Population-based data sources for chronic disease surveillance

**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 $\kappa$ 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.\(^6\)-\(^8\) The congruence between the two data sources has been investigated in a few studies using subject-specific record-linkage techniques.\(^9\)-\(^11\) 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.\(^6\)-\(^8\),\(^10\),\(^12\) 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
Table 1: Survey questions, ICD-9-CM diagnosis codes, and ATC prescription drug codes for chronic disease case ascertainment

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

1. Individuals with congestive heart failure were excluded from the “heart disease” category
2. 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). 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. Figure 1 depicts the variation in estimates of the κ coefficient for all investigated case definitions. For arthritis, κ 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, κ 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, κ ranged from 0.27 to 0.58, which represents fair to moderate agreement.

TABLE 2

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

* 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.
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.
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. 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,” 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. As well, some diseases may not be accurately captured via self-report. For example, Kriegsman et al. argue 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, although the magnitude of the differences in $\kappa$ 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 cause.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 narrower set of diagnostic codes (i.e. 430, 431, 434, 345, 436), estimates of the \( \kappa \) 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 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.42-45 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


