient identifiers in these collections may increase the risk for individual subjects to be identified (or re-identified) in at least two ways.

Firstly, individual UPIs may be matched with transparently identifying information such as names or addresses. This risk can be managed by business rules governing the link between patient's names and their UPIs, supported by technical barriers such as encryption restricting the ability of users to make this linkage.

Secondly, the UPI may be used to link data relating to the same individual in two or more data sets in such a way that the individual, although still 'anonymous', is more easily identifiable through a combination of data items that together uniquely describe the individual. An individual may become recognisable through a combination of data items that may or may not include the UPI. For example, the UPI may be used to link morbidity from different hospitals in different jurisdictions and a file of combined patient level data may be released to a researcher after the UPI has been removed. However, if the file contains dates of admission and discharge for every hospital episode in a person's lifetime there may still be an unacceptable risk that users may be able to identify the individual concerned, especially if some of the hospitals are small or if too much geographical data about the person's place of residence are released; this could have adverse impact not just on the person's privacy but on their standing in the community and willingness to seek further health care. This risk can also be managed by business rules governing the level of aggregation or disaggregation required for data to be released for research and planning purposes.

As the technical scope for linkage increases, these issues will need to be addressed in the context of both existing statistical data sets, such as the hospital morbidity collections and data collected under Medicare and the Pharmaceutical Benefits Scheme, and future data collections that may be established as a result of initiatives such as Health*Connect*.

In fact, three classes of data collections may be need to be defined:

- the UPI register containing clients' UPIs, names and other demographic information—this would be used for client registration and for resolution of possible duplicates;
- statistical unit record data sets containing individual client records each of which would include the client's UPI or a statistical linkage key, but not the client's name;
- data sets for research (possibly linking data across more than one statistical data set) that include encrypted UPIs as an additional safeguard against identification of individual clients, especially where the user may be able to access the UPI register.

PRINCIPLES FOR USE OF IDENTIFIED RECORDS

Patient master indexes

Patient master indexes are already maintained by most hospitals and health care providers. Hospital indexes generally include each patient's name, address, date of birth, and so on, as well as a hospital unit record number and some clinical or service-related information such as service dates, diagnoses and medical alerts. Such systems are mainly used to link clinical information within a single hospital; however, there is an increasing tendency to extend linkage across service providers by establishing consolidated patient master indexes at the regional level. This type of index may be maintained by multi-hospital agencies such as an area health service (in New South Wales) or a metropolitan health service (in Victoria). Some State and Territory health authorities are also developing statewide patient indexes with the potential to cover their entire public hospital systems. (This has already been achieved in the Northern Territory.)

Population registers

Population registers are designed to cover an entire population or sub-population without restriction to a particular group of service providers. At the population level the most obvious example is the register of Medicare card numbers and internal personal identification numbers maintained by the Health Insurance Commission. A more restricted example would be the register assigning 'DVA numbers' to persons entitled to benefits from the Department of Veterans' Affairs.

There are also national and State/Territory registers relating to specific health issues, some of which contain names. For example, the Australian Institute of Health and Welfare maintains the National Death Index and the National Cancer Statistics Clearing House, both of which contain explicitly identified information that is protected under the *Australian Institute of Health and Welfare Act 1987*. There are also an increasing number of specific health issues registers that do not contain names but that may contain some form of UPI.

Each of these indexes and registers is an example of a UPI system and in each case access to the names and UPIs contained in the system is governed by business rules or in some cases by legislation (or both). These rules are primarily designed to protect individual privacy while facilitating the clinical and administrative purposes of the information system.

With the growing use of electronic health records and electronic messaging, however, the general trend is:

- for selected agencies to be provided with access to the names and numbers in the index for approved clinical or administrative purposes, but on the other hand
- for this access to be governed by privacy principles or legislation that may include requirements for individual consent (either on an 'opt in' or 'opt out' basis) thus making it difficult to use the UPI for statistical purposes.

For example, the *Medicare and Pharmaceutical Benefits Programs Privacy Guidelines* issued under section 135AA of the *National Health Act 1953* place limits on data linkage between Medicare benefits data and pharmaceutical benefits data. Even with patient consent, the use of the Health Insurance Commission's internal personal identification number to link such data is prohibited except in specific instances such as the Coordinated Care Trials conducted by the Commonwealth Department of Health and Ageing. However, section 2.3 of the Guidelines permits the routine provision of the Medicare card number in an encrypted form and the internal personal identification number to the Department in conjunction with de-identified or anonymised claims data for a range of public policy purposes some of which may involve linking records relating to the same (unidentified) individual. The use of the data by the Department is then governed by Part B of the Guidelines, in particular section 5, which includes safeguards against the re-identification of the claims data.

PRINCIPLES FOR USE OF RECORDS CONTAINING STATISTICAL LINKAGE KEYS

Statistical linkage keys that consist of date of birth and some of the characters of the client's name were developed to facilitate linkage within and between relatively small or specialised data sets where duplicate keys were unlikely. These keys were intended for linkages for statistical purpose only and were never intended to be used in clinical or client management settings. In addition, if the key is not encrypted the risk of direct identification or re-identification of clients from their SLK is greater than from a numeric UPI. Thus it is recommended that systems containing SLKs adopt the rules described above for name-identified records.

PRINCIPLES FOR THE USE OF UNIQUE PATIENT IDENTIFIERS

Agencies managing or acting as custodian for statistical collections that include UPIs need to adopt business rules and technical barriers that restrict the capacity of users to match the UPI to the individual's name. The key issue to be addressed by these business rules concerns access to the system or systems containing both UPIs and patient names and addresses (e.g. the patient master index or the population register).

While the precise business rules may differ from collection to collection, the basic principle is that, where the UPI is used for clinical or administrative purposes, as well as to link records for statistical purposes, the personnel who use the UPI for clinical or administrative purposes should not normally be able to access additional information on identified clients who have not consented to this access. This could be achieved by encrypting the UPI before it is used for statistical linkage.

PRINCIPLES FOR THE USE OF ENCRYPTED UNIQUE PATIENT IDENTIFIERS

The encrypted UPI could be used in the same way as the statistical linkage key (SLK) has been used to match records in statistical data sets. The result using UPIs should, however, be more accurate than linkage using an SLK because of the possibility of fewer missed matches and mismatches due to the increased discriminating power of UPIs.

The basic principle is that de-identified information used for statistical, research or planning purposes is not used or disclosed in such a way that an individual's identity can be ascertained. Provided that it remains de-identified, information used in this way does not fall within the definition of 'personal information' incorporated in all current privacy legislation.

However, it is essential that patient privacy is maintained in any de-identified statistical collection, even if there is a possibility that a small number of records may be identifiable by particular users due to the unusual nature of the records. This places a responsibility on the custodians and users of data sets to rigorously manage data to minimise the risk of identification and ensure ongoing ethical handling and disposal of all unit record data. It also provides them with a clear specification of reasonable steps to manage risk of identification. Ethical data handling practices also need to be specified and assured to guide users of data in situations where potential or actual recognition occurs as a result of unpredictable circumstances or a conscious attempt to breach the spirit of the privacy principles.

Appendix A

GUIDELINES FOR THE USE OF UNIQUE PATIENT IDENTIFIERS FOR DATA LINKAGE

The following guidelines are a first step towards a model code of practice for custodians of health data collections that are de-identified (i.e. they do not contain explicitly identifying information such as names and addresses) but which include unique patient identifiers (UPIs). The focus is on managing the capacity to match records in different data collections using the UPI. Mechanisms to control access to the matched information are also proposed.

These guidelines are intended to help data custodians to ensure that the data used in health statistical collections and research projects are de-identified and remain de-identified at all stages of their use, storage and eventual destruction. They illustrate 'best practice' in compliance with and the application of the Federal Information Privacy Principles, the National Privacy Principles and the section 95 and section 95A guidelines approved by the National Health and Medical Research Council under the *Privacy Act 1988.* This legislative framework is technologically neutral and must be complied with in the electronic environment.

These guidelines relate to the handling of de-identified statistical data rather than the collection of such data in clinical and administrative situations. However, privacy breaches can be avoided if organisations which manage data advise individuals about what data they collect and why, and ensure that the organisations and individuals have shared expectations in relation to directly related secondary uses and disclosures of the data including the fact that de-identified data may be used for research or statistical collections.

MINIMISATION OF POTENTIALLY IDENTIFYING INFORMATION

Any use of statistical data resulting from linkage of records from more than one collection must be accompanied by steps to prevent individuals being identified or recognised by users of the data. As a general guide, the following principles should be considered and exceptions documented:

- When a UPI is used to create a data set by linking data from two or more sources, the UPI should be removed from the data set or encrypted before it is made available to the research team.
- Other potentially identifying data items such as the unit record number assigned to the patient by the hospital or other health care provider should be removed or encrypted before the data set is made available to the research team. It may also be necessary to ensure that the hospital cannot be identified, especially for small hospitals or those that serve small communities.
- Detail in data items should be reduced to the level necessary for the research. For example, age would normally be computed from date of birth and length of hospital stay would normally be computed from dates of admission and discharge.
- Where possible, data items should be aggregated to the level that is needed for the research project. For example, Statistical Local Area or postcode of residence should normally be aggregated to larger geographical units such as the Statistical Division or health region unless the focus is on a specific small area. Similarly, country of birth or language should normally be restricted to major groups or specific countries or languages of interest rather than used in a form that identifies every country or language (however uncommon) identified in the collection. In accordance with standard statistical practice, tabulations with less than five individuals in a single cell should be avoided in research work and should never be published.
- Diagnosis and procedure codes should only be released with a three-digit ICD-10-AM level of detail unless there is a specific need for greater detail.
- In addition, cross-tabulations of data items should be limited to those that are strictly necessary for the research. For example, while Indigenous status, place of residence, country of birth and preferred language may all be relevant to a health research project, a four-dimensional cross-tabulation of these variables would usually be unnecessarily cumbersome and would often include an unacceptable number of cells with only one or two individuals.

SUPERVISION OF THE USE OF DATA

The following general principles should be applied to most research projects using data sets that either have been linked or are capable of linkage:

- Projects involving the linkage of client level data should be considered by an institutional or departmental ethics committee established in accordance with the guidelines issued by the National Health and Medical Research Council.
- There should be a clearly documented and agreed method for overseeing the project and monitoring linkage and the use of UPIs. This should include explicit procedures and sanctions designed to ensure confidentiality and adherence to best practice as well as relevant legal obligations.
- Security measures and technical protective measures should be specified. This would include details of precautions taken to ensure the physical security of data and prevent unauthorised access to computer systems. Agreed minimum standards should be specified.
- Regular audit procedures designed to identify unauthorised or inappropriate access to data should be adopted. All access requests and uses of data should be logged to provide audit trail information.

DATA EDITING

Research projects using linked data may need to incorporate consistency checks to detect errors in the original unlinked data sets (e.g. there may be inconsistencies between the dates recorded for hospital episodes or vital events in two data sets which may only become apparent after the data sets have been linked). As far as possible this should be applied before data sets are linked to minimise the backtracking from the linked records to the original data sets.

SUBSEQUENT USE AND DESTRUCTION OF DATA SETS

- Rules governing the retention or destruction of data files or data sets after the analyses have been completed need to be implemented, allowing for time for results to be checked and research reports to be refereed.
- Restrictions need to be placed on linkage to data sets other than those that have been approved.
- A register of data releases, termination and destruction should be maintained and methods for regular reporting on progress of long-running research projects should be incorporated.

Conditions of this type are often imposed by data custodians but may not always be rigorously enforced. For this reason, custodian agencies that handle a large number of data requests may need to adopt proactive procedures to ensure that the use of data sets is terminated on or before an agreed date, including a specified period to destroy or de-identify data and related audit procedures. Typically a data set would be made available for a specific number of months or years after which the custodian agency responsible for custody of the data would contact the recipient if necessary in order to satisfy itself that the research had been completed without any breaches of privacy and that the data had been archived, returned or destroyed in a satisfactory manner. Further research projects or extensions of time could then be considered on their merits rather than taken for granted.

While the guidelines would need to be tailored individually for each project, the following standard conditions of release used by one State health authority (Victoria) provide a useful model:

- The data must not be used, published or disseminated in a way that might enable the identity of individual patients or the service profiles of individual doctors or private hospitals to be ascertained.
- The data file is provided solely to the recipient and must not be communicated to other persons or organisations, or linked with files of personal information of other sources, without the prior agreement of the health authority.
- The data will only be used for the purpose(s) outlined by the recipient in requesting the data or for purposes approved by the health authority's ethics committee.
- Data files are to be maintained and stored in a secure manner in an environment where they cannot be linked (either electronically or by personal inspection) with other patient records or patient—level data or personal information.
- When no longer required, or by an agreed date, the data files are to be destroyed or returned to the health authority and the authority is to be notified of such destruction.

If data files are made available to consultants engaged by the recipient then the consultants must also agree to these conditions and the health authority must be provided with written evidence of such agreement.

Appendix B

MODEL FOR CROSS-JURISDICTIONAL DATA LINKAGE FOR MEDICAL RESEARCH

The approach for cross-jurisdictional data linkage was developed in 2000 by staff at Department of Health and Ageing, the Health Insurance Commission, the Health Department of Western Australia, the Department of Public Health at the University of Western Australia and the Australian Institute of Health and Welfare in response to a request for data for an approved public health research project. It may be regarded as a model of 'best practice' in the utilisation of administrative data for the production of de-identified linked data files for specific approved purposes.

PROPOSED APPROACH

The process involves two separate stages. The first stage is a memorandum of understanding to share data for an agreed purpose. The second stage includes the production of linked, de-identified data files for approved projects. Each project is covered by its own agreement and the various data custodians supply the data for the project directly to the researchers. For each research project, a unique set of project identifiers is generated by the custodian of the linkage keys and provides the only way of combining the data files into a single linked de-identified file. These project identifiers, being unique to each project, cannot be later used to link additional data from subsequent projects.

This two-stage process ensures that data custodians have full control over the distribution and usage of their data, as each project must be well defined and then individually approved before proceeding. No research can be undertaken without the written approval of every data custodian supplying data to the project. Linked data files are provided only to the individually identified researchers doing the analysis for each project, and must be destroyed when the analyses are complete.

LINKAGE KEY FILE

The linkage key file is produced by a small technical team specialising in data matching, preferably including personnel from all participating institutions. All people involved in the actual linkage and therefore requiring access to the data used in the linkage process must sign confidentiality agreements and be named on a list provided to the steering committee. Any changes to this list will be reported in writing to this committee. No other personnel are allowed access to the files used in this process, as they contain private and confidential information. The work is done on an isolated computer, with all personal demographic data destroyed as soon as the linkage is complete. Transfer of these data files is only done via tape, diskette or CD-ROM personally carried by those personnel taking part in the data matching. The linkage personnel are not permitted to take any part in the analysis of the linked data, or to have any communication about these data with the researchers.

LINKED DE-IDENTIFIED DATA

The linkage key file contains no actual data but does provide coded keys to the data sets involved. Every custodian supplies the approved records from their databases, together with the project identifiers, directly to the nominated researchers for that project. These researchers are also required to sign confidentiality agreements. They can link the data together using the project identifiers, and are the only people granted access to the de-identified linked information. They are specifically forbidden to disseminate copies of the data files, and are required to destroy these files on completion of the analysis.

ETHICS APPROVAL

Ethics approvals from the researchers' institution as well as signed approval from the CEOs of each of the participating institutions are mandatory.

AIHW cat. no. HWI 39
ISSN

Designed and Typeset by Di Walker Design, Canberra