Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Record Linkage between the 2006 Census of the Population and the Canadian Mortality Database

Mohan B. Kumar, Rose Evra[1]

Abstract

Vital statistics datasets such as the Canadian Mortality Database lack identifiers for certain populations of interest such as First Nations, Métis and Inuit. Record linkage between vital statistics and survey or other administrative datasets can circumvent this limitation. This paper describes a linkage between the Canadian Mortality Database and the 2006 Census of the Population and the planned analysis using the linked data.

Key Words: record linkage; hierarchical deterministic; vital statistics; mortality rates.

1. Introduction

1.1 Description

The Canadian Mortality Database (CMDB) is a vital statistics dataset that contains information on deaths that occurred in Canada (Statistics Canada, 2015). However, it lacks identifiers for certain populations of interest such as First Nations, Métis and Inuit. This can be circumvented by linking CMDB records to survey or other administrative records. For identifying First Nations, Métis and Inuit, one option is to link to the Census of the Population (Census). This has been done previously with the 1991 Census (Wilkins, 2001); however, because that Census data did not include names, it had to be first linked to tax data to get nominal linkage information. In the 2006 Census, for the first time, names were retained thus allowing a direct link between it and the CMDB. Such a link will not only enable the identification of individuals who self-identified as First Nations, Métis and Inuit in the long-form 2006 Census, it will provide access to the socio-demographic information in the Census for analytical purposes.

This project aims to link the 2006 Census and the CMDB to enable identification of First Nations, Métis and Inuit deaths. The linkage will be enhanced by drawing upon tax data from another linked dataset, the Amalgamated Mortality Database. The linkage data will be used to estimate and compare mortality rates for above-mentioned and the non-Aboriginal populations of Canada.

The record linkage was initiated by the Social and Aboriginal Statistics Division at Statistics Canada in 2015. Formal approval for the record linkage was granted in 2015-16, and the linkage was completed in Q1 of 2016-17. It was conducted with assistance from methodologists in the Household Survey Methodology Division (HSMD). Quality assessment and validation of the record linkage was started in Q1 of 2016-17.

[1] Mohan B. Kumar, (mohan.kumar@canada.ca); Rose Evra, , (rose.evra@canada.ca); Social and Aboriginal Statistics Division, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

2. Methods

2.1 Datasets

2006 Census of the Population: The Census of the Population in Canada is carried out every five years; it collects information on the demographic, social and economic characteristics of Canadians. It consists of the short- and long-form census. In 2006, the latter was administered to one in three households with the rest receiving the short-form. The 2006 Census dataset has 30.1 million records which includes some duplicate records. Some undercoverage exists in the Census, specifically relating to young males (Statistics Canada, 2010).

Canadian Mortality Database (CMDB): The CMDB is an administrative database containing medical (cause of death) and demographic information collected annually from all provincial and territorial vital statistics registries on all deaths in Canada. For the study time-period, there are 1.4 million in-scope records of those who died between 2006 (2006 Census year) and 2011 (last year available in the dataset). The CMDB has some undercoverage and overcoverage (Statistics Canada, 2015): the former because only deaths occurring in Canada are registered in the Vital Statistics, therefore deaths of Canadians, including those in the military or living abroad are not registered, and due to late registration. For overcoverage, it may be the result of inclusion of non-Canadians who arrived after Census day and who died in Canada.

2.2 Linkage methods

The records in the two datasets were linked using a hierarchical, deterministic method (Rotermann, 2014). This entails matching records between the two datasets using common variables. Following an initial quality assessment for use for record linkage, the following variables were used: name, sex, postal code and date of birth. The quality and availability of these variables differed in the two datasets. For example, while sex distributions were similar in the two datasets, postal code information was missing in a small fraction of CMDB records (0.9%) with this increasing in the last year of available CMDB records (2.1%).

The linkage was performed in waves with the linkage criteria becoming less stringent in each subsequent wave (Table 3.1-1). Twenty waves of record linkage were carried out. Waves 1 to 16 used different combinations and stringencies of linking variables. In the last four waves, tax information was used for record linkage. As examples, in Wave 1, records had to match exactly on all four linkage variables; in Wave 4, criterion for the name variable was relaxed using the MixMatch software; and, in Wave 9, criteria for both name and postal code variables were relaxed (by using MixMath for names and using only first character of the postal code). Only non-linked records are used in subsequent waves.

3. Preliminary Findings

3.1 Linkage rates

Of the 1.4 million in-scope CMDB records, 89% were linked to Census records. The linkage rate[2] – as defined by the per cent of CMDB records linked – differed by wave. Waves 1, 2, 4 and 9 contributed 81% of all links (Table 3.1-1).

[2] Linkage rate, here, is defined as the proportion of CMDB records that were successfully linked to Census records.

Table 3.1-1: Per cent of total linked records and per cent of CMDB records linked by wave

Waves	Variables	% links	CMDB
1	Name, DOB, SEX, PC	26.28%	23.29%
2	Name, DOB, SEX	12.87%	11.41%
3	Name, DOB, PC	0.15%	0.13%
4	Last name, first name (MM*)DOB, SEX, PC	27.42%	24.30%
5	Name + DOB + PC3 or PC1	0.05%	0.04%
6	Name + DOB6 + SEX ± PC3	1.84%	1.63%
7	Name + YOB + SEX + PC	0.62%	0.55%
8	Name(MM) + DOB + SEX + PC	7.71%	6.83%
9	Name(MM) + DOB + SEX + PC1	14.30%	12.67%
10	Name(MM) + DOB + PC3	0.24%	0.22%
11	Name + Sex + PC	1.06%	0.94%
12	Name(MM) + DOB(at least 5 common digit) + PC + SEX	3.26%	2.89%
13	Name(MM) + YOB/MOB + SEX +PC3	0.31%	0.27%
14	Name and fname + DOB(allowing one digit difference) + SEX	0.59%	0.52%
15	Name(MM) + DOB + SEX	1.45%	1.29%
16	Father's or mother's surname included in name + DOB + PC	0.13%	0.11%
17	Tax name + tax DOB + Tax PC + SEX	1.09%	0.96%
18	Tax name + tax DOB + SEX	0.24%	0.21%
19	Tax name (MM) + tax DOB + Tax PC + SEX	0.17%	0.15%
20	Tax name (MM) + Tax PC3 + Tax DOB	0.21%	0.19%
Total		**100.00%**	**88.60%**

Notes: DOB: Date of birth, YOB: year of birth; MOB: month of birth, tax DOB: date of birth from tax files, PC: Postal code, PC1: only first character of postal code used; PC3: only first three characters of postal code used; fname: first name, * MM: refers to linkages where the restriction on the name variable was eased using MixMatch version 1.3 software, ±PC3 in wave 6 refers to linkage with and without first three characters of the postal code.

4. Quality assessment of record linkage

4.1 Internal validation

For the internal validation of the record linkage, duplicate links, links to out-of-scope records and patterns in linkage rates were examined.

Quality of linkages: for those records with a valid Census date of birth, fewer than ten had a death date prior to the birthdate. The sex of individual differed in the two datasets for 0.5% of the links.

In 2006, of all the deaths that occurred before Census day, 5% were linked to Census records. Of these, 40% occurred were in the month of May; these could be legitimate links since early enumeration occurred on reserves and in the North. Of those deaths which related those who had a birth day after Census day, 1% linked to Census records. These could be false links; however, there was some late enumeration, which may suggest that some of these may be true links.

Duplicate links: some CMDB records linked to more than one Census record (within a wave). Almost all of these appeared to be duplicates in the Census dataset. Also, some Census records linked to more than one CMDB record; some perhaps due to duplications in the CMDB, and others due to linkage error.

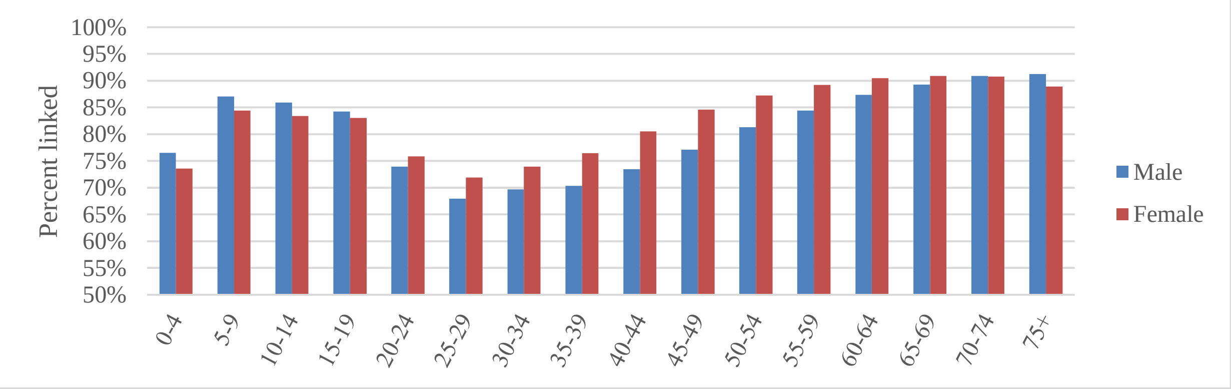
Patterns in linkage rates: linkage rates were examined by categories of different variables such as age group, cause of death, neighbourhood income quintiles, etc. to see if expected patterns emerged.

When linkage rates were examined by sex, contrary to expectations because females are more likely to change last names upon marriage, linkage rates were higher than in males. However, it should be noted that the use of tax names allowed the use of multiple versions of individual's names and parents' names; this may have improved the linkage rates among females. In addition, the undercoverage of males in the Census may have also contributed to the low linkage rates for them.

When linkage rates were examined by year of death, as expected due to mobility, they decreased slightly over time from 89% in 2007 to 88% in 2011.

Linkage rates differed by age group with those aged 25 to 44 having the lowest rates (77%) compared to other groups. Those aged 65 to 84 years had the highest linkage rate (90%). The low linkage rates in the former age group could be explained by their lower Census coverage rates, perhaps due to higher mobility. Within the 25-to-44 age group, among both males and females, linkage rates were lowest among 25-29 year olds (Fig. 4.1-1).

Fig. 4.1-1. Linkage rates by age group and sex

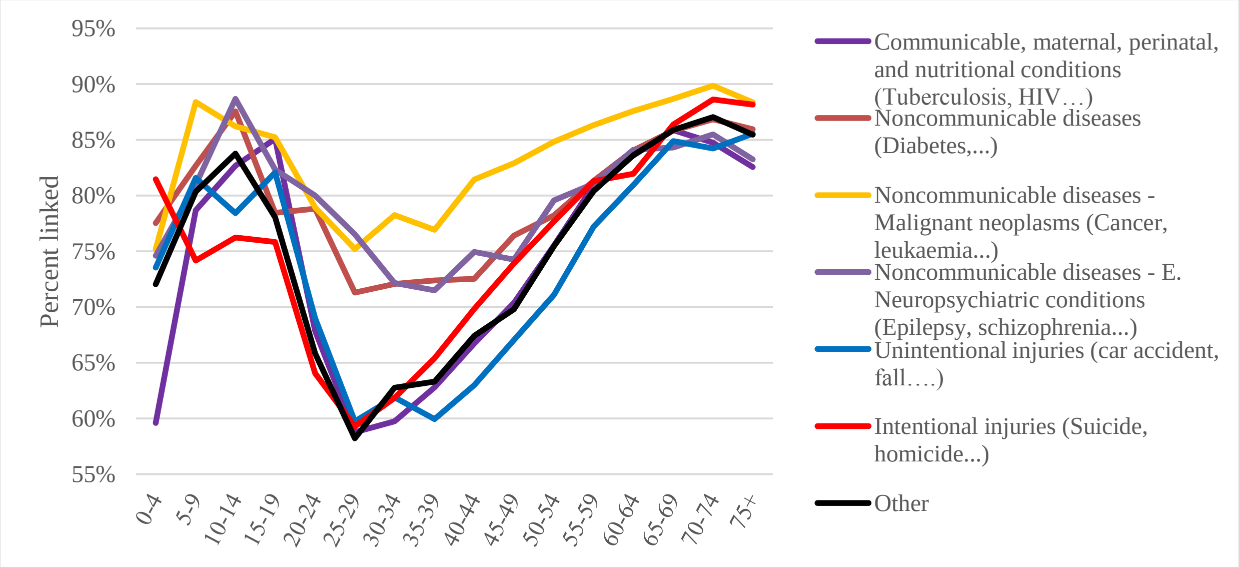


Linkage rates were lower among those who lived in low income neighbourhoods, based on postal code information in the CMDB, than those in high income neighbourhoods.

Also, as anticipated, linkage rates varied by marital status with married people having the highest (93%) linkage rates and single people having the lowest (82%).

Linkage rates were also examined by cause of death: those who died as a result of intentional injuries were least likely to link (78%) and those who died due to cancers were more likely to link (91%; Fig. 4.1-2). The linkage variations by age group, described above, when examined by cause of death, revealed that 20-to-35 year olds who died as result of intentional or unintentional injuries had the lowest linkage results. As expected, these are the causes of death that are more likely among this age group.

Fig. 4.1-2. Linkage rates by age group and cause of death

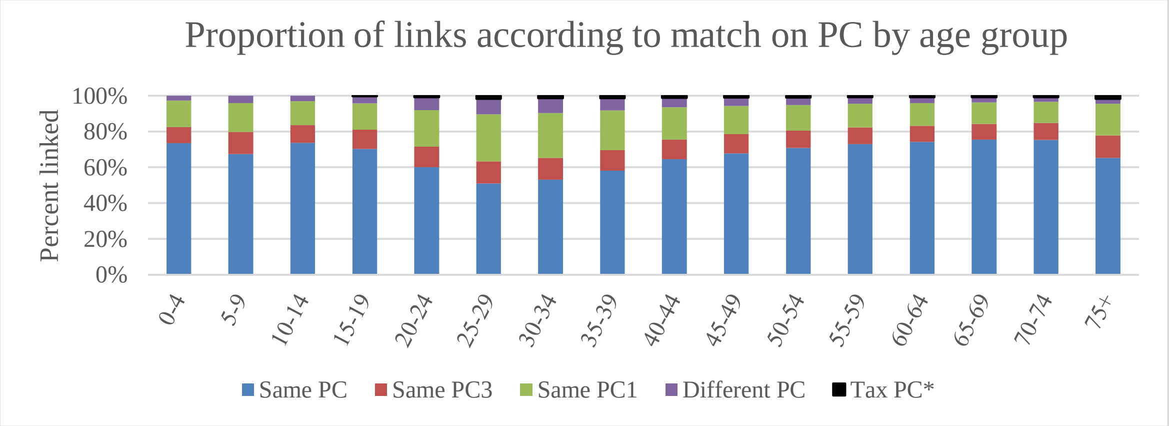


Among the provinces and territories, the territories had the lowest linkage rates – in the low 80%s. Linkage rate was highest in New Brunswick: 91%. The lower linkage rates were found in provinces with higher proportion of young adults.

People who were born outside of Canada were less likely to link with linkage rates being particularly poor among those under 35 years of age (61% or lower). Some of these individuals may have immigrated to Canada after Census day or may have been visitors and temporary workers, thus, potentially explaining the poor linkage.

Much of the weaker linkage rates in the 20-39 year olds can potentially be explained by higher mobility of individuals in this age group. To explore this further, linkage rates were examined by waves where stringency of matching criteria for postal code varied. These waves included those where records had to match exactly on the full postal code ("Same PC" in Fig. 3), match exactly on the first three characters ("Same PC3"), match exactly on only the first character ("Same PC1") and where the postal code was allowed to differ in the two records (Fig. 4.1-3).
For those in the age groups 25-29, 30-34, 30-34 years, a smaller per cent linked in those waves where they had to match exactly on the entire postal code suggesting that mobility played a role in the poor linkage rates for these age groups.

Fig. 4.1-3. Linkage rates by waves of varying stringencies of the postal code variable



Other validation exercises and clerical reviews are in progress, and the findings will be used to assess suitability of using the dataset for analytical purposes.

5. Analyses of linked data

5.1 Analysis aims

The linkage of approximately 90% of CMDB records to the Census records enables the identification of self-reported First Nations, Métis and Inuit in the mortality records, which in turn, allows the estimation of mortality rates in these populations and the non-Aboriginal population. It also offers the potential for several other analyses in these and other populations of interest.

The specific aims of the analysis are: (1) generate population-based estimates of mortality rates for First Nations including Registered Indians, Inuit, Métis and non-Aboriginal people from all causes combined and from specific causes, including cardiovascular disease, cancer, respiratory diseases and suicide, (2) estimate life expectancies at birth, age 21 and 65 for the populations outlined above, and (3) assess the impact of socioeconomic and demographic characteristics on the risk of death from all and different causes for First Nations, Métis and Inuit compared to non-Aboriginal people.

Other potential projects include: (1) estimation of life expectancies for First Nations, Métis, Inuit, and non-Aboriginal populations, (2) examination of avoidable mortality among those who prematurely died (before age 75), (3) estimation of potential years of life lost for those who prematurely died, (4) validation of results of mortality among Registered Indians using Indian Register information, (5) examination of role of socio-economic/demographic factors on rates of death or disparities between mortality rates of many populations of interest and non-Aboriginal population, (6) development of similar analyses for immigrants and visible minorities for reports, fact sheets and data tables, and (7) extension of these analyses with additional years of CMDB over time.

References

Rotermann M, Sanmartin C, Carrière G, Trudeau R, St-Jean H, Saïdi A, et al (2014), "Two approaches to linking census and hospital data", Health Reports, 25(10), pp. 3–14.

Statistics Canada (2010), "Table 1.2.2 Estimated population net undercoverage and standard errors for various characteristics, 2006 Census", available from: http://www12.statcan.gc.ca/census-recensement/2006/ref/rp-guides/rp/coverage-couverture/tbl/tbl-01_2_2-eng.cfm

Statistics Canada (2015), "Vital Statistics – Death Database", available from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3233

Wilkins R, Tjepkema M, Mustard C, Choinière R. (2008), "The Canadian census mortality follow-up study, 1991 through 2001", Health Reports,19(3), pp. 25–43.