

## Analysis of Immigration Data: 1980-1994

*Adam Probert, Robert Semenciw, and Yang Mao, Health Canada  
Jane F. Gentleman, Statistics Canada*

---

### **Abstract**

*This paper describes the record linkages being carried out at Statistics Canada to link data for all immigrations to Canada from 1980 through 1994 to income tax files (for live follow-up) and then to data for all deaths in Canada. The files involved are very large. As an example, the number of immigrants in 1980 was 143,432, and the number in 1994 was 222,538.*

*Numerous studies of immigrants have been published around the world, but the vast majority of them are missing information on the entry date to the country. Our immigration data will not have this limitation. They will be used to follow up immigrants to see how living in Canada has impacted their health and how this is affected by the length of time they have lived in Canada. This project will study how cause-specific mortality varies with country of origin and length of residence in Canada, to aid in disease control and prevention. The study of disease patterns in persons from different geographical areas is an epidemiological technique that can provide important clues to the causes of disease. Such studies can show the potential for preventive actions if a risk pattern from one population can be transposed to another.*

### Introduction

Record linkages are presently being carried out at Statistics Canada to link data for all immigrants to Canada from 1980 through 1994 to income tax data (for live follow-up) and then (for the earliest years) to mortality data. This paper is a description of the data bases and the rationale for the project.

This project will study how cause-specific mortality varies with country of origin and length of residence in Canada, to aid in disease control and prevention. The study of disease patterns in persons from different geographical areas is an epidemiological technique that can provide important clues to the causes of disease. Such studies can show the potential for preventive actions if a risk pattern from one population can be transferred to another. One of the earliest of these studies involved Japanese migrants to Hawaii. From the differences in stomach cancer rates among the immigrant and native populations the researchers were able to implicate diet as a risk factor for stomach cancer.

### Immigration Data

Numerous studies of immigrants have been published around the world, but the vast majority of them are missing information on the entry date to the country. Without this information, the amount of exposure to life in the new country is unknown, so an “exposure-response” relationship cannot be studied. Our immigration data, with the landing date, will not have this problem.

Immigrants comprise a large proportion of the Canadian population. For example, the number of immigrants in 1980 was 143,432, and the number in 1994 was 222,538. According to the 1991 Census,

there were approximately four million immigrants in Canada, or 16% of the population. The health status of such a large segment of the population should be investigated.

The Immigration Data Base has existed in machine-readable form since 1980. It contains information on every landed immigrant to Canada, as of the actual date of landing. It contains data on education level, intended occupation, medical class (a summary variable providing a baseline medical status), language ability and, of course, name, sex and date of birth. It also contains a couple of unique identifiers: visa number, which is unique for every landed immigrant, and family identification number, which is given to all members of a family who immigrate on the same date. The database, for the most part, is complete. The least complete variable for 1980 immigrants is date of birth, which is missing in approximately 1,000 out of 143,476 records (0.7%).

The present study will use probabilistic record linkage to the Canadian Mortality Data Base (CMDB), maintained at Statistics Canada, to link to almost three million immigrant records. If this linkage proves successful, then future linkages to the cancer incidence and tuberculosis data bases will be considered. The linked data will be used to follow up immigrants to see how living in Canada has impacted their health and how this is affected by the length of time they have lived in Canada.

Preliminary analysis will be performed on the immigration data before the linkage to the mortality data. Trends in immigration over the 15-year period will be examined. Specifically, the number of immigrants by country of birth, age, sex, education, medical class and intended occupation will be described over the 15 years. The analyses to be performed on the linked data involve Poisson or logistic regression models of outcome (mortality, cancer or tuberculosis) and exposure variables (length of time in Canada, age, country of birth, medical class at arrival, etc.).

## Data Limitations

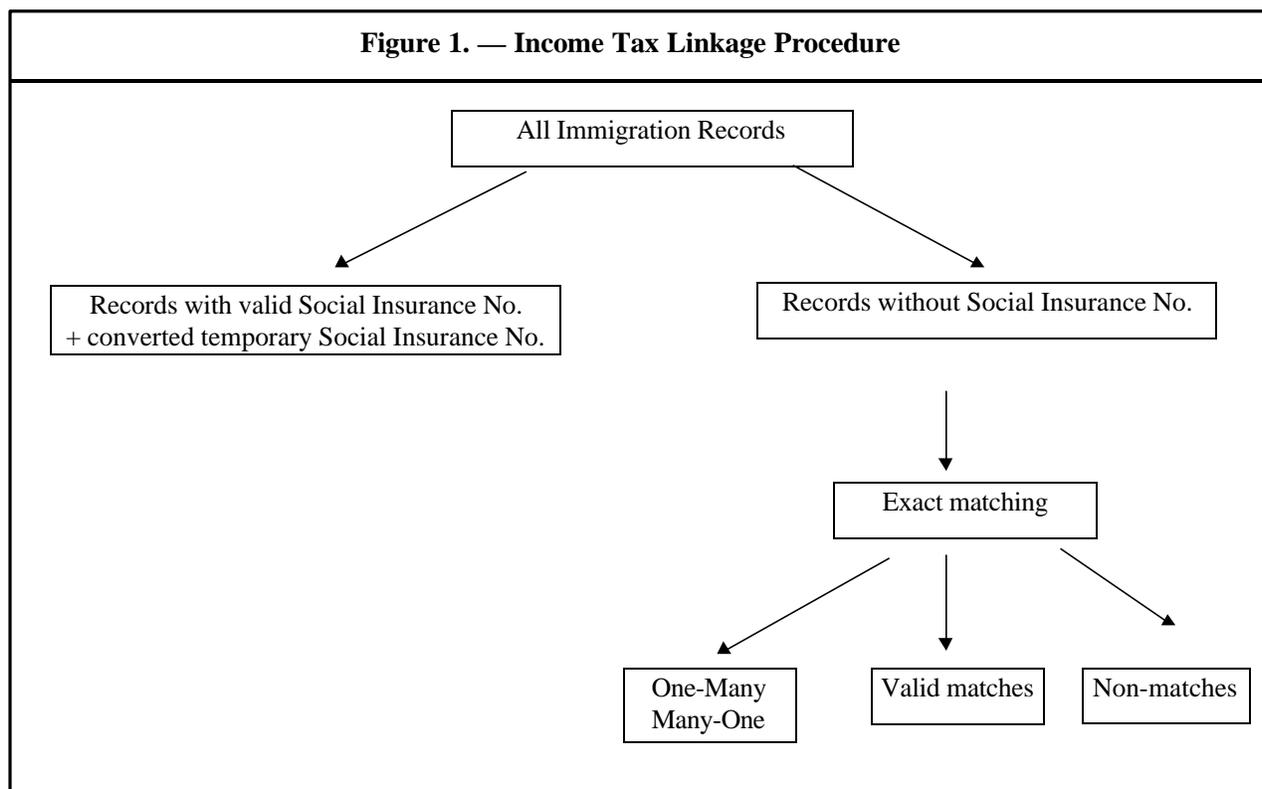
Three challenges have to be dealt with in analyzing these data. Studies of immigrants must deal with what is termed re-migration, i.e., immigration followed by emigration. Of all the landed immigrants to Canada, about 30% will later emigrate, most likely to the United States or return to their home country. The second challenge pertains to immigrants who will not link to the mortality database, that is, who have not died in Canada. If they do not link to the mortality database then it is not known whether they are alive in Canada or deceased in another country. Based on the initial data, one would not be able to estimate the time spent in Canada if there was no record of death. A third issue concerns female immigrants. For those who have changed their name through a change in marital status, it may prove to be extremely difficult to find a match to the mortality database. Whether or not data for both sexes can be analyzed, the 1980 immigration data are expected to yield the most useful results, as the follow-up period during which mortality could occur is longest for this group.

## Record Linkage

To address these problems a record linkage to the income tax files was suggested, not necessarily to obtain tax information. The main purpose of linkage to tax files is live follow-up, i.e., accumulation of evidence that the immigrant remained in Canada. This is critical information because a significant number of immigrants leave Canada subsequent to immigration, as mentioned previously. An extremely useful by-product of this linkage will be the date of death for those immigrants who died since 1980; this will facilitate the second stage of linkage -- to mortality data.

As the first stage of linkage, the immigration data for each year have been linked to all income tax files for 1980-1994 (including tax files for years before immigration to Canada, because it is possible to file before immigrating). Of all the 3 million immigration records, only half had a valid Social Insurance Number

(SIN). This included those who had a temporary SIN, which was later converted to a permanent number. The SIN is the Canadian counterpart of the American Social Security Number. For each immigrant, exact matching was used to find that person on the tax file for any year, based on surname, first four characters of given name, date of birth (year, month, day), and sex. Once the immigrant was found on any tax file for any year, the SIN was known and could be used to find the same person on tax files for other years. As part of the regular income tax form, immigration date, emigration date and date of death all appear in addition to the regular tax information, if applicable. See Figure 1 for a diagram of the tax linkage procedure.



At the stage of exact matching, duplicates are created. For those records where there is a prefix to the surname (e.g., De La or Von), there is a duplicate record created (and flagged) for the surname without the prefix. From this procedure there were approximately 1,000 records that yielded many-to-one or one-to-many linkages; these were ignored for this linkage. All records with a SIN are then linked using that SIN to all the income tax files. It is at this step where the linear record of a landed immigrant's stay in Canada will be found. Regardless of the outcome of the income tax linkage procedure, all immigration records will be incorporated in the mortality linkage.

## Results

Table 1 contains some of the results of the tax linkage. Examining the 1980 data, one can see that there were 143,432 landed immigrants, of which 44,486 linked to the 1980 tax files and 67,782 linked to the 1994 files. Note that the tax files mostly contain data for filers aged 15-65 (about 100,000 out of 143,432 1980 immigrants). In other words, about 68% of those who landed in 1980 filed a tax return in 1994. Also, 152 people who filed a tax return in 1980 did not become landed immigrants until 1994. It is possible, for example, for people present in Canada on business or student visas to pay tax before becoming landed immigrants.

Landing Year	Number of Immigrants	Found 1980 Tax Form	Found 1994 Tax Form
1980	143,432	44,486	67,782
1981	128,735	4,648	62,956
1982	121,253	2,776	60,771
...			
1993	255,087	222	118,795
1994	222,538	152	86,943

## Next Steps

Initially, only the 1980 immigration-tax data (1980 immigrant files linked to 1980-1994 tax files) will be linked to the CMDB using the commercially available Automatch linkage software. The CMDB is a record of all deaths since 1950. The database is mostly complete, with coverage varying between 98% and 100% for most variables. The completeness differs over time and among provinces. Linkage to the CMDB will be done using probabilistic methods. The variables Surname, Given name(s), Date of Birth, Sex and Other Name will be used for the linkage. Marital status and Country of Birth may also be used depending on the success of the previous pass. All of the names will be converted to NYSIIS format to aid in the name-matching process. With foreign names there may be more spelling/typographical errors that NYSIIS coding can alleviate. Reversing of the name and birthdate fields will be allowed to control for those errors where the first and last name or day of birth and month are switched.

As mentioned previously, linkage problems may be encountered for females because of name changes subsequent to immigration and because of the decreased propensity of females to file tax forms. To increase the chances of finding females on the mortality database, all surnames of females (maiden names and married names) found on the income tax records will be captured and used in the death linkage.

Analyzing the output from the linkage will involve many steps. First, we will examine the risk of disease in immigrants compared to their country of birth, controlling for age and sex and examining trends over time. Second, we will examine the Canadian rates of disease for Canadian-born persons. Third, unique to analyses of data of this nature, risk of death by disease and by duration of residence in Canada as well as age at migration will be analyzed. This analysis will also examine the differences by country of birth, occupation, education and other factors.

## Conclusion

To summarize, the immigration database offers the opportunity for new research into immigrant health. By linking to Canadian income tax records, we could know when a landed immigrant is no longer a resident of Canada, something that is not available in most immigrant studies. We should also be able to account for some name changes that occur in female immigrants. From the linkage to the mortality database, we will be able to examine the risk of death by country of birth, length of stay, age, sex, education and other demographic variables. All of these analyses will aid in identifying trends and the etiology of specific diseases.