Population-based data sources for chronic disease surveillance

LM Lix, PhD (1); MS Yogendran, MSc (2); SY Shaw, MA (3); C Burchill, MSc (2); C Metge, PhD (2,4); R Bond, MA (2)

ABSTRACT

This study estimated agreement between population-based administrative and survey data for ascertaining cases of arthritis, asthma, diabetes, heart disease, hypertension and stroke. Chronic disease case definitions that varied by data source, number of years and number of diagnosis or prescription drug codes were constructed from Manitoba's administrative data. These data were linked to the Canadian Community Health Survey. Agreement between the two data sources, estimated by the κ coefficient, was calculated for each case definition, and differences were tested. Socio-demographic and comorbidity variables associated with agreement were tested using weighted logistic regression. Agreement was strongest for diabetes and hypertension and lowest for arthritis. The case definition elements that contributed to the highest agreement between the two population-based data sources varied across the chronic diseases. Low agreement between administrative and survey data is likely to occur for conditions that are difficult to diagnose, but will be mediated by individual socio-demographic and health status characteristics. Construction of a chronic disease case definition from administrative data should be accompanied by a justification for the choice of each of its elements.

Key words: *chronic disease, case ascertainment, record linkage, population health, surveillance*

Introduction

Population-based data about chronic disease prevalence are essential to describe the burden of disease and to plan and evaluate disease prevention, treatment and management strategies.[1] Administrative databases and self-report responses from health surveys are two key sources of population-based data to estimate chronic disease prevalence.[2-5] Both sources have limitations – the former because of concerns about the accuracy of diagnostic information, and the latter because of concerns about the validity of self-reports of disease diagnosis.[2,6-8] The congruence between the two data sources has been investigated in a few studies using subject-specific record-linkage techniques.[2,9,10] Record linkage also allows for investigation of the individual characteristics that modify agreement between the two data sources.

Systematic investigations of the effect that the choice of an administrative case definition has on agreement between the two population-based data sources are largely absent from the literature. Chronic disease case definitions are constructed by selecting specific combinations of the following administrative data elements: source of data, diagnosis or treatment codes, number of years of data and number of contacts in administrative records with the selected code(s).[9-11] There is no consensus about the optimal case definition, and the choice of case definitions is often based on the availability of data in one's jurisdiction.

This study investigates multiple definitions for ascertaining cases of arthritis, asthma, diabetes, heart disease, hypertension and stroke from administrative data and compares them with self-reports of chronic disease from population-based survey data. The specific objectives are to: 1) test the agreement between administrative and survey data while the elements of a chronic disease case definition are systematically manipulated; and 2) examine individual demographic, geographic, socioeconomic and health status characteristics that may affect agreement between these two data sources.

Methods

The study was conducted using population-based data from Manitoba, a centrally located province in Canada with a universal health care system. The Research Data Repository housed at the Manitoba Centre for Health Policy contains administrative records provided by the provincial health ministry under the Manitoba Health Services Insurance Plan (MHSIP). The Repository also houses population-based survey data from the national Canadian Community Health Survey (CCHS), and the two sources can be directly linked via a unique, anonymized personal health identification number (PHIN).

Hospital, physician and prescription drug databases were selected to construct the case definitions. These administrative databases have been used in other studies to ascertain chronic disease cases.[6,7,10,12] A hospital abstract is completed when a patient is discharged from an acute care facility. Each record includes up to 16 diagnosis codes from the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). All 16 codes were searched to identify disease cases in this study. Physician claims are submitted to the provincial ministry of health by all physicians who are paid on a fee-for-service basis. These claims capture

Author References

1 School of Public Health, University of Saskatchewan, Saskatoon, SK
2 Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB
3 Department of Community Health Sciences, University of Manitoba, Winnipeg, MB
4 Faculty of Pharmacy, University of Manitoba, Winnipeg, MB
Correspondence: Lisa M. Lix, School of Public Health, University of Saskatchewan, Health Sciences Building, 107 Wiggins Road, Saskatoon SK S7N 5E5, Tel: 306-966-1617, Email: lisa.lix@usask.ca

all outpatient services, including those for hospital emergency and outpatient departments, and for residents of long-term care facilities. While some physicians are salaried (approximately 7% of family physicians in Manitoba[13]), approximately 90% of these physicians also submit parallel billing claims for administrative purposes. Accordingly, the billing claims database captures almost all contacts with physicians in Manitoba. Physician billing claims contain a single ICD-9-CM code. Outpatient prescription drug records are captured in the Drug Programs Information Network (DPIN), a centralized, electronic, point-of-sale database connecting all retail pharmacies in Manitoba. DPIN collects a variety of information for each dispensation, including the date, drug name and drug identification number (DIN). DINs are linked to Anatomic Therapeutic Chemical (ATC) codes developed by the World Health Organization; these codes classify prescription drugs into groups at each of five levels according to the organ or system on which they act and/or their therapeutic and chemical characteristics.[14] DPIN data have been deemed to be accurate both for capture of drug dispensations as well as the prescription details.[15] The prescription drug database does not contain information about any medications that are obtained without a prescription, including over-the-counter medications, complimentary medication samples distributed directly from physicians to patients, or medications dispensed to hospitalized inpatients or residents of long-term care facilities.

The CCHS provides cross-sectional estimates of health status, health determinants, and health system use for residents of 136 health regions in Canada, including 11 Manitoba regions.[16] The survey uses a multi-stage stratified clustered design. It represents approximately 98% of the Canadian population aged 12 years or older. In cycle 1.1, collected between September 2000 and November 2001, there were 8120 Manitoba respondents.

Anonymized linkage between administrative and survey data was conducted for those survey respondents who provided their consent (n = 7560; 93.1%). After removing invalid or missing PHINs, direct linkage between administrative and survey data was achieved for 6812 respondents (i.e. 83.9% of respondents). A cohort of survey respondents with at least five years of continuous coverage under the MHSIP prior to their interview date was created (n = 6422). The study excluded respondents with less than five years of coverage because preliminary analyses revealed that discontinuous health insurance coverage was associated with lower agreement between survey and administrative data. The administrative data were from fiscal years 1997/98 to 2001/02.

The survey interview schedule included the following directions: "*Now I'd like to ask about certain chronic health conditions which you may have. We are interested in 'long-term conditions' that have lasted or are expected to last six months or more and that have been diagnosed by a health professional*". Table 1 lists the relevant questions.

A total of 152 case definitions were investigated (the complete list is available from the corresponding author upon request), and were developed using previous research as a guide.[5,9-11,17-27] Some definitions were based only on a single administrative data source, while others were based on multiple data sources. The case definitions varied by the number of years of administrative data (one, two, three and five), in accordance with previous research.[9] Some case definitions required only a single contact in administrative data, while other definitions required more than one contact. The constants in the construction of the case definitions were the diagnosis (i.e. ICD-9-CM) and prescription drug (i.e. ATC) codes (Table 1); these were identified from a comprehensive literature review.[28]

Cohen's kappa coefficient (κ) was used to quantify agreement between the two data sources; 95% confidence intervals (95% CIs) were also computed. The estimates and CIs were computed using the sample weights from the survey data. Kappa is a commonly adopted measure because it corrects the agreement between two sources by taking account of the proportion of agreement expected by chance. The magnitude of agreement was assessed as follows:[29] poor agreement: $\hat{\kappa} < 0.20$; fair agreement: $0.20 \leq \hat{\kappa} < 0.40$; moderate agreement: $0.40 \leq \hat{\kappa} < 0.60$; good agreement: $0.60 \leq \hat{\kappa} < 0.80$; and very good agreement: $\hat{\kappa} \geq 0.80$.

Differences between selected κ coefficients (i.e. H_0: $\kappa_1 = \kappa_2$) were tested using a goodness-of-fit statistic,[30] which approximately follows an χ^2 distribution. The tests were conducted for case definitions based on different sources or years of data or number of contacts. All comparisons were identified *a priori*, and the per-comparison error rate was set at $\alpha = 0.05$.

For each chronic disease, a single case definition with the highest estimated agreement was selected to further investigate the respondent characteristics associated with agreement between population-based administrative and survey data. Weighted logistic

TABLE 1
Survey questions, ICD-9-CM diagnosis codes, and ATC prescription drug codes for chronic disease case ascertainment

Disease	Survey Questions	ICD-9-CM Codes	ATC Codes[b]
Arthritis	Do you have arthritis or rheumatism, excluding fibromyalgia?	714, 715, 446, 710, 720, 274, 711-713, 716, 717, 718 719, 721, 725-729, 739	A07, J01, L01, L04, M01, P01, N02, R05, H02
Asthma	Do you have asthma?	493	R03, R06
Diabetes	Do you have diabetes?	250	A10
Heart Disease	Do you have heart disease? [a]	410-414	C01, C07, C08, C09
Hypertension	Do you have high blood pressure?	401	C02, C03, C07, C08, C09
Stroke	Do you suffer from the effects of a stroke?	430-438	B01

[a] Individuals with congestive heart failure were excluded from the "heart disease" category

[b] Not all drugs in each ATC category were selected. For a comprehensive list of prescription drug inclusions and exclusions please refer to the following URL: http://mchp-appserv.cpe.umanitoba.ca/reference/chronic.disease.pdf

regression analyses modeled agreement as a function of socio-demographic and disease variables, including age group (12 to 18, 19 to 49, 50 to 64, 65 to 74, 75+ [reference]), sex (male, female [reference]), region of residence (rural, urban [reference]), income adequacy quintile (highest income group was the reference) and presence/absence of comorbid conditions (presence was the reference).[31,32] CCHS methodologists developed income adequacy quintiles using self-reported total household income and number of persons living in the household. The following comorbid conditions were selected: allergies, emphysema or chronic obstructive pulmonary disease for asthma; heart disease or hypertension for diabetes; diabetes or hypertension for heart disease; diabetes or heart disease for hypertension; and heart disease or diabetes for stroke. Odds ratios (ORs) and 95% CIs are reported. All analyses were conducted using SAS software, version 9.1.

Ethics approval was obtained from the University of Manitoba Health Research Ethics Board. Approval for data access was from the Manitoba Health Information Privacy Committee.

Results

Only adult (19 years and older) respondents (n = 5589; 87.0%) were the subject of the investigation on agreement between population-based administrative and health survey data for diabetes, heart disease, hypertension, arthritis and stroke, while both adult and youth respondents were selected for the asthma analyses. Almost half (46.1%) of individuals in the adult study cohort were 50 years of age or older. Slightly less than half of the entire cohort (45.3%) was male, and almost one-quarter of the individuals (23.1%) were urban residents. The CCHS oversampled rural residents so that estimates could be generated for individual health regions.

The number and percent of the study cohort that reported each of the chronic diseases is given in Table 2. Weighted estimates of chronic disease prevalence were computed from the survey data (see Table 2), and ranged from 1.5% for stroke to 18.6% for arthritis. These provincial estimates are consistent with estimates reported in self-report survey data from other provincial and national studies.[33-35]

Figure 1 depicts the variation in estimates of the κ coefficient for all investigated case definitions. For arthritis, $\hat{\kappa}$ ranged from 0.15 to 0.34, indicating poor to fair agreement. For asthma, the estimates ranged from 0.27 to 0.59, which represents fair to moderate agreement. Agreement for diabetes was good or very good, and ranged from 0.65 to 0.83. For heart disease, $\hat{\kappa}$ ranged from 0.29 to 0.55, indicating fair to moderate agreement. For hypertension, the range was from 0.53 to 0.72, indicating moderate to good agreement. Finally for stroke, $\hat{\kappa}$ ranged from 0.27 to 0.58, which represents fair to moderate agreement.

Table 3 reports $\hat{\kappa}$ for selected case definitions along with the results of the inferential analyses. Tests of the differences in agreement between case definitions that required only a single contact in physician claims, and case definitions that required two or more contacts in physician claims, were conducted. When one year of administrative data was used, significantly lower values of $\hat{\kappa}$ were observed for the latter case definition than for the former for all chronic diseases. However, when two or more years of administrative data were used to construct the case definition, the pattern of differences in agreement varied. For arthritis and diabetes, estimates of agreement were always significantly higher for the two-contact case definition than the one-contact case definition,

FIGURE 1

Weighted $\hat{\kappa}$ for chronic disease case definitions from administrative data

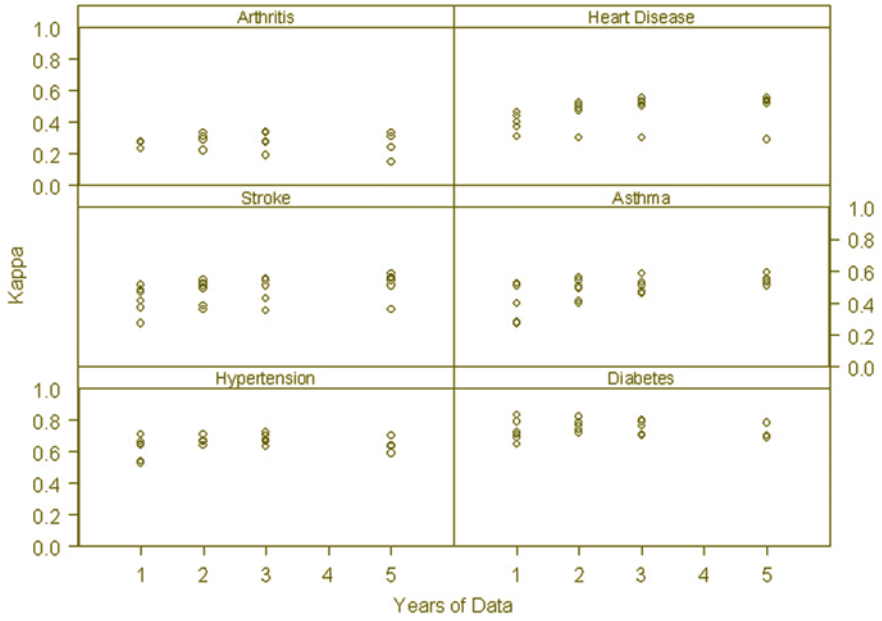

TABLE 2

Frequency (percent) of study cohort with self-reported chronic disease and prevalence estimates from the Canadian Community Health Survey cycle 1.1

Disease	Adult Cohort (n = 5589)	Youth Cohort (n = 833)	Prevalence (%) (95% CI)[a]
Arthritis	1344 (24.0)	–	18.6 (17.4 to 19.8)
Asthma	418 (7.5)	111 (13.5)	8.6 (7.6 to 9.6)
Diabetes	337 (6.0)	–	4.5 (3.9 to 5.2)
Heart Disease	371 (6.6)	–	5.1 (4.4 to 5.7)
Hypertension	1033 (18.5)	–	15.3 (14.2 to 16.5)
Stroke	108 (1.9)	–	1.5 (1.1 to 1.8)

[a] Prevalence estimates were calculated for the population 19 years and older, except for asthma prevalence, which was calculated for the population 12 years and older. Prevalence estimates are based on survey weights, and confidence intervals were calculated using bootstrap variance estimation methods

TABLE 3
Weighted κ (95% CI) for selected chronic disease case definitions from administrative data[a,b,c,d]

Source and Years of Data	Arthritis	Asthma	Diabetes	Heart Disease	Hypertension	Stroke
Physician (P)[a]	**1 P**	**1 P**	**1 P**	**1 P**	**1 P**	**1P**
1	0.27 (0.26,0.27)	0.40 (0.35,0.44)	0.69 (0.68,0.69)	0.44 (0.39,0.49)	0.64 (0.64,0.64)	0.47 (0.47,0.48)
2	**0.29** (0.28,0.29)	0.49 (0.45,0.54)	**0.72** (0.72,0.73)	0.51 (0.46,0.55)	**0.66** (0.66,0.67)	**0.54** (0.53,0.54)
3	0.27 (0.27,0.28)	0.53 (0.49,0.57)	0.70 (0.69,0.70)	**0.52** (0.48,0.56)	**0.66** (0.66,0.67)	0.51 (0.50,0.52)
5	0.24 (0.24,0.24)	**0.55** (0.51,0.59)	0.69 (0.69,0.69)	**0.52** (0.48,0.57)	0.63 (0.63,0.64)	**0.54** (0.54,0.55)
	2 + P	**2 + P**	**2 + P**	**2 + P**	**2 + P**	**2 + P**
1	0.23 (0.23,0.23)*	0.27 (0.23,0.32)*	0.65 (0.65,0.66)*	0.37 (0.31,0.42)*	0.53 (0.52,0.53)*	0.41 (0.40,0.42)*
2	0.31 (0.30,0.31)*	0.40 (0.35,0.44)*	0.76 (0.75,0.76)*	0.47 (0.42,0.52)*	0.64 (0.64,0.65)*	0.49 (0.48,0.50)*
3	**0.34** (0.33,0.34)*	0.46 (0.42,0.51)*	**0.78** (0.78,0.79)*	0.50 (0.46,0.55)*	0.68 (0.68,0.68)*	0.55 (0.54,0.56)*
5	0.33 (0.33,0.34)*	**0.54** (0.50,0.58)	**0.78** (0.78,0.79)*	**0.54** (0.49,0.58)*	**0.70** (0.70,0.70)*	**0.58** (0.57,0.59)*
Hospital (H), Physician(P)[b]	**1 + H or 2 + P**	**1 + H or 1 + P**	**1 + H or 1 + P**	**1 + H or 2 + P**	**1 + H or 2 + P**	**1 + H or 2 + P**
1	0.23 (0.23,0.24)	0.40 (0.36,0.45)	0.72 (0.72,0.73)†	0.40 (0.34,0.45)†	0.54 (0.54,0.54)†	0.48 (0.47,0.48)†
2	0.31 (0.31,0.31)	0.50 (0.46,0.54)	0.74 (0.73,0.74)†	0.49 (0.44,0.54)†	0.66 (0.65,0.66)†	0.52 (0.51,0.53)†
3	**0.34** (0.34,0.34)	0.53 (0.50,0.57)	**0.70** (0.70,0.71)†	0.53 (0.49,0.58)†	0.70 (0.69,0.70)†	0.55 (0.54,0.55)
5	0.33 (0.33,0.34)	**0.55** (0.51,0.59)	**0.70** (0.69,0.70)†	**0.55** (0.51,0.59)†	**0.71** (0.71,0.71)†	**0.56** (0.55,0.57)†
Hospital (H), Physician(P), Prescription(Rx)[c]	**1 + H or 2 + P or (1 + P & 2 + Rx)**	**1 + H or 1 + P or 1 + Rx**	**1 + H or 1 + P or 1 + Rx**	**1 + H or 2 + P or (1 + P & 2 + Rx)**	**1 + H or 2 + P or (1 + P & 2 + Rx)**	**1 + H or 2 + P or (1 + P & 2 + Rx)**
1	0.28 (0.28,0.28)‡	0.52 (0.48,0.56)‡	**0.79** (0.79,0.80)‡	0.46 (0.41,0.51)‡	0.66 (0.66,0.67)‡	0.51 (0.50,0.52)‡
2	**0.33** (0.33,0.33)‡	**0.54** (0.50,0.57)‡	0.76 (0.76,0.77)‡	0.52 (0.48,0.57)‡	0.71 (0.71,0.71)‡	0.51 (0.51,0.52)
3	**0.33** (0.33,0.34)‡	0.51 (0.48,0.55)	0.71 (0.71,0.72)‡	**0.55** (0.50,0.59)‡	**0.72** (0.71,0.72)‡	0.54 (0.53,0.55)
5	0.31 (0.31,0.31)‡	0.48 (0.45,0.51)‡	0.69 (0.69,0.70)	**0.55** (0.51,0.60)	0.70 (0.70,0.71)‡	**0.55** (0.54,0.55)‡

[a] Numeric values in bold typeface represent the highest kappa value(s) in each set of case definitions

[b] Star (*) denotes a statistically significant difference between an algorithm based on one contact in physician claims and an algorithm based on two or more contacts in physician claims ($p < 0.05$); all comparisons are based on the same number of years of data

[c] Dagger (†) denotes a statistically significant difference between an algorithm based on physician data (either one contact or two contacts) and an algorithm based on hospital and physician data ($p < 0.05$); all comparisons are based on the same number of years of data

[d] Double Dagger (‡) denotes a statistically significant difference between an algorithm based on hospital and physician data and an algorithm based on hospital, physician and prescription drug data ($p < 0.05$); all comparisons are based on the same number of years of data

while for asthma the reverse was true. For heart disease, hypertension and stroke, agreement could be significantly higher or lower depending on the number of years of data used to construct the case definition.

Tests of the differences between case definitions based on diagnoses in physician records (either one or two contacts, depending on the disease) and case definitions based on both hospital and physician records were conducted for single and multiple years of administrative data (Table 3). When hospital data were included, statistically significant improvements in agreement were observed for all diseases with the exception of arthritis and asthma. For the other conditions, agreement almost always improved regardless of whether single or multiple years of data were used to construct the case definition.

Finally, Table 3 reports tests of the differences between case definitions based on diagnoses in hospital and physician records and case definitions based on both diagnosis and prescription drug codes. For all chronic diseases, the combination of diagnostic and prescription drug data resulted in a statistically significant improvement in agreement when one year of data were used to construct the case definitions. This was not consistently the case when two or more years of data were used to construct the case definition. In fact when five years of data were used, agreement either did not improve or got worse for all diseases.

The ORs for the weighted logistic regression analyses are reported in Table 4. Age was statistically significant in all models ($p < 0.0001$). The odds of agreement between administrative and survey data were almost always higher for individuals in younger than in older age groups. Sex was also associated with agreement for all chronic diseases, although the magnitude and direction of the effect varied. It was strongest for diabetes and stroke. For arthritis, asthma, diabetes and stroke, the odds of agreement were significantly higher for males than females, while the converse was true for heart disease, hypertension and stroke.

Region of residence was statistically significant for all chronic diseases ($p < 0.0001$) with the exception of asthma. The strength of the association was weakest for arthritis and hypertension. The odds of agreement were higher for rural than urban residents for arthritis, diabetes, heart disease and stroke, while the converse was true for hypertension.

TABLE 4
Odds Ratios* (95% CIs) for predictors of agreement between administrative and survey data for chronic diseases

	Arthritis	Asthma	Diabetes[a,b]	Heart Disease[a,b]	Hypertension	Stroke[a]
Age						
12 to 18 years	–	1.38 (1.33 to 1.43)	–	–	–	–
19 to 49 years	3.67 (3.62 to 3.77)	2.17 (2.11 to 2.24)	–	–	6.00 (5.83 to 6.17)	–
50 to 64 years	1.57 (1.53 to 1.60)	1.86 (1.79 to 1.92)	2.34 (2.28 to 3.46)	8.46 (8.21 to 8.72)	2.42 (2.35 to 2.48)	13.47 (12.75 to 14.23)
65 to 74 years	1.39 (1.35 to 1.42)	1.85 (1.78 to 1.93)	0.85 (0.81 to 0.89)	1.77 (1.71 to 1.82)	1.93 (1.88 to 1.99)	1.90 (1.81 to 1.99)
75+ years	Ref	Ref	Ref	Ref	Ref	Ref
Sex						
Males	1.09 (1.08 to1.11)	1.05 (1.03 to 1.07)	1.40 (1.35 to 1.45)	0.83 (0.81 to 0.85)	0.81 (0.79 to 0.82)	0.44 (0.42 to 0.46)
Females	Ref	Ref	Ref	Ref	Ref	Ref
Residence						
Rural	1.06 (1.04 to 1.07)	1.01 (0.99 to 1.02)	1.18 (1.14 to 1.22)	1.31 (1.28 to 1.34)	0.89 (0.88 to 0.91)	1.28 (1.23 to 1.33)
Urban	Ref	Ref	Ref	Ref	Ref	Ref
Income quintile						
Lowest	0.66 (0.68 to 0.71)	0.33 (0.31 to 0.34)	–	–	0.84 (0.80 to 0.89)	0.07 (0.06 to 0.09)
Low-middle	1.06 (1.03 to 1.09)	0.71 (0.68 to 0.74)	1.07 (1.00 to 1.13)	0.87 (0.83 to 0.91)	1.06 (1.02 to 1.11)	0.10 (0.09 to 0.13)
Middle	0.98 (0.96 to 1.00)	0.59 (0.57 to 0.60)	1.04 (0.99 to 1.09)	0.66 (0.63 to 0.68)	1.10 (1.07 to 1.13)	0.19 (0.17 to 0.21)
Upper-middle	0.91 (0.90 to 0.93)	0.66 (0.64 to 0.68)	1.32 (1.26 to 1.38)	0.78 (0.76 to 0.82)	1.15 (1.12 to 1.18)	0.18 (0.16 to 0.20)
Highest	Ref	Ref	Ref	Ref	Ref	Ref
Comorbid conditions						
Absent	–	1.90 (1.82 to 1.94)	3.09 (2.97 to 3.21)	2.73 (2.66 to 2.80)	2.96 (2.89 to 3.03)	2.97 (2.84 to 3.10)
Present	Ref	Ref	Ref	Ref	Ref	Ref

* Statistically significant at $\alpha = .05$

[a] The 19 to 49 and 50 to 64 years age groups were combined because of small cell sizes for the former category

[b] Low-middle and middle income quintile categories were combined because of small cell sizes for the former category

Income quintile was also associated with agreement between administrative and survey data ($p < 0.0001$). The odds of agreement were generally lower for individuals in poorer income quintiles than for those in wealthier quintiles, except for diabetes and hypertension, where the converse was true.

Finally, the comorbidity variable was statistically significant in all models ($p < 0.0001$); the odds of agreement were consistently higher when a comorbid condition was absent than when it was present. The association between comorbidity and agreement was of a similar magnitude for all chronic diseases except asthma, for which the relationship was weakest.

Conclusions

This record-linkage study compared chronic disease case ascertainment in population-based administrative and self-report survey data while the elements of the case definition in administrative data were systematically manipulated. It also investigated the individual characteristics that moderate agreement between these two data sources. The results show that regardless of the case definition adopted, agreement between these two data sources was highest for diabetes and hypertension, and lowest for arthritis. These findings are consistent with previous research that has compared administrative case definitions and self-reported chronic disease in population-based and clinical samples.[2,9,10,19,22,31] Okura et al. theorize that although diabetes and hypertension are not usually characterized by distinct and dramatic clinical presentations, they are "…chronic and require ongoing repeated engagement with the medical care system."[31(p. 1101)] which increases their likelihood of identification in administrative data. For arthritis, the selection of non-specific diagnostic codes by practitioners, potentially inaccurate diagnoses by non-specialized practitioners and the low probability that this condition will contribute to a hospital stay may all be factors contributing to the lack of concordance between the two data sources.[36] As well, some diseases may not be accurately captured via self-report. For example, Kriegsman et al.[37] argues that because of the non life-threatening nature of arthritis, it may be overreported in surveys, but underreported in administrative data, contributing to the lack of agreement between the two sources. Questionnaire wording is also an important factor in assessing the accuracy of survey data.

Agreement between administrative and survey data varied significantly as a function of the elements of type of data, number of years of data and number of contacts,[9,10] although the magnitude of the differences in κ estimates among the case definitions was sometimes quite small. Multiple types and/or years of data usually

resulted in improved agreement between administrative and survey data, but this observation cannot be generalized across all of the investigated diseases. Improvements in agreement were not always evident when three or five years of data were used to construct the case definition. Using prescription drug data in addition to hospital and physician data to ascertain disease cases had mixed effects on agreement. While case ascertainment for asthma benefited from the use of both diagnostic and prescription drug information when the definition was based on one or two years of administrative data, improvements in agreement were less substantial for diabetes. For hypertension, there was also some improvement in agreement associated with using both diagnosis and prescription drug data for case ascertainment, but not for other diseases. A specific set of prescription drugs are used to treat asthma and diabetes; for other chronic diseases such as hypertension or arthritis, the drugs prescribed for an individual may be used to treat more than one chronic disease, and therefore may not be helpful for identifying cases.

The moderating effect of demographic, geographic, socioeconomic and health status variables on agreement between survey and administrative data has been observed in other studies.[38-40] Simpson et al.[39] found, however, that the absence of comorbid conditions was associated with both increased and decreased agreement, depending on the disease, while this study found that agreement consistently improved when comorbid conditions were absent.

One potential limitation of this research is that a fixed set of diagnosis and prescription drug codes was selected. Some studies have compared case ascertainment results for different sets of diagnosis codes. For example, studies of stroke case ascertainment have sometimes excluded non-specific diagnostic codes such as 437, which represents stroke of undetermined causes.[24,41] The exclusion of non-specific diagnostic codes might improve estimates of agreement between administrative and survey data. However, when our analyses for stroke were repeated and restricted to a narrow set of diagnostic codes (i.e. 430, 431, 434, 345, 436), estimates of the κ coefficient were never higher than the estimates obtained using the full set of codes. This study was conducted using administrative data coded with ICD-9-CM. In the 2004/05 fiscal year, a change was made to ICD-10-CA in Manitoba's hospital databases. Future research will investigate the effect of this change on chronic disease case ascertainment. Agreement between the two data sources may have been moderated by the survey interview date, a factor that was not considered in this study. Finally, this study was limited to a set of conditions for which there were sufficient data available to investigate agreement; absent from this list are other forms of chronic respiratory disease, such as chronic obstructive pulmonary disease, musculoskeletal conditions such as osteoporosis, and gastro-intestinal conditions such as inflammatory bowel disease.

The strengths of this study are that it assesses agreement using three administrative databases that are available in many jurisdictions. It focuses on the incremental benefits of using different data elements in constructing a case definition, and it systematically examines the effect of using both diagnosis and prescription drug information. Finally, it uses record-linkage techniques for individuals to investigate the variables associated with agreement between two population-based data sources. Researchers who construct case definitions from administrative data should be cognizant of the effect that the choice of administrative data elements has on case ascertainment. The selection of a case definition should be justified by describing the potential effects associated with manipulating each data element.

The results of this study suggest a number of further opportunities for research on constructing case definitions from administrative data. The first is to stratify the population on important socio-demographic and/or health status variables and then specify and test case definitions for each of these strata. For example, an algorithm based on a single contact in physician claims in one year of data might result in high agreement for youth with asthma, whereas an algorithm based on two or more contacts in physician claims might result in higher agreement for older adults with asthma. Researchers might also consider a model-based approach that uses classification techniques to construct a case definition.[42-43] A model-based approach can accommodate a large set of variables (i.e. data features), including comorbid conditions, contacts with specialists, socio-demographic variables and indicators of disease diagnosis and treatment, and test the relative contribution of these variables to improving case ascertainment. Finally, additional validation studies that use data repositories in other provinces or that adopt a "gold standard" data source, such as laboratory test results or clinical data about disease diagnosis, will aid in understanding the strengths and limitations of administrative data for monitoring chronic disease. In particular, studies that adopt a gold standard can provide estimates of the sensitivity and specificity of case definitions derived from administrative data. However, because an unbiased gold standard does not exist for some chronic diseases, such as arthritis and irritable bowel disease, further research is also needed about validation techniques in the presence of measurement error.[44-45]

Acknowledgements

This research was supported by funding from Manitoba Health & Healthy Living and a grant from the Canadian Institutes of Health Research (#PPR79240). The authors are indebted to Manitoba Health & Healthy Living for the provision of data under project #2004/05-01. The results and conclusions are those of the authors, and no official endorsement by Manitoba Health & Healthy Living is intended or should be inferred. The authors have no competing interests to declare.

References

1. Thacker SB, Stroup DF, Rothenberg RB, Brownson RC. Public health surveillance for chronic conditions: a scientific basis for decisions. Stat Med. 1995;14:629-641.

2. Cricelli C, Mazzaglia G, Samani F, Marchi M, Sabatini A, Nardi R, Ventriglia G, Caputi AP. Prevalence estimates for chronic diseases in Italy: exploring the differences between

self-report and primary care databases. J Public Health Med. 2003;25:254-257.

3. Powell KE, Diseker RA, Presley RJ, Tolsma D, Harris S, Mertz KJ, Viel K, Conn DL, McClellan W. Administrative data as a tool for arthritis surveillance: estimating prevalence and utilization of services. J Public Health Manag Pract. 2003;9:291-298.

4. Umphrey GJ, Kendall O, MacNeill IB. Assessing the surveillance capability of Canada's national health surveys. Chronic Dis Canada. 2001;22:50-56.

5. Mayo NE, Chockalingam A, Reeder BA, Phillips S. Surveillance for stroke in Canada. Health Rep. 1994;6:62-72.

6. Saydah SH, Geiss LS, Tierney E, Benjamin SM, Engelgau M, Brancati F. Review of the performance of methods to identify diabetes cases among vital statistics, administrative and survey data. AEP. 2004;14:507-516.

7. Maio V, Yuen E, Rabinowitz C, Louis D, Jimbo M, Donatini A, Mall S, Taroni F. Using pharmacy data to identify those with chronic conditions in Emilia Romagna, Italy. J Health Serv Res Policy. 2005;10:232-238.

8. Mahonen M, Salomaa V, Brommels M, Molarius A, Miettinen H, Pyorala K, Tuomilehto J, Arstila M, Kaarsalo E, Ketonen M, Kuulasmaa K, Lehto S, Mustaniemi H, Niemela M, Palomaki P, Torppa J, Vuorenmaa T. The validity of hospital discharge register data on coronary heart disease in Finland. Eur J Epidemiol. 1997;13:403-415.

9. Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease. Comparing administrative data and self-reports. Med Care. 1997;35:932-947.

10. Rector TS, Wickstrom SL, Shah M, Greenlee NT, Rheault P, Rogowski J, Freedman V, Adams J, Escarce JJ. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare plus Choice health plans that have chronic medical conditions. Health Serv Res. 2004;39:1839-1861.

11. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care. 2002;25:512-516.

12. Virning BA, McBean M. Administrative data for public health surveillance and planning. Annu Rev Public Health. 2001;22:213-230.

13. Watson DE, Katz A, Reid RJ, Bogdanovic B, Roos NP, Heppner P. Family physician workloads and access to care in Winnipeg: 1991 to 2001. CMAJ. 2004;171:339-342.

14. World Health Organization. WHO Collaborating Centre for Drug Statistics Methodology: ATC classification index with DDDs and Guidelines for ATC classification and DDD assignment. Oslo, Norway: Norwegian Institute of Public Health, 2006.

15. Kozyrskyj AL, Mustard CA. Validation of an electronic, population-based prescription database. Ann Pharmacother. 1998;32:1152-57.

16. Statistics Canada. Canadian Community Health Survey (CCHS) – Cycle 1.1. Ottawa: Statistics Canada. 2005. URL: http://www.statcan.ca/english/concepts/hs/index.htm.

17. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol. 2004;57:131-141.

18. Kozyrskyj AL, Mustard CA, Becker AB. Identifying children with persistent asthma from health care administrative records. Can Respir J. 2004;11:141-145.

19. Huzel L, Roos LL, Anthonisen NR, Manfreda J. Diagnosing asthma: the fit between survey and administrative database. Can Respir J. 2002;9:407-412.

20. Macy E, Schatz M, Gibbons C, Zeiger R. The prevalence of reversible airflow obstruction and/or methacholine hyperreactivity in random adult asthma patients identified by administrative data. J Asthma. 2005;42:213-220.

21. Morgan CL, Currie CJ, Stott NC, Smithers M, Butler CC, Peters JR. Estimating the prevalence of diagnosed diabetes in a health district of Wales: the importance of adjustment for death and migration. Diabet Med. 2000;17:141-145.

22. Muhajarine N, Mustard C, Roos LL, Young TK, Gelskey DE. Comparison of survey and physician claims data for detecting hypertension. J Clin Epidemiol. 1997;50:711-718.

23. Leibson CL, Maessens JM, Brown RD, Whisnant JP. Accuracy of hospital discharge abstracts for identifying stroke. Stroke. 1994;25:2348-2355.

24. Tirschwell DL, Longstreth WT. Validating administrative data in stroke research. Stroke. 2002;33:2465-2470.

25. Rawson NS, Malcolm E. Validity of the recording of ischaemic heart disease and chronic obstructive pulmonary disease in the Saskatchewan Health care data files. Stat Med. 1995;14:2627-2643.

26. Shah BR, Hux JE, Zinman B. Increasing rates of ischemic heart disease in the native population of Ontario, Canada. Arch Intern Med. 2000;160:1862-1866.

27. Bullano MF, Kamat S, Willey VJ, Barlas S, Watson DJ, Brenneman SK. Agreement between administrative claims and the medical record in identifying persons with a diagnosis of hypertension. Med Care 2006;44:486-490.

28. Lix LM, Yogendran M, Burchill C, McKeen N, Metge C, Moore D, Bond R. Defining and validating chronic diseases: An administrative data approach. University of Manitoba: Manitoba Centre for Health Policy, 2006.

29. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991.

30. Donner A, Shoukri MM, Klar N, Bartfay E. Testing the equality of two dependent kappa statistics. Stat Med. 2000;19:373-387.

31. Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, Rodeheffer RJ. Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. J Clin Epidemiol. 2004;57:1096–1103.

32. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Identifying hypertension-related comorbidities from administrative data: what's the optimal approach? Am J Med Qual. 2004;19:201-206.

33. Heart and Stroke Foundation of Canada. The Changing Face of Heart Disease and Stroke in Canada 2000. Ottawa: Health Canada, 2000.

34. Health Canada. Diabetes in Canada. 2nd ed. Ottawa: Health Canada, 2002.

35. Wolff HK, Andreou P, Bata IR, Comeau DG, Gregor RD, Gephart G, MacLean DR, Sketris I. Trends in the prevalence and treatment of hypertension in Halifax county from 1985-1995. CMAJ. 1999;161:699-704.

36. Harrold LR, Yood RA, Andrade SE, Reed JI, Cernieux J, Straus W, Weeks M, Lewis B, Gurwitz JH. Evaluating the predictive value of osteoarthritis diagnoses in an administrative database. Arthritis Rheum. 2000;43:1881-1885.

37. Kriegsman DM, Penninx BW, van Eijk JT, Boeke AJ, Deeq DJ. Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. J Clin Epidemiol. 1996;49:1407-1417.

38. Haapanen N, Miilunpalo S, Pasanen M, Oja P, Vuori I. Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. Am J Epidemiol. 1997;145:762-769.

39. Simpson CF, Boyd CM, Carlson MC, Griswold ME, Guralnik JM, Fried LP. Agreement between self-report of disease diagnoses and medical record validation in disabled older women: factors that modify agreement. J Am Geriatr Soc. 2004;52:123-127.

40. Verbrugge LM, Lepkowski JM, Imanaka Y. Comorbidity and its impact on disability. Milbank Q. 1989;67:450-484.

41. Benesch C, Witter DM, Wilder AL, Duncan PW, Samsa GP, Matchar DB. Inaccuracy of the International Classification of Disease (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. Neurology. 1997:49:660-664.

42. Hirsch S, Shapiro JL, Turega MA, Frank TL, Niven RM, Frank PI. Using a neural network to screen a population for asthma. Ann Epidemiol. 2001;11:369-376.

43. Shanker M, Hu MY, Hung MS. Estimating probabilities of diabetes mellitus using neural networks. SAR QSAR Environ Res. 2000;11:133-147.

44. Ladouceur M, Rahme E, Pineau CA, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. Biometrics 2007;63(1):272-279.

45. Cole SR, Haitao C, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006;35:1074-1081.