Sources of Data for Rural Health Research: Development of an Inventory of Canadian Databases

D. Heng, R. W. Pong, J. R. Pitblado, C. Lagacé, M. Desmeules

ABSTRACT. Secondary data sources can often be used to help address questions about the health status, health behavior, health resources allocation, and utilization of health services of rural Canadians. But the task of deciding which Canadian databases are amenable to rural health research remains a challenge. As part of a larger research project titled “Canada’s Rural Communities: Understanding Rural Health and Its Determinants,” an inventory of 51 Canadian databases that have the potential of being used for rural health research was compiled, and it continues to be maintained and updated. The websites maintained by two of Canada’s leading statistical data centers were systematically searched, along with other published articles and national reports, to produce this inventory. The criteria used to determine which data sources to include in this inventory are: (1) databases containing data at the national level that can be accessed by researchers, (2) databases containing data that are relevant to a variety of rural health issues, and (3) databases containing data that could be partitioned into rural and non-rural geographies. Detailed information is available by searching the inventory of national rural health research-related databases through the internet (www.cranhr.ca) or by contacting the lead author of this article. This article examines some of the issues in developing this resource and demonstrates the usefulness of its contents to Canadian and other rural health researchers.

Keywords. Database, Internet, Research, Rural health, Secondary data sources.

Population health research and planning activities are very much dependent on the availability of secondary data, which provide raw material for analysis. Data sources can be categorized into those that are created from data routinely collected for administrative purposes, usually for reasons other than research, and other data sources created from survey responses collected from a sample of respondents. Regardless of the method of data collection, these large-scale, population-level databases can be described as basic health research tools and are essential to answering questions geared towards the improvement of the health of a population (Wolfson, 1996). As more emphasis is being placed on understanding the broader determinants of health, rather than simply focusing on medical interventions (Frank, 1995), secondary data are being used more frequently in health research. For example, health indicators have been developed using...
secondary data in an effort to help policymakers and health planners make informed decisions (Statistics Canada and Canadian Institute for Health Information, 2001). The availability of such databases makes them ideal for analyzing a variety of health issues at the population level.

The case for characterizing rural health research in Canada as underdeveloped has been made in the past (Pitblado et al., 1999; Pong, 2000; Watanabe and Casebeer, 2000), and various recent Canadian health commissions have identified the need to support rural health research (Commission on the Future of Health Care in Canada, 2002; National Advisory Council on Rural Health, 2002; Standing Senate Committee on Social Affairs Science and Technology, 2002). Questions such as “What is the health status of rural Canadians and how do they compare to urban Canadians?” remain central to the development and evaluation of rural health programs and policies. The impetus for constructing this inventory was based on our frustration in finding national-level studies that adequately provided analyses of the health status of rural Canadians. Rural health research based on secondary data can help address questions about the health status, health determinants, health behaviors, health resource allocation, and utilization of health services. The development of rural health indicators can be used to describe the health conditions of rural populations, relative to non-rural populations (Pitblado et al., 1999). Examples of health indicators that have been included in recent rural health profile reports include life expectancy, mortality rates for certain diseases, obesity rates, activity limitation, hospital discharge rates, and average length of hospital stay (Australian Institute of Health and Welfare, 2003; Eberhart et al., 2001; Statistics Canada and Canadian Institute for Health Information, 2001).

Many secondary databases are available in Canada, but few researchers know how to efficiently locate and use such databases. Rural health researchers have additional obstacles to overcome, such as defining or operationalizing “rural” and addressing issues of confidentiality (arising mostly from concerns in relation to small numbers) before they can effectively use such sources of secondary data for research purposes. Given the need for further rural health research in Canada and the information that may be obtained from rural health indicators constructed from secondary data, we hope to increase the research capacity of rural health researchers through our work on this inventory of secondary databases. Our development of this inventory represents, to the best of our knowledge, one of the first attempts to help rural health researchers identify a broad range of national databases that have the potential to support rural population health research in Canada.

Methods

An initial inventory of data sources used for the development of rural health indicators was created in 1999 as part of a study titled “Assessing Rural Health: Towards Developing Health Indicators for Rural Canada” (Pitblado et al., 1999). Its goal was to determine whether health indicators could be readily developed, given the types and nature of health and related data sources at the researchers’ disposal. It represents one of the first attempts in Canada at examining the feasibility of developing rural health indicators at a national level, in a systematic manner. Using the Statistics Canada and Organization for Economic Cooperation and Development (OECD) definitions of rural, this project proposed a classification scheme that consisted of five types of health indicators (health status indicators, health determinant indicators, health behavior indicators, health resources indicators, and health services utilization indicators). The researchers constructed a small number of rural health indicators to illustrate the approaches adopted and some of the difficulties encountered.
In 2002, as part of a larger research project titled “Canada’s Rural Communities: Understanding Rural Health and Its Determinants,” an expanded team of researchers reviewed and updated the initial list of databases compiled in 1999. New data sources were added to more accurately reflect the availability of data sources that can be used for rural health research in Canada. The criteria used to determine which data sources to include in this inventory are: (1) databases containing data at the national level that can be accessed by researchers, (2) databases containing data that are relevant to a variety of rural health issues, and (3) databases containing data that could be partitioned into rural and non-rural geographies.

For the purpose of this study, data sources were considered to be national if they reported data for at least nine of the ten Canadian provinces, even if they did not collect data in any of the three territories. Given the very broad nature of rural health research, it was decided to not just incorporate data sources that provided information specifically about health, but also to include data sources with information about the broader determinants of health such as education, labor force participation, income, and environment. In order to determine whether the third criterion was satisfied, the operational definitions of rural used in this inventory were taken from Definitions of “Rural” (du Plessis et al., 2002), which outlines six ways of defining rural in Canada, five of which were adopted by the inventory (table 1). The sixth method of defining rural is based on the Canadian postal code. Areas with “0” as the second character of the six-character postal codes are considered by some as rural. This method was not included in this inventory due to the many changes to the Canadian postal code system since 1996, which have made this method unreliable.

Canada’s two leading organizations of statistical data are Statistics Canada and the Canadian Institute for Health Information (CIHI). Statistics Canada is Canada’s central statistical agency that is legislated to provide statistical information that helps Canadians better understand their country (Statistics Canada, 2004). Part of CIHI’s mandate is to provide and coordinate the provision of accurate and timely data and information required for establishing sound health policy, effectively managing the Canadian health system, and generating public awareness of the factors affecting good health (Canadian Institute for Health Information, 2004b).

Two strategies were used to identify and select databases for this inventory: (1) identifying websites that contained information on the various data sources, and (2) consulting with individuals who have worked extensively with these data sources. The websites maintained by Statistics Canada and CIHI were systematically searched and reviewed with a view to identifying the types and nature of health and related databases that are available and that are potentially useful to rural health research. Other published reports produced by these two agencies, various health research centers, health planning agencies, and ministries of health were screened to provide additional information and to highlight difficulties encountered by rural health researchers who have used specific data sources. Finally, knowledgeable individuals from these organizations were consulted for their input.

<table>
<thead>
<tr>
<th>Definition of Rural</th>
<th>Building Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census “Rural Areas”</td>
<td>Enumeration area (EA)</td>
</tr>
<tr>
<td>Rural and small town MIZ</td>
<td>Census subdivision (CSD)</td>
</tr>
<tr>
<td>OECD “rural” communities</td>
<td>Census consolidated subdivision (CCS)</td>
</tr>
<tr>
<td>OECD “predominantly rural regions”</td>
<td>Census division (CD)</td>
</tr>
<tr>
<td>Ehrensaft’s “Beale codes”</td>
<td>Census division (CD)</td>
</tr>
</tbody>
</table>

Table 1. Alternative definitions of rural and their corresponding geographical building blocks (adapted from du Plessis et al., 2002).
organizations were consulted to help identify, as well as to provide details on, appropriate data sources for this inventory.

Results

During the study period (September 2002 to April 2003), close to 700 data sources were reviewed, including approximately 600 databases from Statistics Canada (Statistics Canada, 2003) and 21 databases from CIHI (Canadian Institute for Health Information, 2004a). As of October 2003, the inventory contained 51 databases, and this number is expected to increase as new databases become available in the future. The inventory is available in two web-based formats and can be found on the website of the Centre for Rural and Northern Health Research (www.cranhr.ca) under “Resources for Rural Health Research.” A complete PDF version of the inventory can be downloaded from this site, as well as PDF versions of the individual data source sheets.

Table 2 provides a breakdown of the data sources included in the inventory according to nine different themes. Data pertaining to labor activity, health outcomes, and the level of education of respondents were the most commonly collected by databases included in the inventory. Of the 51 databases included in the inventory, 31% were assigned to four or more of the themes, with only 10% of databases assigned to six or more of the themes. These themes were chosen to reflect the range of data elements found within the different databases, and data sources were assigned to multiple themes to show the areas where data collection has focused and where data gaps still exist. For example, there are few databases that deal with the physical environment as a health determinant. There are also few databases concerning transportation, which has been identified as a major problem for accessing health care in rural areas. Eighty percent of the inventory databases contain information that has been collected within the past five years.

In order to make the inventory more user friendly, information on each database was summarized in a data sheet. Each data sheet in this inventory describes the main database characteristics and assesses whether the data are collected, analyzable, and can be released at appropriate geographical levels for rural health research (fig. 1). In particular, the data sheet identifies whether the data can be partitioned into rural and non-rural, or into finer geographical categories to describe differing levels of rurality. Each data sheet uses a common format that includes such information as: thematic coverage, originating agency, availability, dates of data collection and data release, geographical coverage, lowest geographical level for data collection and data release, geographical level used in

<table>
<thead>
<tr>
<th>Theme</th>
<th>Data Source Examples</th>
<th>Total No. of Data Sources[^a]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor activity</td>
<td>Labor force survey</td>
<td>26</td>
</tr>
<tr>
<td>Health outcomes</td>
<td>Discharge abstract database</td>
<td>23</td>
</tr>
<tr>
<td>Level of education</td>
<td>Census of population</td>
<td>23</td>
</tr>
<tr>
<td>Social surveys</td>
<td>General social survey</td>
<td>20</td>
</tr>
<tr>
<td>Health behaviors</td>
<td>Canadian tobacco use monitoring survey</td>
<td>16</td>
</tr>
<tr>
<td>Activity limitation / injuries</td>
<td>Participation and activity limitation survey</td>
<td>10</td>
</tr>
<tr>
<td>Health human resources</td>
<td>Southam medical database</td>
<td>6</td>
</tr>
<tr>
<td>Environmental</td>
<td>National pollutant release inventory</td>
<td>4</td>
</tr>
<tr>
<td>Aboriginal surveys</td>
<td>Aboriginal peoples survey</td>
<td>3</td>
</tr>
</tbody>
</table>

[^a] Number of data sources does not sum to 51 because multiple responses were permitted.
<table>
<thead>
<tr>
<th>Database name:</th>
<th>Discharge Abstract Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract:</td>
<td>This database collects data on hospital discharges. These data are also used to evaluate patient expected length of stay and resource consumption. In addition to inpatient hospitalizations, some jurisdictions / facilities provide information on same day surgery discharges.</td>
</tr>
<tr>
<td>Maintained by:</td>
<td>CIHI</td>
</tr>
<tr>
<td>Availability:</td>
<td>Data disclosure is determined by CIHI's Privacy and Confidentiality Policy (see <a href="http://www.cih.ca">www.cih.ca</a>). If privacy and confidentiality criteria are met, data can be released at the sub-provincial level. Costs are dependent on the level data required, plus programming and processing fees.</td>
</tr>
<tr>
<td>Start Date:</td>
<td>1963</td>
</tr>
<tr>
<td></td>
<td>1979/80 - Present form</td>
</tr>
<tr>
<td>Release Date:</td>
<td>Approximately 5 months after reference period for general release</td>
</tr>
<tr>
<td>Frequency of collection:</td>
<td>On-going</td>
</tr>
<tr>
<td>Data Collection:</td>
<td>CIHI receives data directly from participating hospitals.</td>
</tr>
<tr>
<td>Sample size:</td>
<td>About 4.3 million records per annum</td>
</tr>
<tr>
<td>Geographic coverage:</td>
<td>Coverage has been increasing over time. Currently, about 85% of hospital inpatient discharges in Canada are included, consisting of all hospitals in every province and territory except Quebec and parts of Manitoba (Manitoba only submits 40% of their data) for both acute/inpatient data. It also contains day surgery data for all provinces except Quebec, Alberta, and parts of Manitoba. If comprehensive geographic coverage is required, the Hospital Morbidity Database may be used instead for national reporting.</td>
</tr>
<tr>
<td>Lowest geographic level collected:</td>
<td>Postal code of patient. Collection and completeness of postal codes varies (e.g., Quebec reports only the Forward Sortation Area - the first 3 digits of the postal code)</td>
</tr>
<tr>
<td>Lowest geographic level of release:</td>
<td>Forward Sortation Area or Enumeration Area depending on request</td>
</tr>
<tr>
<td>Existing rural variable:</td>
<td>No</td>
</tr>
</tbody>
</table>

Rural definitions that can be constructed from this database include (building block):  
- Census "rural areas" (Enumeration Area)  
- Rural and small town definition (Census Subdivision)  
- Metropolitan area and census agglomeration influenced zones (Census Subdivision)  
- OECD "rural" communities definition (Census Consolidated Sub.)  
- OECD "predominantly rural regions" definition (Census Division)  
- Ehrensaft's "Beale codes" (Census Division)  

Data Elements:  
- Selected data elements include:  
  - Most responsible physician/diagnosis  
  - Principal procedure  
  - Patient gender, date of birth, postal code  
  - Institution/hospital number.  
  - Admission/discharge, (e.g., admission category, length of stay).  

Notes:  
The advantages of the Discharge Abstract Database include its timeliness, value-added data elements (e.g., Case Mix groups and Resource Intensity Weights), and coverage of same-day surgery for selected jurisdictions/facilities.  
Data are reported based on the region of the patient's residence, not region of hospitalization. Consequently, these figures reflect the hospitalization experience of residents of the region wherever they are treated as opposed to the comprehensive activity of the region's hospitals. It represents how frequently residents of a given area received hospital care rather than the volume of services provided by hospitals.  
For more information, contact dad@cihi.ca

Figure 1. Sample data sheet from the inventory of rural health research-related databases.
standard reports, existing “rural flags” and how they may be constructed, data elements, and potential problems in using the data.

Of the 51 databases included in the inventory, only 15 explicitly included some sort of a “rural flag.” These flags are mostly based on Statistics Canada’s “Rural and Small Town” definition of rural, which refers to population living in towns and municipalities outside the commuting zone of larger urban centers (i.e., outside the commuting zone of centers with population of 10,000 or more people). Given the inclusion criteria of this study, it was possible to include the remaining 36 data sources in this inventory even though they did not possess a “rural flag” because they had collected the necessary geographic information (usually the six-character postal code) that would allow rural health researchers to differentiate between rural or non-rural by using the postal code conversion files.

Discussion

There are many examples, both Canadian and international, of inventories of databases that have been compiled to facilitate research efforts. Examples of disease specific inventories in Canada include the Inventory of Canadian Cardiovascular Disease Databases (Centre for Chronic Disease Prevention and Control, 2003) and the Inventory of Injury Data Sources and Surveillance Activities (Centre for Surveillance Coordination, 2004). Initiatives such as the Database for Rural Health Research in Progress in the U.S. (Maine Rural Health Research Center, 2004), the Rural Canada Research Inventory in Canada (Centre for Rural Studies and Enrichment, 2004), the Rural Health Research Register in Australia (McGrail et al., 2003), and the Rural Research Project Inventory in Canada (Canadian Rural Partnership, 2004) are efforts to catalogue the various ongoing or recently completed rural health research projects in various countries. But none of these registries and inventories has mentioned the potential of using secondary data for rural health research. Thus, the research project described in this article has filled a major information and knowledge gap. Additionally, several important lessons have been learned in the process of developing this inventory of rural health research-related databases.

Although our study focuses on Canada, it has international implications. Given the similarity in geography and dispersion of population and quality of health information collected in countries such as Canada, Australia, and the U.S., there are lessons that can be learned from our experiences. First, secondary data are generally not collected for the purpose of doing rural health research. In order to conduct studies using such data, it is necessary to define or operationalize “rural.” But there is not a universally agreed-upon definition of rural in Canada. Pong and Pitblado (2001) have identified some definitions of rural commonly used and have discussed their strengths and limitations. Similarly, the definitions suggested by du Plessis et al. (2002) allow Canadian rural health researchers to choose the definition that best fits their research purpose. But a pre-determined notion of rural (available in some databases as a “rural flag”) may not be deemed suitable by all researchers. By using the six-character Canadian postal code and the postal code conversion files, researchers have greater flexibility in differentiating between rural and non-rural based on their own understanding of rural. Agencies that collect health and related data should be encouraged to collect postal code information and to make such information available when releasing the data. Without the postal code information, 70% of the data sources listed in this inventory would have to be excluded because there would be no way of extracting data from rural residents or locations.
Second, definitions of rural are typically based on geographic boundaries defined for administrative purposes. However, some determinants of health such as climatic conditions, air quality, and water pollution are not constrained by geopolitical borders. Ecozones may be more appropriate geographic units for studying the effects of such factors on health (Pong et al., 2002). Further research is required to determine the best ways to effectively link data on the environment and health data in order to gauge the impact of the environment on human health.

Third, given the large amount of resources needed to collect high-quality data from individuals across Canada, it is expected that this inventory includes the majority of data sources of health that could be useful to rural health research in Canada after having canvassed the two principal providers of national data in Canada. As noted earlier, though, data on certain health determinants are inadequate in Canada. Examples include data on the physical environment and transportation, which are considered by rural health experts as important factors in influencing rural health and rural health care. Much work is needed in Canada to remedy such deficiencies. Some discontinued databases were included in the inventory because they represent the only in-depth examination of a particular topic at a national level. Alternatively, some discontinued databases, such as several Canadian smoking surveys, have been replaced by newer databases but were still included in this inventory to provide historical continuity.

Fourth, even when data can be partitioned into rural and non-rural, there are other practical difficulties arising from data suppression due to small numbers and undersampling of rural populations in surveys. An inherent characteristic of rural regions is sparse population. Unfortunately, although some recent Canadian health surveys provide reliable estimates at the health region level (which allows researchers to distinguish between rural health regions and non-rural health regions), the sample sizes of many surveys are determined with a view to providing reliable estimates only at the provincial level. This means that very often, rural and non-rural comparisons cannot be made. A related problem is the fact that the release of data based on small numbers is not permitted due to concerns about statistical reliability and privacy protection. While we recognize the importance of these two issues, we feel that many database holders are being overly cautious, and this is hindering rural health research.

Conclusion

The Inventory of National Rural Health Research-Related Databases represents an extensive search for national data sources that are relevant to and useful for rural health research in Canada and a review of databases using a common framework of assessment. This effort is intended to further rural health research, particularly rural population health research, by letting researchers know what data sources are available and how they can be accessed and used for rural health research. It will be an on-going effort, as the inventory will continually be updated as new data sources become available. Because it is posted on the internet, the inventory is widely accessible and freely available.

Although this study focuses on Canada, it has international implications. Because there are some similarities between Canada and countries like Australia and the U.S. with respect to geography, population distribution, health care systems, and health information, similar inventories of secondary data sources could be compiled in those countries, even though different countries may understand rural differently. The development of health indicators and the expansion of rural population health research based on or utilizing secondary data in different countries would allow international comparative rural health studies to take place.
Acknowledgement

The authors wish to thank the Canadian Institute for Health Information (CIHI) for funding from the Canadian Population Health Initiative program for the research program “Canada’s Rural Communities: Understanding Rural Health and Its Determinants.” As a component of this research program, the Inventory of National Rural Health Research-Related Databases (Heng et al., 2003) is an effort to support and promote rural population health research.

References


