Catalogue no. 11-633-X — No. 024
ISSN 2371-3429
978-0-660-32654-2

Analytical Studies: Methods and References

Longitudinal Immigration Database (IMDB) Technical Report, 2018

Release date: December 16, 2019
Revised: July 20, 2020

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.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Longitudinal Immigration Database (IMDB) Technical Report, 2018

Diversity and Sociocultural Statistics

11-633-X — 2019005 — No. 024

ISSN **2371-3429**

ISBN **978-0-660-32654-2**

March 2020

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

The initial publication of the IMDB Technical Report was co-authored by **Rose Evra** and **Elena Prokopenko**.

We would like to mention the special contribution of the following people: **Laetitia Martin** of the Diversity and Sociocultural Statistics (DSS), who wrote sections 3.3.1 to 3.3.3 of this report; **Tristan Cayn**, **Kristen James**, **Caroline Li**, **Ian Marrs**, **Scott McLeish**, and **Eric Mongrain**, members of the Administrative Data team, who produced the IMDB and contributed to the content of several sections of the report.

The IMDB is the result of a partnership between Statistics Canada, Immigration, Refugees and Citizenship Canada (IRCC), and the provinces. IRCC has been funding the IMDB since the beginning, and has been continuously collaborating with Statistics Canada to expand the content of the IMDB as well as contributing to the development of new analytical tools.

On the provincial side, the following departments have been part of the consortium funding the IMDB:

Department of Advanced Education, Skills and Labour (Newfoundland and Labrador)
Island Investment Development Inc (Prince Edward Island)
Nova Scotia Office of Immigration (Nova Scotia)
Population Growth Division (New Brunswick)
Ministère de l'Immigration, de la Francisation et de l'Intégration (Québec)
Ministry of Children, Community and Social Services (Ontario)
Manitoba Labour and Immigration
Ministry of Immigration and Career Training (Saskatchewan)
Alberta Labour
Labour Market Information Office (British Columbia)

These provincial partners were also part of consultations to redesign the IMDB and its analytical tools.

Many thanks to the following people for reviewing the report prior to its publication: Margareta Dovgal, Benoît St-Jean, Winnie Chan, Hélène Maheux, and Tiana Major (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 35 years.

Immigration, Refugees and Citizenship Canada (IRCC) administrative records contain exhaustive information about immigrants who were admitted to Canada since 1952 and non-permanent residents who have been issued temporary resident permits since 1980.

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
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
LSIC	Longitudinal Survey of Immigrants to Canada
LCF	Linkage Control File
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
SDLE	Social Data Linkage Environment
SIN	Social Insurance Number
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 35 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).

IRCC administrative records contain exhaustive information about immigrants who were admitted to Canada since 1952 and non-permanent residents issued permits since 1980. Tax records for 1982 and subsequent years are available for taxfilers.

The IMDB was designed to provide detailed and reliable data on the performance and impact of immigration programs. Being a database of immigrants and non-permanent residents (both taxfilers and non-taxfilers alike), the IMDB can be used to answer both broad and very specific research. 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). It is to be noted that yearly updates of the IMDB are independent from one another. From year to year, there have been changes to data processing, including updates to the unique person identifier (IMDB_ID).

As created, the IMDB includes multiple files: one file for each tax year since 1982 (e.g. IMDB_T1FF_2017), two files containing immigration characteristics at the person level (e.g. PNRF_1952_1979 and PNRF_1980_2018), one permit file for non-permanent residents (NRF_PERMIT_1980_2018), and the NRF_PERSON_1980_2018, which includes summary information at the person level 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.

Recently, the content of the IMDB has been expanded. The IMDB now includes citizenship acquisition since 2005, as well as express entry data. There have been several new modules integrated into the IMDB: wages, children, and settlement. As well, there has been a file structure change- the taxfilers and the non-taxfilers were merged in one file called PNRF_1980_2018. Non-taxfilers can be filtered by using the FIRST_TAX_YEAR variable.

The IMDB files are available only to the members of the Research Data Centres (RDC), 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 and published by Statistics Canada. Additionally, external researchers may request ad hoc tables and analyses; Statistics Canada provides these services on a cost-recovery basis.

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 that are possible with the IMDB (Section 9).

1. Formerly Citizenship and Immigration Canada (CIC).

2 Data sources

Several files are included in the IMDB, including the PNRF, the NRF. These files, which will be described in this section, consist of immigration data, immigrant tax files, and auxiliary files covering information available for immigrants admitted since 1952, and non-permanent residents since 1980.

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 Integrated Permanent and Non-permanent Resident File (PNRF 1980-2018)

Every year, admission data is added to create the Immigrant Landing File (ILF). This file contains information such as date of admission, date of birth, and immigration category. 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 admission. This file, however, is not directly available to IMDB users. Admission data for these immigrant taxfilers is available in the Integrated Permanent and Non-permanent Resident File (PNRF). This file also contains information on non-taxfilers for identification; however, the first tax year and last tax year information will be missing.

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 admission. 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.

Researchers can access the Integrated Permanent and Non-permanent Resident File (PNRF), which combines information from the Immigrant Landing File (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-admission experience in Canada. The PNRF covers all the admission data (except emigration and mortality) as well as detailed information 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 admission. This file contains the number of permits for each non-permanent resident who became a permanent resident, and includes admission dates. However, it is to be noted that this file does not include the records of non-permanent residents (temporary residents) who have not become permanent residents. 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, a file named PNRF_EXTRA_1980_2013 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.

In the past, the PNRF used to separate taxfilers from non-taxfilers (e.g. PNRF_2016 and PNRF_NONFILERS_2016). As of the 2018 IMDB release, the taxfilers have been merged with the non-taxfilers and it is called PNRF_1980_2018.

2.1.2 Admissions Prior to 1980: Integrated Permanent and Non-permanent Resident File (PNRF) 1952-1979

Prior to the 2018 IMDB release, the data on immigrants in the IMDB were limited to admissions from 1980 onwards. As a part of the 2018 IMDB release, the PNRF will now include data from 1952, expanding the IMDB universe. The new file (PNRF_1952_1979) contains the immigrants admissions from 1952-1979. However, PNRF 1952-1979 has fewer variables than for the people admitted after 1980 (see section 2.1.1), since it is older data. Its major categories include: Gender, Country, Birth year, Landing year and month.

2.1.3 Non-permanent Resident File (NRF)

The Non-permanent Resident File (NRF_Permit) is created from the data of individuals who have been granted non-permanent resident permits since 1980. This file includes the type of permits (work or study, for example) and the last valid date of a permit for example. The file is updated each year with new annual non-permanent permits data. For the 2018 IMDB release, the Permit NRF is called NRF_PERMIT_1980_2018.

A given person can have multiple permits over time. These permits include Work Permits, Study Permits, Refugee Claims, and Other Permits issued, as well as the date when they were issued and the date that they expire. The NRF Person, called NRF_PERSON_1980_2018, stores information at the person level such as the number of permits and the first year of temporary residence permit.

The data can be linked to the PNRF by means of the IMDB unique person identifier (IMDB_ID). For variables common between the PNRF and NRF, in cases of discrepancies refer to the PNRF values. For more details on the variables included on these files, please refer to the immigration component of the IMDB dictionary.

2.1.4 New: Express Entry (EE)

The Longitudinal Immigration Database (IMDB) includes data on immigrants admitted through the Express Entry[2] (EE) application management system. Express entry (an extension of the PNRF) is an application process for economic immigrants wanting to settle in Canada permanently and wanting to take part in our economy. This selection process was launched on January 1, 2015, and the first draw (to select qualified permanent residents) was on January 31, 2015.

The IMDB contains data on 200,300 individuals (principal applicants and their family members) admitted through EE. These individuals can be identified using the variable EXPRESS_ENTRY_IND from the IMDB's Integrated Permanent and Non-permanent Resident Files called PNRF_1980_2018.

Detailed data on principal applicants admitted through EE are now available. For example, transferability of skills and highest level of education are available.

To obtain more information, there is a detailed technical report on Express Entry.

2.2 New IMDB Modules

The IMDB is composed of several modules on Children, Wages, and Settlement.

The IMDB was released in stages. What follows initially referred to the 2016 IMDB, but was updated to refer to the 2018 IMDB modules. Note that the IMDB's unique person identifier (IMDB_ID) has not changed from last cycle.

2.2.1 Children Data Module

This is a brief introduction to the Longitudinal Immigration Database (IMDB) Children Data module, which includes a file named PNRF_CHILD_1980_2018 with children immigration records and T1 Family Files (T1FF) since 1982, named IMDB_CHILD_T1FF, for immigrant children during their childhood. In these tax files, the parents of children are identified with IMDB_ID_PARENT, which is equal to the parent's IMDB_ID if parents are present on the immigration files (e.g.: had permanent or non-permanent permit(s)).

Since 1980, over 2 million immigrants who were admitted to Canada were aged less than 18 years old at their time of admission. This represents 24.8% of immigrants admitted during that timeframe. These children will most likely receive all or part of their education in Canada and will have different challenges than adult immigrants. Little information is available about immigrant children during their childhood in the Longitudinal Immigration Database (IMDB), as they are likely not tax-filers.

In order to increase the analytical capability of the IMDB, a children module was produced. The ability to study the impact of the childhood socioeconomic condition on adulthood economic outcome is an added value to the IMDB.

2. For more details visit: https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry.html

Different methods were used in order to add tax information for immigrant children. One method consisted in using the immigration application number to identify a parent. The second method used the Statistics Canada Dependant Register, which is the result of record linkages, to identify children's guardians. Once a child-parent connection is made, the remaining task was to produce the tax files; tax files during the years of childhood are created for immigrant children admitted since 1980.

In order to determine the parent-child connection, information from a DIN-SIN (Dependent Identifier Number - Social Insurance Number) connection was prioritized. When this information was not available, the immigration application number was used to identify children's parents. Once parent-child connection is made, tax files during the years of childhood are created for immigrant children admitted since 1980.

Tax files related to children's parent include a subset of the variables included in the IMDB_T1FFs. Only main income variables (such as employment income), tax benefits and deductions provided to families and parents were kept (such as child tax benefits and education amount and tuition fees transferred from a child). The PNRF_ CHILD_1980_2018 file includes a subset of variables available in the PNRF and information about the children's parent(s), such as IMDB_ID and first and last year of filing during the children's childhood.

To obtain more information, there is a detailed technical report on IMDB Children.

2.2.2 Wages and Salaries Data Module

This is a summary of the linkage between the Longitudinal Immigration Database (IMDB) and the Statement of Remuneration (T4) Supplemental file. The Preliminary Wages and Salaries tax files are derived from the T4 Supplemental tax files, which contain tax employment information as provided by the individual's employer. T4 Supplemental files are used to report salary, wages, and taxable benefits paid to employees for services rendered during the year, as well as pension adjustment, amounts of pay for employees who accrued a benefit for the year under a registered pension plan or a deferred profit sharing plan. Variables extracted from these files include province of employment, province of employee, T4 earnings per by tax year, and number of T4 slips per tax year. The preliminary wages and salaries tax data are available from 1997 to 2018.

There are three main reasons for integrating the T4 tax files to IMDB:

1. To better understand the actual coverage of the IMDB with regards to temporary residents working in Canada by using temporary SINs as a basis for analysis.
2. To have a more comprehensive coverage and understanding of temporary residents working within Canada by using T4 slips rather than relying on T1 Filers, in particular those temporary residents who do not transition toward permanent residency.
3. To understand the feasibility of disseminating IMDB findings earlier using the T4 tax file, as the T4 files are available approximately six months before the T1FF files.

Integrating the T4 does provide some benefit to the T4, particularly additional coverage of temporary foreign workers. The values provided by the linkage to the T4 were validated against T1FF values, matching 94.2% of the time, while seeing an average overall difference in T4 earnings of 1.0%.

To obtain more information, there is a detailed technical report on Wages and Salaries.

2.2.3 Settlement Services Data Module

The IMDB's Integrated Permanent and Non-permanent Resident File (PNRF) and other IMDB files can be integrated to the settlement services module. The non-confidential person identification number (IMDB_ID) is included in all the files, it should be used to integrate immigrants and non-permanent residents to their records in other IMDB files. The files DOM_CLIENT_ SETTLEMENT (recipients of domestic services) and FRN_CLIENT_ SETTLEMENT (recipients of foreign services) have information on type and number of services received at the person level. Then a series of files, by type ofservice, with more details on the services received are available to users. In total, this module includes 15 files. A dictionary is available to users for more details (in English only, information in French available on request).

Several files related to settlement services provided to permanent[3] and non-permanent residents selected to become permanent residents are available at Immigration, Refugees and Citizenship Canada (IRCC). The Immigration Contribution Agreement Reporting Environment (ICARE) is where all settlement data are collected and stored. ICARE is a reporting system used by organizations providing resettlement services to immigrants to report their activities. Annually, Statistics Canada received files generated by ICARE, in order to produce an IMDB settlement services module. The data received covered the services provided from 2013 and onward. According to the data received 1,213,850 people were provided services in Canada since 2013 and 74,320 people were provided pre-arrival services since 2015.

Settlement services are received in Canada or pre-arrival. A variety of services are offered to new immigrants and non-permanent residents, some are related to employment or assessments of needs and others to information and orientation. Support services, such as transportation and childminding are also provided.

Several components of the ICARE data add analytical power to the IMDB that combines immigrant admissions and non-permanent resident permits with their tax files. The data currently available does not allow the addition of data for settlement services received prior to 2013 (and 2015 for foreign services).

The coverage for immigrants admitted prior to 2013 is partial. In cases of multiple admissions, which are rare, the settlement services relate to the most recent admission when admission characteristics (kept in the IMDB) relate to the first admission. Settlement services are not limited to recent immigrants. For example, 280 immigrants first admitted in 1980 had received settlement services between 2013 and August 2019, it was also the case for 950 immigrants admitted in 1990 and 4,000 immigrants admitted in 2000. Settlement data not connected to a recent admission were not removed from the IMDB.

It is to be noted that the module includes data on 64,490 non-permanent residents who received settlement services, these could be people who were admitted in 2018 (the IMDB includes admissions up to 2018) or are in the process of becoming permanent residents. Data from organisations located in Quebec are not collected, so only services provided to immigrants outside Quebec are available.

The settlement data module is comprised of several files, as different types of services are in separate files. Also, a distinction is made between services received pre-arrival (foreign) and post-arrival (domestic). Data about these services are available on different files. It was possible to integrate only 23.9% of foreign services recipients to an IMDB record. The main reason is that these people had not arrived in Canada prior to 2018 or are still not in Canada. The coverage of foreign services start in 2015 where the coverage of domestic services begin in 2013.

In order to be included into the IMDB, some of this data was synthesized at the person level. Numerous variables were derived, such as the number of services received by topic. Similar to the IMDB_ID, a non-confidential service number (SERVIC_NUM) was created.

The Longitudinal Immigration Database (IMDB) Integrated Permanent and Non-permanent Resident File (PNRF) and other IMDB files can be integrated to the DOM_CLIENT_SETTLEMENT (person who received domestic services) and FRN_CLIENT_SETTLEMENT (person who received foreign services). The files DOM_CLIENT_SETTLEMENT and FRN_CLIENT_SETTLEMENT have information on type and number of services received at the person level. Then a series of files, by type of service, with more details on the services received are available to users. In total, this module includes 15 files. The non-confidential person identification number (IMDB_ID) is included in all the files; it should be used to integrate immigrants and non-permanent residents to their records in other IMDB files.

To obtain more information, there is a detailed technical report on Settlement Services

3. For eligibility criteria see Appendix 6.2.

2.3 T1 Family File (T1FF)

The T1 Family File (T1FF). Every year, Statistics Canada uses the annual individual T1 file, T4 Tax file, and the Canada Child Tax Benefit file from the CRA and creates an analytical T1FF. T1FF data is available from 1982 to 2017.

The tax files used to create the **IMDB_T1FF** files are those contained in the T1 Family File[4] (T1FF). Statistics Canada takes the annual individual T1 file, T4 tax file and Canada Child Tax Benefit (CCTB)[5] 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 integrated 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, editing 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.

2.4 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 Social Data Linkage Environment (SDLE) and the Linkage Control File (LCF) were used to facilitate the record linkages.

The SDLE at Statistics Canada promotes the innovative use of existing administrative and survey data to address important research questions and inform socio-economic policy through record linkage. The SDLE expands the potential of data integration across multiple domains, such as health, justice, education and income, through the creation of integrated analytical data files without the need to collect additional data from Canadians.

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.

4. For details on the most recent T1FF, the reader may consult the Statistics Canada website.
5. 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. The 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. In the IMDB, the term "**non-permanent resident**" refers to persons from another country who have a work or study permit or who are refugee claimants. 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*,[6] 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 21.9% in 2011 (7.5 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),[7] 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" 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.[7]

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."[8] Under IRPA, there are three overarching categories 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).

6. See Statistics Canada 2016 in list of references.
7. Immigration and Refugee Protection Act (S.C. 2001, c. 27)
8. Definitions approved as a recommended Statistics Canada standard on March 21, 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 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). 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 that includes:

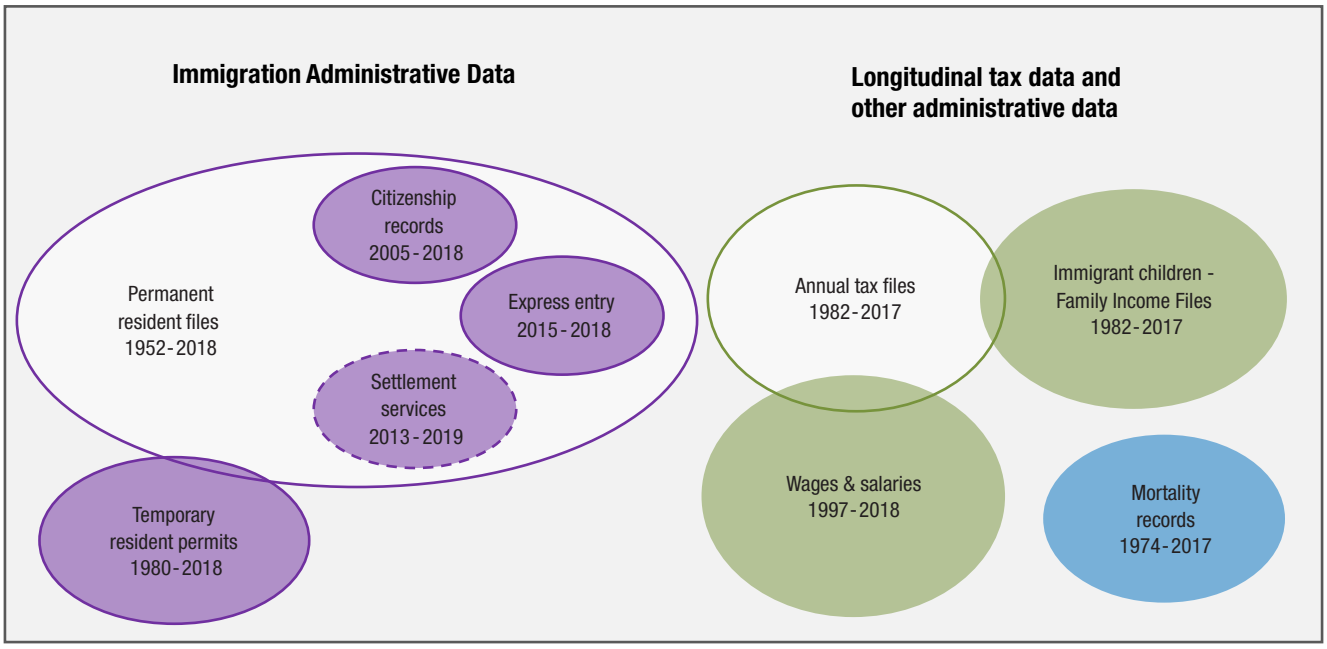
- All immigrants admitted to Canada since 1952;
- All non-permanent resident permits since 1980;
- All tax files since 1982;

The database also provides information on permits for immigrants who were non-permanent residents prior to their admission as permanent residents.

The T1FF covers "all persons who completed a T1 tax return for the year of reference or who received CCTB (Canada Child Tax Benefits), their non-filing spouses (including wage and salary information from the T4 file), their non-filing children identified from three sources (the CCTB file, the births files, and an historical file) and filing children who reported the same address as their parent". Family information is created based on a census family concept: parent(s) and children living at the same address (Statistics Canada 2019).

The IMDB brings together, via probabilistic record linkage (Section 4), administrative data from Immigration, Refugees and Citizenship Canada (IRCC) and the tax files from the T1 Family File (T1FF).

Figure 1
Source of IMDB content



Note: See glossary of terms for definitions of acronyms
Source: Statistics Canada.

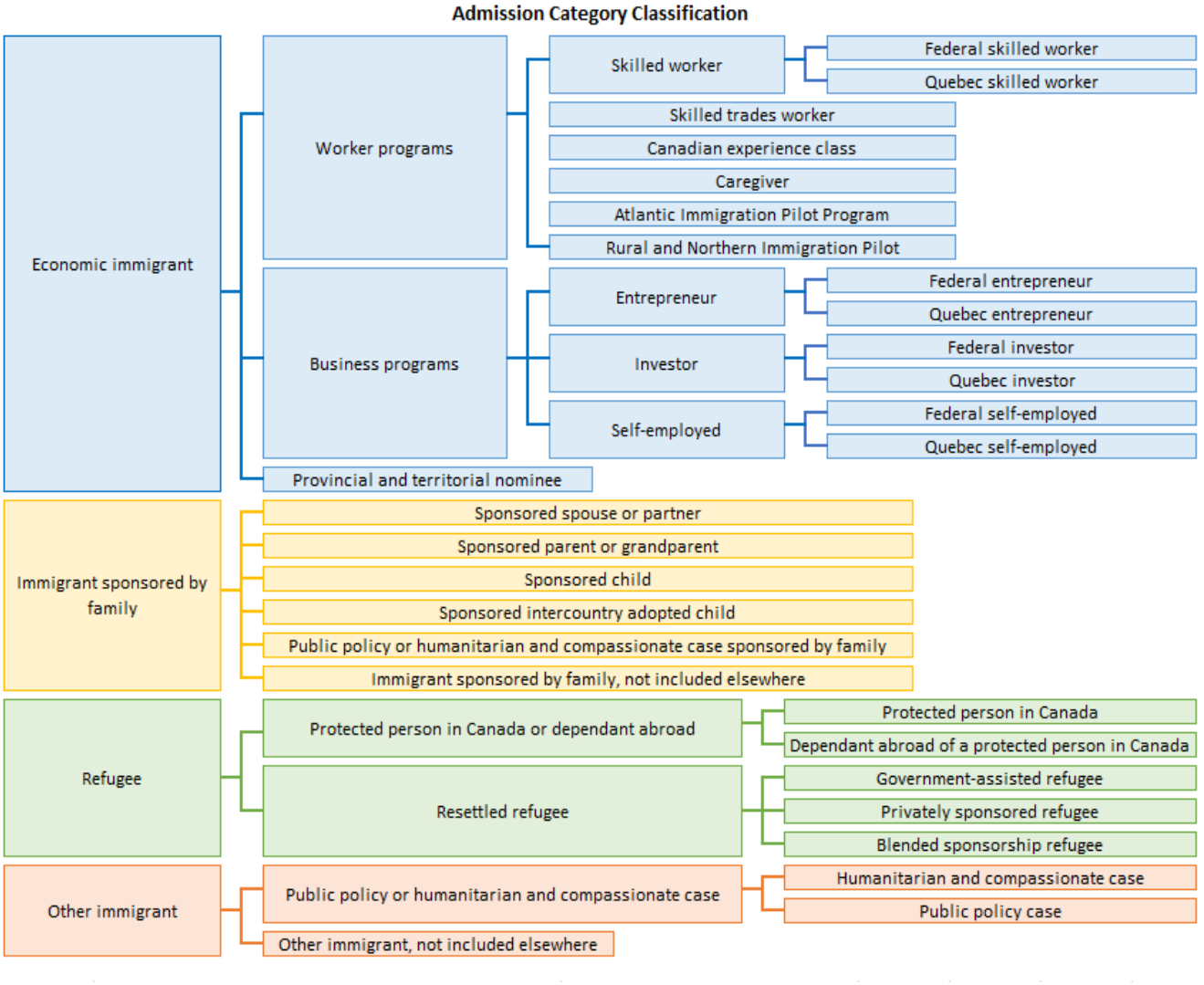
3.3 Admission variables

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.

Figure 2
Admission Category Classification



Note: See glossary of terms for definitions of acronyms.
Source: Statistics Canada.

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.

Economic immigrant

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.
5. Atlantic Immigration Pilot Program (Please see below for a short description)
6. Rural and Northern Immigration Pilot (Please see below for a short description)

Atlantic Immigration Pilot Project

This program helps to find skilled immigrants to immigrate to Atlantic Canada to fill positions that have not already been filled locally. It is an employer-driven programme, meaning that the employers are demanding foreign skilled workers because there is a need for stable skilled foreign workforce in Atlantic Canada. This is path (gateway) to permanent residency for international graduates and for skilled foreign workers wanting to live in one of Canada's four Atlantic Provinces.

It is also open to international graduates wanting to go Atlantic Canada upon their graduation. The candidate can be either living abroad or temporarily in Canada.

There are three sub-programs:

-Atlantic International Graduate Program

-Atlantic High-skilled Program

-Atlantic Intermediate-skilled Program

In order to participate in the aforementioned program requirements must be met, both on the end of the employer as well the candidate.

Rural and Northern Immigration Pilot

The aim of the Rural and Northern Immigration Pilot is to spread the benefits of economic immigration to smaller communities by drawing skilled foreign workers who want to work and thrive in rural communities. Unlike the Atlantic Pilot Project, the Rural and Northern Immigration Pilot is a community-based approach. This means that communities as a whole apply to be hosts and are selected based on three major criteria:

-Whether there is the economic need for immigration

-Whether the resources and the community partners are there to be put into the project

-Whether the federal government already has existing settlement partners and resources in the community

The purpose of the immigration programme is to help fill the local community's labour market needs and support the region's economic development. By welcoming and supporting immigrants, communities help to persuade them to stay in a smaller rural community.

Retention rates should be higher based on:

-Working with their community-based partners

-The involvement of other federal government partners

-The participation of the provincial and territorial governments

The community, as well the candidate, must reach the Rural and Northern Immigration Pilot program's requirements in order to participate in the program.

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.

For additional information Immigrant sponsored by family

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 the 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.

For more information, please visit Refugees

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 discretion 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.

For more information, please visit Other immigrants

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.

3.3.4 Policy Changes over time

The IMDB contains over 35 years' worth of data on immigrants admitted to the country. However, policies and programs have undergone many changes during this time. New programs have been created, others 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 admission 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 of these laws cover the same three key groups (economic immigrants, family sponsorship, and refugees), the administration of these programs has undergone 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 admission 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 when 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.

3.3.5 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. An aggregated version of the information available in the variable IMMIGRATION_CATEGORY is available in the derived variable IMMIGRATION_CATEGORY_CENSUS, which contains fewer categories.

The aggregate variable **IMMIGRATION_CATEGORY_CENSUS**, is a categorization in line with Statistics Canada's standard used in the Census of Population. 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 **IMMIGRATION_CATEGORY_ROLLUP2**, available in the PNRF_EXTRA for immigrants landed up to 2013. 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 **type of applicant** (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 category 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., Immigrant sponsored by family) 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 impacts. 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. New variables were added to the 2016 IMDB, which include Syrian refugee resettlement waves (SYRIAN_RRW), Express Entry (EXPRESS_ENTRY_IND) and the year and month of citizenship (CITIZEN_YEAR and CITIZEN_MONTH).

Variables in the PNRF refer to immigrants' characteristics at admission 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 (MSTCO) 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 admission 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 admission. 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. As a result, the **province of residence** on December 31 of the tax year (PRCO_) may not be the same as the province (PR__) derived from the mailing address. This distinction is important, as using the derived census geography variables (e.g. CMA, CSD) 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"). For more information regarding the quality of geography variables please refer to Section 7.

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 landing should be interpreted with caution. First, about 50% of each landing cohort first files taxes in the landing year (proportion based on taxfilers from integrated immigrants). Secondly, taxes filed in the same year as landing 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 admission 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.

Data quality issues were identified with these education variables since 2011. For example, a significant proportion of individuals did not state their education qualifications or years of schooling and were coded as "0" ("None") on EDUCATION_QUALIFICATIONS and YEARS_OF_SCHOOLING instead of "Missing."

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 to 2018 were imputed to resolve data quality issues. 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 admission and assigns it a classification according to 2011 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 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[9]

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 et al. 2015).

To produce the IMDB the Social Data Linkage Environment (SDLE) was used. It is a highly secure linkage environment that facilitates the creation of integrated population data files for social analysis.

At the core of the SDLE is a Derived Record Depository (DRD or Depot), a national dynamic relational database containing only basic personal identifiers. The DRD is created by integrating selected Statistics Canada source index files for the purpose of producing a list of unique individuals. These files, which contain personal identifiers without analysis variables, are brought into the environment, processed and integrated only once to the DRD. Updates to these data files are integrated to the DRD on an ongoing basis.

In **2018**, the linkage rate to the depot for immigration records was **97.4%** (Cascagnette, 2019).The probabilistic method was used to integrate IRCC's immigration data to CRA's tax data. To perform the record linkage G-Link was used.

The generalized record linkage system used at Statistics Canada, G-Link, is based on the mathematical theory of record linkage developed by Ivan P. Fellegi and Alan B. Sunter. Probabilistic record linkage methodology compares non-unique identifiers (e.g., name and birth date) and estimates the likelihood that records being matched refer to the same entity (e.g. individual). Probabilistic record linkage is especially valuable when the identifiers are prone to change (e.g. surnames of females who get married), error-prone and frequently missing.

Comparisons between records are done field-by-field using comparison rules with outcomes such as exact match, string proximity, missing information or fields disagreement generated by each rule based on the similarity of values in a pair of records. Each pair of records is assigned a comparison result pattern and that pattern is evaluated to classify pairs as linked, possibly linked or not linked.

The theory of probabilistic record linkage is based on the premise that the results of certain comparison result patterns are characteristic of truly linked pairs, while others are characteristic of truly unlinked pairs. Therefore, each rule outcome is assigned a weight based on the ratio of the estimated probability of the outcome occurring for true matches to the estimate probability of the outcome occurring for non-matches.

The composition of the linked set is not known in advance, so the probabilities of result patterns for truly linked records are not known. Linked weight components are estimated from prior knowledge and early iterations of the linkage process, and refined by treating successive iterations of the linkage process.

The unlinked weight components are calculated based on the frequency with which the rule outcomes were observed among record pairs that do not belong together, which is approximately equal to the frequency with which the rule outcomes would be observed among randomly paired records. After repeated iteration of the linkage process, linked weight components stabilize and final weights are ready for use.

9. Most of the content of this section comes from the methodology report (Cascagnette, 2018).

The strategy for the probabilistic record linkage involves the following six steps:

1. Generate potential pairs using initial criterion
2. Develop and apply comparison rules to potential pairs to derive probability ratios
3. Apply frequency weights
4. Assign linkage states to the pairs using probability ratios and thresholds
5. Form groups
6. Resolve conflicts using mapping.

Steps 2 to 4 are repeated iteratively.

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 probabilistic 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.

This year, to improve the record linkage results, the SDLE linkage results were combined with the results of the linkage of the Immigration data to the Linkage control file (LCF) as per the 2015 IMDB instalment.

In order to produce the 2015 IMDB, the hierarchical deterministic method was used to link immigration 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 et 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.

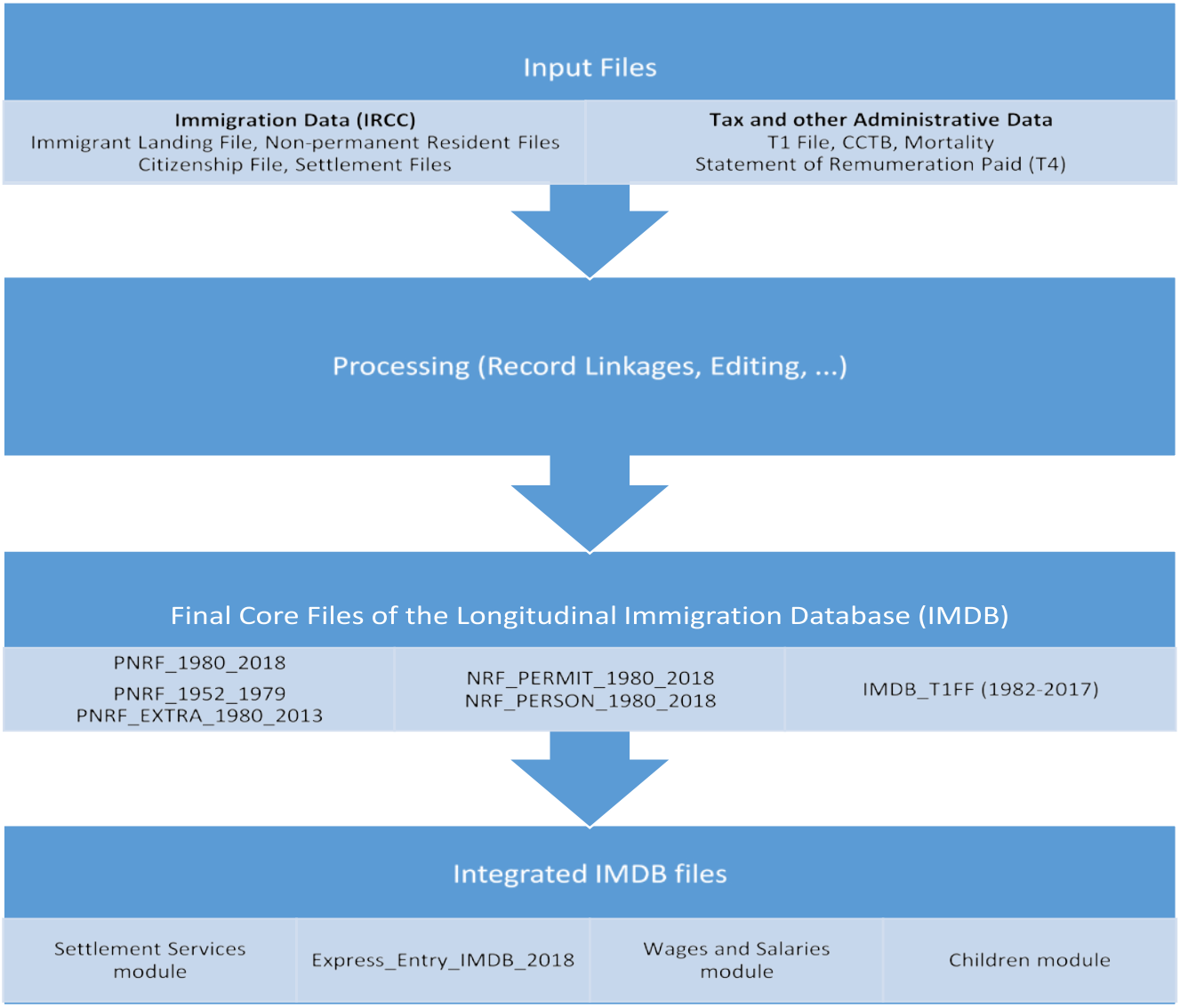
The December 2019 release of the 2018 IMDB included only tax files for Immigrants and non-permanent residents who arrived between 1974 and 2016. The IMDB was updated in January 2020 to include tax files for immigrants who arrived between 1952 and 1973, as well as those who arrived in 2017 and 2018. Tax files for non-permanent residents who arrived in 2017 and 2018 were also added to the IMDB.

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 3
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 3, 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.[10]

On the immigration side, IRCC provides the data on landed immigrants non-permanent residents and citizens 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.

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. 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 admissions since 1980, data from the time of the first admission 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), and Immigration data files (**PNRF_1980_2018, PNRF_EXTRA_1980_2013, PNRF_1952_1979 and NRF_PERMIT_1980_2018**). 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.

10. For more detailed information on T1FF processing and data quality.

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.

In order to identify a taxfiler's immigration status, the admission year (LANDING_YEAR) along with the first effective year, which represents the year that they first obtained a non-permanent residence permit (FIRST_EFFECTIVE_YEAR) have been created. As a result, the presence on the non-permanent resident file indicator (**TR_IND**) has been removed.

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 were admitted in 1980 or thereafter (variable **IMM80F&year**). However, this can be an underestimation as this variable includes only filers and not imputed records, therefore children are under-estimated. 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 7.5).

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. When the FIRST_TAX_YEAR and LAST_TAX_YEAR variables are missing, it is to denote the non-filers or people who have never filed income tax before. This is an update to the 2018 IMDB, since in the past the filers and non-filers have been merged together.

The variable **PREFILER_IND** is used to identify immigrants who have T1FF data prior to their admission 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 related to the outlier flag

	OUTLIER_IND1982	OUTLIER_IND1983	OUTLIER_INDyyyy	OUTLIER_IND2012	OUTLIER_IND2014
IMDB_ID	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 was admitted. 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 (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. Members of the Research Data Centres (RDC) have direct access to administrative microdata files (see Section 6.2). Confidentiality rules are maintained in order to ensure data safety and security (see Section 6.4).

Accessing data via the RDC Program

The Research Data Centres (RDCs) are secure areas in which to view Statistics Canada microdata that are located across Canada. The RDCs are hosted by 29 Canadian universities and are run by Statistics Canada analysts, whereas the FRDC (Federal Research Data Centre), located in Ottawa, has been established to support the analytical needs of federal departments. RDCs offer secure conditions governing all aspects of data usage, from data access to its publication.

6.1 Analytical products

At Statistics Canada, the common repository is an online database which holds data tables that report immigrants' income by various individual characteristics and geographies. From the main page of the Data search engine at Statistics Canada website, the IMDB tables can be accessed by selecting "Immigration and ethnocultural diversity", "Immigrants and non-permanent residents", and then "Longitudinal Immigration Database" under "Survey or statistical program". It should be noted that Statistics Canada has replaced the Canadian Socioeconomic Information Management System (CANSIM) tables with a common repository in June 2018, where the IMDB tables can be found. It is to be noted that yearly updates of the IMDB are independent from one other. From year to year, there may have been changes to data processing. The income measures (averages and median) available on the tables are wages, salaries and commissions, employment insurance, investment incomes, self-employment earnings, and social welfare benefits (for details on how these measures are derived, see Appendix D.8).

For the **2018** IMDB, four tables were released in December 2019 at both national and provincial level, where incomes are in 2017 constant dollars:

Table 1 (43-10-0016)**:** Income of Immigrant taxfilers, by sex, pre-admission experience, knowledge of official languages, immigrant admission category, admission year and tax year, for Canada and provinces, 2017 constant dollars;

Table 2 (43-10-0017)**:** Interprovincial migration of immigrant taxfiler, by province of intended destination, province of residence, age groups at taxation year by sex, pre-admission experience, knowledge of official languages, immigrant admission category, admission year and tax year, for Canada;

Table 3 (43-10-0018)**:** Interprovincial migration of immigrant taxfilers, by age groups at taxation year by sex, knowledge of official languages, immigrant admission category, pre-admission experience, admission year and tax year, for Canada and provinces; and

Table 4 (43-10-0019)**:** Income of asylum claimants, by sex, age groups, birth area, residency status, claim year, and tax year, for Canada and provinces, 2017 constant dollars.

All tables offer provincial breakdown. It is to be noted that the province is based on the province of residence on December 31 of the tax year (variable PRCO).

Additional tables have been released in 2020: 43-10-0020, 43-10-0021, 43-10-0022, 43-10-0023 et 43-10-0024).

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 (**e.g. IMDB_T1FF_YEAR, PNRF_1980_2018, NRF_PERSON_1980_2018 and NRF_PERMIT_1980_2018**) 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 files through the Research Data Centres (RDC) throughout the country. The RDCs provide researchers with direct access to a wide range of population and household surveys, as well as administrative data in a secure university setting. 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 2001 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,[11] 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.

The content of the IMDB has been integrated to multiple data sources, including the Longitudinal Survey of Immigrants to Canada (LSIC), the Longitudinal and International Study of Adults (LISA), and the Canadian Community Health Survey (CCHS). As a result of the IMDB being linked to SDLE, linkages to other statistical programs are now possible.

6.4 Confidentiality[12]

Statistics Canada is committed to respecting the privacy of individuals. For that reason, data safety and security is a top concern. Statistics Canada strives to protect the data of Canadians. 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. The confidentiality of data is enforced through the *Statistics Act*, the *Access to Information Act*, and the *Privacy Act*. For more information, visit Using New and Existing Data for Official Statistics. For additional information on Trust, the data's safety and privacy, as well as transparency and openness, visit Statistics Canada's Trust Centre.

11. Https://www.statcan.gc.ca/eng/microsimulation/demosim/demosim.
12. Source: https://www.statcan.gc.ca/eng/reference/privacy.

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[13] 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.

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).

13. Https://www.statcan.gc.ca/eng/about/pia/lidb.

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 were admitted 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.

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 admission). Both SINs are required in order to have a complete fiscal history from arrival in Canada. The LCF and SDLE (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 admission (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 admissions 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)

7.2.1 2018 IMDB: Linkage rates

This section is based on the **2018 IMDB**. The overall linkage rate between IRCC immigration files and the SDLE Derived Record Depository was **97.4%** (see Section 4). A link does not necessarily mean that a tax file is available since it is possible to link dependents of taxfilers or immigrants who have yet to file taxes. This SDLE theoretical linkage rate mostly informs on how well IRCC files could be associated within a larger repository environment. The IMDB T1FFs for the 1952 to 1973 and 2017-2018 cohorts were not included in the files released in December 2019, the report has been updated after they were added to the 2018 IMDB.

Of the immigrants who landed in any year from 1980 to 2017, 85.0% were linked to at least one T1FF record. This rate represents the effective coverage of immigrant linkage to tax files. As presented in the following statistics, this coverage rate may change according to gender and age.

The proportion of linked taxfilers by age group at admission and sex is shown in Table 5. 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 admission cohort.

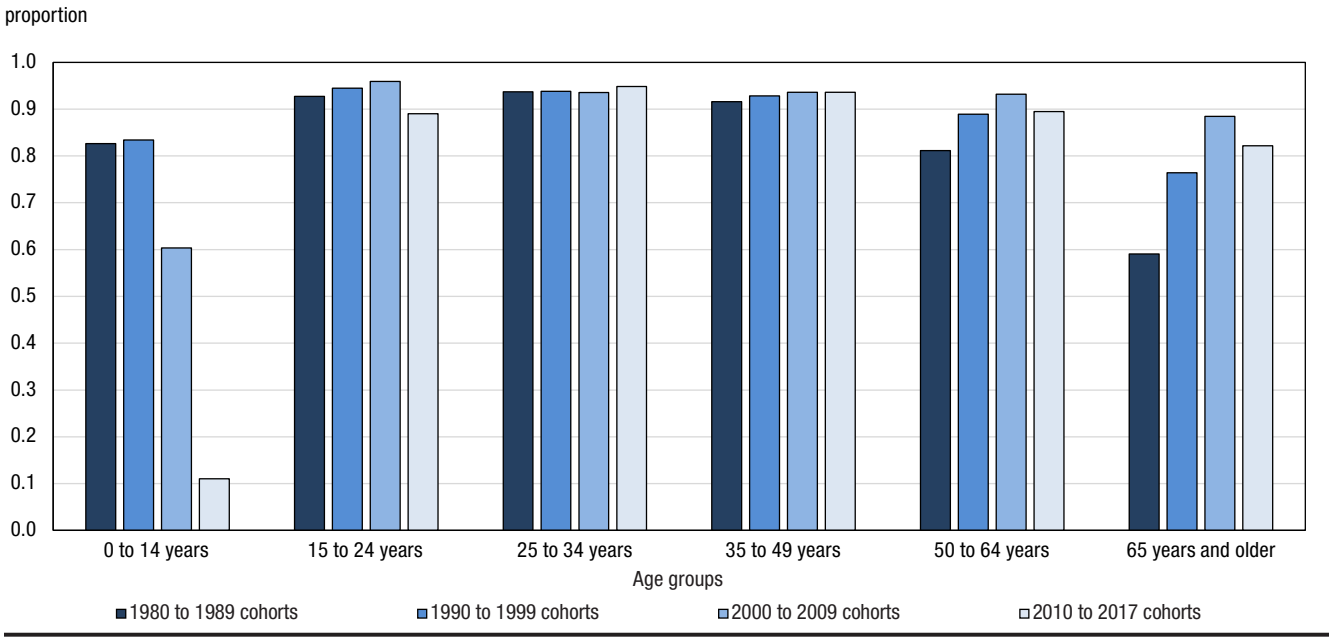
Table 5
Proportion of linked taxfilers by age group at landing and sex

	Age at landing						
	0 to 14	15 to 24	25 to 34	35 to 49	50 to 64	65 and older	Total
	percent						
Male	57.9	93.3	93.8	92.9	89.2	77.5	84.7
Female	57.2	93.4	94.2	93.5	88.0	76.0	85.4
Total	**57.5**	**93.4**	**94.0**	**93.2**	**88.6**	**76.7**	**85.0**

Source: Statistics Canada, 2018 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 admission, sex and admission decade.

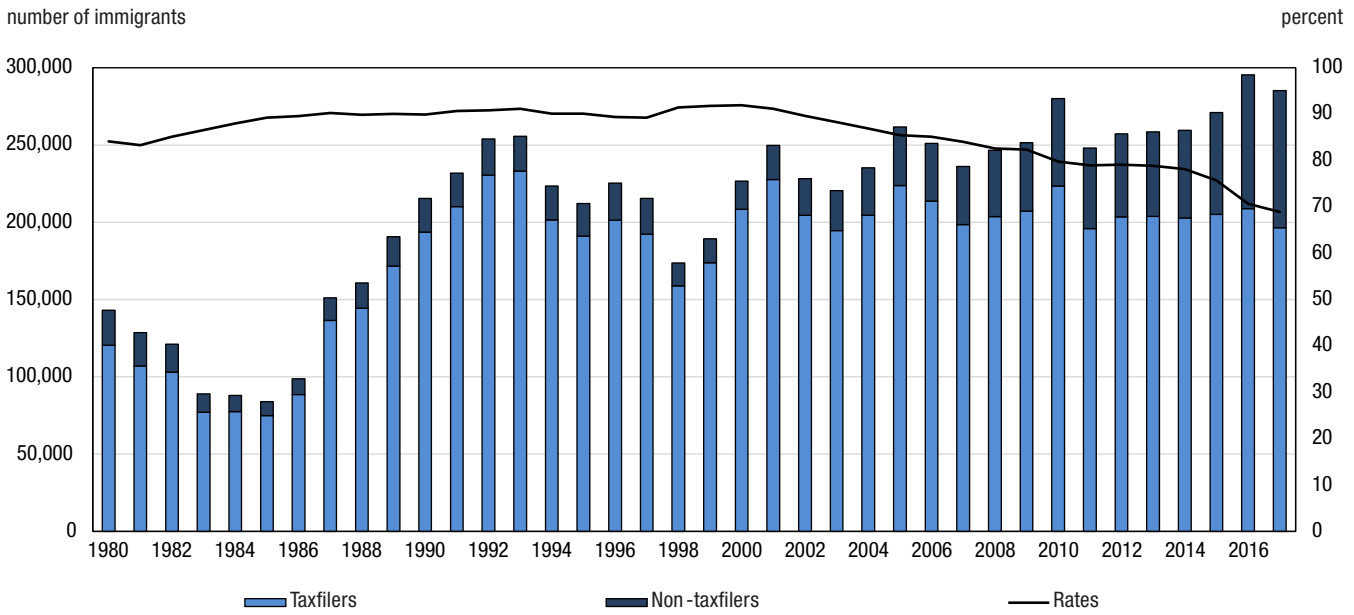
Chart 1
Proportion of linked taxfilers by age groups at landing and landing decade



Source: Statistics Canada, 2018 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 2018 IMDB**, the filing rate varies by landing year, ranging from **68.9%** for those who landed in 2017 to **91.9%** for those who landed in 2000. 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 1990's and early 2000's. See Appendix B, tables 14 and 15, for detailed distribution numbers by landing year.

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



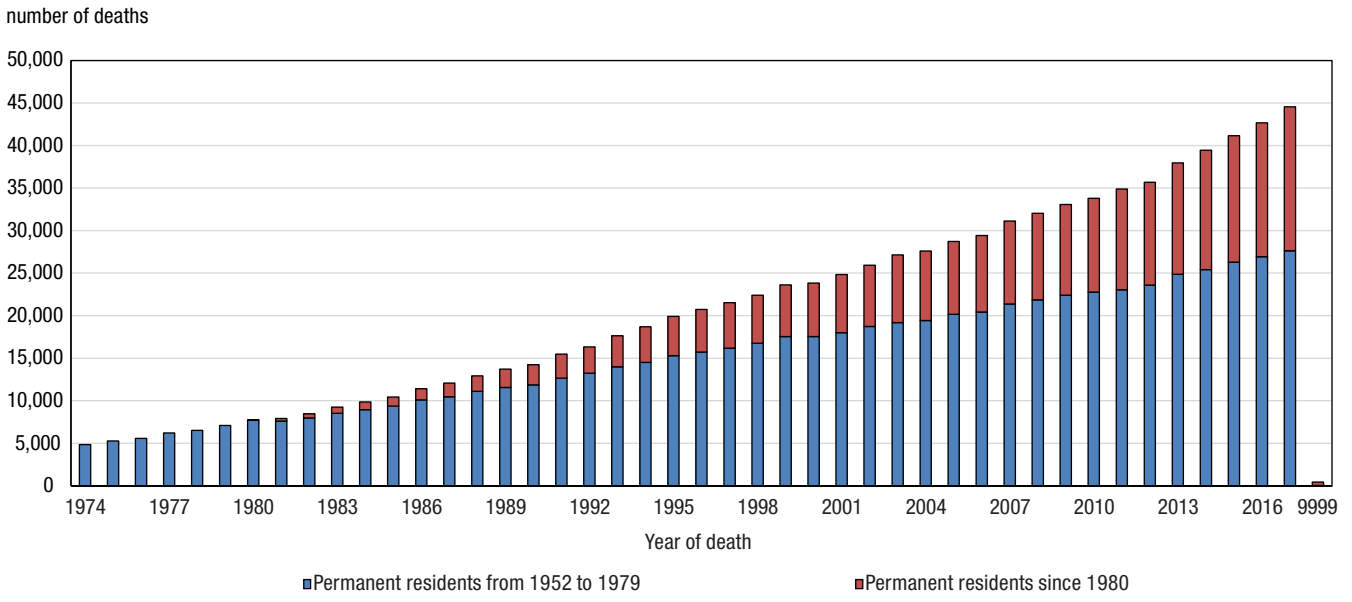
Source: Statistics Canada, 2018 Longitudinal Immigration Database.

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. In the 2018 IMDB, these variables were linked by using the Canadian Mortality Database (CMDB). In the past, these variables were based on Statistics Canada's Amalgamated Mortality Database (AMDB), which is a retired dataset that combined records between CMDB and vital statistics and tax files. The CMDB is an administrative database that collects information on death dates and cause of death from all provincial and territorial vital statistics registries in Canada. Some undercoverage, while minimal, exists in the database as it does not include deaths of Canadians (1) who died outside of Canada, with the exception of United States; (2) who served as members of the Canadian military, or (3) whose bodies were unidentified. Note that the CMDB does not include deaths which were reported in the tax files.

Chart 3 describes the general trend in the number of deaths per year since 1974 for immigrants admitted since 1952. The availability of data for pre-1980 admission was recently added to the IMDB. 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



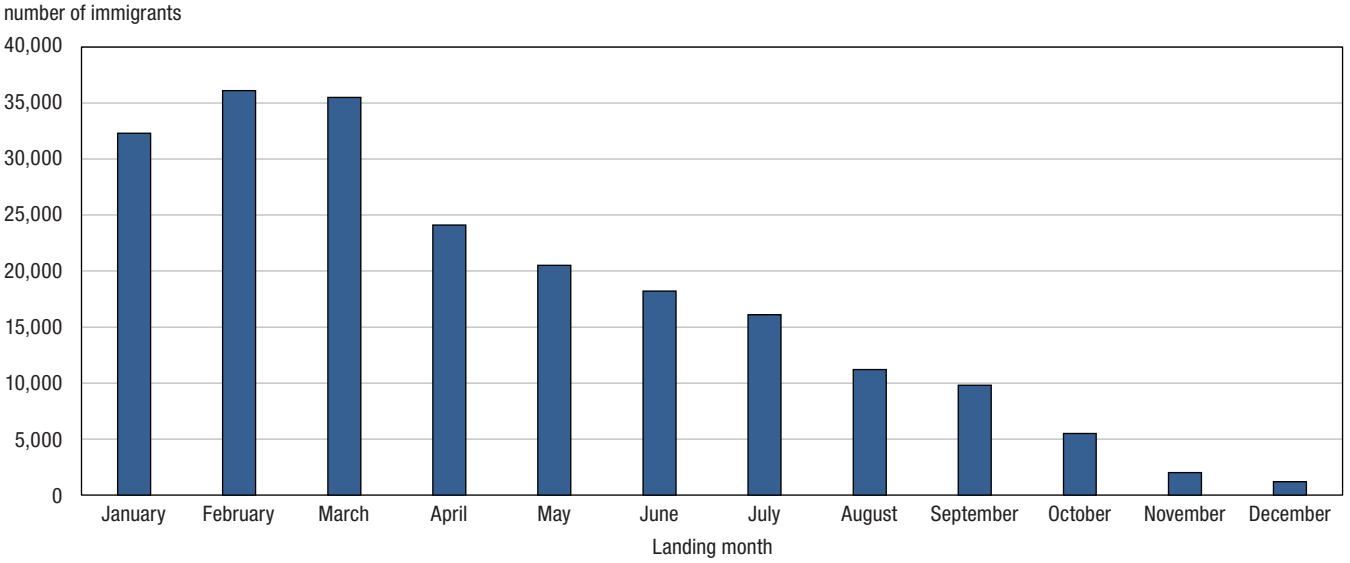
Note: The value 9999 is assigned when the date of death is missing.
Source: Statistics Canada, 2018 Longitudinal Immigration Database.

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

The results included in this section are drawn from a study based on the 2014 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 admission. This section discusses why it is not the case. A total of 1.26 million individuals filed taxes before officially admission 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 admission, and 75% reported no income (96% had no wages). 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 FIRST_EFFECTIVE_YEAR=. from studies on immigrants with pre-admission experience depends on the analysis since FIRST_EFFECTIVE_YEAR=. 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-admission 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 FIRST_ EFFECTIVE_YEAR<>. or the number of temporary resident permits (variable NUMBER_ALL_PERMITS) is more appropriate to study immigrants with pre-admission experience. Prefiler_ind=0 indicates that no tax records have been filed prior to admission, but this does not mean that the individual had no pre-admission 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.

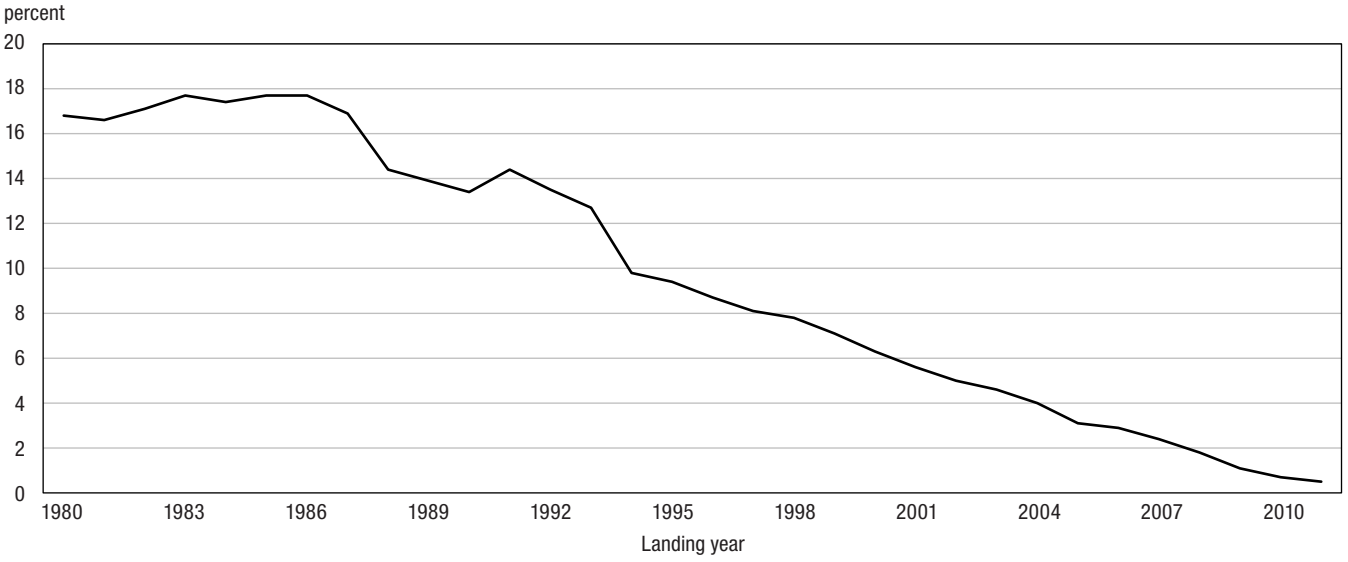
Prior to 1991, the "**single**" category was not available as **marital status** (MSTCO). The "**common-law**" status was made available as of 1992 for all datasets (1982 to 2012). Since 1992, the proportion of IMDB records indicating marital status as "single" has ranged from 20% to 30%. The proportion 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 from 2011 to 2014. 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 admission 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.

For immigrants admitted in 2016, the number of cases where a non-stated education was coded to “0” or “None” instead of “Missing” was reduced. However, a non-negligible number of records had a missing education qualification with a valid number of years of schooling. For these records, years of schooling was used to impute a value for education qualifications.

For principal applicants admitted in 2015 and 2016, under the express entry, the education variables should be set to missing. The year of schooling in most cases are underestimated, which causes the other education variables to be derived improperly.

For the 2018 IMDB, a data quality issue remains due to the increase of missing education qualification and years of schooling. As a result, data on education has been set to missing for admission year 2017 and 2018.

7.4 Coverage

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

The **2018** Integrated Permanent and Non-permanent Resident File (PNRF) contains over 8.0 million records (Table 6); of these, over **6.8** million records (**85.0%**) are linked to at least one tax file. It is to be noted that immigration data belonging to non-taxfilers and taxfilers alike are included in a file named PNRF_ 1980_2018. The following table shows the distribution of records depending on their presence in the different immigration and tax files. About **1.8** million records belong to immigrants who were temporary residents prior to becoming permanent residents; over **1.7** million of these records are linked to at least one tax file. See Appendix B for detailed distribution numbers by landing year.

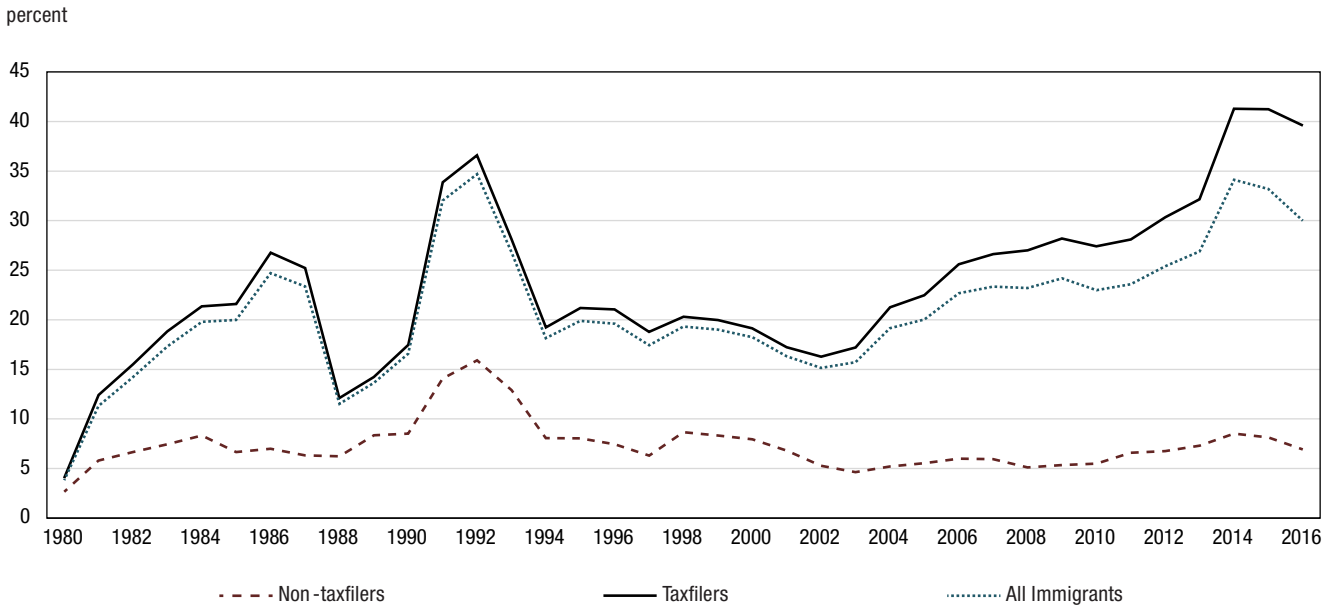
Table 6
Coverage of permanent residents since 1980

	Permanent resident	Permanent resident with non-permanent resident permit	Total
		number	
Total filers	5,100,250	1,714,675	6,814,925
Total non-filers	1,111,240	88,485	1,199,725
Total	**6,211,490**	**1,803,160**	**8,014,655**
		percent	
Percent taxfilers	**82.1**	**95.1**	**85.0**

Source: Statistics Canada, 2018 Longitudinal Immigration Database.

Data on immigrants with non-permanent resident permits are available. The proportion of immigrants with pre-admission experience varies by landing year (Chart 6); it ranges from **3.8%** in **1980** to **37.8%** in **2017**. As a result, the proportion of immigrants with pre-admission experience in the early 1980s is underrepresented. The proportion of immigrant filers with pre-admission experience (solid line) is higher than the overall proportion of immigrants with pre-admission experience (dotted line) because the linkage rate for these immigrants is higher than that for immigrants without pre-admission experience.

Chart 6
Percentage of immigrants with non-permanent resident permits, by admission year



Source: Statistics Canada, 2018 Longitudinal Immigration Database.

7.4.1.2 Coverage of non-permanent residents

This section describes the coverage of individuals who only had non-permanent residents permits since 1980, overall tax records are available for 29.2% of them. Among individuals who have not become permanent residents, asylum seekers have the highest coverage rate, tax records are available for 44.9% of them (Table 6b). There is a wide variety of non-permanent resident permits; some permits are as short as one day.

Table 6b
Coverage of non-permanent residents who never became permanent residents by type of permit

	With work permit	With study permits	Asylum claimants	Total
		number		
Total filers	1,052,635	457,665	131,050	1,235,810
Total non-filers	1,796,965	1,161,415	161,675	3,071,825
Total	**2,849,605**	**1,619,085**	**292,725**	**4,307,640**
		percent		
Percent taxfilers	**36.9**	**28.3**	**44.8**	**28.7**

Source: Statistics Canada, 2018 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 non-permanent and permanent residents. Some permanent residents were non-permanent residents prior to admission. Table 7 gives details on the distribution of linked permanent residents with and without non-permanent permits prior to admission, by tax year. At least one tax file is available for the **79.7%** of permanent residents without a non-permanent permit prior to admission and for the **94.4%** of permanent residents who were non-permanent residents prior to admission. The fact that permanent residents with pre-admission temporary permits have a higher rate of filing taxes than permanent residents without pre-admission 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 7
Permanent and non-permanent residents by tax year

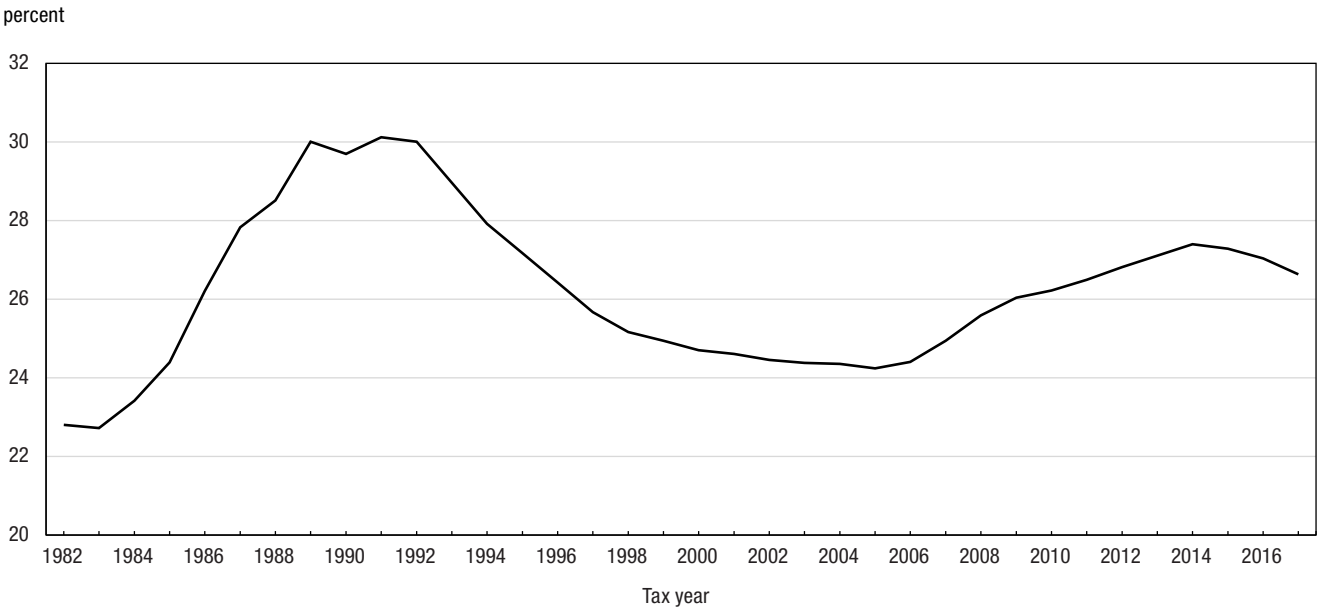
	Permanent resident admitted between 1952 and 1979	Permanent resident since 1980	Permanent resident with non-permanent resident permit	Non-permanent resident only	Number of taxfilers
			number		
1982	1,624,330	184,645	54,550	22,090	1,885,610
1983	1,610,595	221,035	64,990	20,305	1,916,925
1984	1,606,225	260,280	79,595	20,540	1,966,640
1985	1,587,495	294,600	95,050	19,135	1,996,280
1986	1,642,195	352,030	125,035	22,800	2,142,060
1987	1,622,010	411,085	158,520	22,825	2,214,440
1988	1,641,455	502,235	200,305	31,675	2,375,665
1989	1,667,815	614,790	263,595	43,755	2,589,955
1990	1,676,000	736,050	310,885	47,010	2,769,945
1991	1,672,730	834,115	359,485	46,585	2,912,915
1992	1,678,795	940,875	403,330	45,840	3,068,845
1993	1,710,160	1,085,670	442,600	45,635	3,284,060
1994	1,694,930	1,208,490	468,065	45,215	3,416,700
1995	1,678,845	1,322,210	493,320	47,605	3,541,980
1996	1,660,295	1,431,080	514,045	49,515	3,654,930
1997	1,635,695	1,547,135	534,385	51,800	3,769,015
1998	1,611,850	1,649,010	554,500	50,715	3,866,075
1999	1,605,490	1,776,130	590,370	55,355	4,027,340
2000	1,587,875	1,922,900	630,970	63,330	4,205,080
2001	1,574,380	2,083,770	680,170	73,485	4,411,805
2002	1,544,675	2,217,435	717,980	79,960	4,560,055
2003	1,524,870	2,343,060	755,475	85,370	4,708,775
2004	1,509,005	2,477,500	797,605	87,925	4,872,040
2005	1,484,145	2,598,035	831,385	90,640	5,004,205
2006	1,466,480	2,752,040	888,685	95,510	5,202,720
2007	1,447,575	2,879,600	957,125	109,230	5,393,530
2008	1,427,485	3,008,590	1,034,795	133,775	5,604,640
2009	1,407,275	3,127,260	1,101,045	145,045	5,780,625
2010	1,382,720	3,255,905	1,157,385	156,045	5,952,050
2011	1,363,665	3,388,820	1,221,340	167,320	6,141,140
2012	1,337,205	3,509,640	1,285,990	185,515	6,318,355
2013	1,318,860	3,643,910	1,355,230	214,665	6,532,665
2014	1,297,275	3,763,995	1,420,415	250,010	6,731,700
2015	1,270,185	3,890,820	1,460,085	288,640	6,909,735
2016	1,242,100	4,010,320	1,486,020	348,955	7,087,390
2017	1,216,300	4,105,715	1,490,280	352,500	7,164,795
Total taxfilers	**2,031,325**	**5,109,365**	**1,814,480**	**1,234,190**	**...**
Total non-taxfilers	**2,062,340**	**1,302,965**	**107,555**	**3,875,625**	**...**
			percent		
Percent taxfilers	**49.6**	**79.7**	**94.5**	**24.2**	**...**

... not applicable

Note: Permanent residents admitted prior to 1980, in the December 2019 release, included only immigrants admitted between 1974.
Non-permanent residents statistics are for people who obtain their first permits between 1980 and 2017.
Source: Statistics Canada, 2018 Longitudinal Immigration Database.

Chart 7 shows that the proportion of permanent residents who were non-permanent residents prior to admission varies by tax year from a low of 22.7% for the 1983 tax year to a high of 30.1% for the 1991 tax year. Since the 2000s, this proportion has been stable at about 26%.

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



Source: Statistics Canada, 2018 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 **251,100** immigrants who landed in 2006, **102,030 (40.6%)** filed taxes for the first time in 2006, while **16,160 (6.4%)** did so in 2007 and **3,155 (1.3%)** did so in 2015.

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

A validation of the content of the PNRF_1980_2018 was done. While admission and tax data are collected mandatorily from those in scope, some fields may not have been completed. They could be left empty 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 8 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 **2018** PNRF, 25 records had a **birth year** prior to 1880, with 18 records having birth year 1753 with corresponding landing years that are post 1985, even up to 2012.

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 admission records from 2013 (this value is missing for 1195 records, or 0.5% of admissions taking place that year) and 2014, (this value is missing for 5845 records, or 2.3% of admissions taking place that year) and 2015 (missing for 7360 records, or 2.7% of admissions in that year).

The **education variables,** prior to the 2017 cohort, 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 education variables were set to missing for all 2017 and 2018 admissions.

The percentage of valid responses for the **occupation variables** is above 99% for all landing years.

The variables **Family_Status, Mother_Tongue, Official_Language, 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 460 records from the 2011 admissions.

The variable **Official language** has an increasing number of missing values, since 2016, over 7500 per cohort have a missing value.

The variable **Marital_Status** has over 200 missing values per cohort since 2012.

The variables **Destination_CD, Destination_CMA**, and **Destination_CSD** have few missing values; the **2018** IMDB uses the Standard Geographical Classification (SGC) to update the geographical region and code.

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 8
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	8,149,465	97.8	184,900	2.2	0	0.0
Landing_age	8,329,385	99.9	720	0.0	4,260	0.1
Birth_Year	8,334,310	100.0	30	0.0	25	0.0
Gender	8,334,365	100.0	0	0.0	0	0.0
Country_Birth	8,331,255	100.0	3,110	0.0	0	0.0
Country_Citizenship	8,333,425	100.0	940	0.0	0	0.0
Country_Residence	8,318,080	99.8	16,285	0.2	0	0.0
Education_Qualification	7,704,065	92.4	630,300	7.6	0	0.0
Level_of_Education	7,708,750	92.5	625,615	7.5	0	0.0
Years_of_Schooling	7,707,015	92.5	627,350	7.5	0	0.0
Landing_age_6_groups	8,333,645	100.0	720	0.0	0	0.0
Landing_age_9_groups	8,333,645	100.0	720	0.0	0	0.0
Occupation_CD	8,328,070	99.9	6,295	0.1	0	0.0
NOC5-NOC2	8,328,070	99.9	6,295	0.1	0	0.0
Family_Status	8,331,775	100.0	2,590	0.0	0	0.0
Family_Status_rollup	8,331,775	100.0	2,590	0.0	0	0.0
Marital_status	8,330,140	99.9	4,225	0.1	0	0.0
Marital_status_rollup	8,330,140	99.9	4,225	0.1	0	0.0
Mother_Tongue	8,331,610	100.0	2,755	0.0	0	0.0
Official_Language	8,299,580	99.6	34,785	0.4	0	0.0
Skill_level_CD11	8,328,020	99.9	6,345	0.1	0	0.0
Special_Program	1,374,360	16.5	6,960,005	83.5	0	0.0
CSQ_ind	8,334,135	100.0	230	0.0	0	0.0
Destination_CD	8,334,005	100.0	360	0.0	0	0.0
Destination_CMA	8,334,005	100.0	360	0.0	0	0.0
Destination_CSD	8,334,005	100.0	360	0.0	0	0.0
Destination_ER	8,334,005	100.0	360	0.0	0	0.0
Destination_Province	8,334,005	100.0	360	0.0	0	0.0
Permits and NPR-specific variables	1,922,030	23.1	0	0.0	0	0.0
Death_Year	254,730	3.1	390	0.0	0	0.0
Death_Month	254,685	3.1	435	0.0	0	0.0

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, 2018 Longitudinal Immigration Database.

7.6 Quality Assessment of the Province of Residence Variable (PRCO_)

A validation of the geography variables included in the IMDB tax files was done. This section discusses how the variable Province of Residence (PRCO_) was derived and its quality.

The Province of residence (PRCO_) is based on information from tax filers when available. Missing information from the province of residence is replaced by information collected on the postal code of the mailing address either from the individual (PSCO_I), if available, otherwise from the family (PSCO_F).

Table 9
Concordance between PRCO and PSCO_

PRCO	Province and Territories	First character of the postal code (PSCO)
0	Newfoundland and Labrador	A
2	Prince Edward Island	B
1	Nova Scotia	C
3	New Brunswick	E
4	Quebec	G, H, J
5	Ontario	K, L, M, N ,P
6	Manitoba	R
7	Saskatchewan	S
8	Alberta	T
9	British Columbia	V
10	Northwest Territories	X
11	Yukon Territories	Y
12	Non-residents	missing
14	Nunavut	X

Note: The value some postal codes are U or F for blank, respectively U or US is U an Foreign is F.

While the Province of residence (PRCO_) is more reliable than the Taxing province (TXPCO_), some abnormalities were observed mostly on the non-resident code in the reporting for taxation years 1989, 1993, and 1998. These may impact specific provinces.

For the 1993 IMDB_T1FF includes anomalies for the province of Manitoba with an unusual number of residents (48,130 in 1993, compared to 33,650 the tax year before, and 37,365 the tax year after). Similar changes are observed for the Northwest Territories. Additionally, 740 individuals are coded as residing in Nunavut while Nunavut was created in 1998. 725 individuals are coded as residing in multiple jurisdictions. Users can use the information from the variable PSCO_F to diminish the effect of the anomalies on their analyses that include province of residence. However, as stated above, the time are different between PSCO_ (based residence at time of filling) and PRCO_(residence on December 31st).

Non-resident (PRCO_=12) records appear to be overestimated in the 1989 IMDB_T1FF. It includes 79,210 non-residents of Canada, with many of them having a non-permanent residency status. Users can decide to use the postal code of the mailing address (PSCO_ at the individual or family level) to derive the value of PRCO_ or remove the non-residents from their analysis.

In the 1998 IMDB_T1FF, a higher than expected number of records are assigned to Newfoundland and Labrador (PRCO_). In these cases the place of residence of the family at the time of filing is also Newfoundland based on variable PSCO_F.

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-admission 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 admission. 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.

With the 2018 IMDB release, new changes have been made with respect to the persons integrated into the IMDB. First, there has been a file structure change: the taxfilers and the non-taxfilers have merged. As well, there has been a coverage change: The IMDB Universe has changed to include all immigrants since 1952, as well as all non-permanent residents since 1980. The non-permanent (temporary) residents' tax are also now included in the IMDB_T1FFs.

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%.

For the 2015 IMDB, including in the linkage process the Social Insurance Register (SIR) – a database specifically for SIN data – increased the linkage rate to 97%. The Social Insurance Register provides very high-quality data, and about 730,000 Social Insurance Numbers are found exclusively on this register (Diaz-Papkovich 2017).

For the 2016 IMDB, a new record linkage process was used in order to facilitate the linkage of the IMDB to other data sources. Immigration data were linked to tax data via the SDLE (see section 4). From this point the linkage will be to Statistics Canada's Derived Record Depository.

For the 2018 IMDB, in order to improve the record linkage results, a combination of Social Data Linkage Environment (SDLE) linkage results as well as results from the Linkage Control File (LCF) was employed, like the 2015 instalment.

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.

For the 2016 IMDB, a flag was added to identify Express Entry immigration category, along with Syrian refugee resettlement waves and the year and month of citizenship.

For the 2018 IMDB, there has been a coverage change. The IMDB has expanded to include permanent residents admitted from 1952 to 1979, along with the taxfiles of all non-permanent residents since 1980. In addition, several data modules have been added to the IMDB: children, settlement, wages, as well as details on express entry.

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[14] (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 admission. 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-admission information, such as the number of work permits and study permits, is not available in the LAD. Admission 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 of tax filers and the IMDB being limited to immigrant taxfilers. The population of the LAD is estimated by multiplying the records by a weight of 5.

14. For more details on the LAD, please refer to the description available on the Statistics Canada website or enquire about a detailed technical report

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.

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 2012 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 admission. 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. There is a detailed technical report available to obtain more information.

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. Table 12 compares the estimates of immigrant populations by admission 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 estimates between the Longitudinal Immigration Database and the National Household Survey

	NHS estimates	2013 PNRF estimates
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 during their first four years of settlement and to understand the factors that can help or hinder this adjustment. 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 admission, 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 admission while the IMDB is the age at admission. 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 admission. 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 Immigrants 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: IMDB: Longitudinal Immigration Database; NHS: National Household Survey, 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

	LSIC		2013 PNRF	
Age group	number	percent	number	percent
15 to 24	26,730	16.3	27,990	17.9
25 to 34	65,500	39.9	63,050	40.3
35 to 49	53,970	32.9	49,030	31.3
50 to 64	12,890	7.8	12,280	7.8
65 and older	5,100	3.1	4,120	2.6
Total	**164,200**	**100.0**	**156,470**	**100.0**

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 Longitudinal Survey of Immigrants to Canada, Wave 1, 2003.

8.5 Discussion of the IMDB with different linkages

To enhance the analytical capacity of the IMDB, several data sources have been integrated, including the Census, Canadian Community Health Survey (CCHS), Discharge Abstract Database (DAD), General Social Survey (GSS), Longitudinal Administrative Databank (LAD), and Longitudinal Survey of Immigrants to Canada (LSIC). Below a brief overview of each is given.

8.5.1 Census

Conducted every five years, the Census of Population is the primary source of sociodemographic data for specific population groups such as lone-parent families, Aboriginal peoples, immigrants, seniors and language groups. Adjusted population counts from the Census are used as the base for the Population Estimates Program.

The Census is delivered in two questionnaires, the short form and the long form. The short form is used to enumerate all usual residents of all private dwellings in the 2016 Census and residents who are overseas (In 2016, this included Canadian government employees (federal and provincial) and their families, and members of the Canadian Forces and their families). It contains questions on basic demographic information, such as age, sex, knowledge of official languages, household composition, and more.

In 2016, a sample of 25% of Canadian households received a long-form questionnaire. It contains topics ranging from level of education, activity limitations, ethnic origins, and more. Income data were obtained from personal income tax and benefits files. Additional immigration data on admission category were obtained from administrative files from Immigration, Refugees and Citizenship Canada.

The Census undergoes a complex process of frame and sample design, collection, coding, edit and imputation, and certification before dissemination. For more information regarding any aspect of the 2016 Census, please refer to the Guide to the Census of Population, 2016, or the plethora of available reference material.

8.5.2 Canadian Community Health Survey (CCHS)

The Canadian Community Health Survey is a joint project between Statistics Canada and Health Canada. The annual component of the Canadian Community Health Survey (CCHS) collects cross-sectional information about the health, health behaviours, and health care use of the non-institutionalised household population aged 12 or older.

The survey excludes full-time members of the Canadian Forces and residents of reserves and some remote areas, together representing about 4% of the target population. The CCHS was first conducted in 2001 (cycle 1.1), and was repeated every two years until 2005 (cycle 3.1), each time with a sample of size of approximately 130,000. Starting in 2007, the survey was conducted annually (sample size of 65,000). Response rates ranged from 69.8% to 78.9%. Details about the sampling strategy and content are available in the CCHS user guide and data documentation, which are available from your RDC analyst.

The CCHS focus content surveys are designed to provide cross sectional provincial level results on specific focused health topics. Two focus content cycles were used in this linkage project. The CCHS Mental Health and Well-being (2002 and 2012) collected information about mental disorders, mental health system use, and disability associated with mental health problems among the household population aged 15 and older. There is a detailed technical report available to obtain more information.

For more information regarding the CCHS, please refer to the Statistics Canada website.

8.5.3 Discharge Abstract Database (DAD)

The Discharge Abstract Database (DAD) is a national database collecting administrative, clinical, and demographic information on all separations from acute care institutions, including discharges, deaths, sign-outs and transfers, within a fiscal year (April 1 to March 31). With time, DAD has been extended to capture data on day-surgery procedures, rehabilitation, long-term care, and other types of care. Note that DAD is event-based, meaning that there will be more than one record for a person hospitalised more than once in a fiscal year. Collection requirements change by data year and by jurisdiction.

More than 3.2 million abstracts are submitted annually to the DAD, representing approximately 75% of all acute inpatient separations in Canada. Quebec does not submit data to the DAD; Quebec's acute inpatient separations are reported to the Hospital Morbidity Database (HMDB) and usually account for 25% of total inpatient separations in Canada. About 2.4 million day surgery abstracts are submitted to the Canadian Institute for Health Information (CIHI) annually; approximately 35% are sent to the DAD and 65% are sent to the National Ambulatory Care Reporting System (NACRS).

The population of reference usually includes all separations from acute inpatient care and day surgery institutions in Canada (excluding stillbirths and cadaveric donor cases) from April 1 to March 31. All acute care data except that from Quebec is submitted to the DAD; Quebec acute care data is submitted via Quebec's ministère de la Santé et des Services sociaux once per year and is included in the HMDB. Day surgery data from Ontario, Alberta and Nova Scotia is submitted to NACRS.

For more information regarding the DAD, please refer to the Discharge Abstract Database on the CIHI website. There is a detailed technical report available to obtain more information.

8.5.4 General Social Survey (GSS)

GSS (Canada's General Social Survey) is an independent annual, cross-sectional survey, thoroughly examining one topic, in order to monitor changes in living conditions. The GSS is a measure of the Canadian well-being and is able to yield information on specific social policy issues. Each survey collects in-depth socio-demographic data age, sex, education, religion, ethnicity, income, etc.

The GSS is a comprehensive look at several essential topics, including families, caregiving, time-use, victimisation, volunteering, etc. Each of the six survey themes is repeated comprehensively approximately every 5 years.

Until 1998, its sample size was set at 10,000. It increased in 1999 to a 25,000 person target. The larger sample size allows the basic estimates to be available at the provincial, national, and certain census metropolitan area levels (CMAs).

For more information regarding the GSS, please refer to the General Social Survey on the Statistics Canada website.

8.5.5 Longitudinal Administrative Databank (LAD)

The LAD is a random, 20% sample of the T1 Family File (T1FF) tax database. Selection for LAD is based on an individual's SIN. There is no age restriction, but people without a SIN can only be included in the family component. Once a person is selected for the LAD, the individual remains in the sample and is picked up each year from the T1FF if he or she appears on the T1 that year. Individuals selected for the LAD are linked across years by a unique non-confidential LAD identification number (LIN__I) generated from the SIN, to create a longitudinal profile of each individual.

The LAD is augmented each year with a sample of new tax filers so that it consists of approximately 20% of tax filers for every year. The 20% sample has increased from 3,227,485 people in 1982 to 5,579,280 in 2016 (an increase of 73%). This increase reflects increases in the Canadian population and increases in the incidence of tax filing as a result of the introduction of the Federal sales tax credit in 1986 and the Goods and Services Tax credit in 1989.

For more information regarding the LAD, please refer to the Longitudinal Administrative Databank or to LAD Data Dictionary.

8.5.6 Longitudinal Survey of Immigrants to Canada (LSIC)

The Longitudinal Survey of Immigrants to Canada (LSIC) is used to capture information to better understand the lives of recent immigrants to Canada. The LSIC is designed to capture the first four years of their settlement in Canada, providing indicators of how immigrants are meeting challenges such as knowing or becoming more fluent in one of both of Canada's official languages, participating in labour market, accessing or education training. The first four years are a time immigrants when build their ties to Canada, economic, social, and cultural.

Objective of the survey is to: study the lives of new immigrants in Canada their adjustment over time and to see what facilitates and hinders their integration into Canadian society.

The target population for the survey consists of immigrants that must fulfil the following requirements:

-Must have been 15 years and older at the time of admission

-Must have arrived in Canada between 1st October 2000 and 30th September 2001

-Must have landed from abroad as well as have applied through a Canadian Mission Abroad

The population of interest to LSIC are those immigrants still living in Canada at the time of the interview

The survey comes in three waves, which are three separate questionnaires that were each put through a rigorous testing process.

For more information regarding the LSIC, please refer to Longitudinal Survey of Immigrants to Canada

9 Possible Analyses 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 admission can be used to establish the pre-admission profile of immigrants. By comparing these populations (with and without pre-admission Canadian experience), it becomes possible to assess the impacts of pre-admission Canadian experience on the economic outcomes and mobility patterns of immigrants. The specific sociodemographic profile at time of temporary resident permit issuance makes it also possible to evaluate economic outcome and mobility prior to admission. 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 admission by immigration category and economic profile.

9.3 Analytical possibilities with citizenship

Adding citizenship information to the IMDB offers more analytical possibilities. The addition of the year and month of Canadian citizenship provides details on whether and when immigrants obtained their citizenship. It also serves as an additional explanatory variable to study socioeconomic outcomes. For example, the citizenship flag provides the uptake rate over time and informs on the characteristics that are associated with seeking citizenship. Please note that citizenship data are available since 2005.

9.4 Analytical possibilities with Children

The addition of the children module to the IMDB allows for further analysis of data on the socioeconomic conditions of immigrant children during their childhood. Children (persons less than 18 years old) represent about 25% of immigrants admitted in Canada since 1980. Immigrant children face different difficulties and challenges from their parents/guardians. With this module, one can analyse the economic outcome of immigrant children based on economic status as children. Immigrant children from lower income families and their future economic outcome relative to immigrant children from higher income families. Data is also available for Syrian refugee children.

9.5 Analytical possibilities with Express Entry

Express entry is an application process for economic immigrants who want to settle in Canada permanently and take part in our economy. Adding the express entry will improve the analytical capacity with respect to detailed selection criteria and wages among immigrants.

As an example, in 2017, immigrants admitted in 2015 as Federal Skilled workers reported the highest wages among immigrants admitted through the EE system. Among their characteristics we can verify the impact of having a job offer at the time of application.

9.6 Analytical possibilities with salaries and wages files

Preliminary wages and salaries tax files contain salary, wages, and taxable benefits paid to employees. Variables extracted from these files include province of employment, province of employee, T4 earnings per by tax year, and number of T4 slips per tax year.

This data allow analysis on immigrants and non permanent residents that did not file a T1 tax form. Additionally, this information is timelier than the annual tax files, allowing the information to be more up to date on the economic outcome and the geographic location of permanent and non permanent residents. Finally, having multiple T4 records informs on the employment stability.

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-admission experience) who landed in Canada in 1952 or thereafter and non-permanent residents since 1980. The IMDB allows for analysis of tax filers and non taxfilers. Readers should keep in mind that the profile of taxfilers can be completely different to that of non-filers. This technical report was produced to give a thorough description of IMDB data quality and recent changes to this database.

With the 2018 IMDB release, the content has expanded with new modules being integrated. These modules include settlement, data on immigrant children, and wages. In addition, the IMDB also includes express entry data, as well as citizenship acquisition since 2005.

As well, there has been a file structure change- the taxfilers and the non-taxfilers have merged. In the past, it was PNRF_YEAR for filers and PNRF_NONFILERS_YEAR for nonfilers; now merged the file is called PNRF_1980_2018.

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 “Immigration and ethnocultural diversity” under ”Subjects” and typing “Longitudinal Immigration Database (IMDB)” under “Keywords”.

It is to be noted that yearly updates of the IMDB are independent from one another. From year to year, there have been changes to data processing, including updates to the unique person identifier (IMDB_ID).

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)

Portal on Immigrants and Non-permanent Residents Statistics**:** The immigrants and non-permanent residents portal brings together the most requested data, tools and reports on a single page.

Historical IMDB:

IMDB releases in The Daily:

Analysis using the IMDB:

Evra, R. and Kazemipur, A. 2019. *The Role of Social Capital and Ethnocultural Characteristics In The Employment Income Of Immigrants Over Time*. Statistics Canada: Insights On Canadian Society.

Huystee, M. 2016. Interprovincial mobility: Retention rates and net inflow rates 2008-2013 admissions.

Ng, E et al. 2019. Tuberculosis-Related Hospital Use Among Recent Immigrants To Canada. Statistics Canada: Health Reports.

Picot, G, and Lu, Y. 2017. Chronic Low Income Among Immigrants in Canada and its Communities. Statistics Canada.

The Consumer Price Index (62-001-X)

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

B. Coverage

The 2018 IMDB was used to produce these counts. Filers are linked immigrants who have filed a tax return at least once since 1982. Statistics below exclude the 2018 admissions.

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

	Taxfilers[1]	Non-taxfilers	Total	
	Immigrants	Immigrants	Immigrants	Taxfilers
		number		percent
1980	120,385	22,740	143,125	84.1
1981	107,115	21,465	128,580	83.3
1982	103,085	18,005	121,090	85.1
1983	77,060	11,975	89,035	86.6
1984	77,450	10,575	88,025	88.0
1985	74,920	9,020	83,940	89.3
1986	88,475	10,295	98,770	89.6
1987	136,460	14,710	151,170	90.3
1988	144,475	16,280	160,755	89.9
1989	171,750	18,910	190,665	90.1
1990	193,715	21,720	215,430	89.9
1991	210,230	21,590	231,820	90.7
1992	230,590	23,345	253,935	90.8
1993	233,135	22,540	255,675	91.2
1994	201,505	22,090	223,595	90.1
1995	191,140	21,010	212,155	90.1
1996	201,410	23,950	225,360	89.4
1997	192,310	23,150	215,460	89.3
1998	158,845	14,840	173,685	91.5
1999	173,820	15,545	189,365	91.8
2000	208,405	18,340	226,745	91.9
2001	227,715	22,050	249,770	91.2
2002	204,535	23,670	228,210	89.6
2003	194,680	25,850	220,530	88.3
2004	204,540	30,805	235,345	86.9
2005	223,775	38,005	261,780	85.5
2006	213,785	37,320	251,110	85.1
2007	198,505	37,670	236,180	84.0
2008	203,720	42,895	246,615	82.6
2009	207,235	44,350	251,585	82.4
2010	223,480	56,595	280,070	79.8
2011	195,955	52,165	248,120	79.0
2012	203,515	53,725	257,240	79.1
2013	203,815	54,690	258,505	78.8
2014	202,800	56,705	259,505	78.1
2015	205,140	65,860	271,000	75.7
2016	208,850	86,545	295,395	70.7
2017	196,600	88,725	285,323	68.9
Total	**6,814,925**	**1,199,725**	**8,014,655**	**85.0**

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

Note: All counts are rounded.
Source: Statistics Canada, 2018 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 49	50 to 64	65 and older	Total
Sex and cohorts	percent						
1980 to 1989 cohorts							
Male	0.83	0.94	0.95	0.92	0.83	0.60	0.89
Female	0.82	0.92	0.93	0.91	0.80	0.58	0.87
Total	**0.83**	**0.93**	**0.94**	**0.92**	**0.81**	**0.59**	**0.88**
1990 to 1999 cohorts							
Male	0.84	0.94	0.94	0.93	0.90	0.77	0.91
Female	0.83	0.95	0.94	0.93	0.88	0.76	0.90
Total	**0.83**	**0.94**	**0.94**	**0.93**	**0.89**	**0.76**	**0.90**
2000 to 2009 cohorts							
Male	0.61	0.96	0.93	0.93	0.93	0.89	0.86
Female	0.60	0.96	0.94	0.94	0.93	0.88	0.87
Total	**0.60**	**0.96**	**0.94**	**0.94**	**0.93**	**0.88**	**0.87**
2010 to 2017 cohorts							
Male	0.11	0.88	0.95	0.93	0.90	0.83	0.75
Female	0.11	0.90	0.95	0.94	0.89	0.82	0.77
Total	**0.11**	**0.89**	**0.95**	**0.94**	**0.89**	**0.82**	**0.76**

Source: Statistics Canada, 2018 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 2017 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 2007 and 2012, living in Ontario between 2015 and 2017, and did not have any earnings appearing on their T4 slips by sex and year (2015 to 2017)"

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 2007 to 2012
        ii. lived in Ontario from 2015 to 2017
        iii. did not have any earnings on their T4 slips

        And generate the number of SA recipients by **sex** and **year** (in this case, 2015 to 2017).
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**, 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 2015 to 2017 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 2015, the value for RRSPSI2015 (contributions to a spouse's RRSP) should be "0" (zero). If that individual did not file in 2015, the value will be missing.

Sample IMDB program

```
*Sample SAS program using the IMDB;

libname source1 'FILEFOLDER1'; * location of IMDB files ;
libname Out 'FILEFOLDER2'; * 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, 2015 to 2017). Data for
provinces and earnings are from the yearly IMDB files whereas the sex variable is from
the PNRF _ 1980 _ 2018 ;

* 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 _ 2015(where=(prco _ i2015 = 5 and outlier _ ind2015=0) in=a
keep=imdb _ id prco _ i2015 saspyf2010 t4e _ _ i2015 outlier _ ind2015)

source1.imdb _ t1ff _ 2016(where=(prco _ i2016 = 5 and outlier _ ind2016=0) in=b
keep= imdb _ id prco _ i2016 saspyf2016 t4e _ _ i2016 outlier _ ind2016)

source1.imdb _ t1ff _ 2017(where=(prco _ i2017 = 5 and outlier _ ind2017=0) in=c
keep= imdb _ id prco _ i2017 saspyf2012 t4e _ _ i2017 outlier _ ind2017)

source1.pnrf _ 1980 _ 2018(keep= imdb _ id gender landing _ year immigration _
category);

by IMDB _ id ;

If a and b and c and (landing _ year>=2007 and landing _ year<=2012);
*person must be taxfiler in all three years, not be flagged as an outlier, and must
have landed between 2007 and 2012 (population of interest);
* We create a flag variable that identifies the SA recipients for each year.

The result is three variables,
flag _ sa2015, flag _ sa2016 and flag _ sa2017, taking a value of either 1 or 0.;
If (t4e _ _ i2015=0 and saspyf2015>0) then flag _ sa2015 = 1 ;
else flag _ sa2015 = 0 ;
if (t4e _ _ i2016=0 and saspyf2016>0) then flag _ sa2016 = 1 ;
else flag _ sa2016 = 0 ;
if (t4e _ _ i2017=0 and saspyf2017>0) then flag _ sa2017 = 1 ;
else flag _ sa2017 = 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 _ sa2015*flag _ sa2016*flag _ sa2017
gender*flag _ sa2015*flag _ sa2016*flag _ sa2017 /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 category. 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 admission cohort about which retention is calculated (the denominator in the retention rate).

The retention rate can be measured as 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 admission), the denominator is the number of taxfilers with the selected province of admission. The numerator is the number of taxfilers with the selected province of admission who are also residing in the province.

To compute retention rates three years after admission for the 2011 cohort, a researcher would prepare a table with all provinces of admission (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					
	Total province of residence	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Quebec
Province of landing	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					
	Ontario	Manitoba	Saskatchewan	Alberta	British Columbia	Other residence
Province of landing	number of immigrants					
Total province of landing	**70,590**	**9,698**	**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, 2014 Longitudinal Immigration Database (table 43-10-0035-01).

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%. Table 19 also provides information on secondary migrants[15]—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.

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

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 sporadic tax filing behaviour or emigrants that left Canada, for which tax file was not available in 2014.

One alternative is a purely longitudinal approach, where a single admission 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 admission 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 admission, it may be necessary to increase the population size for a region by defining the admission cohort as anyone who first filed taxes in the region within two years of admission (i.e., first_tax_year = landing_year or landing_year+1). Allowing individuals whose first tax filing occurred several years after admission to be part of an "admission 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 admission 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 admission 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 admission 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, 2014 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.

With the **2018 IMDB** release, a mobility summary table is available on the Statistics Canada website. The measures for mobility compare the intended destination from immigration files to the province of residence obtained from tax files. For example, table 21 provides the mobility measures based on the differences between the intended province of destination for immigrants admitted in 2010 and their province of residence in 2015 according to their tax files.

Table 21
Mobility measures for 2010 cohort by province, 2015 tax year

	Total destination (a)	Total residence (b)	Out migration (c)	In migration (d)	Stayed in province (e=a-c)	Population growth rate (f=b/a-1)	Retention rate (g=e/a)	Out migration rate (h=1-e/a)	In migration rate (i=d/a)
Canada	200,600	200,600	27,260	27,260	173,340	0.0	86.4	13.6	13.6
Newfoundland and Labrador	525	410	245	130	280	-21.9	53.3	46.7	24.8
Prince Edward Island	1,930	370	1,630	70	305	-80.8	15.8	84.5	3.6
Nova Scotia	1,630	1,405	570	340	1,065	-13.8	65.3	35.0	20.9
New Brunswick	1,535	920	795	180	740	-40.1	48.2	51.8	11.7
Quebec	38,050	33,900	5,955	1,805	32,095	-10.9	84.4	15.7	4.7
Ontario	83,355	84,965	7,725	9,335	75,630	1.9	90.7	9.3	11.2
Manitoba	11,475	9,785	2,420	730	9,055	-14.7	78.9	21.1	6.4
Saskatchewan	5,620	5,410	1,220	1,015	4,400	-3.7	78.3	21.7	18.1
Alberta	24,255	29,850	2,360	7,955	21,895	23.1	90.3	9.7	32.8
British Columbia	31,820	32,790	4,250	5,215	27,575	3.1	86.7	13.4	16.4
Other	405	420	95	115	305	3.7	75.3	23.5	28.4
Not stated	..	365	..	365	..	0.0	0.0	0.0	0.0

.. not available for a specific reference period

Source: Statistics Canada, 2016 Longitudinal Immigration database (IMDB).

The new table provides the following measures of mobility:

- The **total destination** (column a) represents the number of immigrants admitted in 2010 and filling taxes in year 2015, in Canada;
- The **total residence** (column b) represents the number of immigrant taxfilers in 2015 in the province specified;
- The **out migration** (column c) represents the number of immigrant taxfilers originating from the specified province and filing tax in another province, in year 2015;
- The **in migration** (column d) represents the number of immigrants originating from a difference province of destination and filing tax in the specified province in year 2015;
- The **stayed in the province** (column e) represents the number of immigrant taxfilers continuing their residence from the province of destination, in year 2015;
- The **population growth rate** (column f) represents the percentage of immigrant taxfilers gained or lost by the specified province. This takes into account immigrants migrating out and migrating in the specified province;
- The **retention rate** (column g) represents the percentage of immigrant taxfilers continuing their residence from the province of destination, in year 2015. This does not take into account immigrants migrating in from another province of destination;
- The **out migration rate** (column h) represents the percentage of immigrant taxfilers originating from the specified province and filing tax in another province in year 2015;
- The **in migration rate** (column i) represents the percentage of immigrants originating from a different province of destination and filing tax in the specified province in year 2015.

The table 21 shows that 200,600 immigrants were admitted to Canada in 2010 and filed taxes in 2015.

Of the 83,355 immigrant taxfilers who intended to reside in Ontario, 75,630 remained there in 2015, representing a retention rate of 90.7%.

While 7,725 immigrant taxfilers migrated out of Ontario, 9,335 immigrant taxfilers had moved into Ontario from other destination provinces. So, for this 2010 cohort, the total number of Ontario residents in 2015 was 84,965, or 1.9% more than the number of immigrant tax filers who intended in reside in Ontario.

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 wages, salaries and commissions earnings (or, for that matter, any economic variable) requires consecutive information. For example, if a researcher wants to compare the median wages, salaries and commissions earnings of the 2000 cohort of women aged 24 to 54, 1 year after admission and 5 years since admission (Table 22), 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 22
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	Wages income 2001	Wages 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 20). One exception is pre-filers, those who filed taxes in Canada before admission 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, Consumer Price Index (CPI) adjustments should be made (Appendix D.7). This adjustment is made in the IMDB tables.

D.5 Rounding data

Respecting the privacy of Canadians is important to Statistics Canada. Consequently, any tables produced from IMDB_T1FF 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 available to all researchers accessing the microdata in the Research Data Centres (RDC).

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 (Table 23) gives the distribution of the outliers in the tax files for 1982 and subsequent years by type of resident for the **2018 IMDB**. Less than 0.05% 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 23
Distribution of outliers by tax year

	Total	
Year	number	percent
1982	460	0.02
1983	385	0.02
1984	555	0.03
1985	495	0.02
1986	455	0.02
1987	590	0.03
1988	955	0.04
1989	915	0.04
1990	735	0.03
1991	775	0.03
1992	985	0.03
1993	855	0.03
1994	490	0.01
1995	720	0.02
1996	1,470	0.04
1997	1,865	0.05
1998	2,415	0.06
1999	1,760	0.04
2000	1,970	0.05
2001	1,915	0.04
2002	1,995	0.04
2003	2,180	0.05
2004	1,935	0.04
2005	2,175	0.04
2006	2,390	0.05
2007	2,285	0.04
2008	2,430	0.04
2009	2,625	0.05
2010	2,375	0.04
2011	2,355	0.04
2012	2,255	0.04
2013	2,505	0.04
2014	2,280	0.03
2015	2,340	0.03
2016	2,205	0.03
2017	2,525	0.04

Source: Statistics Canada, 2018 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."[16] The adjustment factors for 2017 are available in Table 24. To transform data to constant dollars of a specific year, data users need to multiply the dollar values in all but the reference year by a year-specific adjustment factor. To obtain the adjustment factors, data users need to divide the CPI of the reference year by the CPI of the specific year. In table 24, the year of reference is 2017.

Table 24
2017 Consumer price index adjustment factors

Year	2017 consumer price index adjustment equals 130.4 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
2015	126.6
2016	128.4
2017[1]	130.4

1. In 2018, the CPI was 133.4. In order to transform a price from another year into 2018 dollars, one must multiply the specified year's dollar amount by the inflation ratio, which is the 2018 CPI (133.4) divided by the specified year's CPI.

Source: Statistics Canada, Table 18-10-0005-01.

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

D.8 Calculating key income measures

The IMDB tables contain several income measures. Table 25 describes which variables of the T1FF are included in their calculation.

Table 25
Description of the Longitudinal Immigration Database income main measures

Measure	Components	Formula
Wages, salaries and commissions income	Earnings from T4 slips	T4E__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 + CLKGX
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, 2018 Longitudinal Immigration Database

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 https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X19990015642 (accessed November 5, 2018). September 19, 2019

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.

Cascagnette, P., and SDLE production section. 2019. "*Social Data Linkage Environment (SDLE) Methodology Report – Linkage between the Immigration File (1952- January 2019) and the SDLE Derived Record Depository* (version 29)". Unpublished document. Ottawa: Statistics Canada.

Diaz-Papkovich, A. 2017. I*MDB Linkage Summary 2017*. 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 https://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 https://www.ncbi.nlm.nih.gov/books/NBK253313/ (accessed March 14, 2017). September 19, 2019

Government of Canada. 2016. *Determine your eligibility – Citizenship*. Available at https://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 https://www.irb-cisr.gc.ca/Eng/RefClaDem/Pages/RpdSpr.aspx (accessed January 13, 2016).

Immigration, Refugees and Citizenship Canada. 2018. *Annual Report to Parliament on Immigration*. Available at https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/annual-report-parliament-immigration-2018/report.html (accessed September 19, 2019)

Immigration, Refugees and Citizenship Canada. 2016. *Report on Plans and Priorities 2015-2016*. Available at https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/report-plans-priorities/2015-2016.html (accessed November 5, 2018). September 19, 2019

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 https://www150.statcan.gc.ca/n1/pub/82-003-x/2015010/article/14228-eng.pdf (accessed March 13, 2017). September 19, 2019

Statistics Canada. 2016. *150 years of immigration in Canada*. Canadian Megatrends. Statistics Canada Catalogue no. 11-630-X. Available at https://www150.statcan.gc.ca/n1/pub/11-630-x/11-630-x2016006-eng.htm (accessed March 14, 2017). September 19, 2019

Statistics Canada. 2019. "Annual Income Estimates for Census Families and Individuals (T1 Family File)." Web page. Available at https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4105 (accessed September 19, 2019).

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