ABSTRACT

As the availability of both health utilization and outcome information becomes increasingly important to health care researchers and policy makers, the ability to link person-specific health data becomes a critical objective. The integration of population-based administrative health databases has been realized in British Columbia by constructing an historical file of all persons registered with the health care system, and by probabilistically linking various program files to this 'coordinating' file. The linkages have achieved a high rate of success in matching service events to person-specific registration records. This success has allowed research projects to be proposed which would otherwise not have been feasible, and has initiated the development of policies and procedures regarding research access to linked data. These policies and procedures include a framework for addressing the ethical issues surrounding data linkage. With continued attention to confidentiality issues, these linked data present a valuable resource for health services research and planning.

ABRÉGÉ

L'accès aux données concernant l'utilisation des services de santé et l'impact des interventions devient de plus en plus important pour les chercheurs en services médicaux et les élaborateurs de politiques de santé. La capacité de lier les données d'identification individuelle devient un objectif essentiel. L'intégration de larges banques de données a été réalisée en Colombie-Britannique en constituant un dossier historique (de coordination) de toutes les personnes enregistrées avec le système de services médicaux. Ceci permet de créer un lien entre plusieurs banques de données d'utilisation des services et les données d'identification individuelle. Ce succès a permis la soumission de nouveaux projets de recherche qui ne seraient autrement pas possible, et le développement de politiques et procédures en ce qui concerne l'accès aux données liées. Ces politiques et procédures incluent une structure pour traiter les questions d'éthique concernant les données liées. En tenant compte des questions de confidentialité, les données liées sont présentement une ressource de grande valeur pour la recherche et la planification des services de santé.

Creating a Population-based Linked Health Database: A New Resource for Health Services Research

*Richard Chamberlayne, MA,¹ Bo Green, BMath,¹
Morris L. Barer, PhD, MBA,^{1,2} Clyde Hertzman, MD, MSc,^{1,2}
William J. Lawrence, MA,³ Samuel B. Sheps, MD, MSc^{1,2}*

Increasingly complex challenges facing the Canadian and other health care systems call for the ability to understand these systems with greater sophistication. Examples like understanding the nature of surgical waiting lists, describing and understanding the utilization of resources by those with a specific health problem, and understanding the relationship between drug utilization and patient use of other health interventions, all require the ability to track individual patients through the system. Additionally, data generated from contact with the health care system can be utilized in some circumstances as a proxy measure of the health of populations.

Linking information from different sources at an individual level has long been accepted as a way to conduct population-based health research. The long-time existence and use of cancer and other disease registries,¹⁻³ perinatal databases,⁴ and a lengthy list of single-use study databases^{5,6} demonstrate the effectiveness of combining information, both for epidemiological study, and for health care management. Some have undertaken the task of linking health records from all available sources for an entire population,⁷ thus taking on a much larger logistical challenge, and creating a much more potent research and management tool.

Provincially administered, and largely publicly funded health insurance in BC covers acute and extended care hospitaliza-

tion, in-home care, prescription drugs for those over the age of 65, diagnostic tests, the fees of physicians and, to a limited extent, fees of other health care practitioners. The central administration of these health service programs has provided British Columbia with the opportunity to develop a set of retrospective population-based administrative health data which is linked at an individual level, across the various programs. Doing these linkages once and preserving them makes it unnecessary to incur the cost of repeated linkage work on the same files, for each new project requiring linked data. Moreover, there will be savings in time to carry out studies which investigate trans-program events, making investigations more feasible which might not have been contemplated due to the expense of creating the data linkages.

The task of creating this health information resource is not only large in size (i.e., large file sizes and populations), it is also large in scope. In other words, linking six different data sources presents many more logistical challenges than linking only two or three sources. In addition, large-scale population-based linkage presents complex ethical challenges. Ensuring confidentiality is key to making such a linked resource useful and accessible.

METHODS

The records in six types of health files have been indexed with a code which is unique to each health care recipient. The file types (and approximate number of records per year) which have been linked are medical services (50 million), hospital separations (discharges plus deaths; 700 thousand), drug prescriptions for the elderly (4.5 million), long term care services (16

1. Centre for Health Services and Policy Research, University of British Columbia
 2. Department of Health Care and Epidemiology, University of British Columbia
 3. British Columbia Ministry of Health
- Correspondence and reprint requests: Bo Green, Centre for Health Services and Policy Research, University of British Columbia, 429-2194 Health Sciences Mall, Vancouver, BC, V6T 1Z3

thousand new assessments), deaths (20 thousand), and births (40 thousand). The linkage begins with files generated in the 1985/86 fiscal year and is complete through (in the worst case) 1995.

The project, titled the BC Linked Health Data Development Project, which was funded by the BC Ministry of Health and situated in the Centre for Health Services and Policy Research at the University of British Columbia, has been carried out using generalized linkage software created by project staff. The linkage strategy has been to create a file of all BC health service recipients, called the Linkage Coordinating File (LCF), and then to link as many records as possible in each program file to a record in the LCF. Links are maintained in dynamic index files. These indexes can be updated by adding new links or removing false links without having to alter the much larger program datasets. Unique individual identifiers can be written onto the program file records as needed for specific requests.

The linkage methodology used probabilistic record linkage theory, pioneered in Canada and now commonly known among health researchers.⁸⁻¹⁰ The uncertain links which are created as a result of probabilistic linkage (sometimes called the grey area), which are commonly reviewed manually in standard probabilistic linkage practice, were anticipated to be too numerous for clerical review. A methodology was developed which organized the grey area for easier review, by grouping candidate links by 'comparison outcome string'—that is, the members in each group had similar patterns of agreement and disagreement. Many of the groups were found to contain either all links, or all non-links, despite large variation in total weight within each group. The grouping technique allowed for more expedient, mostly automated processing of the uncertain links, reducing the clerical review to those groups where a mix of links and non-links was found.

As there is a degree of uncertainty in any probabilistic linkage, a major challenge lies in minimizing the number of false positives (bad links) without unduly raising the number of false negatives (missed true links). Since the linkages described here

	Hospital Separations	MSP Payments	Deaths	Births	Long Term Care	Pharmacare Plan A
PHN	1	4			•	•
MSP Identity #	2	•			•	•
MSP Dependent #	2	•			•	•
Sex	•	•	•	•	•	•
Birth Year	•	•	•	•	•	•
Birth Month	•	•	•	•	•	•
Birth Day	•	3	•	•	•	•
Surname			•	•	•	
Given Name			•	•	•	
Middle Initial			•	•	•	
Postal Code	•	3	•	•	•	

1 Replaced Medical Services Plan (MSP) Identity/Dependent numbers April 1, 1990.
 2 Replaced by Personal Health Number (PHN) April 1, 1990.
 3 Not present prior to April 1, 1988.
 4 Not present prior to April 1, 1994.

File Type	Years Completed	Percent Linked
Hospital Separations	1985/86-1995/96	97.0
MSP Payments	1985/86-1995/96	99.8
Deaths	1985-1995*	97.1
Births	1985-1995	96.6
Long Term Care	1985/86-1996/97	96.0
Prescriptions for Elderly	1985-1995	98.5

* 1985 excludes January-September for this file type, due to insufficient identifying information on the LCF.

will be used for many types of investigative study, it was felt that there would be more risk in introducing false positives. Therefore, a relatively conservative approach was taken in declaring matches to be links.

RESULTS

Linkage of the initial six years of all files were completed in the first phase of the project; subsequent years have been added as they were made available. The identifiers which were used in the probabilistic methodology between the LCF and each program file, are described in Table I. Where identifiers are noted as being present, there was still the possibility of any of them being missing on individual records. The extent of missing values for each identifier tended to decrease over time, making it (generally speaking) slightly easier to link to the LCF in the later years.

The percentage of records which were linked, and the time period completed, for each file type is displayed in Table II. As mentioned earlier, the approach was a con-

servative one, with the intent being to minimize false links in the system. At the same time, the system makes many attempts to link records, in order to maximize the number of links found. Links which are ambiguous are reviewed. Even after completion of each linkage, if at any time during use of the system, a false link is identified, it can be removed from the system. Since the data resource has been developed for research and not administrative purposes, this approach does not affect the day-to-day operation of the resource. It is accepted that there may be a small number of false links in the system, and staff are encouraged to identify and eliminate them. Usually, this means checking study extracts for data exceptions (e.g., death before hospitalization). Thus, the resource becomes more accurate over time. The Linked Health Data Maintenance and Updating Project plans to continue to update this resource as data become available.

The creation of the LCF was critical to the design of the linked resource. Without it, linkages could have suffered from 'prop-

TABLE III
Research Request Categories in Ascending Order of Sensitivity

Category	Request Type Description
A	no person-specific information is released
B	person-specific information is included in the data provided to the requestor, but all individual identifiers are removed
C	person-specific information and identifiers are included in the data provided to the requestor, but no contact with individuals will occur, and individuals are not expected to be directly affected in any way by the research
D	person-specific information and personal identifiers are included in the information provided to the requestor, and there is a possibility or likelihood that the information will indirectly affect future patient management for those individuals, although individuals will not be contacted directly
E	person-specific information and personal identifiers are included in the data provided to the requestor, who intends to use the information to contact subjects or their families. The purpose of any contact and follow-up would necessarily be research; no individuals would be referred for administrative follow-up, although Ministry policies and procedures could eventually be modified based on the research conclusions

agation error', where error from one linkage could affect subsequent linkages. This would occur if identifiers were combined from two data sources, in order to maximize the effectiveness of linkage to a third data source. The LCF design eliminated propagation error, as each source was linked independently to the LCF. Another important project design decision was to maintain the independence of each data source. No attempt was made to integrate the datasets into a single 'health database'. Not only could this create enormous challenges in terms of maintaining the links between records, but it could also have made it difficult for the Ministry of Health to administer the data. As it is, the approval of research requests is the responsibility of the programmatic 'steward' of each data set. The Ministry is able to utilize its existing structure to facilitate access to this enhanced resource.

Access and ethical issues

Requests for access to linked data are complex in two ways. First, since they involve more than one Ministry of Health program area, they represent more than one request for authorization. Second, and perhaps more importantly, each programmatic approval is more than simply approval to access stand-alone databases. Each data steward must consider the request in the context of other program datasets to which access is being simultaneously requested and to which that program's files are being linked.

A Data Access Committee (DAC) was created to oversee the approval process. It

includes Ministry of Health staff, faculty from the Centre, and representatives of the Ministry's Information and Privacy Program and the Research and Evaluation Branch. The mandate of the DAC is to develop and administer policies and procedures¹¹ in order to streamline the processing of research requests for linked health data, while ensuring that research uses of the data are in the public interest, and comply with both British Columbia's Freedom of Information and Protection of Privacy Act (the FOIPP Act) and any relevant sections of the various Acts governing Ministry of Health programs.

The DAC is required to consider a number of different types of research requests as they pertain to issues of concern under the FOIPP Act. These issues include the form in which data are released to researchers, the conditions (i.e., location) under which analyses would be undertaken, security arrangements, and the effects (if any) on individuals of their health care records being used for research. In no case, however, would identifiable personal information ever be published by researchers. Requests have been grouped into the five categories listed in Table III. The DAC considers both scientific peer review and ethics committee review in the approval process, and if these are missing, reserves the right to commission an independent review prior to consideration.

DISCUSSION

This set of population-based, person-specific linked data comprises a valuable

resource for health services research and planning. Administrative files for medical services, hospital visits, prescriptions for the elderly, long term care services, births and deaths have been indexed so that they can be easily and inexpensively merged, allowing the tracking of health services use and outcomes across program boundaries. Other data sets are currently being added.

Among the most challenging tasks in undertaking such a massive effort, was designing a software solution which was general enough to successfully link program files with varying degrees of both completeness and accuracy, to the LCF. Based on an evaluation of generalized probabilistic linkage software, the staff decided to write the programs themselves, in order that any features which were desirable for a particular linkage could be easily integrated into the generalized programs. This meant that the software solution was developed concurrently with the development of the linked files.

The ethical issues surrounding access to a resource made up of linked data are more complex than those pertaining to access to a single data source. Care must be taken that all relevant legislation is satisfied, and that data stewards understand the implications of linking the data for which they are responsible, to data from other sources. Having a formal review process in place which is specifically designed to handle linked data requests can make it easier for researchers and data stewards to assess and monitor the many varieties of requests for access to this valuable resource.

The goal of this project has been to bridge the gap between new requirements for timely and accurate information, and the older data systems which preceded, and generally failed to meet, the new needs. As even newer technology facilitates more efficient data organization and more timely delivery of linked information, the roles of the resource described here will be to provide longitudinality, a bridge between old flat-file technology and future real-time on-line query databases, and to act as a model for addressing the competing priorities of research in the public interest versus security of private information.

While the resource, as described, will continue to be a valuable tool for

researchers interested in issues of importance to public health, health care policy, and health care system management, its utility with respect to understanding the determinants of population health will also be enhanced, as other databases are 'linked in'.

REFERENCES

1. Cote TR, O'Brien TR, Ward JW, et al. AIDS and cancer registry linkage: Measurement and enhancement of registry completeness. *Prev Med* 1995;24(4):375-77.
2. Potosky AL, Riley GF, Lubitz JD, et al. Potential for cancer related health services research using a linked Medicare-tumor registry database. *Med Care* 1993;31(8):732-48.
3. Langley JD, Botha JL. Use of record linkage techniques to maintain the Leicestershire Diabetes Register. *Computer Methods & Programs in Biomedicine* 1994;41(3-4):287-95.
4. Stanley FJ, Croft ML, Gibbins J, et al. A population database for maternal and child health research in Western Australia using record linkage. *Paediatric and Perinatal Epidemiol* 1994;8:433-47.
5. Katz SJ, Armstrong RW, LoGerfo JP. The adequacy of prenatal care and incidence of low birth-weight among the poor in Washington State and British Columbia. *Am J Public Health* 1994;84(6):986-91.
6. Dodds L, Marrett LD, Tomkins DJ, et al. Case-control study of congenital anomalies in children of cancer patients. *BMJ* 1993;307:164-68.
7. Roos NP, Shapiro E (Eds.). Experience with a population-based health information system. *Med Care* (Supplement). 1995;33(12).
8. Newcombe HB. *Handbook of Record Linkage*. New York: Oxford University Press, 1988.
9. Newcombe HB. Record linking: The design of efficient systems for linking records into individual and family histories. In: Baldwin JA, Acheson ED, and Graham WJ (Eds.), *Textbook of Medical Record Linkage*. New York: Oxford University Press, 1987.
10. Gill LE, Baldwin JA. Methods and technology of record linkage: Some practical considerations. In: Baldwin JA, Acheson ED, and Graham WJ (Eds.), *Textbook of Medical Record Linkage*. New York: Oxford University Press, 1987.
11. British Columbia Ministry of Health and Centre for Health Services and Policy Research. Access Policy for Research Uses of Linked Health Data. Victoria, 1996.

Received: May 15, 1997

Accepted: January 19, 1998

Letters, from page 257

both human populations and individuals. As noted by the authors, among others,^{2,6} studies including ecological or group-level variables are necessary to examine the causal link between social environments, human behaviour, and disease development.

It was therefore surprising to read the conclusions of this paper, which seem to indicate that the ecological approach has limited utility in understanding disease etiology. Elliott and Dean write: "The difficulties of studying health issues...indicate that caution must be exercised when interpreting these results. Most important, this is ultimately an ecological level analysis. In order to link heart disease and psychosocial stress, the ideal dataset would contain measures of heart disease outcomes and indicators of stress, all preferably measured at *the individual level* [emphasis added], allowing conclusions to be made about *individual level* [emphasis added] relationships."¹ (p. 140) Presumably the basis for this view is the "ecological fallacy", "a logical fallacy inherent in making causal inferences from group data to individual behaviours."² (p.819) In the epidemiologic literature this fallacy has discredited the ecological approach, with ecologic or group-level studies being

viewed as crude attempts to determine individual-level associations.^{2,6}

Ecologic studies, however, are not simply substitutes for individual-level correlational studies.^{2,6} Very often a variable measured at the group level reflects a different construct or phenomenon than the same variable measured at the individual level. To illustrate, Schwartz writes: "The construct referenced on the ecological level may be the context or social environment in which individuals live, distinct from the attributes of those individuals. Thus, poverty as an individual characteristic and poverty as a neighbourhood characteristic may exert different, independent effects on health."² (p. 820) By viewing ecological studies as mere substitutes of individual-level studies, the potential independent effects (or causal influences) of the broader social environment (e.g., families, peer groups, communities, and legislation) are ignored or neglected. The utility of the ecologic approach is that it provides a unique opportunity to ascertain the social or group contexts which contribute to disease risk.

In the case of heart disease (and most other illnesses), it is likely that a multi-level analysis including both individual-level variables (e.g., physiologic risk factors) and

ecologic-level variables (e.g., job characteristics) will lead to a better understanding of etiology, in particular the mechanisms by which environmental stressors influence perceptions of stress and disease risk in individuals. Multi-level analysis will also indicate whether interventions to reduce risk should focus on the individual level, or the ecological level, or some combination of both.

Janice Husted, PhD

Associate Professor, Department of Health Studies and Gerontology, University of Waterloo, Waterloo, ON

REFERENCES

1. Elliott SJ, Dean A. An ecologic analysis of psychosocial stress and heart disease in British Columbia. *Can J Public Health* 1998;89(2):137-41.
2. Schwartz S. The fallacy of the ecological fallacy: The potential misuse of a concept and consequences. *Am J Public Health* 1994;84:819-24.
3. Von Korff M, Koepsell T, Curry S, et al. Multi-level analysis in epidemiologic research in health behaviors and outcomes. *Am J Epidemiol* 1992;135:1077-82.
4. Susser M. The logic in ecological: I. The logic of analysis. *Am J Public Health* 1994;84:825-29.
5. Susser M. The logic in ecological: II. The logic of design. *Am J Public Health* 1994;84:830-35.
6. Scribner R. Paradox as paradigm- The health outcomes of Mexican Americans. *Am J Public Health* 1996;86:303-5.