THE NEW ZEALAND CONFERENCE ON DATABASE INTEGRATION AND LINKED EMPLOYER-EMPLOYEE DATA

Matt Spittal
Knowledge Group
Ministry of Social Development

The New Zealand Conference on Database Integration was held at Te Papa, Wellington, in March 2002. The purpose of the conference was to provide a forum for a wide range of international and local speakers to discuss the added value associated with database integration. Key themes to emerge from the conference were: (1) that international experience has shown that integrated data can provide decision makers with policy-relevant data in a timely fashion without increasing respondent burden; and (2) that issues of privacy and confidentiality must also be addressed when integrating data.

Data integration refers to the linking of two or more data sets, usually to obtain longitudinal information. Broadly, this can take the form of linking administrative data, which is essentially non-research data collected in the course of programmatic activities for client-level tracking, service provision, or decision making; and linking survey data, which is collected for answering specific research questions. Integrated data offers a number of advantages for research and policy advice over cross-sectional data, such as a richer insight into phenomena such as the interactions between employers and employees; between unemployment, training and employment; and between health, education and economic outcomes. Data integration can provide decision makers with information relevant to policy advice without increasing respondent burden, and it offers the possibility of addressing questions that cannot be answered by cross-sectional data.

An important component of research into any source of data is to understand the limitations of the data. With respect to integrated data, several limitations are evident. Administrative records, which form the basis of many integrated data sets, are not developed for the purpose of research and may be unsuitable for various reasons, including:

- a lack of adequate control variables;
- the fact that all outcomes of interest are not measured;
- the data may only be available for the periods that the person is in a programme; and
- the undetermined reliability of the administrative data.

Sometimes administrative data is difficult to access because of confidentiality issues, such as obtaining informed consent. Many of these problems may also occur with the use of survey data (e.g., incomplete records).

The conference provided a forum for international and national speakers to discuss the usefulness and applicability of data integration techniques. James Farber described the United States Census Bureau's experience of developing an integrated database using national administrative records. Factors such as the increasing cost of conducting the decennial census led the Bureau to consider the possibility of using linked administrative data for the purpose of conducting a population headcount.

The Bureau initiated a formal research programme in 1999, with the goal of developing a national research file (called the Statistical Administrative Records System, or StARS) to support such demographic and decennial census applications as frame improvement, survey and census non-response, population estimates, coverage and content evaluation, targeting of special procedures, and so on. Data were combined from a variety of sources, including tax records, housing assistance records, military registration records, health insurance records, and Social Security Numident files. The resultant database contained about 261 million person records with demographic characteristics that was linked to approximately 106 million housing unit records. Comparison with established benchmarks, such as the 2000 Census, showed that StARS undercounted the population by about 24 million people, but produced very similar population distributions on a number of key variables. These results suggest that although the StARS is not yet suitable as a replacement to the census, expanding the scope of administration records will increase the accuracy of the data base, and perhaps offer an alternative to the Census in the future.

An example of the usefulness of integrated data-sets for research in New Zealand was provided by Tony Blakely. As part of the New Zealand Census Mortality Study's aim to investigate the link between mortality and socio-economic status, the 1981, 1986, 1991, and 1996 census records were integrated with three years of subsequent mortality data. Using information gathered from integrated data has a number of advantages over stand-alone epidemiological studies, principally due to:

- the comparatively short time needed to obtain useful data;
- the reduced cost of data collection;
- the relative ease with which databases can be updated; and
- the statistical power gained from using population responses.

Although only the linked data between the 1991 census and the 1991-1994 mortality records have been analysed, already the data has yielded some new and important information. In particular, these include the extent to which mortality data underestimates the number of Māori and Pacific Island deaths, and the relationship between income and mortality.

In addition to providing a forum for the discussion of the applications of integrated data, the conference addressed issues of privacy and confidentiality. The perspectives of Statistics New Zealand and the Privacy Commissioner are pertinent to this.

Statistics New Zealand is concerned that the use of record linking does not diminish its relationship with data providers and the public, who have particular concerns about privacy and how their data is used. For this reason, their policy on linking data is based on several principles. Broadly, these policies relate to the conditions under which data can be integrated, the circumstances where the Privacy Commissioner and under relevant groups should be consulted, the dissemination of results, and the safe storage of linked data. In all cases, the main privacy issue for Statistics New Zealand is not whether a data integration project is possible, rather whether it should be undertaken given the privacy issues involved.

The New Zealand Privacy Commissioner, Bruce Slain, discussed his views on the privacy issues associated with data linking. As a general principle, there are differences in regard to the extent to which people consider information to be "sensitive". What is burdensome or intrusive to one person may be less so to another. It is the view of the Privacy Commissioner that a consideration of this view is central to designing a survey, and to the integration of data from two or more surveys. Data that appears to be of a low sensitivity, in some cases, may become more sensitive when linked.

His view was that if the principles in the Privacy Act were to be distilled to their vital essence, then they would involve two key aspects: openness and purpose. Openness and transparency enhance the accountability of public sector agencies. Clarity of purpose is essential for answering these key questions:

- Why is there a need to collect personal information?
- Who will it be shared with?
- How long should it be kept?

Answers to these questions will be apparent to an agency with a clear view of its purpose. For this reason, particular care is needed in the collection of unnecessary administrative data for interesting statistical purposes, as opportunistic collections are not consistent with the privacy principles and may fall outside an agency's lawful purpose.

The Privacy Commissioner indicated that he believed the Data Integration Protocols developed by Statistics New Zealand provide a valuable reference point for agencies contemplating data integration. He therefore proposed embedding these principles in a mandatory Code of Practice, which would apply to the same public bodies to which the protocols apply. It was his view that a Code of Practice would provide added clarity for agencies embarking on a data integration project.

In sum, the conference provided a timely opportunity to explore the potential for integrating information collected by different government organisations for use in developing better public policies. International and national speakers showed some of the possible uses of integrated data, and examined the key issues of privacy and confidentiality associated with the use of integrated data-sets. Finally, the conference provided a forum to discuss protocols and procedures for the appropriate use of integrated data for research purposes.