Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Evaluation of the Coverage of Linked Canadian Community Health Survey and Hospital Inpatient Records

by Michelle Rotermann

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

Evaluation of the Coverage of Linked Canadian Community Health Survey and Hospital Inpatient Records

Michelle Rotermann[1]

Abstract

Background: Evaluation of the coverage that results from linking routinely collected administrative hospital data with survey data is an important preliminary step to undertaking analyses based on the linked file. Data and methods: To evaluate the coverage of the linkage between data from cycle 1.1 of the Canadian Community Health Survey (CCHS) and in-patient hospital data (Health Person-Oriented Information or HPOI), the number of people admitted to hospital according to HPOI was compared with the weighted estimate for CCHS respondents who were successfully linked to HPOI. Differences between HPOI and the linked and weighted CCHS estimate indicated linkage failure and/or undercoverage. Results: According to HPOI, from September 2000 through November 2001, 1,572,343 people (outside Quebec) aged 12 or older were hospitalized. Weighted estimates from the linked CCHS, adjusted for agreement to link and plausible health number, were 7.7% lower. Coverage rates were similar for males and females. Provincial rates did not differ from those for the rest of Canada, although differences were apparent for the territories. Coverage rates were significantly lower among people aged 75 or older than among those aged 12 to 74.

Key Words: Coverage, Health surveys, Hospital records, Probabilistic linkage.

1. Introduction

Record linkage is used in health studies to obtain more complete information, to fill gaps in existing datasets, and/or to improve data quality (Fair, 2004, Fair and Whitridge, 1997). This study was motivated by the need to assess the coverage of the linkage between the Canadian Community Health Survey (CCHS) and Health Person-Oriented Information (HPOI), *an administrative database of hospital records*. Initial research on the rate of linkage between the CCHS and HPOI estimated the proportion of CCHS respondents who had been hospitalized during the 1994/1995 to 2004/2005 period, but coverage has yet to be assessed (Nadeau et al., 2006). Evaluation of the coverage is essential if the linked file is to be used for epidemiologic research. It is important to know if findings will be biased, that is, if survey respondents with certain characteristics are more likely than others to have been linked.

HPOI and the CCHS are complementary sources of data. HPOI does not have information about non-medical determinants of health, such as socio-economic and lifestyle factors. For example, hospital records do not contain information about smoking status or body mass index (BMI), two important risk factors. The CCHS, by contrast, is a rich source of information about health status and determinants of health, but lacks the detail needed to study hospitalization. Combining HPOI with the CCHS reduces many of the limitations of each source, and thereby facilitates a more complete understanding of what brings Canadians in contact with the health care system and how they fare within the system.

The two main objectives of this study were to: 1) evaluate the coverage of the linked CCHS and HPOI by calculating coverage rates; and 2) identify characteristics of CCHS cycle 1.1 respondents who were less likely to be in the linked file.

[1] Michelle Rotermann, Health Information and Research Division, Statistics Canada, 100 Tunney's Pasture Driveway, K1A 0T6, Ottawa, Canada (Michelle.Rotermann@statcan.gc.ca)

2. Data and methods

2.1 Data sources

2.1.1 Canadian Community Health Survey

The Canadian Community Health Survey is a cross-sectional survey that collects information about health status, health care use and health determinants. It covers the household population aged 12 or older in the provinces and territories, except members of the regular Forces and residents of institutions, Indian reserves and other Aboriginal settlements, and some remote areas. The rate of coverage is 98% in the provinces, 97% in the Northwest Territories, 90% in the Yukon, and 71% in Nunavut. Data for cycle 1.1 were collected from September 1, 2000 through November 3, 2001 from a sample of 131,535 people; the response rate was 84.7%.

Quebec records were excluded from the analysis because Quebec HPOI records cannot be linked to their corresponding CCHS records (22,667) because the hospital records provided to Statistics Canada contain scrambled health numbers (HNs), no date of birth and incomplete postal codes.

All CCHS information, including provincial HNs and postal codes, is self-reported by respondents, and the extent of error in these variables is unknown. However, data capture applications used by interviewers contain features that check for inconsistent answers, out-of-range responses or invalid alpha-numeric sequences. More information about the CCHS is available in a published report (Béland, 2002). CCHS respondents were asked for permission to link information collected during the interview with their provincial health information, including past and continuing use of services such as hospitals, clinics, doctor's offices or other services provided by the province; nine in ten respondents gave permission. The sample used for this study consists of 72,354 (66.5%) respondents aged 12 or older in all provinces and territories except Quebec, who agreed to link and provided a valid (HN) (Table 2.1.1-1).

Table 2.1.1-1
Number and percentage of Canadian Community Health Survey respondents who agreed to have their survey responses linked with their administrative health records and who provided valid HN, by selected characteristics, Canada excluding Quebec, 2000/2001

	Agreed to link		Agreed to link and HN valid	
	Number	%	Number	%
Total	98,450	90.4	72,354	66.5
Province/territories				
Newfoundland and Labrador	3,533	91.3	2,933	75.8
Prince Edward Island	3,238	88.7	2,236	61.2
Nova Scotia	4,938	92.8	4,108	77.2
New Brunswick	4,634	92.8	3,746	75.0
Ontario	35,674	90.8	24,917	63.4
Manitoba	7,653	90.4	5,552	65.5
Saskatchewan	7,417	92.6	6,142	76.7
Alberta	12,757	88.2	9,155	63.3
British Columbia	16,493	90.1	11,990	65.5
Territories	2,113	83.9	1,575	62.6
Sex				
Female	52,865	90.5	40,334	69.1
Male	45,585	90.3	32,020	63.4
Age groups				
12-74	89,927	90.5	65,824	66.3
75+	8,523	89.5	6,530	68.6

Source: 2000/2001 Canadian Community Health Survey.

Survey weights were used so that estimates produced from the CCHS data were representative of the target population, not just the sample itself. The survey weight is the number of people in the population represented by each respondent. Survey weights reflect the differing probabilities of selection and response. Each record is, therefore, weighted by the inverse of the probability of selecting the person and getting a response from him or her (Statistics Canada, 2008a). Additional survey weights are required for record linkage because not all respondents agree to link and not all those who agree to link, provide a valid HN. For this study, survey weights, adjusted for agreement to link and provision of a valid HN, were calculated.

Statistics Canada does not have access to provincial health insurance databases against which the HNs provided by CCHS respondents could be verified. Instead, all provinces and territories provide check-digit formulas that are used to verify that the HNs are at least plausible. Although check-digits are not a substitute for databases that contain first and last names, birth dates, addresses and HNs, they can detect accidental transcription errors, such as the inversion of two numbers, and offer a simple method of distinguishing meaningful numbers from strings of random digits.

2.1.2 Hospital data

The Health Person-Oriented Information (HPOI) database, maintained by Statistics Canada, contains information about inpatient hospital separations (discharges and in-hospital deaths) from virtually all acute-care and some psychiatric, chronic and rehabilitative hospitals.

HPOI is a person-level dataset derived from discharge records (which can reflect multiple discharges of the same person) in the Hospital Morbidity Database (HMDB). Sequential person-level HPOI records can be used to construct each patient’s hospitalization history. During the linkage process, records belonging to the same individual are identified from the patient’s HN and demographic and diagnosis/intervention information (for example, sex, birth date, sex-specific procedures) (Household Survey Methods Division, 2006).

Hospital records pertaining to the past fiscal year are added to HPOI annually. With each additional year of data, the entire HPOI process is rerun to ensure internal consistency of the demographic information at the person-level for patients with multiple hospital discharges.

Reabstraction studies, which validate the accuracy of hospital records, have found that the non-medical administrative data elements (essential for record linkage) are of high quality. For example, 99% of a random sample of discharge records for hospital stays from September through November 2000 had correct HNs, and 91% of postal codes were error-free (Richards et al., 2001).

Statistics Canada has hospital data with HNs for all provinces (except Quebec) and the Northwest Territories from fiscal year 1994/1995 onwards; data for 1992/1993 and 1993/1994 are available for some provinces. While the HPOI database includes the vast majority of records from HMDB, about 3% of records for patients aged 12 or older (the target population of this study) were excluded because of missing or invalid HNs (Household Survey Methods Division, 2006).

From September 1, 2000 through November 3, 2001, there were 2.3 million discharges of 1,624,972 people aged 12 or older from acute-care hospitals outside Quebec. Discharges from non-acute hospitals were excluded from this study because coverage of such hospitals is inconsistent across provinces.

The target populations of the Canadian Community Health Survey (CCHS) and HPOI differ somewhat. The CCHS excludes full-time members of the Canadian Forces and residents of Indian Reserves, of institutions (for instance, nursing homes and prisons) and of some remote areas. HPOI is a census and, therefore, these groups are included among hospitalizations. In an effort to match the target populations of the CCHS and HPOI more closely, hospitalizations that could be identified as pertaining to the on-reserve or the institutionalized population were removed from this analysis.

The on-reserve population is a derived census variable created by identifying census sub-division (CSD) type according to criteria established by Indian and Northern Affairs Canada (INAC), as well as selected CSDs that correspond to northern communities in Saskatchewan, the Northwest Territories, and the Yukon (Statistics Canada, 2002). The postal code conversion file (PCCF+) (Statistics Canada, 2008c) and a list of facilities used by the Residential Care Facility survey (Statistics Canada, 2008b) were used to identify institutional residents. Hospitalizations pertaining to 31,330 residents of Reserves and associated lands were removed from HPOI, as were hospitalizations of 21,299 residents of institutions. Removal of these 52,629 records, which amounted to about 3% of the HPOI patients hospitalized during the study period, brought the population covered by HPOI more in line with the CCHS target population.

2.2 Analytical techniques

2.2.1 Probabilistic record linkage

Probabilistic record linkage was used to identify CCHS respondents who were hospitalized. The linkage between the CCHS and HPOI was done with Generalized Record Linkage software (GRLS) developed at Statistics Canada. The two data sources contain many variables, but only a few fields appear in both and are distinct enough to be useful in matching for linkage. A CCHS respondent was considered to have been hospitalized if a record containing an HN and/or similar demographic characteristics (for example, birth date, sex, postal code) and an admission date to an acute-care facility between September 1, 2000 and November 3, 2001 was found in HPOI.

Probabilistic linkage does not require complete agreement on the matching variables. Rather, the quality of the match between pairs of records is rated with algorithms that evaluate the likelihood of a correct match (Fair and Whitridge, 1997, Fellegi and Sunter, 1969). Points were given or subtracted depending on the similarity of the values between fields. For instance, high positive scores were assigned if the HNs were identical and the issuing province of the HN matched; if the values were similar but not exact, 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 of linking variables reflected their importance as matching variables, which typically was related to uniqueness. For example, because there are only two possible values for the sex of the respondent/patient, matches on this field scored fewer points than if the postal codes or HNs matched.

Total linkage scores for each pair of CCHS-HPOI records were calculated by summing the scores assigned to each pair of linking variables. The higher the total linkage score, the more likely the two records pertained to the same individual. Total linkage scores ideally form a bi-modal distribution. When pairs of records scored above the selected threshold, they were accepted as “true” matches; pairs below the threshold were rejected. To eliminate the need for manual review, the cut-off points chosen for this study were identical, which meant that each pair of records could have only one of two values: match or non-match.

3. Results

To evaluate the coverage of the linkage between cycle 1.1 of the CCHS and HPOI, the number of people admitted to hospital according to each data source was compared. Survey weights, adjusted for agreement to link and HN validity, were applied to the records of CCHS respondents for whom records were also found in the HPOI database. The HPOI count of hospitalizations was regarded as the standard. The coverage rate was calculated by dividing the weighted estimates of CCHS respondents who successfully linked to HPOI by HPOI counts, minus records identified as pertaining to residents of Indian Reserves or associated lands or of institutions and then multiplying by 100. Differences between the HPOI counts and the weighted estimates from the CCHS were examined. Standard errors and 95% confidence intervals were calculated for the coverage rates using the bootstrap technique. Statistical significance was tested using the t-test ($p<0.05$) (Rao et al, 1992, Rust and Rao, 1996).

According to HPOI, from September 1, 2000 through November 3, 2001, 1,572,343 people were admitted to an acute-care hospital (excluding Quebec) (Table 3-1). Weighted estimates from the CCHS, adjusted for agreement to link and valid HN, were 7.7% lower (1,451,272).

Coverage rates were similar for males and females (91.0% and 93.1%). Provincial rates did not differ significantly from the rate for the rest of Canada. However, based on the CCHS, the estimated number of residents of the territories who were hospitalized was considerably higher than the HPOI number. As a result, the coverage rate for the territories exceeded 100%.

Coverage rates for most age groups were similar. The exception was seniors aged 75 or older whose rate (76.2%) was significantly lower than that of people aged 12 to 74 (96.4%).

Table 3-1
Number hospitalized in acute-care hospitals and coverage rates, September 1, 2000 to November 3, 2001, by selected characteristics and data source, aged 12 or older, Canada excluding Quebec

	Health Person-Oriented Information (HPOI)	Canadian Community Health Survey (CCHS)		CCHS/HPOI Coverage rates		
					95% confidence interval	
	Number	Unweighted number	Weighted number	%	from	to
Total	1,572,343	6,785	1,451,272	92.3	88.9	95.7
Province/territory						
NF	41,394	272	40,445	97.7	83.6	111.8
PE	11,784	237	11,061	93.9	79.6	108.1
NS	67,226	348	60,419	89.9	78.0	101.7
NB	67,542	423	62,203	92.1	81.7	102.5
ON	753,970	2,230	694,463	92.1	86.6	97.6
MB	82,386	567	69,739	84.6	73.6	95.7
SK	82,778	659	78,664	95.0	86.4	103.7
AB	202,498	863	186,301	92.0	83.3	100.7
BC	258,883	1,062	241,647	93.3	85.3	101.3
Territories	3,882	124	6,331	163.1*	139.3	186.9
Sex						
Females†	971,087	4,343	904,318	93.1	88.8	97.5
Males	601,249	2,442	546,955	91.0	85.4	96.5
Age group						
12-74†	1,252,336	5,299	1,207,392	96.4	92.4	100.4
75+	320,007	1,486	243,881	76.2*	70.2	82.2

† reference category

* significantly different from reference category (p<0.05); for provincial comparison, significantly different from rest of Canada, for example, Ontario compared with Canada minus Ontario
Source: Canadian Community Health Survey, 2001 and Health person-oriented information 2000/2001 to 2001/2002.

4. Discussion

The significantly lower coverage rate for seniors aged 75 or older was anticipated because the two data sources did not pertain to exactly the same populations. The CCHS excludes residents of institutions, but they are included in the hospital data (HPOI). Institutionalization is considerably more common among seniors than among younger people: overall, fewer than 2% of Canadians live in an institution, but at age 75 or older, the figure is 16% (Turcotte and Schellenberg, 2007).

In the absence of direct information in HPOI records about patients' place of residence, the postal code in combination with the PCCF+ and the Residential Care Facilities list was used to determine if patients lived in an institution. More than 20,000 institutional residents were identified and subsequently removed from HPOI using the

PCCF+. Nonetheless, the coverage rate for seniors aged 75 or older remained significantly below the rates for younger people.

Use of the PCCF+ and the Residential Care Facilities list to identify institutions based only on the postal code is not ideal. Institutions that accounted for the majority of the population sharing a postal code had a higher chance of being identified and subsequently removed from the HPOI counts. As well, institutions in urban areas have more precise postal codes, and therefore, residents of such institutions were more likely to have been removed from HPOI. Rural and outlying suburban areas and smaller towns often have the same postal code for multiple enumeration/dissemination areas. Consequently, the coding is far less precise than for centralized urban postal codes, which are usually linked to a single enumeration/dissemination area. Therefore, residents of institutions in rural and outlying suburban areas and smaller towns likely remained in the HPOI counts.

The coverage rate in the territories is also problematic, in that the linked CCHS-HPOI estimates exceeded the standard (HPOI). This, however, is less of a concern, because the small number of CCHS records linking to HPOI (124) precludes future analyses featuring this subpopulation. Before the removal of on-reserve residents from the HPOI count, the coverage rate for the territories was 113%; after their removal, the rate was 163%. It is unclear why the linked HPOI-CCHS estimate is so much higher than HPOI. Records of CCHS respondents identified as living in the territories were reviewed to determine if some had high survey weights, which might explain the discrepancy between the HPOI and HPOI-CCHS counts. No discrepant weights were found; the average weight was 51, with weights ranging in value from 11 to 178. In addition, hospitalizations pertaining to military personnel could not be identified and removed from HPOI. Full-time members of the Armed Forces are excluded from CCHS, and their inclusion may affect the coverage rate.

5. Conclusion

The value of record linkage is well established in epidemiological studies of population health. Linking information from routinely collected administrative health data such as HPOI with survey data like the CCCHS holds promise for discoveries about health determinants, different types of health care use and health outcomes. Coverage evaluation is a fundamental pre-requisite to analyses that integrate health-related information from multiple sources based on the CCHS-HPOI linked file.

This evaluation shows that the overall coverage rate is high, often over 90%, although some CCHS respondents, notably seniors, had lower rates. Even this limitation is manageable, however, as long as users of the file explicitly acknowledge that findings pertain only to the general household population (the target population of the CCHS), and not to the total population, particularly residents of institutions.

Acknowledgements:
The author thanks Claude Nadeau for his assistance and thoughtful comments.

References

Béland, Y. (2002). Canadian Community Health Survey – Methodological Overview, *Health Reports* 13, 9-14.

Fair, M. (2004). Generalized Record Linkage System - Statistics Canada's Record Linkage Software, *Austrian Journal of Statistics*; 33, 37-53.

Fair, M.E. and Whitridge, P. (1997). Tutorial on Record Linkage, *Federal Committee on Statistical Methodology*. www.fcsm.gov/working-papers/RL_chap12.html.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-210.

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

Nadeau, C., Beaudet, M.P. and Marion, J. (2006). Deterministic and Probabilistic Record Linkage, *Proceedings: Symposium 2006, Methodological Issues in Measuring Population Health*. Statistics Canada: Ottawa, Canada.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992), Some Recent Work on Resampling Methods for Complex Surveys, *Survey Methodology*, 18, 209-17.

Richards, J., Brown, A., and Homan, C. (2001), The Data Quality Study of the Canadian Discharge Abstract Database, *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistics Canada: Ottawa, Canada www.statcan.ca/english/freepub/11-522-XIE/2001001/session16/s16a.pdf.

Rust, K.F. and Rao, J.N.K. (1996), Variance Estimation for Complex Surveys Using Replication Techniques, *Statistical Methods in Medical Research*, 5, 281-310.

Statistics Canada. (2001), 2001 Census Dictionary, Statistics Canada: Ottawa, Canada.
http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=92-378-X

Statistics Canada (2008a). Canadian Community Health Survey — Estimation, Statistics Canada: Ottawa, Canada. www.statcan.gc.ca/cgin/imdb/p2SV.pl?Function=getSurvey&SurvId=3226&SurvVer=0&InstaId=15282&InstaVer=1&SDDS=3226&lang=en&db=imdb&adm=8&dis=2#b7

Statistics Canada (2008b). Residential Care Facilities (RCF). Statistics Canada: Ottawa, Canada.
www.statcan.ca/bsolc/english/bsolc?catno=83-237-X .

Statistics Canada (2008c). Postal Code Conversion File (PCCF): Update. , Statistics Canada: Ottawa, Canada.
www.statcan.ca/bsolc/english/bsolc?catno=92-153-UCB .

Turcotte M. and Schellenberg G. (2007). *A Portrait of Seniors in Canada*, Statistics Canada: Ottawa, Canada.