Catalogue no. 11-633-X — No. 007
ISSN 2371-3429
ISBN 978-0-660-08653-8

Analytical Studies: Methods and References

Longitudinal Immigration Database (IMDB) Technical Report, 2014

by Rose Evra and Elena Prokopenko

Release date: June 16, 2017

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2017

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Longitudinal Immigration Database (IMDB) Technical Report, 2014

by

Rose Evra and **Elena Prokopenko**

Social and Aboriginal Statistics Division

11-633-X — No. 007

ISSN 2371-3429

ISBN 978-0-660-08653-8

June 2017

Analytical Studies: Methods and References

Papers in this series provide background discussions of the methods used to develop data for economic, health, and social analytical studies at Statistics Canada. They are intended to provide readers with information on the statistical methods, standards and definitions used to develop databases for research purposes. All papers in this series have undergone peer and institutional review to ensure that they conform to Statistics Canada's mandate and adhere to generally accepted standards of good professional practice.

The papers can be downloaded for free at www.statcan.gc.ca.

Table of contents

Acknowledgements

We would like to mention the special contribution of the following people: **Laetitia Martin** of the Social and Aboriginal Statistics Division (SASD), who wrote sections 3.3.1 to 3.3.3 of this report; **Alexandr Diaz-Papkovich** of the Social Survey Method Division (SSMD), who validated the sections of this report relating to the methodology behind the production of the IMDB; and **Tristan Cayn**, **Derek Cho**, **Scott McLeish**, **Ian Marrs** and **Trevor Smith**, members of the Administrative Data team, who produce the IMDB and contributed to the content of several sections of the report.

Many thanks to the following people for reviewing the report prior to its publication: Margareta Dovgal, Benoît St-Jean, Winnie Chan and Hélène Maheux (Statistics Canada); Yoko Yoshida (Department of Sociology and Social Anthropology, Dalhousie University); Michael Haan (Canada Research Chair in Migration and Ethnic Studies and the Department of Sociology of Western University); and Ian Clara (Manitoba Research Data Centre).

Abstract

The Longitudinal Immigration Database (IMDB) is a comprehensive source of data that plays a key role in the understanding of the economic behaviour of immigrants. It is the only annual Canadian dataset that allows users to study the characteristics of immigrants to Canada at the time of admission and their economic outcomes and regional (inter-provincial) mobility over a time span of more than 30 years.

The IMDB combines administrative files on immigrant admissions and non-permanent resident permits from Immigration, Refugees and Citizenship Canada (IRCC) with tax files from the Canadian Revenue Agency (CRA). Information is available for immigrant taxfilers admitted since 1980. Tax records for 1982 and subsequent years are available for immigrant taxfilers.

This report will discuss the IMDB data sources, concepts and variables, record linkage, data processing, dissemination, data evaluation and quality indicators, comparability with other immigration datasets, and the analyses possible with the IMDB.

Key words: Administrative Data, Immigration, IMDB, longitudinal data, non-permanent residents, taxfilers

Glossary of terms

Following are the description of acronyms that will be used several times in the report.

Acronym	Definition
AMDB	Amalgamated Mortality Database
CPI	Consumer Price Index
CRA	Canada Revenue Agency
ILF	Immigrant Landing File
IMDB	Longitudinal Immigration Database
IRCC	Immigration, Refugees and Citizenship Canada
LAD	Longitudinal Administrative Databank
LCF	Linkage Control File
LSIC	Longitudinal Survey of Immigrants to Canada
NHS	National Household Survey
NPR	Non-permanent resident
NRF	Non-permanent Resident File
NRF, Permits	Non-permanent resident, Permit File
PNRF	Integrated Permanent and Non-permanent Resident File
PR	Permanent resident
T1FF	T1 Family File

1 Introduction

The Longitudinal Immigration Database (IMDB) is a comprehensive source of data that plays a key role in the understanding of the economic behaviour of immigrants. It is the only annual Canadian dataset that allows users to study the characteristics of immigrants to Canada at the time of admission and their economic outcomes and regional (inter-provincial) mobility over a time span of more than 30 years.

The IMDB combines administrative files on immigrant admissions and non-permanent resident permits from Immigration, Refugees and Citizenship Canada[1] (IRCC) with tax files from the Canadian Revenue Agency (CRA). Information is available for immigrant taxfilers admitted since 1980. Tax records for 1982 and subsequent years are available for immigrant taxfilers.

The IMDB was designed to provide detailed and reliable data on the performance and impact of immigration programs. Being a database of immigrant taxfilers, the IMDB can be used to answer both broad and very specific research questions. The database also provides information on pre-admission experience, such as work or study permits. Its major strength is that it allows for the analysis of socio-economic outcomes over a period long enough to assess the impact of immigrant characteristics upon admission, including admission category, education, and knowledge of French or English, on outcomes. Moreover, annual information on place of residence allows for the investigation of secondary migration (immigrants' subsequent relocation in Canada).

As created, the IMDB includes multiple files: one file for each tax year since 1982, two files containing immigration characteristics at the person level, and one permit file for non-permanent residents. The IMDB is updated annually via record linkage techniques described in this report. Each year an additional tax year is added, and new admission and non-permanent resident permit data are added to the database.

The IMDB files are available only to Statistics Canada researchers and deemed employees. This is to ensure that proper confidentiality measures are taken to protect privacy and ensure confidentiality. Information from the IMDB is available to the public through annual aggregated summary tables produced by Statistics Canada and published on its website. Additionally, external researchers may request ad hoc tables and analyses; Statistics Canada provides these services on a cost-recovery basis. Users who want information pertaining only to the Immigrant Landing File should contact IRCC, as the IMDB is designed for analysis on immigrant taxfilers only.

This report will discuss the IMDB data sources (Section 2), concepts and variables (Section 3), record linkage (Section 4), data processing (Section 5), dissemination (Section 6), data evaluation and quality indicators (Section 7), comparability with other immigration datasets (Section 8), and the analyses possible with the IMDB (Section 9).

1. Formerly Citizenship and Immigration Canada (CIC).

2 Data sources

Several files are required in order to produce the IMDB. These files, which will be described in this section, consist of immigration data, immigrant tax files and auxiliary files covering many years.

2.1 Immigration data

Every year, Statistics Canada (StatCan) receives admission data on new recipients of permanent residency permits and non-permanent residency permits from IRCC.

2.1.1 Immigrant Landing File (ILF)

Every year, landing data is added to create the Immigrant Landing File (ILF). This file contains information such as landing date, date of birth, and immigration class. The ILF could be seen as a census of the people who have immigrated to Canada as permanent residents since 1980; it holds information on their characteristics at landing. This file, however, is not directly available to IMDB users. The IMDB covers only immigrants who have filed taxes at least once since 1982. Landing data for these immigrant taxfilers is available in the Integrated Permanent and Non-permanent Resident File (PNRF).

Because it is an administrative record of permanent residency, the ILF overestimates the number of immigrants currently living in Canada. This overestimation occurs for two reasons. First, the ILF does not identify the individuals who have left the country. Immigrants who landed in Canada may have left Canada since landing. Second, the death of immigrants who landed in 1980 and thereafter is only partially reported. Further information on mortality data can be found in Section 7.2.

2.1.2 Non-permanent Resident File (NRF)

Data on immigrants with pre-landing experience who landed in 1980 and thereafter and who hold non-permanent resident status have been used to create the Non-permanent Resident File (NRF). The file includes the number of permits, the type of permit (work or study, for example), and the first year of temporary residence. It is updated each year with new annual non-permanent permits data.

A given person can have multiple permits, since the same person can be issued many permits at different times. The NRF stores such information at the person level. In addition to the NRF, a permit file (Section 2.1.4) has been created for information at the permit level. The NRF is not available to data users; however, pertinent variables from this file are available in the Integrated Permanent and Non-permanent Resident File (PNRF) (see Section 2.1.3).

2.1.3 Integrated Permanent and Non-permanent Resident File (PNRF)

Researchers can access the Integrated Permanent and Non-permanent Resident File (PNRF), which combines information from the ILF and the NRF at the person level. The PNRF provides users with the ability to follow the migration history of immigrants, including their pre-landing experience in Canada. The PNRF contains detailed data on the sociodemographic characteristics of immigrants who landed in Canada in 1980 or thereafter, making it possible, for example, to determine whether a person was a non-permanent resident prior to landing. It is to be noted that records of non-permanent residents who have not become permanent residents are not included in this file. This file contains the number of permits for each non-permanent resident who became a permanent resident, and includes landing dates. The PNRF also includes a date of death when a link to a death record has been made (see Section 7.2.2). For more details on the content of this file, please refer to the immigration component of the IMDB dictionary, in sections 3.3 and 3.4 of this report.

In addition to the PNRF, a file named **PNRF_EXTRA** is available to data users; it includes variables that have been retired, have little analytical value, or for which no metadata are available. The complete list of variables can be found on the IMDB immigration data dictionary.

2.1.4 Non-permanent Resident File, Permits (NRF, Permits)

To complement the PNRF, a file containing permit-level information has been created for non-permanent residents who have obtained a permit since 1980. A given person can have multiple entries on this file, that is, one per permit issued. Dates related to each permit, such as the effective date and the valid-date range, are included in this file. Data of non-permanent residents who became permanent residents are available to IMDB users. They can be linked to the PNRF by means of the IMDB unique person identifier (IMDB_ID). For more details on the variables included in this file, please refer to the immigration component of the IMDB dictionary.

2.2 T1 Family File (T1FF)

The tax files used to create the **IMDB_T1FF** files are those contained in the T1 Family File[2] (T1FF). Statistics Canada takes the annual individual T1 file, T4 tax file and Canada Child Tax Benefit (CCTB)[3] file from the CRA and creates the T1 Family File for that year. Processing consists of many steps, ranging from geographical coding to the formation of families (for example, when the taxfiler mentions a spouse and this spouse is also a taxfiler, the spouse is linked via a common identifier to the original taxfiler). T1FF data go back to the 1982 tax year. With the experience gained from many years of T1FF processing, edit rules have been created to reduce the number of inconsistencies in the database and ensure that data quality continues to improve.

The availability of the tax variables depends on the information collected in a given year. The T1FF produced annually for the IMDB includes individual and family incomes as well as family composition variables, such as the number of kids and the spouse identification number. The IMDB contains IMDB_T1FFs for 1982 and subsequent years for immigrant taxfilers. The creation process of these files is described in Section 5.1. For more details on variables available on the IMDB_T1FFs, refer to the tax component of the IMDB dictionary.

2.3 Auxiliary files

To create the IMDB, it is necessary to use auxiliary files that facilitate record linkage and add variables to the database. These auxiliary files are not available to IMDB users.

The **Linkage Control File (LCF)** links immigration data to tax data, enabling the production of the IMDB_T1FF files. The LCF is a database of personal identification numbers containing information on taxfilers for 1981 and subsequent years. The LCF is derived from, among other things, information provided on T1 forms and Canadian Child Tax Benefit (CCTB) forms.

Variables related to mortality are also available in the IMDB. These variables are produced from information contained in Statistics Canada's **Amalgamated Mortality Database (AMDB)**. The AMDB combines records from vital statistics and tax files to produce a mortality dataset.

2. For details on the most recent T1FF, the reader may consult the Statistics Canada website.
3. On July 1, 2016, the Canadian Child Tax Benefit was replaced by the Canada Child Benefit.

3 Concepts and variables

3.1 Immigrant status in Canada

The IMDB provides data on a subset of the immigrant population as described in Section 2. Following are the Statistics Canada definitions of the terms "immigrant" and "non-permanent resident."

The term "**immigrant**" refers to persons who are, or who have been at any time, landed immigrants or permanent residents. Such persons have been granted the right to live in Canada permanently by immigration authorities. Immigrants who have obtained Canadian citizenship by naturalization are included in this category.

"**Non-permanent residents**" are not considered immigrants, although they are a population of interest for the IMDB as described in Section 2. The term "**non-permanent resident**" refers to persons from another country who have a work or study permit or who are refugee claimants, and who have family members sharing the same permit and living in Canada with them.[4] They are allowed to be in Canada for the period of time indicated on their permit.

3.1.1 Immigration to Canada: An overview

A Canadian Megatrends article, *150 years of immigration in Canada*,[5] released in 2016, summarizes the fluctuation in immigration levels and source countries over the past century. Migration to Canada has been continuous since the country's foundation. More than 17 million immigrants have settled in Canada since 1867. The number of landed immigrants has been increasing from the low 200,000s in the 1990s to over 250,000 in the early 2010s. The proportion of Canadians who are foreign-born has increased from 14.7% in 1951 (2.06 million people) to 20.6% in 2011 (6.78 million people).

As per section 95 of the *Constitution Act, 1867* federal and provincial governments have shared jurisdiction over immigration. Additional guidelines are set out in the 2002 *Immigration and Refugee Protection Act* (IRPA),[6] which provides the goals and strategic direction for immigration policy adopted by the Government of Canada and administered in part by Immigration, Refugees and Citizenship Canada (IRCC). Prior to 2002, the *Immigration Act, 1976* served as the primary legislation regulating Canadian immigration.

Under IRPA, the Government of Canada is in charge of "establishing admission requirements, setting national immigration levels, defining immigration categories, determining refugee claims within Canada, reuniting families and establishing eligibility criteria for settlement programs"[6] in all provinces and territories except Quebec. The province of Quebec has full responsibility of its immigration levels, programs, and policies under the *Canada-Quebec Accord Relating to Immigration and Temporary Admission of Aliens*. However, the federal government continues to select and process immigrants sponsored by family and protected persons in Canada and refugee claimants to Quebec.[6]

Permanent residents are defined as "persons who have been admitted to live in Canada on a permanent basis and who have the right to work and study in Canada, but have not become Canadian citizens."[4] Under IRPA, there are three overarching classes of immigrants: economic immigrants, family members, and refugees.

Permanent residents are eligible to become citizens of Canada when they meet certain requirements. The first is a residency requirement, whereby the permanent resident must have been physically in Canada for a set period of time. Permanent residents must also be older than 18 years of age; in the case of minors, the application must be made simultaneously (concurrent) with one or both parents or after one or both parents have become a Canadian citizen (non-concurrent). Permanent residents must have fulfilled their tax filing obligations to Canada. Permanent residents aged 14 to 64 years must also show proof of proficiency in at least one of Canada's official languages and must pass a citizenship test (IRCC 2016; Government of Canada 2016).

The IRPA stipulates that all foreign nationals, except permanent residents, who enter Canada must have a temporary resident visa. Temporary resident visas are issued to workers and students "in a way that maximizes their contribution to Canada's economic, social and cultural development and protects the health, safety and

4. Definitions approved as a recommended Statistics Canada standard on March 21, 2016.http://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=139756&CVD=139756&CLV=0&MLV=1&D=1
5. See Statistics Canada 2016 in list of references.
6. *Immigration and Refugee Protection Act* (S.C. 2001, c. 27) http://laws-lois.justice.gc.ca/eng/acts/I-2.5/index.html

security of Canadians" (IRCC 2015, p. 7). Non-permanent residents are able to apply for permanent residency through different programs, and may have an advantage over applicants abroad if they have Canadian education credentials and / or work experience.

As regards **refugee claimants**, "the Refugee Protection Division (RPD) is the division of the Immigration and Refugee Board of Canada (IRB) that hears claims for refugee protection made in Canada and decides whether to accept them" (IRB 2015 p. 16). In the IMDB, these claimants are classified as non-permanent residents with a refugee claimant permit.

3.2 Target population and coverage period

The IMDB is a database of immigrants admitted to Canada since 1980 for whom at least one T1FF record is available for any year since 1982. The database also provides information on permits for immigrants who were non-permanent residents prior to their admission as permanent residents.

The IMDB brings together, via hierarchical deterministic record linkage (Section 4), administrative data from Immigration, Refugees and Citizenship Canada (IRCC) and the tax files from the T1 Family File (T1FF). Immigration data included in the IMDB comes from the Immigrant Landing File (ILF), which contains the records of immigrants who landed in Canada in 1980 or thereafter, and data from the Non-permanent Resident File (NRF), which contains records of non-permanent residents who obtained a temporary resident permit in 1980 or a subsequent year.

3.3 Admission variables

Immigrants are admitted into Canada under a number of programs, each of which has specific objectives. These programs specify the conditions under which immigrants are admitted into the country and the type of settlement assistance they may receive. Consequently, analyses that guide public policy should usually take this information into consideration. To answer a variety of research questions, the IMDB comprises a number of variables related to the admission of immigrants, which are all derived from two main concepts: admission category and type of applicant.

3.3.1 Admission category

The **admission category** refers to the name of the immigration program or group of programs under which an immigrant was first granted the right to live in Canada permanently by immigration officials since 1980. Over the years, immigrants have been admitted into the country under several dozen different programs. In an effort to make these data easier to use, the IMDB provides users with a number of variables that comprise aggregate programs with similar objectives. The highest level of aggregation is based on the three main objectives of Canada's immigration policy: contribute to the country's economic development, reunite families, and protect refugees.

3.3.1.1 Economic immigrants

The purpose of admitting economic immigrants is to help achieve the first immigration policy objective stated above: contribute to the Canadian economy. Economic immigrants are covered under three main program groups: worker programs; business programs; and provincial and territorial nominee programs.

Immigrants selected for their ability to participate in the labour market are admitted under **worker programs**. Once their skills and professional experience have been evaluated, they are divided into four main categories:

1. Skilled workers selected based on their skills and experience working in management or professional positions, in technical jobs, or in skilled trades.
2. Skilled tradespeople selected based specifically on their skills and work experience in an eligible skilled trade. This category differs from the skilled workers category as applicants are required to have a valid offer of employment from a Canadian employer or a certificate of qualification from a Canadian provincial or territorial organization.

3. Immigrants admitted under the Canadian Experience Class differ from the two first groups in that they are required to have work experience in Canada acquired in a managerial or professional position, a technical job, or a skilled trade.
4. Live-in caregivers and caregivers can obtain permanent resident status if they have provided in-home care in Canada for a given period to children or people with special needs such as the elderly, people with a physical handicap, or someone suffering from a chronic illness.

Economic immigrants admitted into Canada under a **business program** are divided into three main categories:

1. Entrepreneurs selected for their skills and their ability to either own and manage a business, or to establish an eligible business in Canada. Some have a minimum net worth, while others are required to have the backing of a designated organization for their business idea.
2. Investors given permanent resident status provided they make a significant investment in Canada. These investments are allocated to participating provinces and territories in order to stimulate economic development and create jobs.
3. Self-employed workers who are given permanent resident status provided they have the ability—and the intention—to create their own job in Canada and to make a significant contribution to the Canadian economy. This is a broad category that also includes people who intend to make an important contribution to the country's sporting or cultural landscape (i.e., as an artist, actor, writer, or professional athlete).

The final main category under which economic immigrants are admitted into Canada are **provincial and territorial nominee programs**. As the name implies, this category is for immigrants selected by a province or a territory for their ability to contribute to the local economy by meeting specific labour needs. They are assessed based on selection criteria relating to education, work experience, and their specific skills. All participating provinces and territories have their own selection criteria for their fields of interest (students, business people, skilled workers, or semi-skilled workers).

3.3.1.2 Family sponsorship

The admission of immigrants sponsored by family members is intended to reunite families; this allows Canadian citizens and permanent residents to sponsor their relatives. Immigrants admitted under these programs can be given permanent resident status on account of their relationship as spouse, partner, parent, grandparent, or child.

Under certain conditions, immigrants admitted under these programs can also be sponsored by reason of another family relationships, such as young siblings, nieces and nephews, and orphaned grandchildren. Canadian citizens and permanent residents living in Canada can also sponsor someone on the basis of a relationship other than the ones listed above.

Finally, there are cases in which immigrants who would not otherwise have qualified under any other program were sponsored by a Canadian citizen or a permanent resident living in Canada, and who were exceptionally granted permanent resident status on humanitarian grounds.

3.3.1.3 Refugees

The third and final objective of Canada's immigration policy is the protection of refugees, or people who have a well-founded fear of returning to their country of origin. This category includes people who have good reason to fear persecution based on race, religion, nationality, membership in a particular social group, or political opinions (refugees as defined by the Geneva Convention). It also includes people who have been seriously and personally affected by civil war, armed conflict or a massive violation of human rights. Some refugees were already in Canada when they applied for refugee status for themselves and for family members who were with them in Canada or abroad. Others were abroad and were referred for resettlement to Canada by an office of the United Nations High Commissioner for Refugees (HCR) or another referral organization. Referred immigrants receive resettlement support from government sources, organizations, individuals, or private sector groups, or combined support from the Government of Canada and private sector stakeholders.

3.3.1.4 Other immigrants

In addition to the three key objectives listed above, Canada's immigration policy gives immigration officials a certain degree of discretionary to grant permanent resident status under a program for people who are neither economic immigrants, sponsored by a family member nor refugees. This program is for applicants such as immigrants who are exceptionally granted permanent resident status on humanitarian grounds or on the basis of public interest considerations.

3.3.2 Type of applicant

In addition to the admission category, the IMDB gives users access to information on applicant types. This information indicates whether the immigrant is listed as principal applicant, spouse or dependent on the application for a permanent resident visa.

As a general rule, information on the type of applicant is used for analyses to study economic immigrants. Since the principal applicants admitted under these programs are selected on the basis of their ability to contribute to the Canadian economy, it is helpful to separate them from their spouse and dependents, who were not assessed for this ability. Isolating principal applicants from other types of applicants makes it possible to study the efficiency of these programs more directly.

However, with regard to family reunification and refugee protection, the purpose of the immigration policy is the same for all applicants, regardless of type. In the case of immigrants admitted under these two objectives, the concept of "applicant type" takes on more of an administrative value.

This value is particularly pronounced for immigrants with principal applicant status, which does not systematically depend on the legal relationship between the applicants requesting permanent residence. For instance, for the "sponsored spouses and partners" admission category, the spouse is listed as the principal applicant, although "spouse" does not appear as the type of applicant on the application for residence. In addition, for the "sponsored children" admission category, principal applicant status is assigned to one of the children, while the others are listed as dependents. Finally, in certain circumstances, applications for permanent residence can be processed on two fronts: from Canada for the principal applicant and from abroad for the other family members. This type of process exists for live-in caregivers and protected persons in Canada. In these cases, a family member applying from abroad is given principal applicant status, even if he or she is the spouse of an immigrant whose application submitted in Canada has been previously approved.

3.3.3 Changes over time

The IMDB contains over 30 years' worth of data on immigrants admitted to the country. However, policies and programs underwent many changes during that time. New programs were created, others were abolished, and in some cases, selection and eligibility criteria were changed. Therefore, a person admitted as a skilled worker in 1980 was not necessarily assessed on the same criteria as a skilled worker admitted in 2000. Although every effort was made to create aggregate programs that are as similar as possible, IMDB users should be aware of these differences when drawing conclusions about various landing cohorts.

The most striking change implemented during the period covered by the IMDB is undoubtedly the replacement of the *Immigration Act, 1976*, by the *Immigration and Refugee Protection Act*, which came into force in 2002. While both these laws cover the same three key groups (economic immigrants, family sponsorship, and refugees), administration of these programs program underwent many changes under these laws. In addition, program administration was also modified based on sociodemographic needs and priorities set by successive governments within these two legislative frameworks. As a result, it is strongly recommended that data users with an interest in a specific program or a number of landing cohorts find out more about policy and program changes relevant to their field of study.

It should be noted that it may take a few years for the impact of an administrative change to be observed in the database. For instance, when a new program is created, it may take several months or years from implementation (i.e., the date on which applicants can apply) to the time immigrants are first admitted into the country under the new program. The same can be said about abolished programs. There may well be a delay between the time all the applications have been studied and all eligible applicants have entered the country, and the time when abolished programs vanish completely from the statistics on annual admissions.

3.3.4 PNRF admission category variables

A variety of admission category variables exist in the PNRF. These are described in the immigration component of the IMDB dictionary. This section provides additional information on some of these variables.

The most detailed is **IMMIGRATION_CATEGORY**, which includes over 100 categories that existed at any point from 1980 to the present IMDB. Aggregated versions of the information available in the variable IMMIGRATION_CATEGORY are available in the derived variables IMM_CATEGORY_STC_ROLLUP1 and IMM_CATEGORY_STC_ROLLUP2, which contain fewer categories.

The aggregate variable **IMM_CATEGORY_STC_ROLLUP1** is a categorization in line with Immigration, Refugees, and Citizenship Canada's official publication *Facts and Figures*. However, it does not make clear that some immigrants were admitted through the **Backlog Clearance and Administrative Review programs**. These programs expedited the processing of immigrants in the late 1980s, in response to geopolitical crises abroad that affected temporary residents' ability to return to their countries (e.g., Tiananmen Square protests and dissolution of the USSR and Yugoslavia). The result of not separating these categories is that these individuals, processed quickly and with distinctive criteria, are not comparable to other immigrants processed in the same categories. To identify immigrants admitted through these programs, users should refer to the variables BACKLOG_CLEARANCE_IND and ADMINISTRATIVE_REVIEW_IND (available in the PNRF_extra, for landing years prior to 2014).

The user may also use the immigration aggregate information from the **IMM_CATEGORY_STC_ROLLUP2**. This variable was designed to provide consistent reporting across different policy / regulation changes (i.e., *Immigration and Refugee Protection Act* (2002) and *Immigration Act, 1976*) and to maintain specific immigration programs (i.e., skilled workers) over time. This variable offers the detailed information on backlog clearance and administrative review. Detailed grouping information for derived variables is available in the IMDB immigration data dictionary.

Another consideration with the admission category variables is their relation to **applicant type** (PNRF variable FAMILY_STATUS). As a general rule, the principal applicants are the individuals being assessed on admission criteria under each of the categories, while their accompanying spouse and dependents are admitted automatically with the principal applicant (although the spouse's language skills can be an asset to economic class immigrants' applications as well). In the rollup variable, some of the admission categories explicitly state whether they represent (1) principal applicants or (2) spouses and dependents, while other categories (i.e., Family Class) must be cross-referenced with the FAMILY_STATUS variable to determine an individual's status as a principal applicant or as a spouse / dependent.

Two categories constitute exceptions to the above: Live-in Caregiver Dependents and Refugee Dependents. When cross-referenced with FAMILY_STATUS_ROLLUP, these variables contain principal applicants as well as dependents. This can happen when the principal applicant is already in the country and his or her dependents submit a separate application for permanent residence from abroad. As each separate application must have a principal applicant, even a nominal one, one of the dependents (usually the spouse) is considered the "principal applicant" for the dependents' application. There is, however, no difference in processing between the principal applicants, spouses, and dependents in these two admission categories.

3.4 Variables of interest

The **IMDB** is an extensive database, providing researchers with a myriad of variables to study outcomes related to immigrant characteristics and various long-term impact. The number of variables exceeds 600 variables on the largest tax files (with roughly half at the individual level and half at the family level of aggregation). The Integrated Permanent and Non-permanent Resident File (PNRF) contains over 50 variables. While the exact definitions of these variables are covered in the immigration component of the IMDB dictionary, some of the more nuanced concepts warrant elaboration in this report. The following sections discuss geography, time, education and intended-occupation variables to provide further insight into the meaning and use of these variables. More detailed information on income variables can be found in the tax component of the IMDB dictionary.

Variables in the PNRF refer to immigrants' characteristics at landing or upon reception of a temporary resident document, while variables in the tax files refer to characteristics at taxation year. Whereas some variables are available in both files, the taxation variables are subject to changes over time. For example, age is available in

both files and is expected to change in the tax file each year. Immigrants' marital status (MARITAL_STATUS) and destination province (DESTINATION_PROVINCE) upon application for permanent residence can also be different from the marital status (MTSCO) and province of residence (PRCO) when tax returns are filed. For variables not expected to change through time, the PNRF should be used for consistency.

3.4.1 Geography variables

The IMDB enables the study of immigrant taxfiler mobility and retention in Canada over time. It is to be noted that complete outmigration cannot be captured, as there is no requirement for immigrants or filers to declare that they have left, or will be leaving, the country. Both the PNRF and tax files contain various measures of geographic location that allow researchers to establish an intended destination at admission and subsequent area of residence for immigrants. In the PNRF, **intended destination** is measured at the provincial, census metropolitan area, census division, and census subdivision levels. These variables originate from a self-reported destination at landing on the immigration application. Unlike the T1FF geography variables, the landing file variables are available only for the geographies defined in the latest available census; this means that they reflect only the most recent census boundaries.

The other geography available on the landing file is **province of nomination**, available for provincial nominees. The province indicated is the one under whose criteria the applicant has been admitted; however, it does not necessarily correspond to the province-of-destination variable.

Under **geographies of origin**, the country variables on the landing file indicate the individual's country of birth, country of citizenship, and last residence at the time of landing. It should be noted that these geographies may not be comparable over time, as national boundaries change from year to year. Some examples include the dissolution of the USSR, Yugoslavia, and Czechoslovakia; the union of Sikkim and India and of Vietnam and North Vietnam; and the creation of South Sudan.

Some individuals in the landing file report their country of birth as Canada. Normally, those who are born in Canada are granted citizenship at birth and do not need to apply for permanent residency. Those on the landing file who are born in Canada are most likely individuals born to foreign diplomats while residing in Canada who later chose to apply for permanent resident status.

A number of geographic variables in the T1FF datasets refer to slightly different notions of geographical location from the landing file. The most detailed geography in the T1FF is available at the census tract level; it is derived from the **postal code of the mailing address** (PSCO_). The postal code generally indicates the address of residence at tax filing in the spring of the following year. The mailing address may also refer to a business, such as an accounting firm or a law firm, and is not necessarily the person's current address. The **province of residence** on December 31 of the tax year (PRCO_) may not be the same as the province in the mailing address. This distinction is important, as using the derived census geography variables may not correspond to the province of residence on December 31 (PRCO_; however, it should correspond to the **province code** (PR__). The PRHO_ variable indicates an alternative to the mailing address and exists only for 2008 and subsequent years. Moreover, while the variable named taxing province code (TXPCO_) is, by definition, the same as the province of residence on December 31 (PRCO_), the **taxing province code** (TXPCO_) is less reliable (a known data quality issue exists with this variable, where both missing values and Newfoundland and Labrador are coded as "0").

Using the tax file variables to study geographic mobility amongst immigrants requires careful consideration of timing in making inferences about relocation and location of work. A researcher's guide to studying mobility and retention is included in Appendix D.

3.4.2 Time variables

"Landing year" and "tax year" are time variables often used to produce tables and perform analyses using the IMDB. The landing year is the year when the immigrant was granted permanent resident status, while the tax year is the tax filing year.

It is recommended that "landing year + 1" be counted as the first year of income, as it is the first full year in which the person will be in the country. Taxes filed in the year of admission should be interpreted with caution. First, about 50% of each landing cohort first files taxes in the landing year (proportion based on taxfilers from linked

immigrants). Secondly, taxes filed in the same year as admission may not represent a full year of income. An individual who landed in October 2010 will have only three months of income to declare in the spring of 2011, while an individual who landed in January 2010 will have 12 months of income to declare.

It is also possible to see taxes filed for individuals after their year of death, for example, in cases where the deceased person's relatives file taxes on his / her behalf. The variable Family Type (FCMP_) from the T1FF would be used in such cases. Please refer to the data dictionary for more details.

For example, Table 1 illustrates possible scenarios and describes which records should be included in a study to evaluate the socio-economic outcomes of the 1995-to-2000 immigrant cohort five years after landing. In order for a record to be included, the immigrant must have landed in any year from 1995 to 2000 and filed taxes five years after landing. This analysis would include the following IMDB records: IM19952 and IM19963.

Table 1
Example defining a cohort of interest

IMDB_ID	Landing_Year	Available tax years	Included in scope of study
IM19801	1980	1982 to 2013	No, landed prior to 1995
IM19952	1995	1988 to 2011	Yes
IM19963	1996	1996 to 2013	Yes
IM19974	1997	2010 to 2013	Yes, but no tax files available 5 years after landing
IM20095	2009	2011 to 2013	No, landed after 2000

Note: This example is based on fictitious data.
Source: Statistics Canada, example from the Longitudinal Immigration Database.

One of the shortfalls of using administrative tax data is a lack of precision with respect to timing. Apart from the year for which taxes are declared, no timing variables exist in the T1FF. This presents difficulties for studying job and unemployment spells, timing of relocation, and marriage. It also makes it difficult to distinguish self-employed individuals who are also seasonal employees from those who concurrently earn income from both sources. Despite these limitations, decisions about timing can still be informed by keeping in mind the following considerations.

Since the previous year's taxes are typically filed in the spring of the subsequent calendar year, some uncertainty may arise with respect to the specific year in which a change in address occurred. For example, Person A could file 2011 taxes with a mailing address in Toronto, and could then file his or her 2012 taxes with an Ottawa mailing address. It may be inferred that Person A lived in Toronto when filing 2011 taxes in the spring of 2012 and lived in Ottawa in the spring of 2013. Person A could have moved either in 2012, after filing the previous year's taxes, or in the first few months of 2013 prior to filing his or her 2012 taxes. If Person A moved between provinces, the variable for the province of residence on December 31 (PRCO_) could be useful in narrowing down the year of the move, since it relates to this person's province of residence as at December of the tax year. It should be noted that the mailing address does not necessarily correspond to the location of residence.

3.4.3 Education variables

Several variables in the landing file, such as years of schooling and education qualifications, allow education at landing to be measured. The former takes the form of a write-in answer to the question "How many years of formal education do you have?" The latter is phrased as "What is your highest level of completed education?"; options are provided. The derived "Level of Education" variable combines information from the two.

A data quality issue was identified with these education variables for 2011 and subsequent years, as a significant proportion of individuals in these years did not state their education qualifications or years of schooling and were coded as "0" ("None") on EDUCATION_QUALIFICATIONS and YEAR_OF_SCHOOLING instead of "missing." This problem became prevalent in 2011 and subsequent years. In 2011, up to 35% of immigrants stated that they had no education qualifications, compared to roughly 10% in the 1990s. The education variables for 2011 and subsequent years were imputed to resolve this issue. For more details on the imputation, see Section 7.3.

3.4.4 Intended-occupation variables

IRCC collects the intended occupation from the record of landing and assigns it a classification according to National Occupational Classification (NOC) codes in the landing file. These are broadest at the skill level, with a five-digit NOC codes being the most specific (see dictionary appendix for full definitions).

While intended occupation is also considered to be a good proxy of the individual's source-country occupation, caution is recommended. In order to list a specific intended occupation, applicants must prove that they have obtained the necessary education qualifications, as well as at least one year of experience in the field. As a result, this is considered to be a conservative measure for the intended variable after arrival by IRCC, as these requirements are quite stringent. For example, students completing a degree in engineering cannot list their intended occupation as engineer (given their lack of work experience) and are instead classified as students. Additional variables such as **labour market intention** (LM_INTENTION) and **skill level** (SKILL_LEVEL) can be used to obtain information on an individual's source-country field of work. Also, it should be noted that the intended occupation field is mandatory for principal applicant immigrants within the economic categories. For all other immigrants, this information is optional. Therefore, the information may not be as reliable as a measure of their intended occupation.

3.4.5 Other IMDB variables

Only variables that require detailed explanation and can present the most difficulty for analysts were included in Section 3.4. For further details on the variables included in the IMDB, please refer to the IMDB dictionaries. The tax component describes the variables included in the IMDB_T1FF files while the immigration component describes the variables included in the other files. These dictionaries are available to data users, or can be obtained upon request by writing to Statistics Canada at STATCAN.infostats-infostats.STATCAN@canada.ca.

IMDB data users should be aware that data from the immigration files and the tax files are collected at different times and that, in some instances, individuals' characteristics evolve with time. As a result, the marital status and the composition of a person's family might change through the years and consequently differ between the PNRF and the T1FF. The variables to use for analysis depend on the subject of the study.

4 Record linkage

As described in this document, the IMDB is the product of numerous record linkages. It was created for the purpose of providing statistical information in an anonymous format. This section gives an overview of the record linkage methods used to create the IMDB. For more details regarding data processing related to record linkage, see Section 5.

Record linkage is the process of matching records between or within databases. This approach is commonly used to fill data gaps and create a dataset with broad applications (Rotermann and al. 2015).

The hierarchical deterministic method was used to link ILF and NRF records to the Linkage Control File (LCF), a database of personal identification numbers (see Section 2 for the descriptions of these files). This method consists of matching records between multiple files (or within a given file) by means of common variables (Dusetzina and al. 2014). Over the course of waves of matches, the linkage criteria become less and less stringent. The LCF is not available to researchers; it is used only to produce the IMDB.

Users of a dataset created as a result of record linkage need to be aware that linkage errors are possible. Record linkages will have one of four outcomes: true matches correctly classified as matches, true matches falsely classified as non-matches, true non-matches falsely classified as matches, or true non-matches correctly classified as non-matches (Winkler, W.E. 2009). As shown in the example in Table 2, where records from file 1 are linked to records from file 2, the result of the record linkage between two variables will be either a match or a non-match. A good record linkage will maximize the proportion of true matches correctly classified as matches and the proportion of true non-matches correctly classified as non-matches, and minimize the other record linkage outcomes.

Table 2
Example of record linkage outcomes

		File 2			
	Record	**A**	**B**	**D**	**Type of Outcome**
File 1	A	**Match**	Non-match	Non-match	**True match**
	C	Non-match	**Match**	Non-match	**False match**
	D	Non-match	Non-match	Non-match	**False non-match**
	E	Non-match	Non-match	Non-match	**True non-match**

Source: Statistics Canada, example of record linkage outcomes.

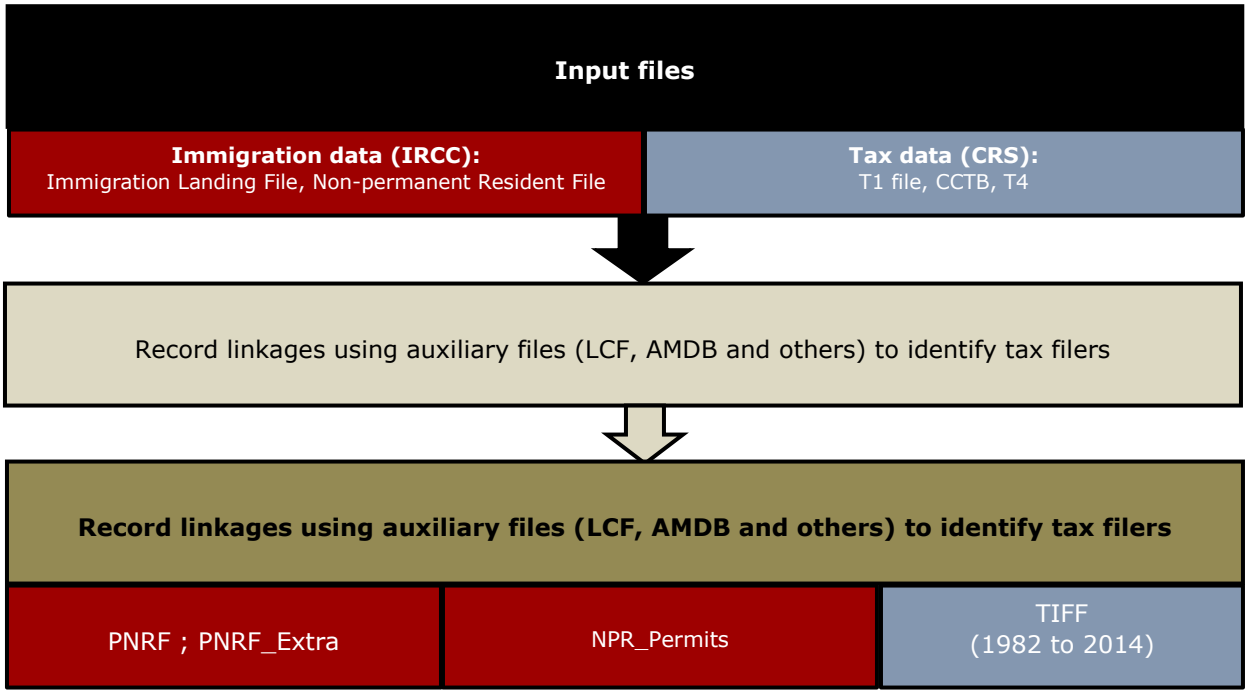
The results of deterministic record linkage are dependent on the quality of the linkage variables. For example, misspelled names or typos in the date of birth can create missed or erroneous matches. A non-match does not necessarily mean that the person did not file taxes. The record linkage rates for the most recent IMDB are available in Section 7.2.1.

5 Data processing

5.1 Processing

A number of government agencies are involved in the creation and processing of the IMDB. From initial data collection, to processing and dissemination, their cooperation is required to ensure the high standard of data quality that data users expect from Statistics Canada. At each step in the processing sequence, thorough manual and automated data quality checks are performed, and feedback loops are in place to correct any detected errors at the source. The following section briefly describes the annual processing that updates the IMDB.

Figure 1
Summary of the IMDB process flow



Note: See glossary of terms for definitions of acronyms.
Source: Statistics Canada, description of the IMDB process flow.

As shown in Figure 1, Statistics Canada first receives from the Canada Revenue Agency (CRA) the T1 data, in a file called “Personal Master File” (PMF), and other tax files. The tax files are then used to create the T1 Family File (T1FF), where individuals are linked to spouses and children via a common identifier, and geographic variables are created. Statistics Canada performs manual quality checks, and compares estimates from the T1FF with other data sources, such as the census (in census years) and the Survey of Labour and Income Dynamics, as well as annual income statistics produced by the CRA.[7]

On the immigration side, IRCC provides the data on landed immigrants and non-permanent residents used to produce the IMDB. These data serve to create the Immigrant Landing File (ILF) and the Non-permanent Resident File (NRF). The ILF and NRF are assumed to be complete censuses of permanent and temporary resident permits issued by IRCC since 1980.

Records in the ILF and NRF are linked to the Linkage Control File (LCF) by means of key linking variables, such as name, date of birth, and sex. Hierarchical deterministic linkages are made over the course of multiple waves with decreasing strictness of criteria with each wave. Nonetheless, even the least-strict waves are conservative in nature, in order to avoid false matches (linking two records that do not belong to the same individual), while generating the fewest false non-matches (missed links). The resulting linkage file is used for the final processing step, which links immigration data with the T1FF.

7. For more detailed information on T1FF processing and data quality, see http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=240717.

In addition to adding the information for the most recent tax year, a full back-sweep of previous years is done in order to add tax information for any new individuals that have been linked. Annual improvements in methodology, as well as the availability of new linkage information, lead to better year-to-year linkage rates, and thus an increase in linked records even for previous years. This could mean that a landed immigrant's or non-permanent resident's filed tax records are not linked in the IMDB one year but that their subsequent tax filings could still be linked in a later year. As methodology improves, the back-sweep could ensure that all their previous tax filings, if they are on the T1FF, can become linked as well. This is how, after the processing of the most recent tax data, individuals who had landed and filed taxes many years earlier could still be added to the IMDB. For individuals with multiple landings since 1980, data from the time of the first landing are retained.

Although taxes for a given year are usually filed in the spring of the following year (i.e., claiming 2013 income in 2014), there are exceptions. At times, someone may have filed taxes later in the year, and would not be included in that year's T1 processing done by Statistics Canada. When that file is handed down for IMDB processing, these late-filers are excluded and will not be included in the next year's processing, as the T1FF is not updated. Similarly, individuals who file taxes for previous years are not added to the IMDB for those years, as previous years' T1FF is not updated. In that case, a person's first on-time filing will show up as their first year in the database.

At this point, a series of programs are run to assess the data quality and linkage rates, ensuring that there are no duplicates and flagging outliers. Once the database is linked, it is deemed complete and dissemination is ready to take place.

In the end, the database consists of SAS files, one tax file per year since 1982 (IMDB_T1FF_&year_OUTLIERS), and Immigration data files (PNRF_&year, PNRF_EXTRA_&year and NRF_PERMIT). All these files are described in Section 2. The IMDB Unique Person Identifier (IMDB_ID) is used to connect all these files (see Appendix D.1 for programming tips).

5.2 Non-permanent Resident File (NRF) linkage

The Non-permanent Resident File (NRF), provided by IRCC, covers records of temporary resident permits issued for 1980 and subsequent years. It provides some demographic information about non-permanent residents as well as detailed information regarding their permits, such as permit type and the valid-date range.

The NRF contains millions of observations. These, however, include duplicates, whereby a single individual may have a number of different IDs. This issue is due mainly to records from the late 1980s where the original person identification number was lost. These records have been removed by linking the NRF to itself. This has resulted in approximately 220,000 records (roughly 400,000 observations) being identified as duplicates. In cases where both non-permanent resident records had their own landing record, the duplication link has been nullified (applicable to fewer than 1,000 records), as it is assumed that the landing file contains unique identifiers. After cleaning, only distinct non-permanent residents remain.

Both immigration files (ILF and NRF) contain some demographic information. However, the demographic information contained in the two files may not always be consistent. This is the case when more than one source is available or when there is a conflict. It has been decided that information in the ILF on the Integrated Permanent and Non-permanent Resident File (PNRF) shall be retained in light of data quality issues with the NRF in its earlier years.

5.3 Derived variables included in T1FF

Once record linkages have been performed, immigration-specific variables for immigrants and temporary residents are added to the T1FF.

A flag that identifies a taxfiler's immigration status on the basis of the linkage to IRCC's immigration data is created. To determine whether a tax record belongs to a taxfiler who was a non-permanent-resident, the flag **TR_IND** is available (1: yes; 0: no).

Derived variables that identify and describe families are also created. In each annual T1FF, it is possible to have an estimate of the number of immigrants in a family who landed in 1980 or thereafter (variable **IMM80F&year**). It is also possible to determine whether the immigrant has a spouse (in the given taxation year) and whether this

spouse is an immigrant or a non-permanent resident (variable **SP_IDI&year**). Data users can identify immigrants in the same family, each tax year, by using the variable **Family Identification Number** (FIN_). All members of a family have the same value for this variable, namely the IMDB_ID of the oldest family member who landed in 1980 or thereafter. The quality of these variables depends on the quality of the record linkage and the T1FF files, since only linked individuals will be counted (see Section .2.5).

The variables with the prefix **TNK** are counts of the number of claimed children of a given age in the families of immigrants and non-permanent residents (see the tax component of the data dictionary for more details). The term “children” (“child”) is defined as any person who is single and living with one or two parents; a child can be of any age. For example, in Table 3, the family of immigrant identified as IM19801 has two children aged 1 in 2011 (TNK01I2011), while family IM19873 has a total of three children in 2011 (TNKIDI2011), one aged 0 (TNK00I2011), one aged 1 (TNK01I2011), and one who is older than 18 years of age (TNK19I2011). The immigrant IM20105 has no children in 2011.

Table 3
Example on variables related to number of children in family

	TNK00I2011	TNK01I2011	TNKxxI2011	TNK19I2011	TNKIDI2011
IMDB_ID			number		
IM19801	0	2	0	0	2
IM19802	0	1	0	0	1
IM19873	1	1	0	1	3
IM19994	0	0	0	1	1
IM20105	0	0	0	0	0

Note: Not all variables are presented in this table. This example is based on fictitious data.
Source: Statistics Canada, example from the Longitudinal Immigration Database.

Another variable added to the T1FF is **OUTLIER_IND** (1: outlier; 0: no). It is a flag added to identify records with extreme incomes (see Section 5.5 for more details) and to be removed from any tables or calculation. Records identified as outliers have some extreme incomes that could bias analysis results.

5.4 Derived variables included in PNRF

When the PNRF is produced, some variables relating to tax filing patterns are derived and included in the file. The variable **FIRST_TAX_YEAR** indicates the first year for which a tax record was available for a given individual, while **LAST_TAX_YEAR** indicates the last year for which a tax file is available. It is to be noted that a tax record does not necessarily exist for every year between the first tax year and the last tax year. For example, a case where First_tax_year=1982 and Last_tax_year=2012 does not necessary indicate that the taxfiler has filed taxes continuously, as the tax file for 2006 may be missing, for example.

The variable **PREFILER_IND** is used to identify immigrants who have T1FF data prior to their landing year. Most have been linked to a non-permanent resident record, as expected (see Section 7.2.4 for more details).

5.5 Outlier detection

After creating the IMDB_T1FFs, outlier detection is performed on all tax files to identify outlier records. A record is deemed to be an outlier when it is determined to contain one or some extreme income values compared to other records. The criteria used to identify the outliers are confidential. The variable **OUTLIER_IND** is created to identify the records with extreme values.

The outlier flag, OUTLIER_IND, is in the tax files, but is not in the PNRF. A given person’s record may be flagged as an outlier in a specific year without necessarily being found to be an outlier for all years for which the person filed taxes. All outliers are to be removed from analysis. As shown in Table 4, for person IM19802, only the 1983 record has been flagged as being an outlier, while person IM19801 has no tax files flagged as outlier. No outlier flag is available in 2012 for IM19994 because no tax records are available for that person in 2012.

Table 4
Example on outlier flag

IMDB_ID	OUTLIER_IND1982	OUTLIER_IND1983	OUTLIER_INDyyyy	OUTLIER_IND2012	OUTLIER_IND2014
	number				
IM19801	0	0	0	0	0
IM19802	0	1	0	0	0
IM19873	1	1	1	0	...
IM19994	0	0	0	...	0

... not applicable
Note: Not all variables are presented in this table. This example is based on fictitious data.
Source: Statistics Canada, example from Longitudinal Immigration Database.

The outliers are removed from tabulations and any analysis. The IMDB excludes (the very few) large incomes as they would skew averages and give users an incorrect impression of the income situation for certain types of immigrants. Consider a fictitious example where the average income of Czech-Canadians is $40,000 in a given year and, the next year, it suddenly jumps to $500,000 because, by chance, a Czech hockey player landed. This would bias the “real” income situation for Czech-Canadians. For that reason, the “un-representative” Czech hockey player’s income would be removed from calculations. There is a confidentiality component to this example, as well. If such a jump in average income were observed, one could deduce the Czech hockey player’s income, which would be a breach of confidentiality. Incidentally, in some IMDB products, median income, which is more resistant to the changing influence of large individual values, is also provided as a measure.

When one is producing tables or analyzing data, records deemed to be outliers for a given year have to be removed from calculations relating to the year in question for the reasons mentioned above. For further details, see Appendix D.6.

6 Dissemination

Once the linkage is complete, the data files (including a synthetic dataset, see Section 6.3) are stored on Statistics Canada servers for data users to create customized tables and model output. Statistics Canada disseminates output via tabular and analytical products while maintaining strict adherence to the confidentiality of the data. Users who would like to obtain information pertaining only to the immigrant landing file (ILF) should contact IRCC, as the IMDB has data on immigrant taxfilers only.

6.1 Analytical products

At Statistics Canada, the key analytical products are the **Canadian Socioeconomic Information Management System (CANSIM) tables** that report immigrants' income by various individual characteristics and geographies. These tables can be found on the Statistics Canada website under series number 054. The income measures (averages and median) available on the CANSIM tables are employment earnings, employment insurance, investment incomes, self-employment earnings, and social assistance (for details on how these measures are derived, see Appendix D.8).

In 2014, four CANSIM tables based on IMDB data were produced at the national level:

Table 054-0001: Income of immigrants, by sex, landing age group, immigrant admission category, years since landing, and landing year (2014 constant dollars);

Table 054-0002: Income of immigrants, by world area, sex, immigrant admission category, education qualifications, knowledge of official languages, and landing year for the 2014 tax year;

Table 054-0003: Number of immigrant taxfilers, by province of landing, province of residence, sex, landing age group, immigrant admission category, and landing year for the tax years from 1996 to 2014; and

Table 054-0018: Income of immigrants by sex, landing age group, immigrant admission category, period of immigration, family status, and tax year (2014 constant dollars).

All of these tables, except table 054-0003, have a series of tables at the provincial level as shown in Table 5. Retention rates can be calculated from CANSIM table 054-003 (see Appendix D.4).

At the provincial level, all Atlantic Provinces are included in a single table; there are no individual tables at the territory level. It is to be noted that the province is based on the province of residence on December 31 of the tax year (variable PRCO).

Table 5
List of Longitudinal Immigration Database CANSIM tables

National table	Provincial table
054-0001	0004, 0006, 0008, 0010, 0012, 0014, 0016
054-0002	0005, 0007, 0009, 0011, 0013, 0015, 0017
054-0003	None
054-0018	0019, 0020, 0021, 0022, 0023, 0024, 0025

Source: Statistics Canada, CANSIM.

In addition, several analytical articles related to the IMDB have been written over the years (see Appendix C). Moreover, Statistics Canada analysts take ad hoc data requests from researchers and data users. These are filled on a cost-recovery basis.

6.2 Requesting analytical files

Once the IMDB has been released, all the analytical files described in this report (PNRF, PNRF_Extra, T1FF, and NRF_Permit) are also available to on-site researchers, who are granted access once they have deemed employee status with Statistics Canada. These individual micro-data are stripped of all identifying information (such as exact date of birth, landing date, Social Insurance Number (SIN), and name). Researchers unable to be physically present at Statistics Canada's headquarters can access a **synthetic file**, complete with variables and variable formats, but containing fictitious data. These researchers can use the synthetic file to write their desired programs,

and send the syntax to Statistics Canada to run. This file should be used in conjunction with the most up-to-date dictionaries, as some variables on the synthetic file have been dropped since its creation and some have been added. IMDB users can request custom tabulations from Statistics Canada; such requests are filled on a cost-recovery basis, and cost will vary according to the nature and type of each request.

Before any output can be released, results are vetted for confidentiality by Statistics Canada. Minimum cell size requirements and rounding minimize the risk of breach of confidentiality.

6.3 Other statistical programs using IMDB data

IMDB data are used in many Statistics Canada programs for a variety of purposes. The **Longitudinal Administrative Databank** (LAD) uses IMDB data to include a sample of 20% of IMDB records in its sample. The LAD also uses IMDB records to add immigrant-specific variables, such as landing year, to its databank.

The **Canadian Employer-Employee Dynamics Database** (CEEDD) is a set of longitudinal analytical data files maintained by Statistics Canada to provide matched data between employees and employers of the Canadian labour market for 2000 and subsequent years. The CEEDD files cover all individuals that can be identified from the T1 and T4 files as well as employer or self-employment information that individuals can be linked to. The IMDB is one of the component files of CEEDD, and this linkage allows researchers to conduct analysis related to labour market outcomes and job dynamics with respect to the immigrant population in Canada.

The **2013 General Social Survey (GSS) on Social Identity** (SI) collects detailed information on the social networks and civic participation and engagement of Canadians. The 2013 GSS on SI was linked to the IMDB for the purpose of selecting a representative sample of the immigration population to support and evaluate immigrant policies and programs. In particular, Immigration, Refugees and Citizenship Canada (IRCC) used this linked data file to develop a descriptive profile of the social connections and civic engagement of immigrants across admission categories.

DEMOSIM[8], a Statistics Canada microsimulation model, uses the IMDB-LAD for population projections for the provinces, territories, census metropolitan areas, and selected smaller geographies, on the basis of a number of characteristics. **Census programs** use the database for certification of immigration data.

6.4 Confidentiality[9]

Statistics Canada is committed to respecting the privacy of individuals. All personal information created, held, or collected by Statistics Canada is protected by the Privacy Act, as well as by the Statistics Act in the case of respondents to the agency's surveys.

In view of its unique mandate as the national statistical agency in collecting personal information solely for statistical and research purposes, Statistics Canada has prepared privacy impact assessments that address privacy issues associated with its survey activities.

Statistics Canada initiated a privacy impact assessment[10] following approval by its Policy Committee (the agency's senior executive committee, chaired by the Chief Statistician) of significant changes to the Longitudinal Immigration Database. The purpose of this assessment was to determine whether there were any privacy, confidentiality or security issues associated with these changes and, if there were, to make recommendations for their resolution or mitigation.

This assessment concluded that, given existing Statistics Canada safeguards as well as the additional measures put into place for the Longitudinal Immigration Database, the risk of inadvertent disclosure is extremely low. The importance of the data to public policy outweighs the privacy implications. The governance mechanisms in place constitute safeguards against inappropriate use of the data. Through the periodic review by its Policy Committee, Statistics Canada regularly assesses the continued relevance of the IMDB and the value of the information against the implied privacy invasion.

8. Http://www.statcan.gc.ca/eng/microsimulation/demosim/demosim.
9. Source: http://www.statcan.gc.ca/eng/reference/privacy.
10. Http://www.statcan.gc.ca/eng/about/pia/lidb.

The agency's statistical work involves record linkage projects that bring together information about individual respondents for research purposes. This is a recognized source of valuable statistical information, but the linkage must always serve a public good. To address possible privacy intrusions from this type of research, Statistics Canada not only has a directive in place, but also practices a well-defined review and approval process for all record linkages.

To ensure confidentiality, it is mandatory to round tabular and descriptive output when producing tables with IMDB data (see Appendix D.5).

7 Data evaluation and quality indicators

7.1 Error sources

Because the IMDB is the product of several record linkages, it is subject to different sources of errors, including record linkage errors, measurement errors and coverage errors. In this section, the sources of errors are explained and the prevalence of some of these errors is presented.

It is to be noted that, given that it is a census of immigrant taxfilers who landed in 1980 or thereafter, no weights are created in the IMDB. No adjustments are made for the missing tax years of filers or for linkage errors; no sampling is performed; and every linked taxfilers is kept in the final dataset. However, the linkage itself presents a form of sampling error when links are missed.

7.1.1 Record linkage errors

Datasets produced from the results of record linkages are subject to record linkage errors. Two **types of errors** are possible—false positives (false matches) and false negatives (false non-matches). A link is considered a false positive when two records not belonging to the same person are deemed a match. A link is considered a false negative when two records belonging to the same person are deemed a non-match. Following a manual review of the 2014 IMDB record linkage result, the false positive rate was estimated at 0.27% (Diaz-Papkovich 2016a). Currently, no estimate of the false negative rate is available.

The **linkage rate** gives an indication of the quality of the database—the linkage rate is 89% (Diaz-Papkovich 2016b) for immigrants who landed after 1980. Matching employs variables common to both the Linkage Control File (LCF) and immigration datasets (e.g., first and last name, sex and date of birth). Most matches occurred in the first waves which allowed for little or no difference in the value of these variables. As well, this means that most matches are of high quality and explains the low rate of false positives. For more details on the linkage rates, please refer to Section 7.2.1.

Furthermore, it is possible to miss part of an immigrant's fiscal history since some immigrants have more than one social insurance number (SIN) through time (a temporary SIN assigned at arrival to the individual as a non-permanent resident, and later a permanent SIN assigned after landing). Both SINs are required in order to have a complete fiscal history from arrival in Canada. The LCF (described in Section 2.3) allows for identification of these SINs. It is possible that, in a few instances, some SIN connections are missed or false connections are made.

7.1.2 Measurement errors

Measurement error is the difference between a variable's measured value and its true value. This type of error can be attributed to a number of factors, including data capture (e.g., typos) and respondent error (e.g., misinterpretation of the question asked). This type of error was taken into account in the creation of the Integrated Permanent and Non-permanent Resident File (PNRF) to avoid conflicting information for any individual. For example, when a person has a record on both the ILF and the NRF, and the sociodemographic variables have inconsistent values, the values at landing (in the ILF) are kept. See sections 7.2 and 7.5 for some counts.

7.1.3 Coverage errors

Coverage errors are the result of omissions, erroneous additions, duplicates, and errors of classification of records in the database. Coverage errors can result from inadequate coverage of the population. They can create biased estimates, and the impact can vary for different sub-groups of the population. These errors often result in undercoverage. **Undercoverage** in the IMDB is in part the result of the exclusion of tax files of immigrant taxfilers from the database. Immigrants who do not file taxes for a given year or who file late would not have an IMDB_T1FF record although linked to tax and part of the population of interest. If, for any reason, an immigrant record was not included in the Immigrant Landing File (ILF), it would not be part of the IMDB. **Overcoverage** is the result of the addition to the database of records excluded from the target population. An immigrant could have more than one ILF record as a result of multiple landings not identified as such, for example. Please refer to Section 7.4 and Appendix B for more information on IMDB coverage.

7.2 Data accuracy

This section will discuss the accuracy of the immigration data. For details on the accuracy of the T1 Family File (T1FF), please refer to the T1FF entry (record number 4105).

The accuracy of the IMDB is dependent on the representativeness of the population included in it. A study conducted in the first years of the IMDB concluded that the IMDB "appears to be representative of the population most likely to file tax returns. Therefore, the results obtained from the IMDB should not be inferred to the immigrant population as a whole, but rather to the universe of tax-filing immigrants" (Carpentier and Pinsonneault 1994).

The reasons for the differences between taxfilers and the entire foreign-born population are explained in an article by Badets and Langlois (2000) describing the challenges of using the IMDB:

> The characteristics of the immigrant taxfiler population will differ from those of the entire foreign-born population because the tendency or requirement to file a tax return will vary in relation to a person's age, family status, and other factors. One would expect a higher percentage of males to file a tax return, for example, because males have higher labour force participation rates than females. The extent to which immigrants are "captured" in the IMDB will also be influenced by changes to the income tax. For example, the introduction of federal and provincial non-refundable tax credit programs encourage individuals with no taxable income to file a return to qualify for certain tax credits. (Badets and Langlois 2000)

These statements on the representativeness of the IMDB are still relevant. As shown in the following section, females have lower linkage rates than males, for example.

7.2.1 2014 IMDB: Linkage rates

This section is based on the 2014 IMDB. The overall linkage rate for the 2014 IMDB was 89% with differences by gender and age. A link does not mean that a tax file is available since it is possible to link dependents of taxfilers or immigrants who have yet to file taxes. Of the immigrants who landed in any year from 1980 to 2014, 82.5% were linked to at least one T1FF record.

The proportion of linked taxfilers by age group at landing and sex is shown in Table 6. Immigrant females have lower rates than males for all age groups. The lower rates for the 0-to-14 age group are expected since those in this age group are not of working age. See Appendix B for rates by sex, age group and landing cohort.

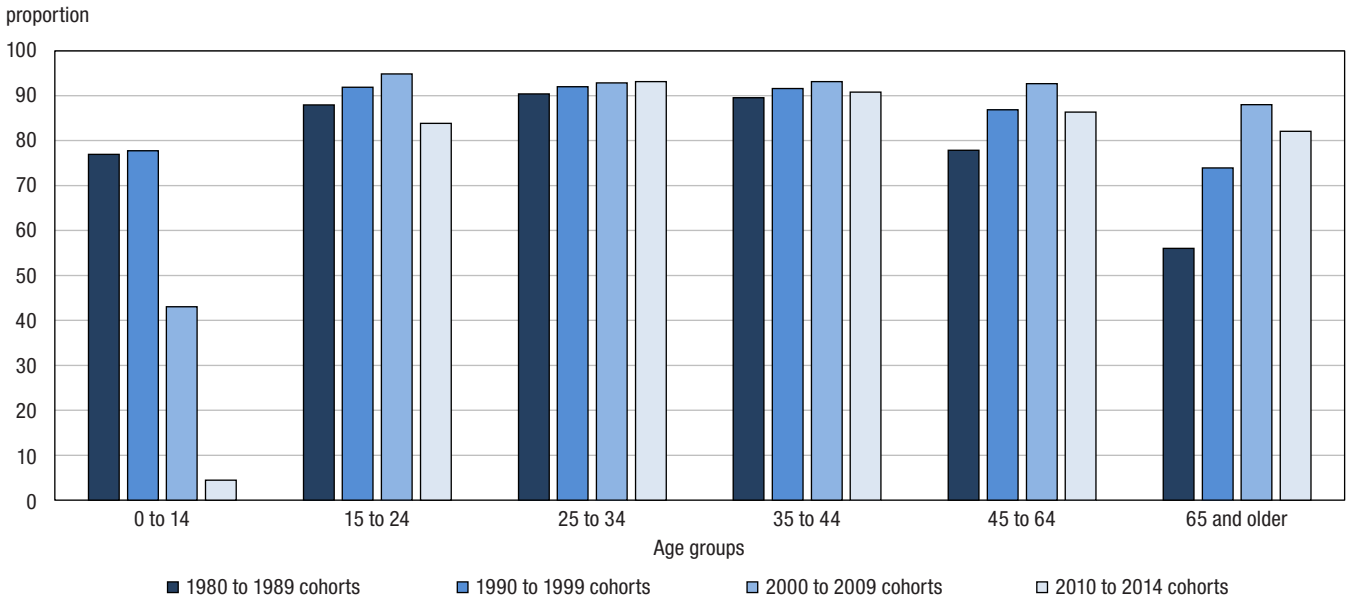
Table 6
Proportion of linked taxfilers by age group at landing and sex

	Age at landing						
	0 to 14	15 to 24	25 to 34	35 to 44	45 to 64	65 and older	Total
				percent			
Male	53.5	91.8	92.8	92.0	87.8	75.3	82.8
Female	52.6	89.8	91.7	91.5	85.2	73.6	82.3
Total	**53.1**	**90.7**	**92.3**	**91.8**	**86.4**	**74.3**	**82.5**

Source: Statistics Canada, 2014 Longitudinal Immigration Database.

As immigrants become older, they start filing taxes and are included in the IMDB. Chart 1 shows that, among immigrants who landed at any age from birth to age 14, the proportion of linked taxfilers is higher for immigrants who landed prior to 2000 than for immigrants who have landed since 2000. Recent immigrants also have lower linkage rates. See Appendix B for table showing the proportion of linked taxfilers by age group at landing, sex and landing decade.

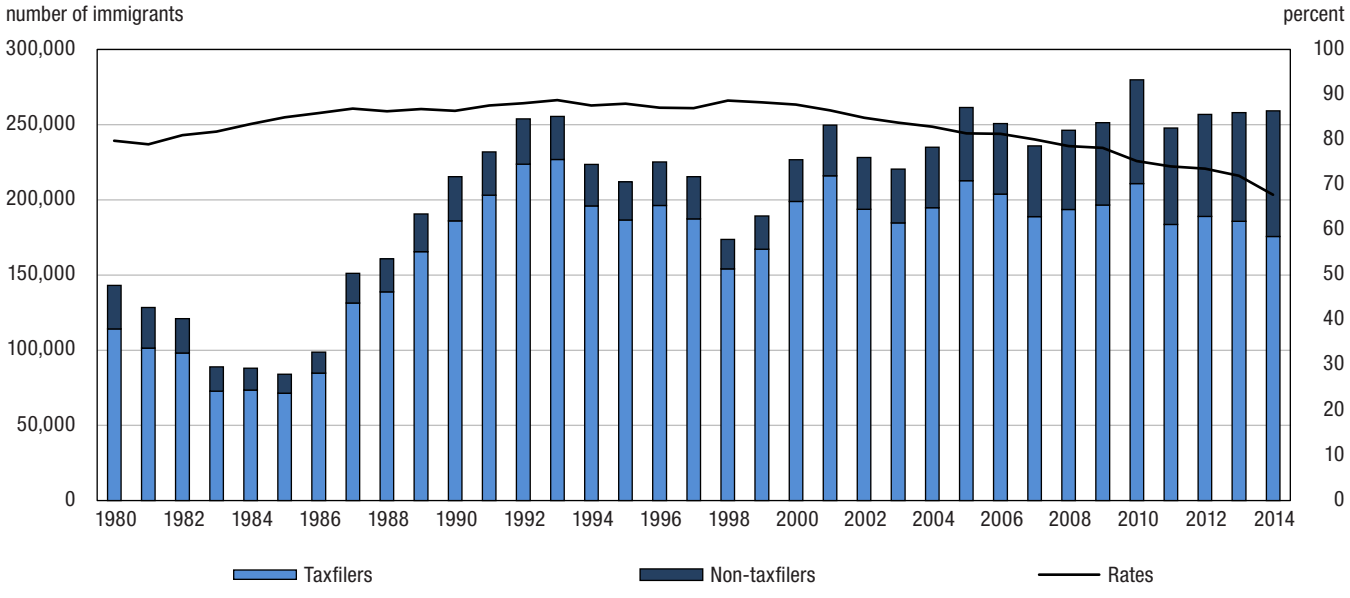
Chart 1
Proportion of linked taxfilers by age groups at landing, sex and landing decade



Source: Statistics Canada, 2014 Longitudinal Immigration Database.

Chart 2 illustrates the proportion of filers, and the number of filers and non-filers by landing year, where the term "non-filer" means that no T1FF records are available. For the 2014 IMDB, the filing rate varies by landing year, ranging from 67.8% for those who landed in 2014 to 88.8% for those who landed in 1993. The filing rates increase with the number of years that immigrants stay in Canada; this may explain why the linkage rates are higher for those who landed in the 1980s and 1990s. See Appendix B, tables 15 and 16, for detailed distribution numbers by landing year.

Chart 2
Distribution of taxfilers compared to non-taxfilers, by landing year



Source: Statistics Canada, 2014 Longitudinal Immigration Database.

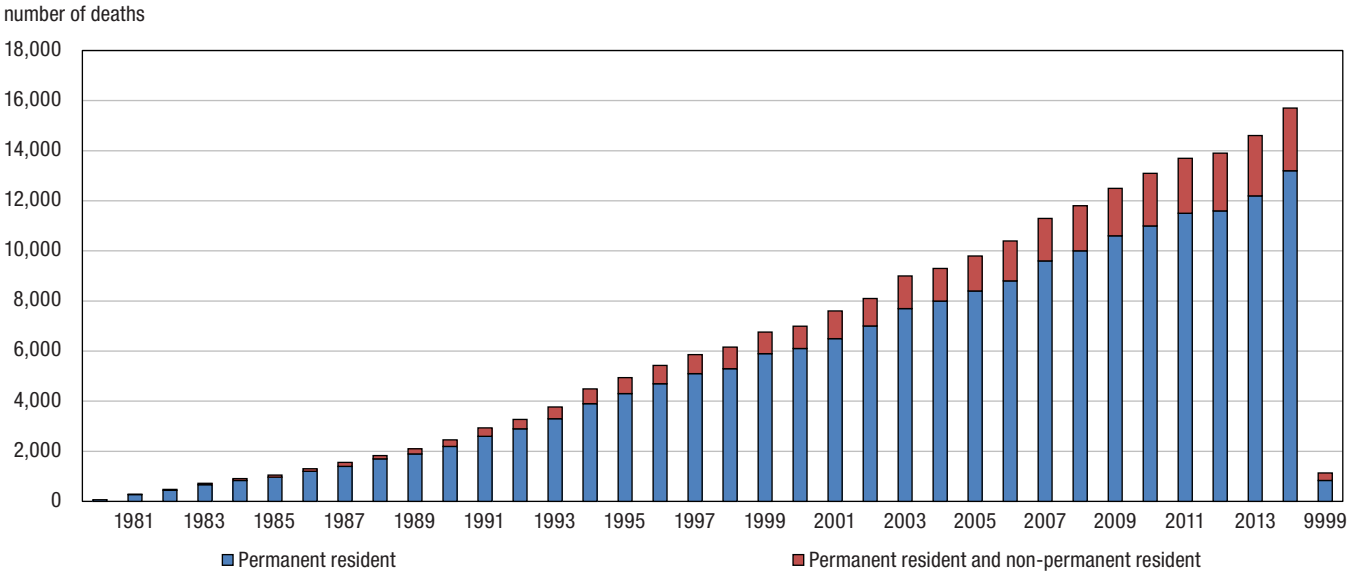
7.2.2 Availability of date of death

The year and month of death, as well as a death flag, are included in the PNRF. These variables can be added by using information from Statistics Canada's Amalgamated Mortality Database (AMDB). The AMDB combines records from vital statistics and tax files to produce a mortality dataset.

This information is the result of a record linkage. It is likely that a small proportion of links will be false; that is, that record linkages will yield false positive links. All links with a date of death prior to landing have been removed (fewer than 200 records). Individuals who died prior to 1981 or never filed taxes are less likely to be linked as a result of the files used for the linkages.

Chart 3 describes the general trend in the number of deaths per year since 1980. "Permanent resident" deaths refer to the deaths of all immigrants who landed in 1980 or thereafter and who did not have pre-landing experience. "Permanent resident and non-permanent resident" deaths refer to the deaths of all immigrants with pre-landing experience. The value "9999" represents the records of deceased immigrants for which the year of death is not available.

Chart 3
Permanent and non-permanent residents, by year of death



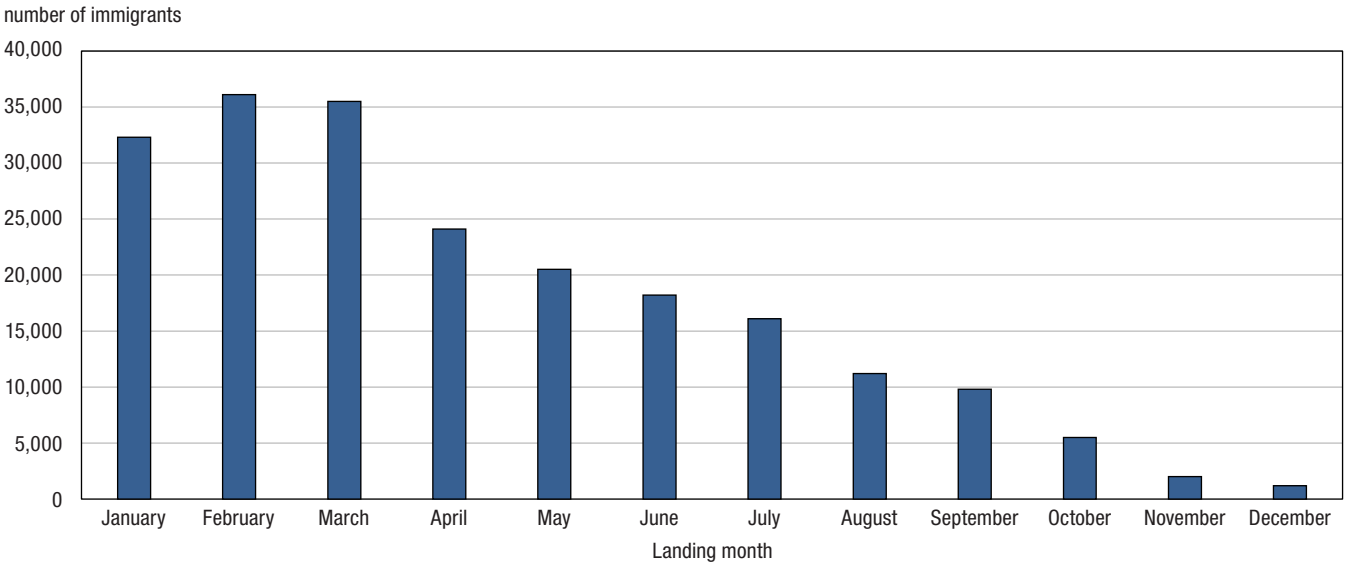
Note: The value 9999 is assigned when the date of death is missing.
Source: Statistics Canada, 2014 Longitudinal Immigration Database.

7.2.3 Prefilers compared to record on the Non-permanent Resident File (NRF)

The results included in this section are drawn from a study based on the 2013 IMDB. **Prefilers** are immigrants who filed taxes prior to their landing year. It is sometimes assumed that all prefilers are immigrants who were non-permanent residents prior to landing. This section discusses why it is not the case. A total of 1.26 million invididuals filed taxes before officially landing in 1980 or a subsequent year—of these, 212,500 are not linked to a non-permanent resident record as may otherwise be expected. Upon further investigation, it has been discovered that most permanent resident prefilers not linked to a non-permanent resident record are likely immigrants who have filed taxes when not required: 96% of these prefilers filed taxes only for the year prior to landing, and 75% reported no income (96% had no employment income). As shown in Chart 4, most of these prefilers landed in the first months of the year, prior to the deadline to file taxes for the previous year. It appears some immigrants who landed prior to the month of May filed taxes for the year prior to their landing year, for which they were not required to file.

Given these findings, whether it is appropriate to remove records with Prefiler_ind=1 and TR_IND=0 from studies on immigrants with pre-landing experience depends on the analysis since TR_IND=0 means no record is available on the non-permanent permit file.

Chart 4
Distribution of prefilers without a non-permanent resident permit, by landing month



Source: Statistics Canada, 2014 Longitudinal Immigration Database.

Not all immigrants with pre-landing experience are identified as prefilers: 478,100 immigrants have non-permanent resident records with Prefiler_ind=0. Depending on the subject of interest, using the TR_IND=1 or the number of temporary resident permits (variable NUMBER_ALL_PERMITS) is more appropriate to study immigrants with pre-landing experience. Prefiler_ind=0 indicates that no tax records have been filed prior to landing, but this does not mean that the individual had no pre-landing Canadian experience.

7.2.4 Spouse indicator

The IMDB contains variables that enable data users to obtain information on marital status and spouses. The following section contains results of a study done on the 2012 IMDB. No major changes have occurred since then in the marital status codes or family flag.

The spouse identification number (SP__IDI) is derived from tax files. This information can be derived only when the respondent claims his or her spouse or common-law partner while filing taxes; this causes an underestimation of couples as compared to the marital status declared in the tax files. From the T1FF, it is also possible to obtain the marital status at time of filing.

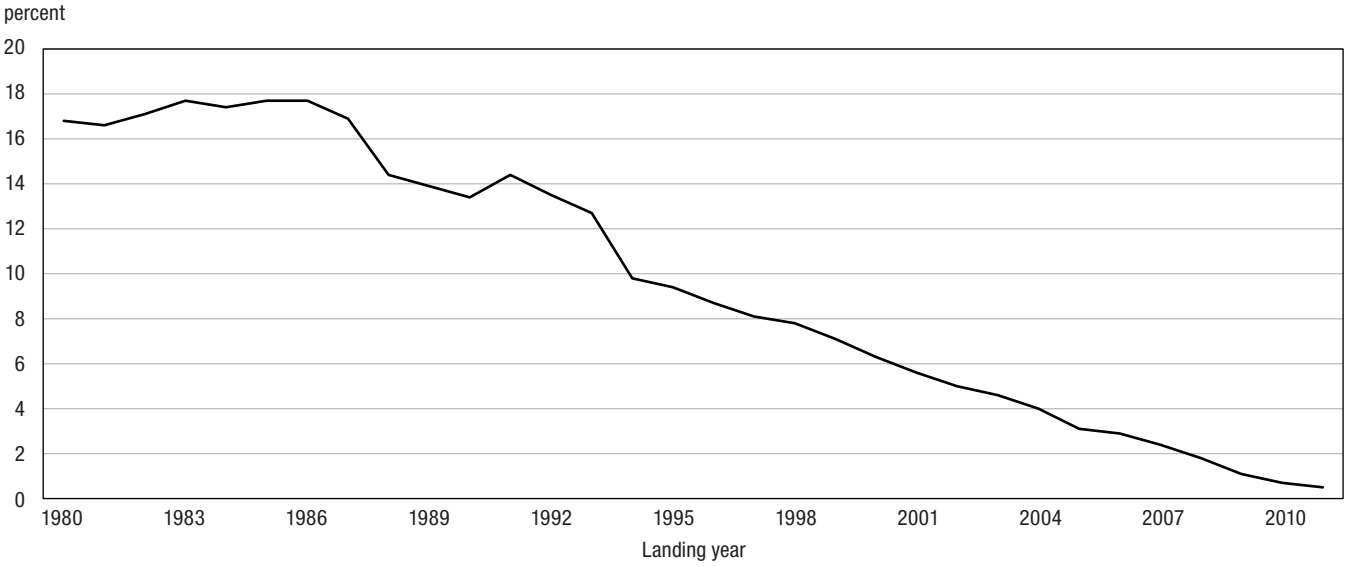
Prior to 1991, the "**single**" category was not available as **marital status** (MSTCO). The "**common-law**" status was made available as of 1992 for all datasets (1982 to 2012). Since 1992, the proportion of IMDB records indicating marital status as "single" has ranged from 20% to 30%. The proportion of "**separated**" has declined from 30% prior to 1992 to 4% after. The other marital status categories have not been affected by pattern changes.

Analysis done on the distribution of marital status (MSTCO from tax files) and the spouse ID (SP__IDI) shows differences between the two variables. This is because values for marital status are missing for some records. In a perfect situation, the records of all married persons would have spousal information, and the records of all single persons would have no spousal information. This analysis shows data quality to be better after 1992, when separate statuses for "common-law" and "single" were introduced.

Presence of spouse reporting gaps

Further to a review of the longitudinal history of immigrants on the 2012 IMDB, some cases where the spouse or common-law partner is missing (or different) for a given year and the same spouse is declared two or three years later have been found. The Chart 5 gives a summary of these gaps.

Chart 5
Proportion of cases with inconsistent spouse identification number, (SP__IDI) by landing year



Source: Statistics Canada, 2012 Longitudinal Immigration Database.

Most immigrants on the file have one or no spouse in the years from 1980 to 2012 according to the IMDB_T1FF files. It is to be noted that no marital status (and no spouse info) is available for 1.2 million immigrants out of approximately 6 million immigrants.

7.3 Imputation of education variables

A data quality issue regarding the variables for education qualifications and years of schooling was identified. A non-negligible proportion of individuals who did not state their education qualifications or years of schooling were coded as “0” or “None” instead of “Missing” on **EDUCATION_QUALIFICATIONS** and **YEARS_OF_SCHOOLING**. This problem was prevalent in 2011 and subsequent years. In 2011, 35% of immigrants stated that they had no education qualifications, compared to roughly 10% in the 1990s.

This issue was resolved by imputing the education variables by means of values for education variables from 2008 to 2010 to model the most recent year’s education variables. For the imputation, variables such as landing age, immigration_category_rollup2, intended occupation, gender and country of last permanent residence were used. The nearest-neighbour imputation method was employed. The variable **Education_imputation_ind** (0: no; 1: yes), available in the PNRF, was created to identify records with imputed education variables.

7.4 Coverage

7.4.1 Coverage of the Integrated Permanent and Non-permanent Resident File (PNRF)

The 2014 Integrated Permanent and Non-permanent Resident File (PNRF) contains over 7.1 million records (Table 7); of these, close to 5.9 million records (82.5%) are linked to at least one tax file. It is to be noted that immigration data belonging to non-taxfilers are included in a file named PNRF_NONFILER_2014 while the taxfilers immigration data are in a file named PNRF_2014. The following table shows the distribution of records depending on their presence in the different immigration and tax files. About 1.5 million records belong to immigrants who were temporary residents prior to becoming permanent residents; close to 1.4 million of these records are linked to at least one tax file. See Appendix B for detailed distribution numbers by landing year.

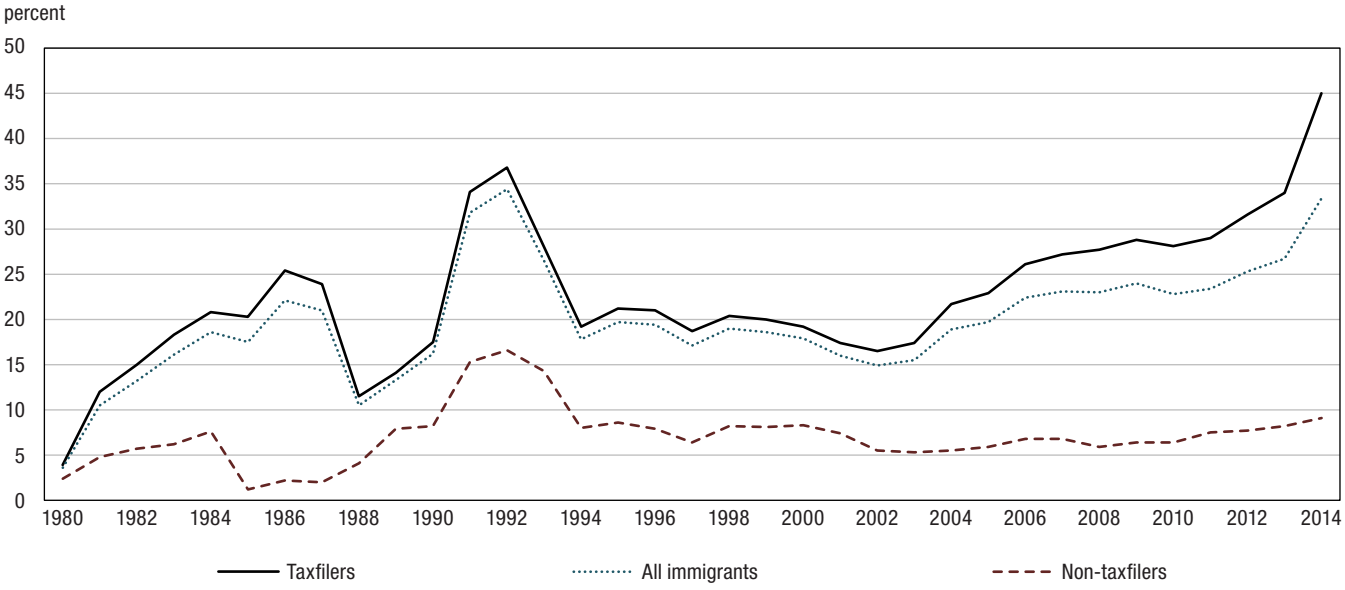
Table 7
Coverage of permanent residents

	Permanent resident	Permanent resident with non-permanent resident permit	Number of taxfilers
	number		
Total filers	4,515,500	1,392,900	5,908,400
Total non-filers	1,159,000	91,700	1,250,700
Total	5,674,500	1,484,600	7,159,100
	percent		
Percent of taxfilers	79.6	93.8	82.5

Source: Statistics Canada, 2014 Longitudinal Immigration Database.

Data on immigrants with non-permanent resident permits are available. The proportion of immigrants with pre-landing experience varies by landing year (Chart 6); it ranges from 4% in 1980 to 33% in 2014. As mentioned in Section 2.1.2, non-permanent permits issued prior to 1980 are not available in the IMDB. As a result, the proportion of immigrants with pre-landing experience in the early 1980s is underrepresented. The proportion of immigrant filers with pre-landing experience (solid line) is higher than the overall proportion of immigrants with pre-landing experience (dotted line) because the linkage rate for these immigrants is higher than that for immigrants without pre-landing experience.

Chart 6
Proportion of immigrants with non-permanent resident permits, by landing year



Source: Statistics Canada, 2014 Longitudinal Immigration Database.

7.4.2 T1 Family File (T1FF) size and coverage by year

Tax files for 1982 and subsequent years are available for linked permanent residents. Some permanent residents were non-permanent residents prior to landing. Table 8 gives details on the distribution of linked permanent residents with and without non-permanent permits prior to landing, by tax year. At least one tax file is available for the 79.6% of permanent residents without a non-permanent permit prior to landing and for the 93.8% of permanent residents who were non-permanent residents prior to landing. The fact that permanent residents with pre-landing temporary permits have a higher rate of filing taxes than permanent residents without pre-landing permits can be explained by a requirement in the permanent resident application process with respect to non-permanent residents. Non-permanent residents who apply for permanent residency are required to fulfil their obligation to file tax in Canada. The number of taxfilers on the IMDB_T1FF increases as the years pass since the size of the in-scope population increases.

Table 8
Permanent and non-permanent residents by tax year

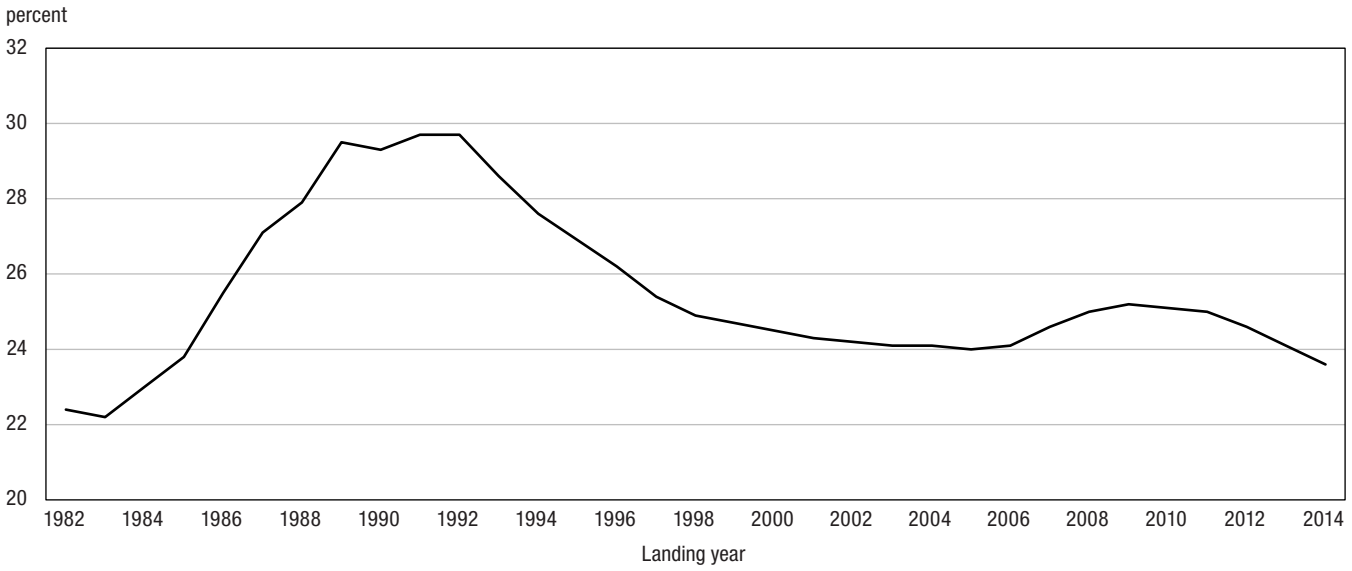
	Permanent resident	Permanent resident with non-permanent resident permit	Number of taxfilers
		number	
1982	183,200	52,900	236,100
1983	218,200	62,400	280,600
1984	255,700	76,200	331,900
1985	289,700	90,700	380,400
1986	346,700	118,800	465,500
1987	404,700	150,600	555,300
1988	493,600	191,000	684,600
1989	603,600	252,800	856,400
1990	721,900	298,800	1,020,700
1991	818,000	346,400	1,164,400
1992	922,800	389,100	1,311,900
1993	1,064,000	427,100	1,491,100
1994	1,184,700	452,000	1,636,700
1995	1,296,500	476,900	1,773,400
1996	1,404,300	497,300	1,901,600
1997	1,518,900	517,500	2,036,400
1998	1,619,300	537,400	2,156,700
1999	1,744,000	572,100	2,316,100
2000	1,888,600	611,200	2,499,800
2001	2,047,500	659,000	2,706,500
2002	2,180,100	695,900	2,876,000
2003	2,304,400	732,400	3,036,800
2004	2,438,000	773,300	3,211,300
2005	2,557,800	806,200	3,364,000
2006	2,710,700	860,900	3,571,600
2007	2,837,600	923,800	3,761,400
2008	2,965,900	989,300	3,955,200
2009	3,084,300	1,040,200	4,124,500
2010	3,212,400	1,077,600	4,290,000
2011	3,344,200	1,112,500	4,456,700
2012	3,463,000	1,130,800	4,593,800
2013	3,595,000	1,138,900	4,733,900
2014	3,701,100	1,140,300	4,841,400
Total taxfilers	**4,515,500**	**1,392,900**	**5,908,400**
Total non-taxfilers	**1,159,000**	**91,700**	**1,250,700**
		percent	
Percent of taxfilers	**79.6**	**93.8**	**82.5**

Source: Statistics Canada, 2014 Longitudinal Immigration Database.

It is possible for an immigrant to be linked without having any T1FF records, since different auxiliary files are used to link immigrants and non-permanent residents. For example, children are not expected to be taxfilers, but could be linked to the Linkage Control File (see Section 2.3 for description of the LCF). The linkage rate for the immigrants who landed in 1980 or thereafter was 89%, while at least one T1FF was available for 82.5% of these immigrants.

Chart 7 shows that the proportion of permanent residents who were non-permanent residents prior to landing by tax year varies from a low of 22.2% for the 1983 tax year to a high of 29.7% for the 1991 and 1992 tax years. Since the late 1990s, this proportion has been stable at about 25%.

Chart 7
Percentage of permanent residents who were non-permanent residents prior to landing, by tax year



Source: Statistics Canada, 2014 Longitudinal Immigration Database.

An immigrant who filed taxes for a given year will not necessarily file taxes the next year. For example, if Person A landed in 1983, this individual might be found on tax files from 1984 to 1999, but not be found on the 2000 file, and again be found on the 2001 to 2013 files. For example, 34% of filers from the 1980 cohort had tax files available for all years. Out-migration, death and late filing are some of the reasons immigrant filers might stop filing permanently or for some years.

Most immigrants file taxes for the first time in the year they land or one year before or after. For example, of the 250,800 immigrants who landed in 2006, 101,530 (40.5%) filed taxes for the first time in 2006, while 16,010 (6.3%) did so in 2007 and 3,135 (1.25%) did so in 2014.

7.5 Quality assessment of the Integrated Permanent and Non-permanent Resident File (PNRF)

A validation of the content of the 2014 PNRF was done. While landing and tax data are collected mandatorily from those in scope, some fields may not have been completed. They could be left empty because the information was not deemed to be of high importance (voluntary, as opposed to mandatory, fields), because the response was unknown, or for other reasons, unbeknownst to database users (e.g., refusal) (McLeish 2011). Item non-response can present issues when one is considering the IMDB for statistical purposes, including the following:

1. If the database user is interested in producing a sample based on characteristics for which there are missing records, there will be coverage error (i.e., those being included in the sampling frame may not be representative of the target population).
2. If the non-response is non-ignorable (i.e., the fact that information is missing is not a random occurrence; the fact that there is no response is indicative of what the response would have been), any analysis using those variables would be biased.

The presence of missing variables and invalid values was assessed. The numbers presented in this section are rounded. Invalid values are either inconsistent or not listed in the metadata tables available to users (see the immigration component of the data dictionary appendix). Most of the quality issues listed in Table 9 are for data collected in the 1980s and 1990s. It should be noted that some seemingly valid values may be erroneous as well.

The variable **Case Identification Number** (CASE_ID) has item response rates generally in the high 90% range (usually over 99%). However, for some landing years, the response rate drops significantly (to as low as 80% in 1991 and 1992). Therefore, any analysis using this variable for all landing years will under-represent those years where the item non-response is higher (e.g., 1986, 1987, 1990, 1991, 1992, 1993). No detection of invalid values was performed for the variable Case Identification Number (CASE_ID).

The variable **Landing_age** was defined as invalid when it was greater than 99, although it is possible in some instances that these values are accurate. It should be noted that, according to the values for this variable, the number of immigrants who landed after age 99 was much higher between 1986 and 1994 than the other landing years. This could be the result of a data capture issue.

In the 2014 PNRF, 10 records had a **birth year** prior to 1880, and 5 records had a landing year that preceded the birth year. The **gender** is missing for 570 immigrants, mainly for individuals who landed in the 1980s.

The variables related to country have quality issues as well. The **country of birth** is missing for some records in almost all landing years. For example, values are missing for over 100 records in each of the years from 1985 to 1993. The **country of citizenship** is missing for fewer than 10 records per landing year for most years (with the exception of 2004, 2005 and 2006, where over 40 records were missing per landing year). The **country of residence** is missing for many landing records from 2013 (this value is missing for 920 records, or 0.5% of landings taking place that year) and 2014 (this value is missing for 3500 records, or 2% of landings taking place that year). This can be explained by the fact that this field became non-mandatory in 2013.

The **education variables** after imputation (see Section 6.3) have over 150 missing values per landing year from 1980 to 1984; this translates as a rate of missing values per landing year of less than 0.5%.

The percentage of valid responses for the **occupation variables** ranges from 98% to 99.9% for all landing years.

The variables **Family_Status, Mother_Tongue, Official_Language, Point_of_Service, Point_of_Entry, CSQ_IND and Destination_Province** have most of their missing values for records with a landing year prior to 1999. **Mother_Tongue** is missing for 450 records from the 2011 landings.

The **year and month of death** was missing for some individuals identified as deceased (Death_Indicator=1). The value “9999” was assigned to Death_Year and the value “99” was assigned to Death_Month in cases where the year and month of death were unknown.

Table 9
Quality assessment of the Integrated Permanent and Non-permanent Resident File

	Valid responses		Blanks		Invalid responses	
PNRF variables	number	percent	number	percent	number	percent
Case_ID	5,737,500	97.11	170,900	2.89	0	0.00
Landing_age	5,906,500	99.97	10	0.00	1,900	0.03
Birth_Year	5,908,400	100.00	0	0.00	15	0.00
Gender	5,907,900	99.99	570	0.01	0	0.00
Country_Birth	5,906,100	99.96	2,400	0.04	0	0.00
Country_Citizenship	5,908,000	99.99	430	0.01	0	0.00
Country_Residence	5,903,200	99.91	5,200	0.09	0	0.00
Country_residence_world_area	5,903,200	99.91	5,270	0.09	0	0.00
Education_Qualification	5,906,800	99.97	1,700	0.03	0	0.00
Level_of_Education	5,907,000	99.98	1,400	0.02	0	0.00
Years_of_Schooling	5,906,800	99.97	1,600	0.03	0	0.00
Landing_age_6_groups	5,908,400	100.00	10	0.00	0	0.00
Landing_age_9_groups	5,908,400	100.00	10	0.00	0	0.00
Occupation_ID	5,904,000	99.92	4,500	0.08	8	0.00
NOC5-NOC2	5,858,300	99.15	50,200	0.85	0	0.00
LM_Intention	5,859,100	99.16	49,400	0.84	0	0.00
Family_Status	5,906,100	99.96	2,300	0.04	0	0.00
Family_Status_rollup	5,906,100	99.96	2,300	0.04	0	0.00
Marital_status	5,907,800	99.99	670	0.01	0	0.00
Marital_status_rollup	5,907,800	99.99	670	0.01	0	0.00
Mother_Tongue	5,906,100	99.96	2,300	0.04	0	0.00
Official_Language	5,906,300	99.96	2,100	0.04	0	0.00
Point_of_Service	5,906,900	99.97	1,500	0.03	0	0.00
Point_of_Entry	5,907,200	99.98	1,200	0.02	0	0.00
Special_Program	1,131,400	19.15	4,777,100	80.85	0	0.00
CSQ_ind	5,908,300	100.00	180	0.00	0	0.00
Destination_CD	5,820,500	98.51	87,900	1.49	0	0.00
Destination_CMA	5,820,500	98.51	87,900	1.49	0	0.00
Destination_CSD	5,820,500	98.51	87,900	1.49	0	0.00
Destination_Province	5,906,300	99.96	2,100	0.04	0	0.00
Permits and NPR-specific variables	1,345,000	100.00	0	0.00	0	0.00
Death_Year and Death_Month	208,900	99.51	1,030	0.49	0	0.00

Notes: PNRF: Integrated Permanent and Non-permanent Resident File. NPR: non-permanent resident Only variables with missing or invalid values were included in the table. All numbers are rounded.
Source: Statistics Canada, Longitudinal Immigration Database.

8 Comparability

8.1 Historical coverage changes

Over the years, the coverage and content of the IMDB has evolved. The original IMDB_T1FF files included only data on immigrants who landed in Canada in 1980 or thereafter. Since the 2013 IMDB release, for the 1982 tax year and subsequent tax years, non-permanent resident filers were added to the IMDB_T1FF files. As a result of this change, it is now possible to have temporary resident permit information for immigrants with pre-landing experience in Canada.

In 2012, the IMDB underwent a redesign. Coverage of the IMDB was modified to include in the database immigrants who obtained landed immigrant status in 1980 or thereafter and have filed at least one tax return since 1982, regardless of whether or not they filed taxes after landing. The IMDB initially included only individuals who obtained landed immigrant status in 1980 or subsequent years and had filed at least one tax return after becoming landed immigrants. Prior to this cycle, the IMDB included up to the first 16 years of tax files belonging to a given permanent resident (Dryburgh 2004). This cap on the number of tax files for a given individual no longer applies.

The tax data included in the IMDB initially came mainly from T1 forms, and only a select number of key tax variables at the person level were retained. For the 2006 IMDB and subsequent iterations of the IMDB, files in the T1FF for 1982 and subsequent years were used and resulted in an initial linkage rate of 80%. From this point in time, the IMDB excluded the 1980 and 1981 tax files since information for these years is not available in the T1FF.

The Field Operations Support System (FOSS) was initially used to gather the immigration data included in the IMDB. For the 2013 immigration year and subsequent immigration years, the Global Case Management System (GCMS) will be used. As a result some variables have ceased to be provided by IRCC. These legacy variables will be available on the file PNRF_extra; they are listed in the immigration component of the IMDB data dictionary.

8.2 Methodological changes

The methodology used to perform the record linkage has been modified over the years.

The initial IMDB linkage rate was 55% for the 1995 IMDB (Langlois and Dougherty 1997), but the tools and methods used to perform the record linkages have evolved. This explains the improvement in linkage rates through the years.

In the late 2000s, the linkage rate was approximately 81%. For the 2012 IMDB, information on dependents was used to perform the record linkage; this allowed for linking a greater proportion of immigrant children. This information was available from the Canadian Child Tax Benefit (CCTB) file. It is to be noted that the addition of children does not improve the taxfiler rate. As a result of the methodological changes, the linkage rate of the 2014 IMDB was 89%.

8.3 Historical database content changes

Please refer to IMDB dictionaries (immigration and tax components) for a complete description of file content. Some key recent IMDB content modifications are listed below.

In 2012, the 2009 IMDB underwent a redesign; a flag was added to identify outliers on the T1FF files being created. The spouse identification number (SP_IDI) variable was introduced in the 2010 IMDB, allowing for the identification of immigrants with immigrant spouses. The year and month of death were added to the 2013 IMDB; this allowed for the identification of immigrants admitted to Canada in 1980 or thereafter who were deceased. Following the addition of non-permanent resident data to the IMDB, some temporary resident permit details (type, effective dates, etc.) have been available since the 2013 IMDB.

8.4 Comparability with other immigration data sources

The IMDB is one of many statistical programs that can serve to produce estimates pertaining to the immigrant population. In some instances, these estimates will differ as a result of a number of factors, such as coverage and limitations due to the type of data (administrative data versus survey data versus census data). Some of these statistical programs and differences with the IMDB are described in this section. The 2013 IMDB is used in performing the comparisons.

8.4.1 Longitudinal Administrative Databank (LAD)

The Longitudinal Administrative Databank[11] (LAD) consists of a 20% longitudinal sample of Canadian taxfilers. It is linked to the IMDB to include a sample of 20% of the IMDB record and to add immigrant-specific variables, such as landing year, immigration category, and marital status at landing. It contains information about individuals and census families. It is useful for longitudinal analysis, which compares immigrant income and mobility with those of Canadian taxfilers. Any analysis comparing immigrant taxfilers to the Canadian taxfiler population should employ this dataset.

It is to be noted that the LAD contains fewer immigration variables than the IMDB. For example, pre-landing information, such as the number of work permits and study permits, is not available in the LAD. Landing information, including the intended occupation and the destination province, is also not available in the LAD.

Table 10 contains the mean and median total income (XTIRC) from the 2012 tax year of immigrants who landed during the period from 1982 to 2013, by gender, illustrating how comparable the estimates produced from these databases are. The mean and median total income by gender, as expected, are similar for both data sources. The differences can be explained by the fact the LAD is a 20% sample of the Canadian population and the fact that the IMDB is a census of linked immigrant taxfilers admitted to Canada since 1980. The population counts are different, but neither sources should be used for population counts, the LAD being a sample and the IMDB being limited to immigrant taxfilers. The population of the LAD is estimated by multiplying the records by a weight of 5.

Table 10
Comparability of the 2012 total income between the LAD and the IMDB for immigrants who landed in any year from 1982 to 2013

	Male			Female			Total		
	Population	Mean	Median	Population	Mean	Median	Population	Mean	Median
	number	dollars		number	dollars		number	dollars	
Individual									
IMDB	2,776,700	41,900	29,400	2,906,000	28,700	20,600	5,682,690	35,000	24,200
LAD	2,686,300	41,700	29,200	2,803,100	28,700	20,500	5,489,390	34,900	24,100
Family									
IMDB	...	73,700	56,300	...	69,900	51,400	...	71,700	53,700
LAD	...	73,800	56,500	...	69,700	51,500	...	71,700	53,900

... not applicable
Note: IMDB: Longitudinal Immigration Database; LAD: Longitudinal Administrative Databank.
Source: Statistics Canada, 2013 Longitudinal Immigration Database and 2013 Longitudinal Administrative Databank.

In Table 11, the comparability was restricted to the 2012 total income of immigrants who landed in 2011. The estimated differences observed between the IMDB and the LAD for this group are greater than those observed for the immigrant population that landed in any year from 1980 to 2013. This could be explained by the fact that the population of interest is smaller and more specific. The LAD estimates are derived from the records included in the 20% sample of immigrants who landed in 2011. These records do not always correspond to the 20% of the specific population in the IMDB. They are likely to constitute a smaller proportion of the specific population in the IMDB, as the sample was not drawn to be representative of this specific population. The IMDB estimates are derived from the linked immigrant population who landed in 2011 and filed taxes in 2012. Thus, the estimates from LAD may take on slightly different values than the IMDB when subsets of populations are examined.

11. For more details on the LAD, please refer to the description available on the Statistics Canada website.

Table 11
Comparability of mean and median 2012 total income for immigrants who landed in 2011

	Male		Female		Total	
	Mean	Median	Mean	Median	Mean	Median
	dollars					
Individual						
IMDB	30,100	22,400	18,900	14,100	24,300	17,800
LAD	29,500	22,100	18,700	13,900	23,900	17,500
Family						
IMDB	49,900	39,300	48,200	37,100	49,000	38,200
LAD	49,300	39,000	48,000	36,600	48,600	37,800

Note: IMDB: Longitudinal Immigration Database; LAD: Longitudinal Administrative Databank.
Source: Statistics Canada, 2013 Longitudinal Immigration Database and 2013 Longitudinal Administrative Databank.

8.4.2 Census

The census long form and the 2011 National Household Survey (NHS) collect data on immigrants. These data are collected for a proportion of the population (refer to Census Program description for exact proportion, as this value has differed throughout time). The place of birth, place of birth of parents, immigration status, year of immigration, age at immigration, and citizenship are collected. Since the 2016 Census immigration category is also available. The Census collects data on first-, second-, and older-generation Canadians, whereas the IMDB collects only data on newcomers and their families. The Census also contains data on visible minorities, education, housing and language for the census year although, unless the landing year is a Census year, it holds no record of this information at landing. The Census does not allow longitudinal study of the economic outcomes or long-term mobility of immigrants. More details on the Census Program are available on the Statistics Canada website.

The 2011 National Household Survey (NHS) estimated that over 4.6 million immigrants living in Canada in 2011 had landed during the period from 1981 to 2011 (Source: CANSIM table 99-010-X2011026). Table 12 compares the estimates of immigrant populations by landing decades from NHS and the PNRF. The 2013 PNRF should not be used to estimate population counts, even after identified death records are removed. Doing so would result in an overestimation of the immigrant population living in Canada who were admitted during the period from 1981 to 2011 because the PNRF does not take into account emigration. Also, the PNRF is a subset of the immigrant population, as only taxfilers are included in this file. This may account for lower population counts in the PNRF than in the NHS for the most recent cohort of immigrants (2001 to 2011). Deaths shown in Table 12 are based on the Death_indicator (described in Section 7.2.2).

Table 12
Comparability of population counts between the Longitudinal Immigration Database and the National Household Survey

	NHS counts	2013 PNRF counts
Landing decade	number	
1981 to 1990	949,890	1,052,650
1991 to 2000	1,539,055	1,896,235
2001 to 2011	2,154,985	2,120,290
Total	**4,643,930**	**5,069,175**

Note: IMDB: Longitudinal Immigration Database; NHS: National Household Survey, PNRF: Integrated Permanent and Non-permanent Resident File.
Source: Statistics Canada, 2013 Longitudinal Immigration Database and National Household Survey, 2011.

8.4.3 Longitudinal Survey of Immigrants to Canada (LSIC)

The Longitudinal Survey of Immigrants to Canada (LSIC) was designed to provide information on how new immigrants adjust to life in Canada and to understand the factors that can help or hinder this adjustment. The LSIC was designed to examine the first four years of settlement. Data on immigrants aged 15 years and older who landed in Canada from abroad at any time from October 1, 2000, to September 30, 2001, were collected for three waves. The LSIC allows studies on language proficiency, housing, education, foreign credential recognition, employment, health, values and attitudes, the development and use of social networks, income, and perceptions of settlement in Canada. The IMDB contains characteristics such as education and language only at landing, whereas the LSIC allows for the evaluation of changes through time. Additional information on the LSIC is available on the Statistics Canada website.

The LSIC estimated that 164,200 immigrants aged 15 years and older landed in Canada from abroad at any time from October 1, 2000, to September 30, 2001. The estimate for the same population is 156,670 for the 2013 IMDB when calculated according to the PNRF (Table 13). Some of this difference is due to the combination of the exclusion of non-filers from the PNRF estimate and emigration not being captured in the IMDB. Part of the difference is explained by the fact that the LSIC is a survey that introduces variance estimates. As shown in Table 14, the coverage proportions by age group vary across age groups despite the LSIC population being of tax filing age. It is to be noted that the LSIC age is the age approximately six months after landing while the IMDB is the age at landing. Also, the calculation of the LSIC estimates used wave 1 weights, which were designed to estimate the number of immigrants in this cohort still living in Canada six months after landing. The lower proportion of immigrants aged 65 years and older could be due to a lower proportion of filers for this age group. The higher number of immigrants aged 15 to 24 in the IMDB than in the LSIC likely results from emigration not being accounted for.

Table 13
Gender distribution: Longitudinal Immigration Database compared to Longitudinal Survey of Ommigrants to Canada

	2003 LSIC		2013 PNRF	
	number	percent	number	percent
Male	81,550	49.7	77,640	49.6
Female	82,650	50.3	78,830	50.4
Total	**164,200**	**100.0**	**156,470**	**100.0**

Note: LSIC: Longitudinal Survey of Immigrants to Canada; PNRF: Integrated Permanent and Non-permanent Resident File.
Source: Statistics Canada, 2013 Longitudinal Immigration Database and Longitudinal Survey of Immigrants to Canada, Wave 1, 2003.

Table 14
Age group distribution: Longitudinal Immigration Database compared to Longitudinal Survey of Immigrants to Canada

	2003 LSIC		2013 PNRF	
Age group	number	percent	number	percent
15 to 24	26,700	16.3	28,000	17.9
25 to 34	65,500	39.9	63,100	40.3
35 to 49	54,000	32.9	49,000	31.3
50 to 64	12,900	7.8	12,300	7.8
65 and older	5,100	3.1	4,100	2.6
Total	**164,200**	**100.0**	**156,500**	**100.0**

Note: LSIC: Longitudinal Survey of Immigrants to Canada; PNRF: Integrated Permanent and Non-permanent Resident File.
Source: Statistics Canada, 2013 Longitudinal Immigration Database and Longitudinal Survey of Immigrants to Canada, Wave 1, 2003.

9 New analyses possible with the IMDB

The IMDB was created to allow analysis on immigration-related topics, and this section gives an overview of possible analyses with the additional information now available in the IMDB. As described in this report, the content of the IMDB has evolved; this has increased its analytical capabilities. Below are some examples of analyses that could benefit from the IMDB.

9.1 Analytical possibilities with non-permanent resident data

The addition of non-permanent resident data expands the scope of analysis currently possible with the IMDB. The number and type of permits obtained prior to landing can be used to establish the pre-landing profile of immigrants with pre-landing experience. Also, the non-permanent resident data can be used to identify permanent residents with and without the pre-landing experience in Canada. By comparing these populations (with and without pre-landing Canadian experience), it becomes possible to assess the impacts of pre-landing Canadian experience on the economic outcomes and mobility patterns of immigrants. The specific sociodemographic profile at time of temporary resident permit issuance is now available, and makes it possible to evaluate economic outcome and mobility prior to landing. Changes in intended occupation, skill level and level of study through temporary resident permits are also available.

9.2 Analytical possibilities with data on deaths

The addition of the death flag, death year, and death month variables to the PNRF makes it possible to estimate the proportion of records included in the IMDB that belong to deceased immigrants. These variables will complement the year of death (YOD) variable included in the tax files. YOD is available only in instances where a T1 form was filed posthumously on behalf of the deceased, whereas the year and month of death are available for any record linked to the mortality dataset regardless of tax filing profile. New possible analyses may include the evaluation of economic profiles of immigrants prior to their death and the study of life expectancies after landing by immigration category and economic profile.

10 Summary

The IMDB is a dataset combining immigration and tax records created for the purpose of performing socio-economic and mobility analysis on immigrants (with or without pre-landing experience) who landed in Canada in 1980 or thereafter. The IMDB allows for analysis pertaining to taxfilers. This is important given that taxfilers can have a different profile than non-filers. This technical report was produced to give a thorough description of IMDB data quality and recent changes to this database.

How to access the IMDB

As described in Section 6, several products are available to researchers. They can be accessed via the Statistics Canada website by selecting “Ethnic diversity and immigration” under ”Subjects” and Longitudinal Immigration Database (IMDB) under Featured products.

Appendix

A) Links to key IMDB documents and web pages

Dictionaries (tax and immigration component):

Available to data users or upon request by contacting Statistics Canada by email at STATCAN.infostats-infostats.STATCAN@canada.ca)

IMDB CANSIM tables:

http://www5.statcan.gc.ca/COR-COR/COR-COR/objList?lang=eng&srcObjType=SDDS&srcObjId=5057&tgtObjType=ARRAY

Historical IMDB:

http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getInstanceList&Id=7196

IMDB releases in *The Daily*:

http://www5.statcan.gc.ca/COR-COR/COR-COR/objList?lang=eng&srcObjType=SDDS&srcObjId=5057&tgtObjType=DAILYART

Analysis using the IMDB:

http://www5.statcan.gc.ca/COR-COR/COR-COR/objList?lang=eng&srcObjType=SDDS&srcObjId=5057&tgtObjType=STUDIES

The Consumer Price Index (62-001-X):

http://www5.statcan.gc.ca/olc-cel/olc.action?lang=en&ObjId=62-001-X&ObjType=2

Description of the annual Income Estimates for Census Families and Individuals (T1 Family File):

http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4105&lang=fr&db=imdb&adm=8&dis=2.

B) Coverage

The 2014 IMDB was used to produce these counts. Filers are linked immigrants who have filed a tax return at least once since 1982.

Table 15
Distribution of taxfilers and non-taxfilers by landing year

Landing year	Taxfilers[1]				Non-taxfilers				Total			
	Immigrants	PR	PR I NPR	Deaths	Immigrants	PR	PR I NPR	Deaths	Immigrants	PR	PR I NPR	Taxfilers
	number											percent
1980	114,200	109,700	4,500	15,600	28,900	28,200	700	1,700	143,100	137,900	5,200	79.8
1981	101,500	89,300	12,200	13,900	27,000	25,700	1,300	1,400	128,600	115,000	13,500	78.9
1982	98,100	83,400	14,700	12,500	23,000	21,700	1,300	1,300	121,100	105,100	16,000	81.0
1983	72,800	59,500	13,300	10,100	16,200	15,200	1,000	1,100	89,000	74,700	14,300	81.8
1984	73,500	58,200	15,300	9,300	14,500	13,500	1,100	1,000	88,000	71,700	16,400	83.5
1985	71,400	56,900	14,500	8,200	12,600	12,400	150	700	83,900	69,300	14,650	85.1
1986	84,800	63,300	21,500	8,300	13,900	13,700	300	600	98,800	77,000	21,800	85.8
1987	131,400	100,100	31,400	10,300	19,700	19,400	400	700	151,200	119,500	31,800	86.9
1988	138,800	122,900	15,900	9,700	22,000	21,100	900	600	160,700	144,000	16,800	86.4
1989	165,500	142,200	23,300	10,000	25,200	23,200	2,000	600	190,700	165,400	25,300	86.8
1990	186,100	153,500	32,600	10,600	29,300	26,900	2,400	500	215,400	180,400	35,000	86.4
1991	203,100	133,900	69,200	11,600	28,800	24,400	4,400	600	231,800	158,300	73,600	87.6
1992	223,800	141,500	82,300	11,600	30,100	25,100	5,000	500	253,900	166,600	87,300	88.1
1993	226,900	163,300	63,600	10,900	28,700	24,600	4,100	500	255,700	187,900	67,700	88.7
1994	196,000	158,400	37,600	8,800	27,600	25,300	2,200	400	223,600	183,700	39,800	87.7
1995	186,600	147,100	39,500	7,100	25,500	23,400	2,200	400	212,100	170,500	41,700	88.0
1996	196,300	155,000	41,300	6,000	29,000	26,700	2,300	300	225,300	181,700	43,600	87.1
1997	187,400	152,300	35,100	5,000	28,000	26,200	1,800	200	215,400	178,500	36,900	87.0
1998	154,100	122,700	31,400	3,600	19,600	18,000	1,600	200	173,700	140,700	33,000	88.7
1999	167,200	133,700	33,500	3,400	22,100	20,300	1,800	160	189,400	154,000	35,300	88.3
2000	199,000	160,600	38,300	3,400	27,800	25,500	2,300	200	226,700	186,100	40,600	87.8
2001	216,100	178,600	37,500	3,500	33,700	31,200	2,500	180	249,800	209,800	40,000	86.5
2002	193,800	161,800	32,000	3,200	34,400	32,500	1,900	180	228,200	194,300	33,900	84.9
2003	184,700	152,500	32,200	2,600	35,800	34,000	1,900	150	220,500	186,500	34,100	83.8
2004	194,800	152,500	42,300	2,100	40,300	38,100	2,200	130	235,100	190,600	44,500	82.9
2005	212,700	164,000	48,700	1,700	48,800	45,800	2,900	80	261,400	209,800	51,600	81.4
2006	203,800	150,600	53,100	1,800	47,000	43,800	3,200	110	250,800	194,400	56,300	81.3
2007	188,900	137,600	51,300	1,400	47,000	43,800	3,200	90	235,900	181,400	54,500	80.1
2008	193,600	140,000	53,600	1,200	52,700	49,700	3,100	90	246,300	189,700	56,700	78.6
2009	196,600	139,800	56,700	1,000	54,800	51,300	3,500	90	251,300	191,100	60,200	78.2
2010	210,800	151,400	59,300	700	69,000	64,600	4,400	70	279,700	216,000	63,700	75.4
2011	183,700	130,500	53,200	500	64,100	59,300	4,800	40	247,800	189,800	58,000	74.1
2012	189,000	129,300	59,700	400	67,800	62,600	5,200	10	256,900	191,900	64,900	73.6
2013	185,800	122,700	63,100	300	72,200	66,200	5,900	10	258,000	188,900	69,000	72.0
2014	175,700	96,700	79,000	40	83,600	76,000	7,600	10	259,200	172,700	86,600	67.8
Total	**5,908,500**	**4,515,000**	**1,392,900**	**210,340**	**1,250,700**	**1,159,000**	**91,700**	**14,900**	**7,159,200**	**5,674,000**	**1,484,600**	**82.5**

1. Taxfilers are linked immigrants who have filed taxes at least once since 1982.

Notes: PR: permanent resident; NPR: non-permanent resident. All counts are rounded.

Source: Statistics Canada, 2014 Longitudinal Immigration Database.

Table 16
Proportion of linked taxfilers by age group at landing, sex and admission decade

	Age at landing						
	0 to 14	15 to 24	25 to 34	35 to 44	45 to 64	65 and older	Total
Sex and cohorts	percent						
1980 to 1989 cohorts							
Male	77.8	91.9	93.3	91.4	81.0	57.2	86.5
Female	76.1	84.3	87.4	87.7	75.4	55.2	81.2
Total	**77.0**	**88.0**	**90.4**	**89.5**	**77.9**	**56.1**	**83.8**
1990 to 1999 cohorts							
Male	78.5	92.9	92.8	92.1	88.3	74.2	88.5
Female	77.0	91.0	91.2	91.0	85.6	73.7	87.1
Total	**77.7**	**91.9**	**92.0**	**91.6**	**86.8**	**73.9**	**87.8**
2000 to 2009 cohorts							
Male	43.6	95.0	92.4	92.9	92.8	88.6	81.8
Female	42.6	94.7	93.3	93.4	92.6	87.3	83.1
Total	**43.1**	**94.8**	**92.9**	**93.2**	**92.7**	**88.0**	**82.5**
2010 to 2014 cohorts							
Male	4.4	82.8	93.2	90.5	86.7	82.7	71.3
Female	4.6	84.8	93.1	91.1	86.1	81.5	73.8
Total	**4.5**	**83.9**	**93.1**	**90.8**	**86.4**	**82.1**	**72.6**

Source: Statistics Canada, 2014 Longitudinal Immigration Database.

C) Previous analysis

Since its creation, the IMDB has been used to produce several analyses. The following is a summary of some Statistics Canada studies that have made use of the IMDB.

In recent years, several releases in *The Daily* have featured the IMDB. The subjects discussed include changes in the regional distribution of new immigrants to Canada, income and mobility of immigrants, immigrants in the hinterlands, and immigrants who leave Canada. These articles are accessible via the Statistics Canada website.

Papers using the IMDB have been published in the *Perspectives on Labour and Income* publication series (75-001-X) and the Analytical Studies Branch Paper Series. Among the topics covered were the income of immigrants who pursue postsecondary education in Canada, and the earnings advantage of landed immigrants who were previously temporary residents in Canada.

D) Best practices and tips for analysts

D.1 Programming tips

This section provides programming information for individuals who want to have a better understanding of the programming structure used to access data from IMDB files. Please note that individuals may conduct their own programming. There are two types of IMDB files—the yearly IMDB data files and the immigration data (for more details on IMDB files, refer to Section 3). IMDB tax variables are identified with a variable name that consists of three parts: (1) the acronym name as described in the IMDB tax data dictionary, (2) the aggregate level (I or F), and (3) the year (the four-digit year extension exists in most, but not all, cases).

Example: The interest and investment income at the individual level for 2014 would be named INVI_I2014.

Observations in the IMDB files are sorted according to a variable, **IMDB_id** (note that there is no year extension for this variable), which enables users to maintain a link across years. Data access takes place by means of the SAS programming language. A sample SAS program designed to access IMDB data is provided below. The samples below are created to perform the following task:

> "retrieving the number of **Social Assistance (SA) recipients** for immigrants who landed between 2000 and 2005, living in Ontario between 2010 and 2012, and did not have any earnings appearing on their T4 slips by sex and year (2010 to 2012)"

Researchers who are new to the IMDB are encouraged to go through this sample SAS program. There are generally three components in the sample.

1. Library set-up: The library assignments on the first two lines are the locations for the input files (first line) and the output files (the second line).
2. Steps to generate a working dataset:
   a. The input files are stored in SAS format and can therefore be accessed with a **SET** or **MERGE** statement.
   b. This program is aimed at retrieving the number of **Social Assistance (SA) recipients** for immigrants who:
      i. landed at any time from 2000 to 2005
      ii. lived in Ontario from 2010 to 2012
      iii. did not have any earnings on their T4 slips

      And generate the number of SA recipients by **sex** and **year** (in this case, 2010 to 2012).
3. The dataset used to produce the number of the SA recipients: The part, which starts with "proc freq," produces the numbers of interest as they are specified in the rest. At the end of the program, four tables are created from the output data file.

It is generally recommended that programs use the variables available in the PNRF rather than the yearly tax files for consistency. For example, the sample program uses the variable **GENDER_ROLLUP**, a variable found in the **PNRF**, rather than **SXCO_I&year**, the variable found in the yearly IMDB_T1FF. In this program, only individuals who have filed every year from 2010 to 2012 are selected.

When programming in SAS, one should keep in mind the distinction between missing values and zeros in numeric fields. With SAS, most mathematical operations performed with missing values will return missing values. In IMDB, in years that an individual is present, numeric variables not relevant to that individual have a value of "0" (zero). For example, if a person without a spouse filed in 2010, the value for RRSPSI2010 (contributions to a spouse's RRSP) should be "0" (zero). If that individual did not file in 2010, the value will be missing.

Sample IMDB program

```
*Sample SAS program using the IMDB;

libname source1 '\\f8prod05\cic\1.Database'; * location of IMDB files ;

libname Out '\\f8prod05\cic\3.workarea'; * user's directory ;

* This sample program's objective is to use the IMDB to retrieve the number of Social
Assistance (SA) recipients in Ontario that did not have any earnings appearing on
their T4 slips, according to sex and year (in this case, 2010 to 2012). Data for
provinces and earnings are from the yearly IMDB files whereas the sex variable is from
the PNRF _ 2014. ;

* The first step is to create a datafile containing all the information that we need to
produce our tables. This datafile will be called SAOnt and will be saved in the 'out'
directory. The Longitudinal Identifier Number (IMDB _ ID) is used to merge the annual
IMDB datasets. ;

data out.SAOnt;
merge
source1.imdb _ t1ff _ 2010 _ outliers(where=(prco _ i2010 = 5) in=a keep=imdb _ id prco _
i2010 saspyf2010 t4e _ _ i2010)

source1.imdb _ t1ff _ 2011 _ outliers(where=(prco _ i2011 = 5) in=b keep= imdb _ id prco _
i2011 saspyf2011 t4e _ _ i2011)
```

```
source1.imdb _ t1ff _ 2012 _ outliers(where=(prco _ i2012 = 5) in=c keep= imdb _ id prco _
i2012 saspyf2012 t4e _ _ i2012)

source1.pnrf _ 2014(keep= imdb _ id gender _ rollup landing _ year immigration _
category);

by IMDB _ id ;

If a and b and c and (landing _ year>=2000 and landing _ year<=2005);

*person must be taxfiler in all three years and must have landed between 2000 and 2005
(population of interest);

* We create a flag variable that identifies the SA receipients for each year. The
result is three variables,
flag _ sa2010, flag _ sa2011 and flag _ sa2012, taking a value of either 1 or 0.;
If (t4e _ _ i2010=0 and saspyf2010>0) then flag _ sa2010 = 1 ;
else flag _ sa2010 = 0 ;
if (t4e _ _ i2011=0 and saspyf2011>0) then flag _ sa2011 = 1 ;
else flag _ sa2011 = 0 ;
if (t4e _ _ i2012=0 and saspyf2012>0) then flag _ sa2012 = 1 ;
else flag _ sa2012 = 0 ;
run;

* The SAS 'freq' procedure is used to produce our tables. We would also need to make
sure that confidentiality guidelines standards are respected. ;

proc freq data = out.SAOnt;
tables immigration _ category*flag _ sa2010*flag _ sa2011*flag _ sa2012
gender _ rollup*flag _ sa2010*flag _ sa2011*flag _ sa2012 /missing ;
run ;
* End of the sample program;
```

D.2 Creating a cohort

Prior to starting an analysis, the cohort of interest needs to be defined. The cohort can be restricted by landing year, geography, or any other variable of interest (e.g., admission category or gender) according to the researcher's need. A clearly defined single cohort should be followed to allow comparability. For example, a researcher might be interested in women who landed in 2000 and who lived in a family that received social assistance in 2001 (Table 17). A study question regarding this cohort could be "What proportion of this cohort received social assistance in the following two years (2002 and 2003)?" It is worth noting that the Canada Revenue Agency (CRA) requires the spouse with the higher net income to report the social assistance payment. As a result, measurement on social assistance (SASPY_F), even for individuals, is best reported with the family-level information.

Table 17
Example—Women who landed in 2000 and received social assistance (SASPY_F) in 2001

			SASPY_F2001	SASPY_F2002	SASPY_F2003
IMDB_ID	Landing year	Gender		dollars	
IM583	2000	Female	20,500	19,000	14,000
IM145	2000	Female	3,000	0	0
IM548	2000	Female	11,500	13,800	0
IM798	2000	Female	16,000	18,000	8,000
IM961	2000	Female	10,000	0	0
IM967	2000	Female	9,500	0	0
IM110	2000	Female	5,000	2,000	1,000
IM125	2000	Female	1,000	0	200

Source: Statistics Canada, example from Longitudinal Immigration Database (IMDB).

D.3 Calculating retention rates

A key strength of the IMDB is the presence of geographic variables that allow for the study of mobility and retention. No other dataset contains a comparable level of detail on taxfilers annually, especially when it comes to smaller geographies. Having annual provincial, census division (CD), census metropolitan area (CMA), census agglomeration (CA), census subdivision level (CSD), and census tract level updates allows for a broad range of analyses.

Individual mobility trajectories can be studied simply by flagging changes in postal codes, and mobility trends can be calculated by studying relocations at specific levels of geography. For example, CSD-level mobility (year-to-year changes in CSD) and provincial mobility (year-to-year changes in province) significantly vary by a number of immigrant characteristics, such as age and admission class. These geographies are derived from the postal code (IMDB variable PSCO at the individual and family levels). The postal code is a six-character alphanumeric code that locates the point of delivery of mail addressed to post office customers in Canada. See Section 3.4.1 for a description of the geography variables.

In the example below (Table 18), the researcher is interested in mobility until 2002. IM798, IM961, IM967 and IM110 could be excluded from the mobility study because data (or files) are missing.

Table 18
Example—mobility until 2002 of immigrants who landed in 2000

IMDB_ID	Landing year	Destination province	PRCO 2000	PRCO 2001	PRCO 2002
IM583	2000	B.C.	B.C.	B.C.	B.C.
IM145	2000	Alta.	Alta.	Sask.	Sask.
IM548	2000	Alta.	Ont.	Ont.	Ont.
IM798	2000	Ont.	..	Ont.	Ont.
IM961	2000	N.B.	N.B.	N.B.	..
IM967	2000	Ont.	..	Alta.	Ont.
IM110	2000	..	Que.	..	Que.

.. not available for a specific reference period
Note: PRCO is province of residence.
Source: Statistics Canada, example from Longitudinal Immigration Database (IMDB).

While mobility, at the individual level, is fairly straightforward, retention of immigrants in a jurisdiction can be calculated in several ways. How retention is calculated is an analytical decision based on the individual researcher's particular needs. The number of individuals retained is fairly straightforward to define—it is the number of individuals filing taxes in the jurisdiction of interest at a given time. A decision has to be made about what constitutes the initial landing cohort about which retention is calculated (the denominator in the retention rate).

The provincial rates reported in *The Daily* are defined as the proportion of immigrant taxfilers who reside in the province where they landed (defined as the province of intended destination) at a given time. For a given cohort (e.g., landing year) and a given tax year (or years since landing), the denominator is the number of taxfilers with the selected province of landing. The numerator is the number of taxfilers with the selected province of landing who are also residing in the province.

For example, using CANSIM table 054-0003 to compute retention rates three years after landing for the 2011 cohort, a researcher would choose all provinces of landing (i.e., the province of intended destination), all provinces of residence, landing year = 2011, and reference year = 2014. The table would look as follows:

Table 19
Province of residence in 2014 and province of landing, 2011 cohort

	Province of residence					
Province of landing	Total province of residence	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Quebec
	number of immigrants					
Total province of landing	**174,740**	**405**	**330**	**1,365**	**880**	**31,505**
Newfoundland and Labrador	515	325	0	5	0	5
Prince Edward Island	1,245	0	265	25	10	30
Nova Scotia	1,460	10	5	1,080	10	25
New Brunswick	1,340	0	10	35	750	55
Quebec	36,275	10	10	35	15	30,200
Ontario	69,135	35	25	115	70	875
Manitoba	11,190	0	0	15	0	55
Saskatchewan	6,360	0	0	0	0	20
Alberta	21,940	10	0	20	0	95
British Columbia	25,000	5	0	30	5	140
Other	280	0	0	0	0	0

	Province of residence					
Province of landing	Ontario	Manitoba	Saskatchewan	Alberta	British Columbia	Other residence
	number of immigrants					
Total province of landing	**70,590**	**9,695**	**6,120**	**26,965**	**26,390**	**500**
Newfoundland and Labrador	75	5	0	60	30	0
Prince Edward Island	560	0	0	50	295	0
Nova Scotia	185	0	5	90	30	10
New Brunswick	275	0	10	80	120	0
Quebec	3,255	40	75	1,190	1,400	45
Ontario	63,145	275	335	2,815	1,325	115
Manitoba	645	9,170	80	825	380	10
Saskatchewan	295	45	5,370	445	165	10
Alberta	810	65	140	20,170	590	35
British Columbia	1,330	85	100	1,200	22,030	70
Other	15	0	0	35	20	200

Source: Statistics Canada, CANSIM table 054-0003.

Results for Nova Scotia shed some light on the matter. A total of 1,460 individuals landed in Nova Scotia in 2011 and filed taxes in 2014. Of those, 1,080 had Nova Scotia as their province of residence in 2014. Nova Scotia's three-year retention rate would be 1,080/1,460, or about 74%. The CANSIM table also provides information on secondary migrants[12]—1,365 individuals who landed in 2011 resided in Nova Scotia in 2014, of which 1,080 intended to land in Nova Scotia, and 285 had a destination province other than Nova Scotia.

The above definition of retention assumes that the number of taxfilers with the specific province of intended destination is the total population that can be retained in a year (i.e., if all 1,460 individuals who had intended to land in Nova Scotia had filed taxes there in 2014, the province would have 100% retention). This method does not take into account late or sporadic tax filing behaviour. While the total population in the 2014 tax year for Nova Scotia is 1,460, in the 2013 tax year 1,425 immigrants who intended to land in Nova Scotia filed taxes.

One alternative is a purely longitudinal approach, where a single landing cohort is selected (according to the province of intended destination, the province of initial tax filing, or both), and the retention rate is calculated as the proportion of this cohort that is still filing taxes in the province. When the province of initial tax filing is used to define the landing cohort, it is recommended that the first tax file occur in the year the immigrants were admitted (landing year = tax year), to exclude individuals who may have first arrived elsewhere and subsequently migrated to the region before filing taxes for the first time. A further restriction can be made if a researcher is interested in the population whose destination geography matches the geography of the first tax file.

Given that a portion of each annual cohort do not file taxes for their year of landing, it may be necessary to increase the population size for a region by defining the landing cohort as anyone who first filed taxes in the region within two years of landing (i.e., first_tax_year = landing_year or landing_year+1). Allowing individuals whose first

12. Individuals who relocated within Canada after reaching their initial destination in Canada.

tax filing occurred several years after landing to be part of a "landing cohort" is not recommended, as it is possible that they first landed elsewhere but did not file taxes. It is also a good idea to exclude intermittent filers from these analyses, as their place of residence is unknown in the years for which there is no tax data. Retention calculated this way will show a gradual decline in numbers; this decline is due to immigrants who stop filing, out-migration, and death.

If researchers are interested in secondary migrants to a region, this can be found by removing individuals in the defined landing cohort from the total number of immigrants filing taxes in the region at the time of interest. Again, however, these analyses should be restricted to individuals who first filed taxes within the same time period (year 0 or year 1) to avoid mistaking late-filers for in-migrants. If the landing cohort is restricted to immigrants whose destination geography matches the geography of first tax filing, a subsequent distinction should be made between secondary migrants who first filed elsewhere (and subsequently filed in the region of interest) and immigrants who first filed in the region of interest but were subsequently recruited by other jurisdictions (or information on their intended destination is missing altogether).

The following table presents an example of a longitudinal approach to provincial retention using fictitious data, with various definitions of the initial landing cohort.

Table 20
Number of immigrant tax filers within the specified population residing in British Columbia and associated retention rate, by years since landing

Years since landing	Taxfilers who first filed taxes in B.C. in year 0	Retention rate	Taxfilers who first filed taxes in B.C. in year 0 or 1	Retention rate	Taxfilers who first filed taxes in B.C. in year 0 or 1 and province of intended destination was B.C.	Retention rate
	number	percent	number	percent	number	percent
0	20,000	100	20,000	...	17,500	...
1	18,000	90	25,000	100	19,000	100
2	17,000	85	23,000	92	18,000	95
3	16,500	83	22,000	88	17,500	92

... not applicable
Source: Statistics Canada, example from the Longitudinal Immigration Database.

In the above example, retention in British Columbia can be calculated according to three definitions of the population, and the three-year retention rate varies per the definition adhered to. Importantly, all individuals in the sample filed taxes at each point in time.

Finally, analysts should use caution when studying low-level census geographies over a long period of time, as CA and CMA boundaries change and CSDs are dropped and added. If possible, analysts should run the Postal Code Conversion File (PCCF+) program to standardize postal codes to a constant census geography.

D.4 Calculating income trajectories over time

As is the case with retention, calculating year-to-year changes in employment earnings (or, for that matter, any economic variable) requires consecutive information. For example, if a researcher wants to compare the median employment earnings of the 2000 cohort of women aged 24 to 54, 1 year after landing and 5 years since landing (Table 21), records with missing T1FF files could be removed from the analysis. The decision to remove these records would be based on the desire to evaluate the cohort's median income versus the cohort filer's median income.

Table 21
Median employment earnings of the 2000 cohort of women aged 24 to 54, 1 year after landing and 5 years since landing

IMDB_ID	Landing year	Age at landing	Gender	Employment income 2001	Employment income 2005
				dollars	
IM583	2000	34	Female	20,500	49,000
IM145	2000	53	Female	..	56,000
IM548	2000	29	Female	11,500	33,800
IM798	2000	31	Female	36,000	0
IM961	2000	42	Female	10,000	..
IM967	2000	40	Female	..	..
IM110	2000	35	Female	0	59,000

.. not available for a specific reference period
Source: Statistics Canada, example from Longitudinal Immigration Database.

Use caution when calculating the "first year in Canada" income as it might not represent a full year of taxation. For example, someone who landed in November of 2013 and filed taxes for 2013 would have only two months of income in 2013. A best practice is to use the first full year of income (landing year +1, see Table 19). One exception is pre-filers, those who filed taxes in Canada before landing and filed at landing year as well, are most likely reporting income for the entire year.

Over-time income should also be studied in constant dollars. Consequently, Canadian Price Index (CPI) adjustments should be made (Appendix D.7). This adjustment is made in the IMDB CANSIM tables.

D.5 Rounding data

Respecting the privacy of Canadians is important to Statistics Canada. Consequently, any tables produced from IMDB_TIFF files are subject to rounding. The purpose of rounding is to ensure that no small cells are released that may reveal information on specific individuals or small groups of individuals. In general, the macros will take an unrounded input dataset of various statistics (counts, means, medians, etc.) and output a rounded dataset.

The rounding rules are confidential, but the rounding macros are available to all researchers. Documentation describing how to use the macros is available. These macros are applied to the output tables of all researchers, to all external data requests, and to the released CANSIM tables.

D.6 Identifying outliers

The variable OUTLIER_IND was created to identify outliers within the T1FF (see Section 5.5). It should be used to remove outlier data from any calculation (e.g., mean, median, or regression) employing tax data. Outliers differ from one year to another, meaning that a person's data may be identified as an outlier for a given year but not for a subsequent year.

The following table gives the distribution of the outliers in the tax files for 1982 and subsequent years by type of resident for the 2014 IMDB. Less than 0.1% records were identified as outliers per tax year. The proportion of outliers increased from 1995 to 1996 as a result of updates to the outlier detection method applied to tax files for 1997 and subsequent taxation years.

Table 22
Distribution of outliers by tax year

	PR	PR with NPR permit	Total	
	number		percent	
1982	10	10	20	0.01
1983	30	0	30	0.01
1984	40	10	50	0.01
1985	30	0	40	0.01
1986	50	10	50	0.01
1987	60	10	70	0.01
1988	70	20	90	0.01
1989	70	20	90	0.01
1990	60	20	80	0.01
1991	70	20	100	0.01
1992	150	40	180	0.01
1993	150	60	210	0.01
1994	60	20	90	0.01
1995	160	70	230	0.01
1996	340	190	530	0.03
1997	360	220	580	0.03
1998	470	250	720	0.03
1999	390	230	620	0.03
2000	480	260	740	0.03
2001	440	260	700	0.03
2002	510	260	780	0.03
2003	470	240	710	0.02
2004	600	290	890	0.03
2005	660	340	1,000	0.03
2006	610	310	920	0.03
2007	590	340	930	0.02
2008	620	380	1,000	0.03
2009	710	370	1,080	0.03
2010	720	430	1,150	0.03
2011	730	430	1,150	0.03
2012	730	420	1,150	0.02
2013	790	460	1,250	0.03
2014	920	440	1,350	0.02

Note: PR: Permanent resident; NPR: Non-permanent resident.
Source: Statistics Canada, 2014 Longitudinal Immigration Database.

D.7 Adjusting income for the Consumer Price Index (CPI)

In order to take into account the cost of living, all incomes should be adjusted to the Consumer Price Index (CPI) for Canada. "The Consumer Price Index (CPI) is an indicator of changes in consumer prices experienced by Canadians. It is obtained by comparing, over time, the cost of a fixed basket of goods and services purchased by consumers. Since the basket contains goods and services of unchanging or equivalent quantity and quality, the index reflects only pure price change."[13] The adjustment factors for 2104 are available in Table 23. To transform data to constant dollars of a specific year (base year), data users need to multiply the dollar values in all but the base year by a year-specific adjustment factor. To obtain the adjustment factors, data users need to divide the CPI of the base year by the CPI of the specific year. In Table 23, the base year is 2014.

13. Http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2301.

Table 23
2014 Consumer price index adjustment factors

Year	2014 consumer price index adjustment equals 125.2 divided by number
1982	54.9
1983	58.1
1984	60.6
1985	63.0
1986	65.6
1987	68.5
1988	71.2
1989	74.8
1990	78.4
1991	82.8
1992	84.0
1993	85.6
1994	85.7
1995	87.6
1996	88.9
1997	90.4
1998	91.3
1999	92.9
2000	95.4
2001	97.8
2002	100.0
2003	102.8
2004	104.7
2005	107.0
2006	109.1
2007	111.5
2008	114.1
2009	114.4
2010	116.5
2011	119.9
2012	121.7
2013	122.8
2014	125.2

Source: Statistics Canada, CANSIM table 326-0021.

D.8 Calculating key income measures

The IMDB CANSIM tables contain several income measures. Table 22 describes which variables of the T1FF are included in their calculation.

Table 24
Description of the Longitudinal Immigration Database income measures

Measure	Components	Formula
Employment income	Earnings from T4 slips; other employment income	T4E__i + OEI__i
Self-employment income		
Since 1988	Self-employment income from business, profession, commission, farm, and fishing; limited partnership income	SEI__i + LTPI_i
Before 1988	Self-employment income from business, profession, commission, farm, and fishing	SEI__i
Investment income	Interest and investment income; dividends; capital gains/losses, net taxable*	INVi_i + XDIV_i + CLKGLi*
1982 to 1987	Capital gains/losses, net taxable	2 x CLKGLi
1988 and 1989	Capital gains/losses, net taxable	(3/2) x CLKGLi
1990 to 1999	Capital gains/losses, net taxable	(4/3) x CLKGLi
2000	Capital gains/losses, net taxable	(100/64.58) x CLKGLi
Since 2000	Capital gains/losses, net taxable	2 x CLKGLi
Employment Insurance benefits	Employment Insurance benefits	EINS_i
Social welfare benefits	Social welfare benefits (use family-level)	SASPYf
Total income	Sum of all measures described above	

Source: Statistics Canada, Longitudinal Immigration Database, CANSIM table processing.

It is to be noted that all outliers are removed from these calculations (Outlier_ind=1), that the variable Province of Residence at the End of the Year (PRCO_) is used to identify the province, and that all incomes are adjusted according to the Consumer Price Index (CPI) of the year of the most recent T1FF available. “Mean with income” is the mean income of immigrant tax-filers with income of the given type. “Median with income” is the median income of immigrant tax-filers with income of the given type.

References

Badets, J., and C. Langlois. 2000. "The Challenges of Using Administrative Data to Support Policy-Relevant Research: The Example of the Longitudinal Immigration Database (IMDB)." In *Symposium 99 - Combining Data from Different Sources, 1999*. Statistics Canada International Symposium Series: Proceedings. Statistics Canada Catalogue no. 11-522-XPE. Available at http://data.library.utoronto.ca/datapub/data/cst/cst_free/11-522-xie/1999/Symp99e.pdf (accessed March 13, 2017).

Carpentier, A., and G. Pinsonneault. 1994. *Representativeness Study of Immigrants Included in the Immigrant Data Bank (IMDB Project)*. Ministère des Affaires internationales, de l'Immigration et des Communautés culturelles. Government of Quebec.

Diaz-Papkovich, A. 2016a. *Evaluating the quality of the Permanent Resident Linkage*. Unpublished document. Ottawa: Statistics Canada.

Diaz-Papkovich, A. 2016b. *IMDB Linkage Highlights 2016*. Unpublished document. Ottawa: Statistics Canada.

Dryburgh, H. 2004. *The Longitudinal Administrative Databank (LAD) and the Longitudinal Immigration Database (IMDB): Building the LAD_IMDB - A Technical Paper*. Statistics Canada Catalogue no. 89-612-XIE. Available at http://publications.gc.ca/Collection/Statcan/89-612-X/89-612-XIE2003001.pdf http://publications.gc.ca/Collection/Statcan/89-612-X/89-612-XIE2003001.pdf (accessed March 13, 2017).

Dusetzina, S.B., S. Tyree, A.M. Meyer, A. Meyer, L. Green, and W.R. Carpenter. 2014. *Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]*. Prepared by the University of North Carolina at Chapel Hill under contract no. 290-2010-000141. AHRQ Publication no. 14-EHC033-EF. Rockville, MD: Agency for Healthcare Research and Quality. Available at http://www.ncbi.nlm.nih.gov/books/NBK253312/ (accessed March 14, 2017).

Government of Canada. 2016. *Determine your eligibility – Citizenship*. Available at http://www.cic.gc.ca/english/citizenship/become-eligibility.asp (accessed July 20, 2016).

Immigration and Refugee Board of Canada. 2015. "Refugee Protection Division." Web page. Available at http://www.irb-cisr.gc.ca/Eng/RefClaDem/Pages/RpdSpr.aspx (accessed January 13, 2016).

Immigration, Refugees and Citizenship Canada. 2015. *Annual Report to Parliament on Immigration*. Available at http://www.cic.gc.ca/english/pdf/pub/annual-report-2015.pdf (accessed June 10, 2016).

Immigration, Refugees and Citizenship Canada. 2016. *Report on Plans and Priorities 2015-2016*. Available at http://www.cic.gc.ca/english/resources/publications/rpp/2015-2016/#a2.1.1 (accessed June 10, 2016).

Langlois, C., and C. Dougherty. 1997. *The Longitudinal Immigration Database (IMDB): An introduction*. Proceedings of the 1997 Citizenship and Immigration Canada Conference on Immigration, Employment and the Economy.

McLeish, S. 2011. 2008 IMDB Landing File: Data Quality Working Paper. Unpublished document. Ottawa: Statistics Canada.

Rotermann, M., C. Sanmartin, R. Trudeau, and H. St-Jean. 2015. "Linking 2006 Census and hospital data in Canada." Health Reports. Vol. 26, no. 10. Statistics Canada Catalogue no. 82-003-X. Available at http://www.statcan.gc.ca/pub/82-003-x/2015010/article/14228-eng.pdf (accessed March 13, 2017).

Statistics Canada. 2016. *150 years of immigration in Canada*. Canadian Megatrends. Statistics Canada Catalogue no. 11-630-X. Available at http://www.statcan.gc.ca/pub/11-630-x/11-630-x2016006-eng.htm (accessed March 14, 2017).

Winkler, W.E. 2009. "Record linkage." *Sample Surveys: Design, Methods and Applications* 29A: 351–380.