Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Deterministic and Probabilistic Record Linkage

Claude Nadeau, Marie P. Beaudet and Jocelyne Marion[1]

Abstract

Pursuing reduction in cost and response burden in survey programs has led to increased use of information available in administrative databases. Linkages between these two data sources is a way to exploit their complementary nature and maximize their respective usefulness. This paper discusses the various ways we have performed record linkage between the Canadian Community Health Survey (CCHS) and the Health Person-Oriented Information (HPOI) databases. The files resulting from selected linkage methods are used in an analysis of risk factors for having been hospitalized for heart disease. The sensitivity of the analysis with respect to the various linkage approaches is investigated.

KEY WORDS: Heart disease; Sensitivity analysis.

1. Introduction

Record linkage within and across data sources can increase the amount and quality of information available for analysis. Within a database, it can identify duplicate records. Across databases, it can be used to augment the range of measures or to check the degree of agreement between different versions of a construct. Combining data sources allow analyses that would not be achievable otherwise. Furthermore, record linkage may permit the inclusion of individuals who tend to be non-participants in surveys because of time constraints, ill health or refusals, and may help reduce costs and response burden.

The present study, in which data files have been linked using both a deterministic and a probabilistic approach, illustrates the sensitivity of the analysis when selected linkage methods are used. The analysis, based on linkage of the 2001 Canadian Community Health Survey (CCHS) and the Health Person-Oriented Information (HPOI) database focused on the association between personal characteristics and modifiable risks factors and the likelihood of hospitalization for heart disease.

Section 2 provides a short description of the data sources and the number of survey respondents who agreed to have the information they provided linked to administrative records. The results of analyses contrasting the findings when records were obtained with deterministic and probabilistic record linkage methods are presented; limitations associated with the various linkage methods, as well as their impact on statistical power and bias, are also discussed. Closing remarks are found in Section 4.

2. Record Linkage

This section describes the record linkage component of the study. The two data sources that were linked are introduced in Section 2.1. Seven linkage methods are described in Section 2.2, followed by summary results in Section 2.3.

[1] Claude Nadeau, Statistics Canada, Household Survey Methods Division, 16-F R.H. Coats Building, Ottawa, Canada K1A 0T6 (claude.nadeau@statcan.ca); Marie P. Beaudet, Statistics Canada, Health Statistics Division, 2200-H, Main Building, Ottawa, Canada K1A 0T6 (marie.p.beaudet@statcan.ca); Jocelyne Marion, Statistics Canada, Household Survey Methods Division, 16-Q R.H. Coats Building, Ottawa, Canada K1A 0T6 (jocelyne.marion@statcan.ca)

2.1 Data sources and linking variables

The data sources that were linked are the Hospital Person-Oriented Information (HPOI) and the Canadian Community Health Survey (CCHS).

At the time of the analysis, the HPOI contained information on each hospital discharge in Canada from fiscal year 1992/93 through fiscal year 2003/04 (from April 1st 1992 through March 31st 2004) for which a Health Insurance Number (HIN) was available. In addition to general characteristics about the patient such as date of birth, sex and HIN, the HPOI contains diagnostic and treatment information for each hospital discharge.

Every fiscal year, provinces send their hospital discharge data to the Canadian Institute of Health Information (CIHI). CIHI then sends an edited version of the data to Statistics Canada where additional processing is done to check the consistency and quality of the records. Part of this additional processing is assuring that date of birth, sex and discharge disposition are consistent across all records belonging to the same person using a particular HIN. This is how multiple users of the same HIN are identified (e.g. a child using his/her mother's HIN). The database contains about 3 million discharges per fiscal year. Data from the first two fiscal years (1992/93 and 1993/94) were not included in the linkage because discharge information from some provinces or territories was not available. Therefore, only data from fiscal year 1994/95 onward (from April 1st 1994 through March 31st 2004) were used. Information for this period is complete except for the Yukon in fiscal years 1994/95 through 1996/1997. Furthermore, data from Quebec were not included in the linkage for reasons described below.

The CCHS is an annual survey that collects information on the health of Canadians, their use of health care services, and some of the factors that can affect health. It is conducted in two cycles: ".1" and ".2". The regional surveys (".1") have about 130,000 respondents. They have been carried out every other year starting with cycle 1.1 in 2001. The provincial surveys (".2") have about 35,000 respondents and have been conducted every other year starting in 2002. Detailed information about the CCHS can be found in Béland (2002).

At each CCHS interview, respondents were asked for permission to link the information they provided to administrative data. Only those who agreed to such a linkage were retained for this study. As can been seen in Table 1, approximately 90% agreed to link in cycle 1.1. In subsequent cycles, the figure was around 85%.

Some individuals may be surveyed more than once in CCHS. This typically happens across cycles, but on rare occasions, it can occur within a cycle because of the use of dual frames, a listing of household and a telephone frame. The repeated selection of a respondent does not cause a problem for record linkage as long as caution is exercised to ensure that linkage results are consistent for those individuals.

Table 1. Number and percentage of Canadian Community Health Survey respondents who agreed to linkage and number and percentage lacking health insurance number by survey cycle

		Respondents who agreed to link*		Respondents without a HIN*	
Cycle (year)	Number of Respondents *	Number	Percent	Number	Percent
1.1 (2001)	131535 (108868)	119383 (98450)	90.8% (90.4%)	36103 (29767)	30.2% (30.2%)
1.2 (2002)	36984 (31652)	32269 (27370)	87.3% (86.5%)	5114 (4067)	15.8% (14.9%)
2.1 (2003)	135573 (106473)	114287 (89536)	84.3% (84.1%)	32923 (24792)	28.8% (27.7%)
2.2 (2004)	35107 (30327)	30141 (25866)	85.9% (85.3%)	5385 (4386)	17.9% (17.0%)
3.1 (2005)	132947 (103056)	115399 (89108)	86.8% (86.5%)	36239 (26674)	31.4% (29.9%)

* Information in parentheses excludes Quebec

Although the databases contain many variables -- more than 100 for the HPOI and more than 1000 for the CCHS -- only a few were chosen as matching fields for linkage: HIN, date of birth, postal code, province and sex. Names are not available on the HPOI. In the HPOI, key variables for the linkage were missing or incomplete for Quebec

because HINs are scrambled, postal codes are truncated to the first 3 characters, and date of birth is missing. As stated previously, Quebec residents were excluded from the analyses.

For the remaining provinces and territories, the HIN may be missing on the CCHS, but it is always available on the HPOI. The date of birth is almost always available on the HPOI. A date of birth is always available on the CCHS, but depending on the cycle, between 1% and 6% are partial dates, usually year without a month or day of birth. The postal code is seldom missing on the CCHS or HPOI -- fewer than 1% of records are affected on either file. In the HPOI, province refers to the province issuing the HIN and is always available. In the CCHS, two variables indicate province: one refers to the province that issued the HIN, and the other, to the province of residence of the respondent. The former is missing whenever the HIN is missing; the latter is always available. Finally, sex is always available on both files.

2.2 Deterministic and probabilistic linkage methods

In a deterministic linkage, two records are linked if and only if the matching fields are not missing and agree perfectly. Six deterministic linkages were done. The first was performed with province, HIN and sex as linking variables and is referred to as Method 1. In Method 2, year of birth was added to the first set of linking variables. In the third approach, month of birth was included as well. For Method 4, the complete date of birth, (year, month and day) was used in addition to the first set of linking variables, province, HIN and sex. In Method 5, postal code was included in the set of the linking variables used in Method 2. In Method 6, the HIN was excluded, and province, sex, date of birth and postal code were used to link the two data sets. Province, when used as a linking variable, refers to the province issuing the HIN, except for Method 6. In that approach, a match on province was accepted when either of the two values for the CCHS variables indicating province matched the value of province in the HPOI. Table 2 in Section 2.3 provides a summary of the matching fields used in the deterministic linkages.

By definition, Methods 1 through 4 are nested. This means that links made by Method 4 will be made by Methods 1, 2 and 3. Likewise, links obtained by Method 3 will be replicated by Methods 1 and 2, but not necessarily by 4. Method 5 corresponds to Method 2 with the additional requirement of obtaining an agreement on postal code. Methods 1, 2 and 5 are, therefore, also nested. Method 6 is the only method that does not use the HIN as a matching variable. Unlike the other deterministic methods, it is the only approach with the potential to link two records when the HIN is missing from the CCHS database. Since the HIN is such an important matching field, in order to minimize the likelihood of false links, it was ignored only when all the other linking variables were used.

Probabilistic linkage does not require complete agreement on the matching variables. Point systems are devised for all the matching fields. The value of the points may be chosen in various ways, including through probabilities, hence the name probabilistic linkage. When two records are compared, points are given or subtracted based on similarities or differences between the fields being matched. For instance, if the values of the postal code are the same between two records, a positive score is assigned; if the values look similar according to a string comparison algorithm, a lower positive score is assigned, reflecting the partial agreement; if the values on the two records are totally different, points are subtracted. The number of points should reflect the importance of the matching variable, which is usually related to its uniqueness. In our study, the HINs are unique and were given a comparatively high score in absolute value.

The scores are added across matching fields to arrive at a total linkage weight for each potential pair. Based on the distribution of the linkage weight, thresholds are selected. The optimal distribution is bi-modal. Pairs between records above a selected threshold are accepted as true matches, pairs below a selected threshold are rejected as matches, and pairs between the two selected cut-off points are considered potential matches and are usually reviewed manually. For this project, to minimize review work, the two cut-off points were identical. Newcombe (1988) contains a detailed discussion on how points should be attributed and used. The software, Generalized Record Linkage Software (GRLS) developed at Statistics Canada, was used to execute the probabilistic linkage for this study. It follows the probabilistic linkage theory developed by Fellegi and Sunter (1969). The probabilistic linkage approach for this project is referred to as Method 0. A detailed description of this approach to linkage is provided by Nadeau (2007).

2.3 Summary results

Table 2 displays the matching rates of the CCHS sample to the HPOI data base for respondents who agreed to link. The proportion varies by cycle and linking method. Rates are not weighted by the sample weights and exclude residents of Quebec.

As shown in Table 1, the proportion of missing HIN in the CCHS for ".1" cycles is about twice that of ".2" cycles: 30% and approximately 15%, respectively. Therefore, a higher matching rate for ".2" cycles is expected. This is the case for cycle 1.2 (Table 2), except for Method 6, which does not use HIN. However, the matching rate is not higher for cycle 2.2, because children were over-sampled for this cycle and they are less likely to be hospitalized.

Table 2. Proportion of CCHS respondents linked to HPOI records, by method and cycle

Type of linkage	**Probabilistic**	**Deterministic**					
Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Linking variables	Province HIN Sex Birth date Postal code	Province HIN Sex	Province HIN Sex Birth year	Province HIN Sex Birth year Birth month	Province HIN Sex Birth date	Province HIN Sex Birth year Postal code	Province Sex Birth date Postal code
Cycle (year)	%	%	%	%	%	%	%
1.1 (2001)	34.6	29.2	28.1	27.4	25.3	20.1	24.7
1.2 (2002)	37.2	35.1	34.0	33.4	31.0	23.3	24.5
2.1 (2003)	33.4	29.1	27.6	26.6	24.5	19.0	22.4
2.2 (2004)	27.7	26.4	25.2	24.8	22.6	15.6	16.5
3.1 (2005)	30.8	26.4	26.1	25.7	23.6	15.1	18.9

Note: Rates exclude Quebec respondents. Proportions are not weighted by sample weights.

Regardless of method, matching rates in ".1" cycles tended to decrease over time. This is partly attributable to the decrease in the number of records in the HPOI database, presumably owing to a greater reliance on ambulatory care services and day surgery. The decrease between cycle 2.1 to 3.1 is more important than the one between cycle 1.1 to 2.1 because cycle 1.1 had a smaller proportion of respondents older than 65. Since older people are more likely to be hospitalized, having fewer older respondents lowers the matching rates for cycle 1.1.

Method 5 always yielded the lowest matching rate, because it requires agreement on many distinct fields. The only difference between Methods 2 and 5 is the requirement for postal code agreement for the latter. The impact of this requirement can be measured by comparing the matching rates of the two methods.

The decline in matching rates for Methods 1 through 4 was expected because of the nested nature of these linking approaches. The difference in matching rates between Methods 3 and 4 is more substantial, mainly because of some missing values on day of birth in the CCHS and the HPOI databases. Missing information on day of birth in the HPOI is imputed as the first day of the month, and the imputed value has a lower probability of linkage.

Method 6 tends to yield links that differ from the other deterministic methods. Among the 24,292 CCHS records in cycle 1.1 that were matched to HPOI records, 6,373 were not matched by the other deterministic methods; 91.9% of these had no HIN on the CCHS database. Most (97.4%) of the 24,292 links made by Method 6 were replicated by Method 0, the probabilistic method.

3. Risk factors associated with hospitalization for heart disease

Seven linked files between CCHS 1.1 and HPOI were created, one for each record linkage method. They included the variables needed for the analyses. Each file contained 72,493 observations, down from 119,383 because residents

of Quebec and respondents younger than 30 were excluded from the analysis. The prevalence of heart disease is low among people younger than 30. Seven logistic regressions were fitted to estimate the association between the outcome -- being hospitalized for a heart disease-related diagnosis in the two years following a CCHS interview -- and selected socio-demographic characteristics, a history of hospitalization for heart disease and modifiable risk factors available in the CCHS. The seven record linkage methods yielded seven distributions of the outcome and of one independent variable, history of hospitalization for heart disease in the five years before each respondent's interview. To account for the survey design effects, the bootstrap method was used to estimate the standard errors of the coefficients (Rao, Wu & Yue, 1992; Rust & Rao, 1996; Yeo, Mantel & Liu, 1999; Kovacevic & Roberts, 2002). The data and methodology are described in Section 3.1. Results are presented in Section 3.2 and are followed by a section on study limitations.

3.1 Data and methods

Data from the CCHS (cycle 1.1) provided information on the socio-demographic characteristics of the respondents and on selected modifiable risk factors for heart disease. The dichotomous outcome, hospitalization for heart disease in the two years following the date of respondents' CCHS interview, was obtained from the HPOI. Respondents were classified as having been hospitalized for heart disease and assigned a value of 1 if their most responsible diagnosis fell within ICD-9 codes beginning with 402, 404, 41 or 42, or ICD-10 codes beginning with I11, I13, I2, I3, I4, I50 or I51. All others were assigned a code of 0, indicating no hospitalization for heart disease. The HPOI also provided a dichotomous indicator of history of hospitalization for heart disease. Respondents who had a diagnosis of heart disease in any of the diagnosis fields as indicated by the ICD-9 or ICD-10 codes listed above in the five years preceding their CCHS interview were assigned a code of 1, reflecting a history. All others were given a code of 0 for no history.

Personal characteristics include sex, age, education, income, marital status and living arrangements. A set of dummy coded variables was created to reflect each measure. For sex, a code of 0 was assigned to men and a code of 1 was assigned to women. Five age groups were formed: 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70 or older. The 30-to-39 age group was used as the reference category. Four groups were established for educational attainment: less than secondary graduation, secondary graduation, some postsecondary, and post-secondary graduation, which was selected as the reference category. Household income takes into account income of all household members from all sources and household size. Four groups were established: lowest, lower-middle, upper-middle, and highest. The last was used as the reference category. To maximize the sample size, a dummy variable was created to indicate whether information on household income was available. Respondents with a missing value on this derived variable were coded as 1. Marital status was subdivided into three groups: married or living with partner, single, or previously married, which included respondents who were separated, divorced or widowed. The first group was the reference category. Respondents who lived alone were assigned a code of 1 on the living arrangement variable. All others were assigned a code of 0.

A self-perceived measure of stress was included in the analysis. Respondents were asked: "Thinking about the amount of stress in your life, would you say that most days are not at all stressful, not very stressful, a bit stressful, quite a bit stressful or extremely stressful?" Three groups of approximately equal size were created: not at all or not very stressful (the reference group), a bit stressful, and quite a bit or extremely stressful.

Modifiable risk factors available in the CCHS include smoking, leisure-time physical activity, alcohol consumption and body mass index. Smoking status was determined by asking individuals if they smoked cigarettes daily, occasionally, or not at all. Three groups were created: never, former, and current (daily or occasional) smoker. The "never" category was used as the reference category. Two levels of leisure-time physical activity were defined: active or moderate (1.5 or more kilocalories per kilogram of body weight per day), the reference category, and inactive (less than 1.5 kilocalories per kilogram of body weight per day). A moderate level of activity would, for example, be walking for an hour four times per week. Following the methodology of Wilkins (2002), four groups were specified to indicate level of alcohol consumption. Lifetime abstainer, the reference category, was made up of respondents who reported never having had a drink. Former drinkers included respondents who reported that they had not had a drink in the past year, but that they had consumed at least one drink before the past year. The level of alcohol consumption for respondents who reported that they had had at least one drink in the past 12 months was

derived from the number of drinks they reported during the week before their CCHS interview. Occasional drinkers were those who reported no drinks in the past week. Light drinkers were those who reported one drink in the past week. Because men and women metabolize alcohol differently, sex-specific cutoffs were used to classify moderate and heavy drinkers. Moderate drinking was defined as two to nine drinks in the past week for women, and two to fourteen for men. Heavy drinking was defined as 10 or more drinks in the past week for women and 15 or more for men. Body mass index (BMI) is calculated by dividing weight in kilograms by the square of height in meters. Four BMI groups were created. Respondents with a BMI below 20 were classified as having insufficient weight. Those with a BMI greater than or equal to 20 and less than 25 were considered as having a healthy weight and were selected as the reference category. Respondents with a BMI greater than or equal to 25 but less than or equal to 27 were classified as having some excess weight, and those with a BMI greater than 27 were labeled as overweight.

3.2 Results

For each linking method, Table 3 shows, the distribution of the outcome -- hospitalization in the two years following a CCHS interview -- and of one of the independent variables -- history of hospitalization for heart disease. In general, methods with low linkage rates (Table 2) yielded fewer people with a positive response on these two variables. This was expected, since case ascertainment depends on the record linkage method. Results from the logistic regressions are presented in the Appendix, Table 4. Because of lack of space, regressions results for linkage Methods 2 and 3 are not shown. Results from these two methods typically fell between Method 1 and Method 4 and are available upon request from the authors.

Table 3. Distribution of hospitalization for heart disease in the two years following the CCHS interview and history of hospitalization for heart disease, by record linkage method

Variable	Value	Method						
		0	1	2	3	4	5	6
Hospitalization for heart disease post CCHS	yes	1707	1410	1349	1315	1239	1158	1410
	no	70786	71083	71144	71178	71254	71335	71083
Previous hospitalization for heart disease	yes	3937	3345	3221	3142	2905	2674	3145
	no	68556	69148	69272	69351	69588	69819	69348

As can be observed in Table 4, results from the logistic regressions across all the linkage methods showed a high level of consistency and were similar to those reported in the scientific literature (Wilkins, 2002). As expected, a previous hospitalization for heart-related disease increased the odds of a future hospitalization for heart disease. As well, women were less likely than men to be hospitalized for heart disease, even when heart-related history of hospitalization was controlled. Advancing age was associated with a higher likelihood of being hospitalized for heart disease. These results were consistent across all linkage methods with one exception. With Methods 4 and 5, no difference emerged in the likelihood of an heart-related hospitalization between the 40-to-49 age group and the reference category, the 30-to-39 age group, probably the result of lower statistical power.

When controlling for selected factors available in the datasets and listed in Section 3.1, certain modifiable risk factors consistently increased the likelihood of a future hospitalization for heart disease: being a former or a current smoker, being inactive, and being overweight. Some excess weight reached statistical significance with the data of Methods 0 and 6, while being underweight reduced the odds in data derived with Methods 5 and 6.

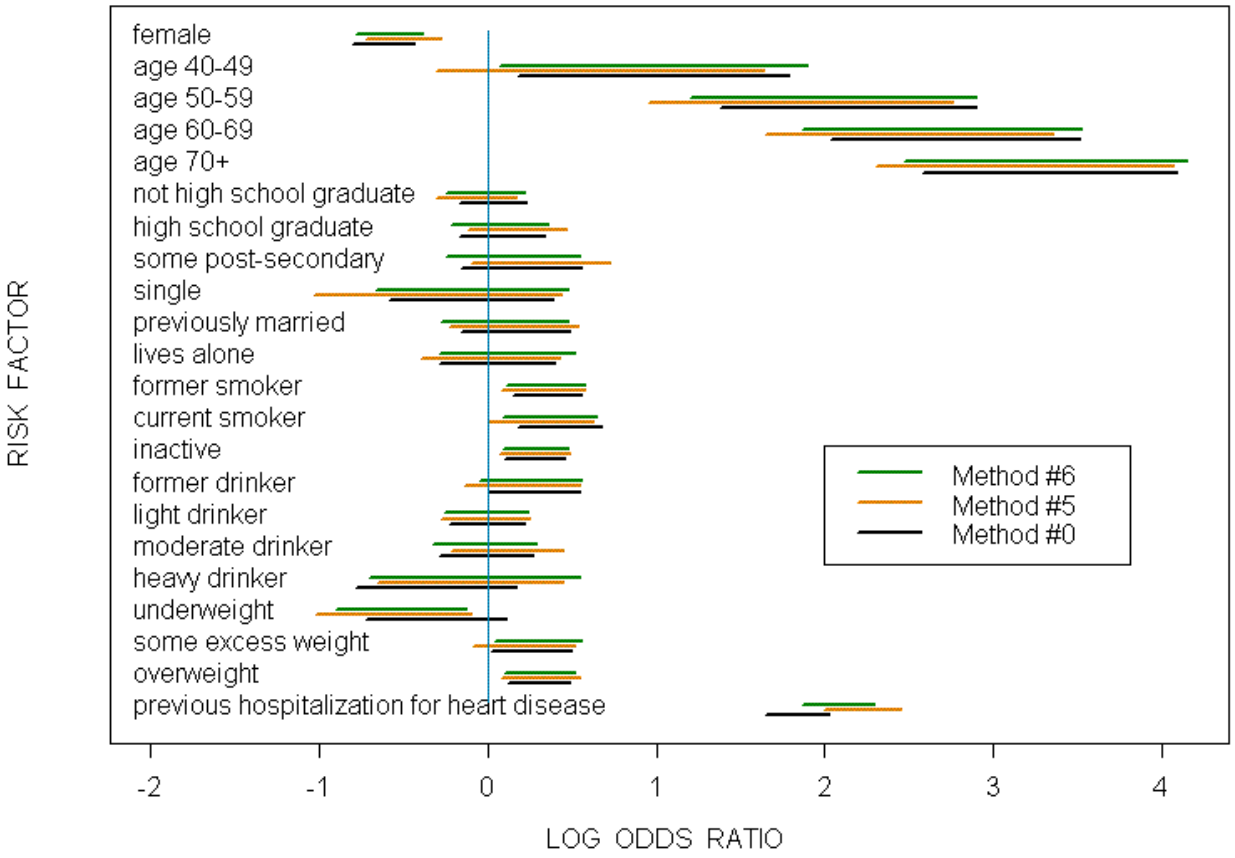
In general, education, income, marital status, living arrangements and perceived stress did not influence the odds of a future hospitalization for heart disease when other factors were controlled. Methods 1, 4 and 5 revealed a higher likelihood of hospitalization for respondents in households in the lower-middle income category, and a significant association was detected for the lowest income group with Method 1.

Figure 1 depicts 95% confidence intervals for the logistic regression coefficients, i.e., the log of odds ratios. Only Methods 0, 5 and 6 were retained because they showed the greatest disparity in confidence intervals. This is not surprising, since Methods 0 and 5 have, respectively, the highest and lowest linkage rates and Method 6 does not use the HIN. Likewise, to improve readability, stress and income are not shown. Confidence intervals from the three

methods tended to overlap substantially. However, alignment was poor for history of heart-related hospitalization. It is possible that the linkage method chosen has an impact on the response variable and history of hospitalization for heart disease, since both were derived from the linkage. In contrast, the other variables selected from the CCHS are based on self-report.

The confidence intervals tended to be smaller for Method 0 and larger for Method 5. Making fewer links leads to a response variable that equals 1 less often. Logistic regression is known to yield parameter estimates with greater variability under such circumstances. Linking errors may also have played a role. Method 0 is the best of the seven methods; it yielded few false links and missed fewer links. Neter, Maynes & Ramanathan (1965) and Krewsky, Wang, Bartlett, Zielinski & Mallick (2001) showed that linkage errors may lead to bias and increased variability. Although the contexts of their papers were, respectively, linear models and survival analysis, it is reasonable to believe that their results may apply to logistic regression. Scheuren & Winkler (1993) introduced a technique that could alleviate bias in analyses caused by linkage errors. It was not implemented here, since the methodology was developed for linear models.

Figure 1. 95% confidence intervals for log of odds ratios



Note: Log of odds ratios are equivalent to logistic regression coefficients.

The analyses were redone to include only respondents who had not been hospitalized for heart disease in the five years before their CCHS interview. Detailed results are not shown here, but they are available upon request from the authors. As with the previous data sets, results were fairly consistent across record linkage methods. As well, many of the associations present in the first set of analyses were replicated.

Women were less likely to be hospitalized with the most responsible diagnosis as heart disease. As age increased, so did the likelihood of being hospitalized. However, the association between age groups and hospitalization was not present for the 40 to 49 age group. Being a former smoker increased the likelihood of hospitalization for heart disease in the two years after a CCHS interview. Smoking and being overweight increased the odds of

hospitalization with all the linkage methods, except Method 5, the method with the lowest linkage rate. Being underweight lowered the odds of hospitalization in data sets derived with Methods 4, 5 and 6. Inactivity was positively associated with future hospitalization, except with Methods 4 and 5. With these data sets, being a former drinker increased the odds of a future hospitalization for heart disease. This is not unexpected, since former drinkers may be in poorer health (Wilkins, 2002). Living alone and reporting a moderate amount of stress were positively associated with hospitalization with, Methods 0 and 6 for the former, and Methods 0, 1, 4 and 6 for the latter.

As was the case in the previous set of analyses, education and marital status were not associated with a future hospitalization for heart disease. Belonging to a lower-middle income household was predictive of a future heart disease-related hospitalization in data sets derived from Methods 0, 1, 4 and 5.

3.3 Limitations

The present study has a number of limitations. The CCHS sample for this study was made up of respondents who agreed to have the information they provided linked to administrative data sources. The amount of bias created by the exclusion of respondents who did not agree to such a linkage is unknown. However, it is expected to be small, since 90% of respondents agreed to the linkage, and the sampling weights were adjusted to ensure that the reduced sample remained representative, as were the results of the logistic regressions.

This study excluded people living in institutions. In general, residents of institutions tend to be in poorer health than their community counterparts. Their exclusion may have weakened the strengths of the reported associations.

The survey information is based on self-report, and its accuracy is unknown. Errors could arise from respondents' unwillingness to disclose certain information, difficulty in recall, and a desire to please the interviewer or to present a positive image. The distortion created by these reporting errors is unknown but is expected to be small since it should be random, i.e., unrelated to the outcome variable.

Two types of errors can occur when records are linked: two records are linked but they do not belong to the same person or two records which refer to the same person are not linked. Both types of error may have occurred in this study. It is possible that a record on the CCHS was linked to multiple records in the HPOI and that these records did not belong to the same person. This was most likely with the probabilistic approach, Method 0, or with Method 6 where the HIN was not a linking variable. Whenever problems were identified, some or all links were rejected.

As well, because of mobility within a province, the use of postal code in Method 5 and 6 may have lowered the number of links obtained, especially when matching was attempted between old hospitalisation records and CCHS records. Mobility between provinces compounds the problem, since matching variables change after an interprovincial move (except for date of birth and gender), and all linking methods, including the probabilistic one, are affected equally. The magnitude of the underestimation of a previous or a future hospitalization for heart-related disease or both depends on the number of moves and their timing. Availability of respondents' names as a matching variable would have reduced the magnitude, but they were not available in the HPOI database.

The availability of names would also have helped in the evaluation of the quality of the links. However a review of the links obtained with Method 6 based on a subsample for whom the HIN was available in cycle 1.1, suggested that it yielded false links about 3% of the time. In other words, whenever it was available, the HIN corroborated about 97% of the links.

The effect on the results caused by the exclusion of Quebec residents is unknown and limits the generalizability of the study findings. The effect of this exclusion is proportional to the extent that Quebec residents have a different personal and risk profile and to the extent that healthcare delivery in that province differs from those in other provinces.

To ensure that people included in the analysis had been hospitalized for heart disease and to reduce the effect of different diagnostic and coding practices across hospitals, only the most responsible diagnosis was chosen for the outcome. Re-abstraction studies indicate that diagnostic and coding practices are fairly consistent across

jurisdictions, especially when the most responsible diagnosis is examined and the condition has a high prevalence (Juurlink, Preyra, Croxford, Chong, Austin, Tu & Laupacis, 2006; CHIMA & CIHI, 2005; CIHI, 2004; CIHI, 2003; CIHI, 2002). Previous hospitalization for heart disease was based on all the available diagnoses to maximize the chance of identifying every respondent with such a history.

Because heart disease was selected as the outcome, people who were hospitalized for a myocardial infarct (AMI) and who underwent a revascularization during their hospital stay may have been missed if their most responsible diagnosis was coded as coronary artery disease. This omission includes a small number of patients and may have contributed to an underestimation of the relationships presented here, since these patients' characteristics should be similar to those for whom the most responsible diagnosis was heart disease.

People who were hospitalized for heart disease in the two years after their CCHS interview, but who were not discharged by the end of the two-year period, were treated as if they had not been hospitalized. Those who died during that two-year period without being hospitalized were also considered as not having been hospitalized. The resulting bias is unknown, but probably weakened the associations of the present study, lowering its statistical power since it is highly probable that these individuals' personal and risk profile was similar to that of the respondents who were hospitalized for heart disease.

4. Concluding remarks

Results were consistent across record linkage methods irrespective of the inclusion of subjects with a history of hospitalization for heart disease. Results were also in accord with previous research that has examined risk factors for heart disease. In the present analysis, women were less likely to be hospitalized for a heart disease-related diagnosis. Advancing age was associated with increased odds of hospitalization for heart disease for age groups 50 or older. Former smokers and persons classified as overweight were at increased odds of hospitalization for heart disease.

Differences between the two sets of analyses, that is, the set which included subjects with a history of hospitalization for heart disease and the set which did not, are noteworthy. In the latter, being classified as a former drinker was positively associated with the outcome. In contrast, inactivity was a predictor in the former. As expected, in the latter set, a history of heart disease-related hospitalization was associated with increased odds of future hospitalization for heart disease. In general, education, household income, living arrangements and marital status were not predictive of hospitalization for heart disease.

Despite the limitations of the present study, including its reduced potential for generalization to other classes of diseases, the probabilistic method appears to yield findings that were replicated by most of the deterministic linkage methods. It is recommended as the method of linkage, since it provides increased power to detect statistically significant associations, yielding few false links and missing fewer links than the deterministic linkage approaches.

Acknowledgements

Helen Johansen, Brad Thomas and Richard Trudeau are thanked for comments and discussions that have improved this paper. As well, thank you to Mary Sue Devereaux who has greatly ameliorated the readability of the article.

References

Béland Y. (2002), "Canadian Community Health Survey - Methodological overview", *Health Reports*, 13(3), 9-14.

Canadian Health Information Management Association (CHIMA) & Canadian Institute for Health Information (CIHI) 2005, "Reabstraction Study of the Ontario Case Costing Facilities for Fiscal Years 2002/2003 and 2003/2004", Toronto.

Canadian Institute for Health Information (CIHI) 2004, "Data Quality of the Discharge Abstract Database Following the First-Year Implementation of ICD-10-CA/CCI – Final report", Toronto.

Canadian Institute for Health Information (CIHI) 2003, "Discharge Abstract Database (DAD)/ CMG/Plx Data Quality", Toronto.

Canadian Institute for Health Information (CIHI) 2002. "Discharge Abstract Database Data Quality Re-abstraction Study. Combined Findings for Fiscal Years 1999/2000 and 2000/2001", Toronto.

Fellegi I.P. and A.B. Sunter (1969), "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64, 1183-1210.

Juurlink D., C. Preyra, R. Croxford, A. Chong, P. Austin, J. Tu and A. Laupacis (2006), "Canadian Institute for Health Information Discharge Abstract Database: A Validation Study", Institute for Clinical Evaluative Sciences, Toronto.

Kovacevic, M and G. Roberts (2002), "Notes on Estimating Equations Bootstrap", unpublished report, Statistics Canada.

Krewski D., Y. Wang, S. Bartlett, J.M. Zielinski and R.Mallick (2001), "The Effect of Record Linkage Errors on Statistical Inference in Cohort Mortality Studies", *Proceedings of Statistics Canada Symposium*, 1-13.

Nadeau C. (2006), "Linking HPOI to CCHS", unpublished report, Ottawa, Canada: Statistics Canada.

Newcombe H. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford Medical Publications, Oxford.

Neter J., E.S. Maynes and R.Ramanathan (1965), "The Effect of Mismatching on the Measurement of Response Errors", *Journal of the American Statistical Association,* 60, 1005-1027.

Rao J.N.K., C.F.J Wu. and K. Yue (1992), "Some recent work on resampling methods for complex surveys", *Survey Methodology* (Statistics Canada, Catalogue 12-001), 18(2), 209-217.

Rust K.F. and J.N.K. Rao (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research*, 5, 281-310.

Scheuren, F. and W.E. Winkler (1993), "Regression Analysis of Data Files that are Computer Matched", *Survey Methodology,* 19, 39-58.

Wilkins K. (2002), "Moderate alcohol consumption and heart disease", *Health Reports*, 14(1), 9-24.

Yeo D., H. Mantel and T.P. Liu (1999), "Bootstrap variance estimation for the National Population Health Survey", American Statistical Association: Proceedings of the Survey research Methods Section, Baltimore, August 1999.

Appendix

Table 4. Adjusted odds ratios for an heart disease-related hospitalization in the two years following a CCHS interview, by selected characteristics including history of hospitalization for heart disease, household population, Canada including the territories but excluding Quebec

	Method 0		Method 1		Method 4		Method 5		Method 6	
Risk factor	Odds ratio	95% C. I.	Odds ratio	95% C. I.	Odds ratio	95% C. I.	Odds ratio	95% C. I.	Odds ratio	95% C. I.
female	0.54***	0.45-0.65	0.57***	0.46-0.69	0.59***	0.48-0.73	0.61***	0.49-0.76	0.56***	0.46-0.68
age 40-49	2.68*	1.20-5.98	2.31*	1.01-5.30	2.26	0.94-5.46	1.95	0.74-5.13	2.68*	1.07-6.69
age 50-59	8.51***	3.97-18.2	6.82***	3.12-14.9	6.41***	2.77-14.8	6.41***	2.60-15.8	7.75***	3.31-18.1
age 60-69	16.1***	7.65-33.8	12.9***	6.12-27.1	11.4***	5.15-25.2	12.2***	5.19-28.6	14.8***	6.46-34.1
age 70 or older	28.2***	13.3-59.9	24.9***	11.5-53.9	23.7***	10.5-53.9	24.3***	10.0-58.8	27.5***	11.9-63.3
not high school graduate	1.03	0.84-1.26	0.95	0.77-1.18	0.93	0.74-1.18	0.94	0.74-1.19	0.99	0.78-1.25
high school graduate	1.09	0.85-1.40	1.14	0.87-1.49	1.10	0.82-1.47	1.19	0.89-1.59	1.07	0.81-1.43
some post-secondary	1.22	0.85-1.74	1.31	0.90-1.90	1.29	0.87-1.89	1.37	0.91-2.06	1.16	0.78-1.72
lowest income	1.33	0.95-1.87	1.55*	1.09-2.22	1.43	0.99-2.06	1.38	0.91-2.10	1.13	0.78-1.64
lower-middle income	1.28	0.93-1.77	1.48*	1.08-2.03	1.47*	1.05-2.07	1.49*	1.05-2.12	1.22	0.85-1.76
upper-middle income	0.89	0.66-1.19	1.02	0.75-1.39	1.01	0.73-1.40	1.01	0.72-1.43	0.85	0.62-1.15
undisclosed income	1.26	0.90-1.78	1.31	0.90-1.90	1.10	0.74-1.64	1.26	0.84-1.90	1.09	0.75-1.58
single	0.91	0.56-1.47	0.76	0.42-1.37	0.71	0.35-1.42	0.74	0.36-1.55	0.91	0.51-1.61
previously married	1.18	0.86-1.63	1.17	0.84-1.63	1.08	0.75-1.55	1.17	0.79-1.72	1.11	0.76-1.61
lives alone	1.06	0.75-1.48	1.01	0.71-1.43	1.07	0.72-1.58	1.02	0.68-1.53	1.12	0.75-1.67
a bit stressful	1.12	0.92-1.35	1.09	0.89-1.34	1.09	0.88-1.34	1.11	0.89-1.39	1.15	0.92-1.43
quite a bit, extremely stressful	1.04	0.84-1.30	1.14	0.90-1.45	1.11	0.86-1.44	1.16	0.89-1.53	1.02	0.80-1.32
former smoker	1.42***	1.16-1.75	1.43**	1.14-1.78	1.44**	1.14-1.82	1.39**	1.09-1.78	1.41**	1.12-1.78
current smoker	1.54***	1.20-1.97	1.53**	1.16-2.02	1.41*	1.05-1.90	1.37*	1.01-1.86	1.45**	1.10-1.91
inactive	1.33**	1.11-1.58	1.31**	1.09-1.59	1.30**	1.06-1.59	1.32*	1.07-1.63	1.33**	1.10-1.62
former drinker	1.31	1.00-1.73	1.30	0.96-1.76	1.27	0.91-1.77	1.22	0.87-1.72	1.29	0.95-1.75
light drinker	1.00	0.79-1.25	1.01	0.80-1.29	1.03	0.80-1.33	0.99	0.76-1.29	0.99	0.77-1.27
moderate drinker	0.99	0.75-1.31	1.06	0.79-1.43	1.05	0.77-1.44	1.12	0.81-1.57	0.98	0.72-1.34
heavy drinker	0.74	0.46-1.18	0.87	0.53-1.42	0.92	0.55-1.53	0.90	0.52-1.57	0.92	0.49-1.73
underweight	0.74	0.49-1.12	0.76	0.47-1.20	0.68	0.42-1.09	0.57*	0.36-0.90	0.60**	0.41-0.88
some excess weight	1.30*	1.03-1.65	1.20	0.92-1.57	1.23	0.93-1.62	1.24	0.92-1.68	1.35*	1.04-1.75
overweight	1.35**	1.13-1.62	1.32**	1.08-1.61	1.35**	1.09-1.67	1.37**	1.09-1.72	1.36**	1.11-1.68
previous hospitalization for heart disease	6.28***	5.20-7.59	7.36***	6.01-9.01	8.61***	6.97-10.6	9.25***	7.37-11.6	8.00***	6.47-9.89

Data Sources: Canadian Community Health Survey, cycle 1.1, 2001 and Health Person Oriented Information Database, 1994/95 through 2003/04.

* p <=.05; ** p<= .01; *** p<= .001

Note: Based on sample respondents who have agreed to have their information linked. The odds for the reference categories are always 1.00 and are not shown. Results are based on weighted data; standard errors used in the calculation of the confidence intervals were estimated with the bootstrap technique.