Evaluation of the coverage of the linked Canadian Community Health Survey and hospital inpatient records

Michelle Rotermann and Claude Nadeau

Statistics Canada
Michelle.Rotermann@statcan.ca Claude.Nadeau@statcan.ca

Abstract

Rationale: The Longitudinal Health and Administrative Data Initiative is a joint project between the provincial Ministries of Health and Statistics Canada. One objective of the Initiative is to address information gaps through research linking person-oriented hospital inpatient (HPOI) and survey data. Evaluation of the coverage as a result of the linkage between theses data sources is an important preliminary step before engaging in further analyses based on the linked file. Methods: To evaluate the coverage of the linkage between the Canadian Community Health Survey (CCHS) and HPOI, the number of individuals admitted to hospital according to each source was compared. CCHS respondents were considered to have been hospitalized if a record containing similar demographic characteristics and/or health number (HN), and an admission to an acute-care facility between September 2000 and November 2001 was found in HPOI. Survey weights were applied to these records to calculate the estimated number of individuals hospitalized. Because HPOI is a virtual census of hospital admissions, it was regarded as the standard; differences between HPOI and weighted CCHS counts indicated possible linkage failure and undercoverage. Quebec records were excluded because the HNs in HPOI are scrambled and other linking variables are not provided/incomplete. Results: According to HPOI for September 2000 - November 2001, 1,612,269 Canadians were hospitalized. Weighted estimates from the CCHS adjusted for agreement to link and agreement and plausible HN were lower (17% and 10%, respectively). The under-coverage rates were similar between men and women; lower among persons living in Manitoba than the rest of Canada; higher among individuals aged 12-74 than among those aged 75+. Conclusion: The number of Canadians (outside Quebec) aged 12+ who were hospitalized, based on HPOI, was higher than the number based on CCHS-HPOI linkage. The discrepancy varied by province/territory and age; therefore, use of the linked file could lead to biased estimates. Nevertheless, the coverage as a result of the linkage was acceptable and use of the file for population-based research is encouraged.

Introduction

The Longitudinal Health and Administrative Data (LHAD) Initiative is a project between the provincial Ministries of Health and Statistics Canada (STC). One objective of the initiative is to address health information gaps through research combining individual-level administrative health care records, vital events data, such as cause and date of death, the cancer registry and information from STC's general health surveys. This research will provide pan-Canadian and comparative information across provinces and will improve understanding of the relationships between risk factors, socio-economic characteristics, health status measures and health care system utilization.

Initial research on the rate of linkage between CCHS and HPOI quantified the proportion of CCHS survey respondents who had been hospitalized during the 1994/95-2004/05 period, but coverage of the linkage was not assessed (Nadeau C, Beaudet MP, Marion J; 2006). Evaluation of the coverage as a result of the linkage between person-level inpatient hospitalization data (HPOI) and survey data is an important preliminary step before engaging in further analyses based on the linked file.

Background

Health care service data such as hospital discharges are collected as part of the billing and payment process of each province and territory. Though the data were not initially collected for research, these data are successfully used to describe the quantity and quality of health care services and health outcomes of inpatients (Rotermann M, 2004; Johansen H, Nair C, Mao L et al, 2002; Wray NP, Ashton CM, Kuykendall DH et al. 1995).

Administrative hospital discharge records are a viable source of information about inpatient hospital stays but do not include information about services provided on an outpatient basis, including services received via the emergency department or day surgeries. Currently, deaths occurring outside of hospital are not routinely linked with hospitalization data. Hospital discharge records also provide limited information about the non-medical determinants of health, such as socio-economic status and behavioural/lifestyle factors, for example smoking status or body mass index (BMI). In contrast, STC's health surveys are a rich source of information about health status, determinants of health and to a lesser extent health care service use.

Combining micro-level administrative data with STC's data holdings can minimize many of the limitations inherent in each individual data source, thereby facilitating a more complete understanding of how individuals fare in their contacts with the health care system.

Methods

Data sources

Canadian Community Health Survey

The CCHS is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. It covers the household population aged 12 or older in all provinces and territories, except members of the regular Forces and residents of institutions, Indian reserves, and some remote areas (approximately 98% of the population). Data for cycle 1.1 were collected between September 1, 2000 and November 3, 2001 from a sample of 131,535 persons and the response rate was 84.7%. The vast majority of the respondents (90.8%; 119,383) gave permission to link their survey records with administrative health records and about three-quarters provided a valid HN (73.1%; 87,268). More information about the CCHS is available in a published report (Béland, 2002). The sample used for this study consists of 98,450 respondents aged 12+ in all provinces and territories except Quebec.

In order for estimates produced from survey data to be representative of the target population, and not just the sample itself, survey weights were used. Generally survey weights reflect the differing probabilities of selection and differential response rates. The survey weight is the number of persons in the population represented by each survey respondent. Each record is therefore weighted by the inverse of the probability of selecting the person and getting a response from him or her.

With record linkage additional survey weights are required because not all survey respondents agree to link and not all respondents who agree to link provide a valid HN. For this study, two sets of survey weights specifically designed for record linkage were derived. The first set adjusted for respondents who agreed to have their survey data linked to their administrative health records; the second set adjusted for agreement to link and HN validity.

Currently, STC does not have access to jurisdictional population registries which would provide a means of verifying the HNs provided by the respondent. Instead all provinces provide check-digit formulas which are used to verify that the HNs provided by the respondents are at least plausible. Check-digits are not a substitute for registries which would provide first and last names, birth dates, addresses and HNs but they can detect accidental transcription errors, such as the inversion of two numbers and provide a simple method of distinguishing meaningful numbers from strings of random digits.

The territories did not provide a check-digit. As a result HN length and other known characteristics of the number, such as expected first alpha or numeric were used to assess HN plausibility.

Table 1: Number of Canadian Community Health Survey respondents who agreed to have their survey responses linked with their administrative health records and the proportion who provided a valid health number, by selected characteristics, Canada excluding Quebec, 2001

	Agreed to link		Agreed to linkage and provided valid HN	
	Unweighted	%	Unweighted	%
Total	98,450	90.4	72,671	73.8
Province/territories				
NF	3,533	90.3	3,037	86.0
PE	3,238	88.7	2,236	69.1
NS	4,938	92.8	4,109	83.2
NB	4,634	92.8	3,746	80.8
ON	35,674	90.8	24,917	69.8
MB	7,653	90.4	5,692	74.4
SK	7,417	92.6	6,142	82.8
AB	12,757	88.3	9,156	71.8
BC	16,493	90.1	11,991	72.7
YK	633	78.2	439	69.4
NT	882	88.1	701	79.5
NU	598	84.6	505	84.4
Sex				
Males	45,585	90.3	32,202	70.6
Females	52,865	90.5	40,469	76.6
Age groups				
12-24	19,246	91.8	13,592	70.6
25-34	14,482	90.9	10,162	70.2
35-44	18,892	90.2	13,942	73.8
45-54	16,036	89.6	11,968	74.6
55-64	11,493	90.0	8,732	76.0
65-74	9,778	90.3	7,713	78.9
75+	8,523	89.5	6,562	77.0
12-74	89,927	90.5	66,109	73.5

Data source: Canadian Community Health Survey, 2001

Note: CCHS respondents were asked if Statistics Canada had their permission to link information collected during the interview with provincial health information, including information on past and continuing use of services such as hospitals, clinics, doctor's offices or other services provided by the province was determined by answering the following question. HN validity was assessed using a provincially provided check-digit formula. HN length and other known characteristics of a particular jurisdiction's HN were used in the absence of a check-digit.

Health person-oriented information database (HPOI)

The Health Person-Oriented Information (HPOI) Database, maintained by Statistics Canada, contains information on inpatient hospital separations (discharges and in-hospital deaths) and is composed of hospital discharge abstracts from virtually all acute-care, and some psychiatric, chronic and rehabilitative hospitals across Canada. Data are available for all provinces and the Northwest Territory from fiscal year 1994/1995 onwards; data for 1992/1993 and 1993/1994 are available for some provinces. Although the HPOI database includes all records from the Hospital Morbidity Database (HMDB), not all records can be

submitted to the person-oriented linkage process. Annually, approximately 13% of the HMDB records are excluded from the additional processing that enables the files to be analyzed at the person level: 10% because they are records for newborns and the remaining 3% because the record contains either an invalid identifier (HN) or is for a person residing outside the province.

During the linkage process, records belonging to the same individual are identified by means of the patient's health number (HN). A critical part of the process is to determine if more than one person is using the same HN (e.g. spouses, mother and child). Other data elements in the HMDB (e.g., sex and birth date) can be used to differentiate among individuals using the same HN. Unlike the HMDB which is event-based, HPOI is a person-oriented database that allows for longitudinal analysis of inpatient hospitalizations. Each record contains demographic (for example, postal code, date of birth), non-medical administrative (such as health number, dates of admission and separation) and clinical information (for example, diagnoses and procedures).

Over the 14 months studied (September 1, 2000-November 3, 2001), there were 2.3 million hospital discharges of 1,612,269 individuals aged 12+ from acute-care hospitals, outside Quebec. Admissions to non-acute hospitals were excluded because coverage of other hospital types varied by province.

Quebec records excluded

Quebec HPOI records cannot be linked to CCHS survey records because the Quebec HPOI records provided to Statistics Canada contain scrambled health numbers (HNs), no date of birth and incomplete postal codes.

Analytical techniques

To evaluate the coverage of Cycle 1.1 of the CCHS and HPOI, the number of individuals admitted to hospital according to each data source was compared (an approach that has not been used before). Probabilistic linkage was used to identify CCHS respondents who were hospitalized. Generalized Record Linkage software (GRLS) developed at Statistics Canada was used to execute the linkage.

Record linkage is a process of bringing together two or more separately recorded pieces of information belonging to the same entity. By combining datasets, new information is learned because each dataset contains some information not available in the other (Winglee M, Valliant R, Scheuren F 2005). The two data sources linked for this study contained many variables –HPOI has more than 100; CCHS has more than 1000, but only a few fields appeared in both and were unique enough to be useful as matching fields for linkage. A CCHS respondent was considered to have been hospitalized if a record containing similar demographic characteristics (for example, birth date, sex, postal code) and/or HN, and an admission date to an acute-care facility between September 1, 2000 and November 3, 2001 was found in the HPOI database.

Probabilistic linkage does not require complete agreement on the matching variables. Instead the quality of the match between pairs of records is rated using sophisticated algorithms to evaluate the likelihood of a correct match between two records (Fellegi and Sunter 1969). Points are given or subtracted depending on the similarity between fields. The higher the score, the more likely the two records pertain to the same individual. Point values are chosen in various ways, including through probabilities, hence the name probabilistic linkage. For this study, high positive scores were assigned if the HNs were identical between two records and the issuing province matched; if the values were similar but not exact, according to a string comparison algorithm, a lower positive score was assigned, reflecting partial agreement; if the values on the two records were totally different, points were subtracted. The number of points assigned to each pair reflected the importance of the matching variable, which was usually related to its uniqueness. Total linkage weights for each pair of CCHS-HPOI records were calculated by summing the scores assigned to each pair of linking variables, including HN, postal code, sex, and date of birth.

Total linkage weights optimally form a bi-modal distribution. Typically pairs of records which score above a selected upper threshold are accepted as "true" matches; pairs below are rejected; pairs between the selected cut-offs are considered potential matches and require manual intervention. To minimize the

amount of manual review the cut-off points chosen for this study were identical which meant that scored pairs could only take one of two values: match or non-match. Pairs with scores greater than or equal to 250 were considered to be matches; pairs below were considered to be non-matches.

Chart 1: Example of how pairs of records were assessed and scored using Generalized Record Linkage Software (GRLS)

Id	Prov	Birthdate	Postal code	HN	Sex	Score calculated by GRLS	Match?	Commentary
A	ON	11/06/1964	L9Y3B9	3512345678	F	447	Y	All fields match
1	ON	11/06/1964	L9Y3B9	3512345678	F			
F	MB	24/07/1927	R0A0T0	55667788	M	-308	N	Nothing matches
1	ON	11/06/1964	L9Y3B9	3512345678	F			
B	MB	21/05/1945	R0A0T0	-	M	268	Y	HN missing; everything else matches
7	MB	21/05/1945	R0A0T0	4624252627	M			
B	MB	21/05/1945	R0A0T0	-	M	-244	N	HN missing; nothing else matches
1	ON	11/06/1964	L9Y3B9	3512345678	F			

CCHS record ▢ HPOI record ▢

To calculate the number of individuals hospitalized using CCHS two sets of survey weights were applied to the records of CCHS respondents for whom records were also found in the HPOI database. The first set was adjusted for agreement to link (8,216); the second was adjusted for agreement to link and HN validity (6,810). Two coverage rates were calculated by dividing each set of weighted CCHS estimates by hospital inpatient counts and then multiplying by 100%. Because HPOI is a virtual census of hospital admissions, the count of hospitalizations from HPOI was regarded as the standard; the difference between the HPOI count and each weighted count from the CCHS was examined. Standard errors and 95% confidence intervals were calculated using the bootstrap technique. Statistical significance was tested using the t-test ($p<0.05$).

Limitations

Coverage of the data sources is not exactly the same. For example, CCHS excludes institutionalized persons; HPOI is a virtual census of acute-care hospitalizations. The inclusion of seniors will help us to quantify the extent of missed hospitalization attributable to institutionalized seniors. Quebec was excluded from the analysis because the HNs of Quebec residents are scrambled and other identifying information is not provided to Statistics Canada. Individuals under age 12 were excluded because the target population of the CCHS (.1 cycles) covers only the 12+ population. Linkage of Manitoba CCHS records with HPOI was compromised by the collection of incomplete birth dates and the provision of family rather than personal HNs (Manitoba residents have both). Probabilistic linkage results in some false links and misses some true links. Both types of errors may have occurred in this study. Intra-provincial and inter-provincial mobility may have also lowered the number of links, although the effect of mobility on the linkage is believed to have been minimized by restricting the study period to hospitalizations occurring in the same 14 month-period overlapping with the CCHS collection. Mobility within a province would have less of an effect on linkage than a move to another province. Moves within a province would only change the postal code, whereas a move to another province/territory may also change the HN since it is provincially/territorially assigned and non-transferable across provinces.

Findings

According to HPOI, between September 1, 2000 and November 3, 2001, 1,612,269 Canadians were admitted to an acute-care hospital (excluding Quebec). Weighted estimates from the CCHS adjusted for agreement to link or agreement and valid HN were lower (1,333,478 and 1,450,211, respectively). Regardless of which CCHS estimate was used, coverage rates were similar between men and women; lower among persons living in Manitoba (68.9 % and 78.1%) than the rest of Canada; higher among individuals aged 12-74 (86.0% and 94.2%) than among those aged 75+ (70.2% and 73.5%). While the coverage rates for the territories are relatively high, the small number of CCHS records linked to HPOI precludes any analysis featuring jurisdictional comparisons.

Table 2: Number hospitalized in acute-care hospitals, September 1, 2000-November 3, 2001, aged 12+, Canada excluding Quebec, by province/territory and data source

Health person-oriented information (HPOI)	Canadian Community Health Survey (CCHS)				Coverage rates	
	Weights adjusted for:					
	Agreed to linkage		Agreed to linkage and provided valid HN		Agreed to linkage	Agreed to linkage and provided valid HN
	Count (N)	Unweighted (n)	Weighted (N)	Unweighted (n)	Weighted (N)	%	%
Canada‡	1,612,269	8,216	1,333,478	6,810	1,450,211	82.7	89.9
Province/territory							
NF	41,455	309	37,430	278	39,769	90.3	95.9
PE	11,840	298	10,155	237	11,071	85.8	93.5
NS	68,109	407	56,650	350	60,799	83.2	89.3
NB	68,099	480	57,839	424	62,487	84.9	91.8
ON	760,982	2,843	635,679	2,230	694,240	83.5	91.2
MB	88,305	672	60,877	581	68,985	68.9*	78.1*
SK	88,571	752	77,338	660	78,755	87.3	88.9
AB	210,192	1,054	171,149	862	185,899	81.4	88.4
BC	268,876	1,255	220,773	1,060	241,838	82.1	89.9
YK	1,830	48	1,827	37	1,992	99.8	108.9
NT	3,015	77	2,884	70	3,323	95.7	110.2*
NU	995	21	878	21	1,051	88.2	105.6
Sex							
Females†	996,926	5,241	843,973	4,356	904,061	84.7	90.7
Males	615,337	2,975	489,504	2,454	546,150	79.6	88.8
Age group							
12-24	180,655	897	155,850	718	165,050	86.3	91.4
25-34	282,609	1,258	242,187	1,047	274,388	85.7	97.1
35-44	215,326	1,003	185,754	826	205,475	86.3	95.4
45-54	187,202	932	159,856	779	175,734	85.4	93.9
55-64	183,314	1,026	156,849	854	165,678	85.6	90.4
65-74	229,522	1,274	198,716	1,089	218,733	86.6	95.3
75+	333,641	1,825	234,267	1,497	245,154	70.2*	73.5*
12-74†	1,278,628	7,318	1,099,212	6,810	1,205,058	86.0	94.2

‡ Reference category is Canada minus each province (Rest of Canada) † Reference category

* Significantly different from reference category ($p<0.05$).

Data source: Canadian Community Health Survey, 2001 and Health person-oriented information 2000/01-2001/02

Notes: Counts, linkage and coverage rates as of July 31, 2007; counts, linkage and coverage rates may change pending final revision.

Conclusion

The value of record linkage is now a well-established technique in epidemiological studies of population health. By linking information from routinely collected administrative health data files with survey data already available at STC much can be learned about the health determinants, different types of health care utilization and health outcomes. Because the survey data are population-based the information to be learned will be generalizable to the household population of Canada, with few exceptions.

Coverage evaluation is a fundamental pre-requisite to further analyses based on the CCHS-HPOI linked file. The number of Canadians (outside Quebec) aged 12+ who were hospitalized, based on HPOI, was higher than the number based on the CCHS-HPOI linkage. The discrepancy varied by province/territory and age. Therefore, use of the linked file could lead to bias. For example, because the coverage rate of persons over the age of 75 is relatively low, use of the linked file underestimates hospitalization in this age group. Nevertheless, the overall estimates produced as a result of this linkage are of an acceptable level. Future research using this linked file is encouraged, bearing in mind the aforementioned limitations.

References

Béland Y. Canadian Community Health Survey -methodological overview. *Health Reports* (Statistics Canada, Catalogue no. 82-003) 2002; 13(3): 9-14.

Canadian Institute for Health Information. Trends in acute inpatient hospitalizations and day-surgery visits in Canada, 1995-1996 to 2005-2006. *Analysis in Brief.* Available on-line at http://www.cihi-ices.ca....

Fellegi IP, Sunter AB. A theory for Record Linkage. *Journal of the American Statistical Association* 1969; 64: 1183-1210.

Johansen H, Nair C, Mao L, et al. Revascularisation and heart attack outcomes. *Health Reports* (Statistics Canada, Catalogue no. 82-003) 2002; 13(2): 35-46.

Nadeau C, Beaudet MP, Marion J. Deterministic and probabilistic record linkage. Proceedings of Statistics Canada Symposium 2006

Rotermann M. Infection after cholecystectomy, hysterectomy, or appendectomy. *Health Reports* (Statistics Canada, Catalogue no. 82-003) 2004; 15(4): 11-24.

Statistics Canada, Household Surveys Methodology Division. External linkage production report: Data years: F1992 to F2004. (Unpublished, 2006).

Winglee M, Valliant R, Scheuren F. A case study in record linkage. *Survey Methodology* (Statistics Canada, Catalogue no. 12-001) 2005; 31(1): 3-11.

Wray NP, Ashton CM, Kuykendall DH. Using administrative databases to evaluate the quality of medical care: A conceptual framework. *Social Science Medicine* 1995; 40(12): 1707-15.