CMAJ OPEN

Research

Self-reported versus health administrative data: implications for assessing chronic illness burden in populations. A cross-sectional study

Martin Fortin MD MSc, Jeannie Haggerty PhD, Steven Sanche MSc, José Almirall MD PhD

Abstract

Background: Various data sources may be used to document the presence of chronic medical conditions. This study examined the agreement between self-reported and health administrative data.

Methods: A randomly selected cohort of participants aged 25–75 years recruited by telephone from the general population of Quebec reported on the presence of 1 or more chronic conditions from a candidate list of 12 conditions: diabetes, hypertension, thyroid disorder, any cardiac disease, cancer diagnosis in the previous 5 years (including melanoma but excluding other skin cancers), asthma, osteoarthritis, rheumatoid arthritis or lupus, osteoporosis, chronic obstructive pulmonary disease, intestinal disease and hypercholesterolemia. We also used health administrative data from Quebec's universal health insurance provider to identify participants' chronic conditions. Unique identifiers allowed linkage of both data sources to the individual participant. The frequencies of the 12 conditions and the prevalence of multimorbidity (≥ 2, ≥ 3 and ≥ 4 conditions) were analyzed for each data source.

Results: We analyzed data for 1177 participants (mean age 53 [standard deviation 12.4] yr; 684 women [58.1%]). We found low (but varied) agreement between the 2 data sources, with the poorest agreement for hypercholesterolemia (κ = 0.04 [95% confidence interval (CI) 0.01 to 0.07]) and the best for diabetes (κ = 0.82 [95% CI 0.76 to 0.88]). Prevalence estimates of multimorbidity obtained with health administrative data were lower than those obtained with self-reported data regardless of the operational definition used. Most participants with multimorbidity were identified by self-report.

Interpretation: We argue for the use of self-reported chronic conditions in the study of multimorbidity, as health administrative data based on the billing system in Quebec seem to underestimate the prevalence of many chronic conditions, which results in biased estimates of multimorbidity.

Various data sources may be used to document the presence of chronic medical conditions in studies of multimorbidity.[1] Self-reporting of diagnoses, as in surveys, is based on the premise that the patient is knowledgeable enough to report on chronic conditions accurately and that the questionnaire is adequate to guide the patient in this process. However, surveys require a high level of planning[2] and are generally resource intensive. Health administrative data offer an efficient alternative to document the presence of chronic conditions for a large population of patients at low cost.[3] However, the validity of the data could be questioned, as studies have shown inconsistent agreement between diagnoses of chronic conditions obtained by self-report and with health administrative data.[4–6]

We examined the concordance between self-reported data and health administrative data for 12 common chronic conditions and evaluated the usefulness of the 2 data sources to estimate the prevalence of multimorbidity.

Methods

Setting and participants

This study was nested in the Program of Research on the Evolution of a Cohort Investigating Health System Effects

Competing interests: None declared.

This article has been peer reviewed.

Correspondence to: Martin Fortin, Martin.Fortin@usherbrooke.ca

***CMAJ Open* 2017. DOI:10.9778/cmajo.20170029**

(PRECISE).[7] The goal of PRECISE was to examine the effect of patient-centred and effective primary health care on the evolution of chronic illness burden and health functioning in 1 population and, in particular, 2 vulnerable subgroups: people with multimorbidity and poor people. As it was a longitudinal study, cohort participants completed a self-administered questionnaire on current health status and health behaviours 2 weeks, 12 months and 24 months after recruitment to enable evaluation of primary health care received in the previous year. In brief, potential participants were randomly selected from the general population of 4 regions of the province of Quebec. Participants had to be aged 25–75 years, be able to respond to written and oral questions in English or French, and reside in 1 of the 4 networks identified. The sample was selected by random-digit dialing from March to April 2010 within the administrative boundaries of the 4 networks. To ensure random selection, the eligible adult selected in the household was the adult with the most recent birthday.

Sources of data

Two weeks after recruitment, participants who provided verbal consent were mailed a self-administered questionnaire. Participants did not receive any compensation for their participation.

Participants reported on their demographic characteristics, use of health care services and health, including responding to specific questions about a list of chronic medical conditions. They were instructed to report on diagnoses only if a physician had confirmed them previously. We collected information about 21 conditions (Appendix 1, available at www.cmajopen.ca/content/5/3/E729/suppl/DC1) but included only 12 in the current analysis: diabetes, hypertension, thyroid disorder, any cardiac disease (including heart failure confirmed by a physician), cancer diagnosis in the previous 5 years (including melanoma but excluding other skin cancers), asthma, osteoarthritis, rheumatoid arthritis or lupus, osteoporosis, chronic obstructive pulmonary disease, intestinal disease and hypercholesterolemia. We limited our analysis to these 12 conditions as some of the 21 referred to symptoms and others were judged to be too vague and difficult to match with health administrative data codes. We also combined 2 diagnoses referring to heart conditions under a single term.

We also used an alternative method based on health administrative data to identify participants' chronic conditions. The data were obtained for each participant from the Régie de l'assurance maladie du Québec (RAMQ), Quebec's universal health insurance provider, and covered a 2-year period before the day we received the participant's completed questionnaire (index date). The RAMQ does not provide data in response to any request for a given time until the data for that period are completed, which generally occurs about 6 months after the end of that period. There were no missing data in the administrative data. We used 2 data sources: 1) MED-ÉCHO (Maintenance et exploitation des données pour l'étude de la clientèle hospitalière), a database containing information on hospital stays in acute care that gathers all hospital discharge data, and 2) fee-for-service billing records, which contain information on services provided by physicians under this remuneration model. Both sources comprised diagnosis codes originating from physicians. For MED-ÉCHO, we used all main and secondary diagnoses (up to 16 diagnoses per hospital stay); admission diagnoses were discarded owing to reliability issues. For fee-for-service billing records, a single diagnosis code is provided per ambulatory visit. We discarded codes originating from medical laboratories, as they contain a large proportion of rule-out diagnoses. Hospital discharge data used International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) codes, and physician billings data used ICD-9 codes. All codes were identified and linked to chronic conditions (Appendix 1). Unique identifiers allowed linkage of each code to the participant involved in the medical act or hospital stay. We assigned a condition to a participant if at least 1 code linked to the condition was observed in the participant's data. We computed 1 index variable per condition along with a variable describing the number of identified conditions. We obtained person-level data by linking the PRECISE and RAMQ databases with the unique identifier used by the RAMQ.

Statistical analysis

We described participants' baseline characteristics using conventional descriptive statistics. The total number of participants with a given chronic condition, identified either by self-report or by health administrative data, determined the frequency (numerator) of the condition in the entire sample. With this number, we estimated the prevalence of chronic conditions among participants. We also calculated the frequencies of the 12 chronic conditions independently for the 2 data sources.

We obtained κ statistic[8] estimates along with 95% confidence intervals (CIs) for all compared measures. We conducted a sensitivity analysis by categorizing participants into 2 age groups: less than 65 years, and 65 years or more. We obtained κ statistics with CIs for either age group.

We estimated the prevalence of multimorbidity using 3 different operational definitions of multimorbidity as the numerator: the presence of 2 or more, 3 or more, and 4 or more of the 12 chronic conditions.

Ethics approval

Approval for PRECISE was obtained from the ethics committees of the Centre de santé et de services sociaux de Chicoutimi and Hôpital Charles-LeMoyne, Longueuil, Quebec.

Results

A total of 2409 participants were recruited. Of the 2409, 1718 (71.3%) responded, 1178 (68.6%) of whom agreed to permit access to their health administrative data. We found missing values in only 1 participant's questionnaire, which was excluded from the analyses. The final sample thus included 1177 participants. The near-absence of missing values was the result of a rigorous process adopted while the project was exe-

cuted: participants were contacted immediately if there were missing values on any question. We collected health administrative data for all 1177 participants.

The average age of the participants was 53 (standard deviation 12.4) years. Their distribution by age group is shown in Table 1. There were more women than men in the sample (684 [58.1%] v. 493 [41.9%]). Overall education levels reflected those of the general Quebec population, with good representation of both extremes (no high school diploma and university degree).

The agreement in identification of chronic conditions between self-report and health administrative data is presented in Table 2. The κ coefficient for diabetes was 0.82. For most of the remaining conditions, the κ coefficient ranged between 0.23 and 0.59. Two conditions (intestinal disease and hypercholesterolemia) showed poor agreement (κ < 0.20). The discrepancy between the 2 methods in the frequency of individual chronic conditions was high for many conditions (Table 2). The frequency of a diagnosis identified by self-report, and therefore the prevalence of the condition, was always higher than the frequency (and prevalence) of the same diagnosis in the health administrative data except for cancer. This difference reached a maximum for hypercholesterolemia.

The prevalence of multimorbidity, according to the 3 operational definitions used, as assessed with the 2 data collection methods is shown in the Figure 1. Estimates obtained with health administrative data were lower than those obtained with self-reported data regardless of the operational definition used. The estimate obtained by combining the 2 methods was closer to the estimate with self-reported data for all 3 operational definitions. Most participants with multimorbidity were identified by self-report.

The concordance between self-report and health administrative data by age group is presented in Table 3. The agreement was similar for the 2 age groups except for rheumatoid

Table 1: Sociodemographic characteristics of participants

Characteristic	No. (%) of participants *n* = 1177
Age group, yr	
25–34	120 (10.2)
35–44	182 (15.5)
45–54	295 (25.1)
55–64	343 (29.1)
65–75	223 (18.9)
Missing	14 (1.2)
Female	684 (58.1)
Education	
No high school diploma	248 (21.1)
High school or college diploma	614 (52.2)
University diploma	310 (26.3)
Missing	5 (0.4)

Table 2: Frequency of chronic conditions by self-report and with health administrative data, and agreement between the 2 data sources

Chronic condition	Prevalence, %	Total no. of participants*	Method; no. (%) of participants			κ (95% CI)
			Self-report	Health administrative data	Both methods	
Diabetes	10.5	124	109 (87.9)	104 (83.9)	89 (71.8)	0.82 (0.76 to 0.88)
Thyroid disorder	13.1	154	147 (95.4)	77 (50.0)	70 (45.4)	0.59 (0.51 to 0.67)
Hypertension	34.9	411	363 (88.3)	260 (63.3)	212 (51.6)	0.57 (0.52 to 0.62)
Any cardiac disease	12.0	141	111 (78.7)	87 (61.7)	57 (40.4)	0.54 (0.45 to 0.62)
Cancer	8.5	100	45 (45.0)	92 (92.0)	37 (37.0)	0.52 (0.41 to 0.62)
Asthma	12.6	148	129 (87.2)	63 (42.6)	44 (29.7)	0.42 (0.33 to 0.51)
Arthrosis	28.6	337	290 (86.1)	127 (37.7)	80 (23.7)	0.28 (0.21 to 0.34)
Rheumatoid arthritis	4.8	57	35 (61.4)	33 (57.9)	11 (19.3)	0.30 (0.16 to 0.45)
Osteoporosis	6.6	78	71 (91.0)	22 (28.2)	15 (19.2)	0.30 (0.18 to 0.43)
Chronic obstructive pulmonary disease	6.3	74	54 (73.0)	31 (41.9)	11 (14.9)	0.23 (0.11 to 0.36)
Intestinal disease	9.9	116	97 (83.6)	34 (29.3)	15 (12.9)	0.19 (0.1 to 0.29)
Hypercholesterolemia	28.6	337	332 (98.5)	17 (5.0)	12 (3.6)	0.04 (0.01 to 0.07)

Note: CI = confidence interval.
*Total number of participants with the condition who were identified either by self-report or by health administrative data.

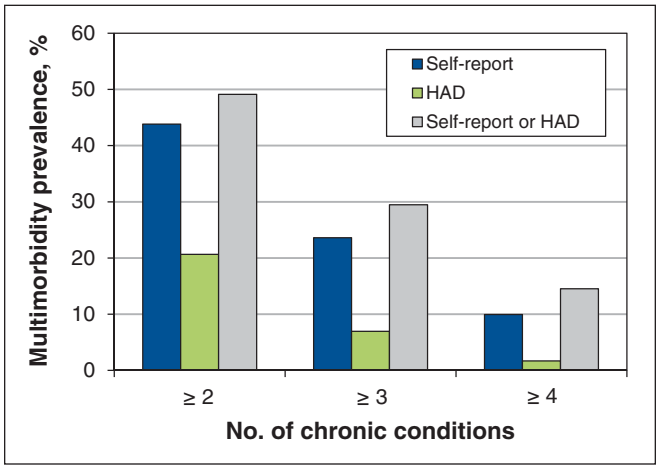

Figure 1: Prevalence of multimorbidity with self-reported data, health administrative data (HAD) and the 2 data sources pooled together, according to number of chronic conditions considered in the operational definition of multimorbidity.

arthritis and osteoporosis, but, even for these 2 conditions, the 95% CIs for the 2 age groups overlapped.

Interpretation

In this study, we found low (but varied) agreement in the identification of chronic conditions between self-reported data and health administrative data, with the poorest agreement observed for hypercholesterolemia, a condition managed mainly in primary care. The frequency (and hence prevalence) of conditions was higher by self-report except for 1 condition (cancer), and the estimated prevalence of multimorbidity was also higher with self-reported data.

As previously reported,[4–6,9,10] there is an important inconsistency between diagnoses of chronic conditions obtained by self-report and by the use of health administrative data, which leads to poor agreement between the 2 data sources most of the time. For some diagnoses, agreement is generally good, whereas other diagnoses may be underreported by 1 method or the other, resulting in low agreement. Both methods of data collection have inherent limitations. An important problem with administrative data is the accuracy and completeness of diagnostic information, and an important concern when using self-report data is recall bias.[9] In our study, diabetes was the only diagnosis for which agreement between the 2 data sources was good. Our finding that diabetes was quite consistently reported by both methods is in keeping with previous reports.[4,5,11] This finding is not surprising, considering that diabetes implies a substantial burden for the patient and the health care system. This condition is unlikely to be forgotten when completing a questionnaire on chronic health conditions, and it is also likely to appear in health administrative data, probably influenced by the fact that the ICD-9 code is easy to remember (250 diabetes mellitus). On the opposite side, hypercholesterolemia was the condition with the most striking difference in reporting between the 2 data sources in our study. Our findings suggested massive underestimation of this condition in the health administrative data. Hypercholesterolemia is likely to be acknowledged and reported by patients because most of those who have this condition receive a medication and/or diet for treating it. However, it is rarely used as the single diagnosis code for ambulatory care visits. It probably does not show up in health administrative data because more importance is given to other conditions associated with hypercholesterolemia, such as heart conditions. Because only 1 diagnosis is recorded in the billing sys-

Table 3: Comparison of agreement by age group

Chronic condition	Age; κ (95% CI)	
	< 65 yr	65–75 yr
Diabetes	0.85 (0.77 to 0.91)	0.74 (0.62 to 0.86)
Hypertension	0.59 (0.52 to 0.65)	0.42 (0.30 to 0.53)
Thyroid disorder	0.60 (0.50 to 0.69)	0.58 (0.43 to 0.72)
Any cardiac disease	0.57 (0.46 to 0.67)	0.44 (0.29 to 0.57)
Cancer	0.52 (0.40 to 0.63)	0.52 (029 to 0.72)
Asthma	0.40 (0.29 to 0.50)	0.49 (0.26 to 0.69)
Arthrosis	0.26 (0.19 to 0.34)	0.26 (0.13 to 0.38)
Rheumatoid arthritis	0.38 (0.20 to 0.54)	0.08 (–0.05 to 0.32)
Osteoporosis	0.39 (0.24 to 0.55)	0.12 (–0.03 to 0.33)
Chronic obstructive pulmonary disease	0.22 (0.07 to 0.37)	0.25 (0.03 to 0.47)
Intestinal disease	0.20 (0.09 to 0.31)	0.19 (0.01 to 0.38)
Hypercholesterolemia	0.05 (0.01 to 0.09)	0.04 (0.00 to 0.08)

Note: CI = confidence interval.

tem in Quebec, it is likely that diagnoses with higher disease burden are predominantly recorded by the treating physician, thus limiting the information about less severe or more stable diagnoses in health administrative data.

We observed that the frequency of 11 of the 12 conditions was higher in the self-reported data than in the administrative data, the exception being cancer. This is at variance with a study by Muggah and colleagues,[5] who reported that the prevalence of 5 of 7 conditions (diabetes, congestive heart failure, myocardial infarction, stroke, hypertension, asthma and chronic obstructive pulmonary disease) was higher when determined with health administrative data than with self-reported data, the exceptions being stroke and myocardial infarction. A possible explanation for this discrepancy is that those investigators used hospital and physician billing codes from Ontario, and differences in the system for recording health conditions are to be expected.

Inconsistencies in the frequency of diagnoses of chronic conditions between the data sources used in our study led to important differences in the prevalence of multimorbidity estimated from either data set. Regardless of the operational definition of multimorbidity used, prevalence estimates of multimorbidity obtained with health administrative data were lower than those obtained with self-report data. In the absence of a gold standard, it is difficult to assess which method performs the best. However, our findings suggest that administrative data were not an appropriate source for estimating the prevalence of the 12 chronic conditions studied in the general population or the prevalence of multimorbidity using that list of conditions, at least with the methods we used to collect administrative data. We used hospital discharge data and fee-for-service billing records. If more administrative data were used, such as drug prescriptions and laboratory data, the value of health administrative data as a source for estimating prevalence might improve.

Recently, Tonelli and colleagues[3] reported a scheme for using validated algorithms with administrative data to identify the presence of 30 chronic conditions. They were able to apply the method to inpatient and outpatient claims and to use data for people residing in Edmonton during 1 fiscal year to estimate the prevalence of the conditions and to identify people with multimorbidity (≥ 2 conditions) in the sample. Their work is a promising study in favour of the use of administrative data. Studies comparing data collected by means of their method to self-report or other data sources (e.g., medical records) are warranted.

Limitations

Data from electronic medical records, to provide a better, unbiased measure of multimorbidity as a gold standard, were not available for comparison. Our estimates of the prevalence of the 12 chronic conditions and, hence, of multimorbidity are probably outdated because data collection was conducted in 2010. However, estimating the actual prevalence of each condition was not the main goal of this study; rather, the agreement between the 2 data sources was. A limitation of the Quebec administrative data is that they do not contain any information on reliability/validity, as the data set is used mostly for billing. Also, we used a 2-year window for health administrative data, and it is possible that the use of additional years of administrative data might have increased the agreement between the 2 data sources. Finally, we used 1 algorithm to ascertain the conditions in the administrative data. The use of more algorithms might have also increased the agreement.

Conclusion

This study revisited the agreement between the self-report of chronic conditions and health administrative data. Given the results, we argue for the use of self-reported chronic conditions in the study of multimorbidity, as health administrative data based on 1 algorithm that uses the billing system in Quebec seems to underestimate the prevalence of many chronic conditions, thus resulting in biased estimates of multimorbidity.

References

1. Fortin M, Bravo G, Hudon C, et al. Prevalence of multimorbidity among adults seen in family practice. *Ann Fam Med* 2005;3:223-8.
2. Dillman DA. *Mail and internet surveys. The tailored design method*. 2nd ed. New York: John Wiley & Sons; 2000.
3. Tonelli M, Wiebe N, Fortin M, et al.; Alberta Kidney Disease Network. Methods for identifying 30 chronic conditions: application to administrative data. *BMC Med Inform Decis Mak* 2015;15:31.
4. Jiang L, Zhang B, Smith ML, et al. Concordance between self-reports and Medicare claims among participants in a national study of chronic disease self-management program. *Front Public Health* 2015;3:222.
5. Muggah E, Graves E, Bennett C, et al. Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. *BMC Public Health* 2013;13:16.
6. Lix LM, Yogendran MS, Shaw SY, et al. Population-based data sources for chronic disease surveillance. *Chronic Dis Can* 2008;29:31-8.
7. Haggerty J, Fortin M, Beaulieu MD, et al. At the interface of community and healthcare systems: a longitudinal cohort study on evolving health and the impact of primary healthcare from the patient's perspective. *BMC Health Serv Res* 2010;10:258.
8. Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley; 1981.
9. Lix LM, Yogendran MS, Shaw SY, et al. Comparing administrative and survey data for ascertaining cases of irritable bowel syndrome: a population-based investigation. *BMC Health Serv Res* 2010;10:31.
10. Singh JA. Accuracy of Veterans Affairs databases for diagnoses of chronic diseases. *Prev Chronic Dis* 2009;6:A126.
11. Lujic S, Watson DE, Randall DA, et al. Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. *BMJ Open* 2014;4:e005768.

Affiliations: Department of Family Medicine and Emergency Medicine (Fortin, Almirall), Université de Sherbrooke, Sherbrooke, Que.; Faculty of Medicine (Haggerty), McGill University; St. Mary's Research Centre (Sanche), St. Mary's Hospital, Montréal, Que.

Contributors: Steven Sanche, Martin Fortin and Jeannie Haggerty contributed to the study conception and design. Steven Sanche and José Almirall analyzed the data under the supervision and guidance of Martin Fortin and Jeannie Haggerty. All of the authors contributed to data interpretation and the writing of the manuscript, gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Funding: This research was supported by the Canadian Institutes of Health Research (CIHR). Martin Fortin was supported by the CIHR and partners (CIHR Applied Chair in Health Services and Policy Research on Chronic Diseases in Primary Care/CIHR Institute of Health Services and Policy Research; Canadian Health Services Research Foundation; and Centre de santé et de services sociaux de Chicoutimi).

Supplemental information: For reviewer comments and the original submission of this manuscript, please see www.cmajopen.ca/content/5/3/E729/suppl/DC1.