UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Sixty-eighth plenary session
Geneva (Switzerland), 22-24 June 2020

GUIDANCE ON THE USE OF LONGITUDINAL DATA FOR MIGRATION STATISTICS

Prepared by the Task force on the use of longitudinal data for migration statistics

Summary

This document presents for your comments the draft *Guidance on the use of longitudinal data for migration statistics*.

The Guidance was prepared by the Task Force on the use of longitudinal data for migration statistics, composed of representatives from Canada (chair), Austria, Belgium, Germany, Italy, Kazakhstan, Mexico, Netherlands, Russian Federation, Spain, Switzerland, Turkey, United Kingdom, Eurostat, OECD and UNECE.

In February 2020, the CES Bureau reviewed the Guidance and requested the UNECE secretariat to send the document to all CES members for consultation.

The deadline for comments is 24 April 2020. Please provide your comments using the attached questionnaire and send to social.stats@un.org. The Secretariat will summarize the feedback received and present it to the CES plenary session (22-24 June 2020, Geneva). Subject to a positive outcome of the consultation, CES will be invited to endorse the Recommendations.

Acknowledgements

The Guidance was prepared by the UNECE Task Force on the Use of Longitudinal Data for Migration Statistics, which consisted of the following members:

Scott McLeish (Canada) – Chair of the Task Force

Stephan Marik-Lebeck (Austria)
Patrick Lusyne (Belgium)
Rose Evra, Tristan Cayn, Elena Prokopenko (Canada)
Gunter Brückner, Jan Eberle (Germany)
Cinzia Conti, Enrico Tucci (Italy)
Bakytkul Uteulina (Kazakhstan)
Edgar Vielma Orozco (Mexico)
Han Nicolaas (The Netherlands)
Liana Vologirova (Russian Federation)
Jorge Vega Valle, Amor Gonzalez Redondo (Spain)
Florence Bartosik (Switzerland)
Alper Acar, Neriman Can Ergan, Şerife Dilek Yilmaz (Turkey)
Nicky Rogers, Louisa Blackwell, Justine McNally, Katherine Worley (United Kingdom)
Giampaolo Lanzieri (Eurostat)
Cécile Thoreau, Christophe Dumont (OECD)
Andres Vikat, Adam Thomas (UNECE)

Contributions from the following individuals are acknowledged as well:

Edward Ng, Hélène Maheux (Canada)
Anna Sára Ligeti (Hungary)
Minja Tea Dzamarija (Norway)

Contents

Chapter 1: Introduction

1. As the number of international migrants continues to grow, it is becoming increasingly important for the public and policymakers to understand migratory flows and the impact of migration on individuals, families, societies and economies. In many cases, the key questions pertain to the process of migrant settlement – how long migrants stay in the receiving country, how they integrate with the receiving societies and how their socioeconomic outcomes change over time.

2. Ultimately, to study migration is to study change. This begins with changes in residence but can expand to include changes in legal or residence status and changes in socio-economic outcomes. Migration, integration and settlement are processes, not states, and outcomes can be short- or long-term. Because of this transient nature, these are topics well suited to be studied using a longitudinal approach. The need for a temporal basis for migration statistics was underscored in the Global Compact for Safe, Orderly and Regular Migration, which calls for data that "allows for effective monitoring and evaluation of the implementation of commitments over time".

3. Traditional methods for longitudinal data collection (e.g. cohort studies and panel surveys) have become more challenging to undertake due to high costs and attrition. Consequently, countries are increasingly turning towards alternative data sources to produce longitudinal results.

4. The production of statistics in many countries has been evolving due to the increased availability and usability of administrative data. More and more, administrative data holdings are improving in terms of the completeness, frequency and quality of the information collected. With the increasingly widespread use of administrative data and data integration for producing migration statistics, more and more countries can construct longitudinal datasets without bearing excessive costs.

5. While in the past longitudinal studies of migration were often standalone or produced on an ad hoc basis, national statistical offices (NSOs) are beginning to incorporate them more and more into their regular production of migration statistics. However, there are currently no international guidelines on how to develop longitudinal data sources from integrated data, how to address various challenges associated with such projects, and how to disseminate key indicators and other findings.

6. The use of a longitudinal approach for measuring integration of migrants has been discussed several times at the joint UNECE-Eurostat Work Session on Migration Statistics (e.g. in 2012, 2014 and 2017). In October 2017, the Work Session recommended pursuing further methodological work on this topic, to review the national practices and develop recommendations that will promote the international comparability of longitudinal data.

7. In February 2018, the Conference of European Statisticians (CES) Bureau established a UNECE Task Force on the use of longitudinal data for migration statistics. The objective of the Task Force was to prepare guidelines on how to incorporate longitudinal data into annual migration statistics and complement the available cross-sectional measurements. This Guidance contains the results of that Task Force.

8. The Guidance builds on recent methodological work by UNECE task forces. The publication "Measuring change in the socio-economic characteristics of migrants" (UNECE 2015) illustrated the benefits of using longitudinal data and recommended that countries develop data linking methodologies to acquire longitudinal data sets. The subsequent "Guidance on data integration for measuring migration" (UNECE 2019b) presented several examples where integration of different datasets led to the compilation of longitudinal data.

9. The Guidance consists of three main substantive chapters:

- Chapter 2: Overview of longitudinal data sources for migration statistics
- Chapter 3: How to develop a longitudinal data set for migration statistics using integrated data
- Chapter 4: Disseminating regular migration statistics from longitudinal data sources

10. The chapters are organized as standalone parts and are not necessary to read in combination with the other parts of this report. What is important is for users to take advantage of the parts most relevant to their work.

11. Chapter 2 provides an overview of longitudinal data sources for migration statistics. It provides guidance on different types of data for migration statistics and offers concrete examples from various countries. A dedicated section features guidance particular to the use of population registers for migration statistics including specific challenges and means to address them.

12. Chapter 3 presents a guide to producing longitudinal data sets for migration statistics using integrated data. With reference to work done by previous task forces, it provides a recipe to follow from statistical design to dissemination including common challenges and concrete examples of best practices from various countries. Phases covered include:

- Statistical design
- Assessment and pre-processing of source files
- Data integration for longitudinal data
- Assignment of longitudinal individual identifiers
- Create final database
- Disseminate results

13. Chapter 4 includes a proposed set of longitudinal migration indicators along with best practices for dissemination of longitudinal results. Each indicator presented includes a summary of challenges and best practices with illustrative examples from various countries.

14. One overarching theme in this Guidance is that the development and use of longitudinal data for migration statistics is complex and challenging. However, the individual chapters illustrate how these limitations can be addressed. Chapter 4 highlights how indicators of value can still be produced even with limitations of the resulting database. To address the emerging and growing needs to better understand the migration, integration, and settlement patterns of international migrants, it is important for NSOs to understand the data sources available, consider how limitations can be mitigated or addressed, and propose approaches to disseminate results that speak to these patterns.

15. Through the increased use of integrated data and statistical registers, many national statistical offices are sitting on a mountain of unexploited potential for longitudinal migration statistics. Chapter 3 provides the recipe for countries to develop new longitudinal data sources. Chapter 4 provides examples of what can be done with the results.

16. Ultimately, the study of migration is a natural application of longitudinal analysis. The increased use of data integration and statistical registers opens new possibilities to develop and disseminate longitudinal data for migration statistics. While there are challenges, the benefits of exploiting these new possibilities are far greater. To study migration is to study change – and to study change, longitudinal data are essential.

Chapter 2: Overview of longitudinal data sources for migration statistics

2.1 Definition

17. **Longitudinal data refers to information which is collected from the same units of analysis, such as individuals or households, over time.** Longitudinal analysis can uniquely and accurately describe individual trajectories through time. Longitudinal data allow us to study and understand life events and transitions, over the life course and inter-generationally. This is particularly useful for the study of international migration, since settlement into a new country is a long-term process. Longitudinal data can reveal the geographic and socio-economic outcomes of the migration experience. Longitudinal data collected over an extended period allows us to understand not only the migrants' experiences but also those of their children.

18. In this chapter, different types of longitudinal data are described for studying international migration. Cross-sectional data are also examined for how they can be used to provide a longitudinal perspective on migration and the experiences of migrant groups. Complementarity and comparability between longitudinal and cross-sectional measures are also considered, and designs that combine both approaches are discussed.

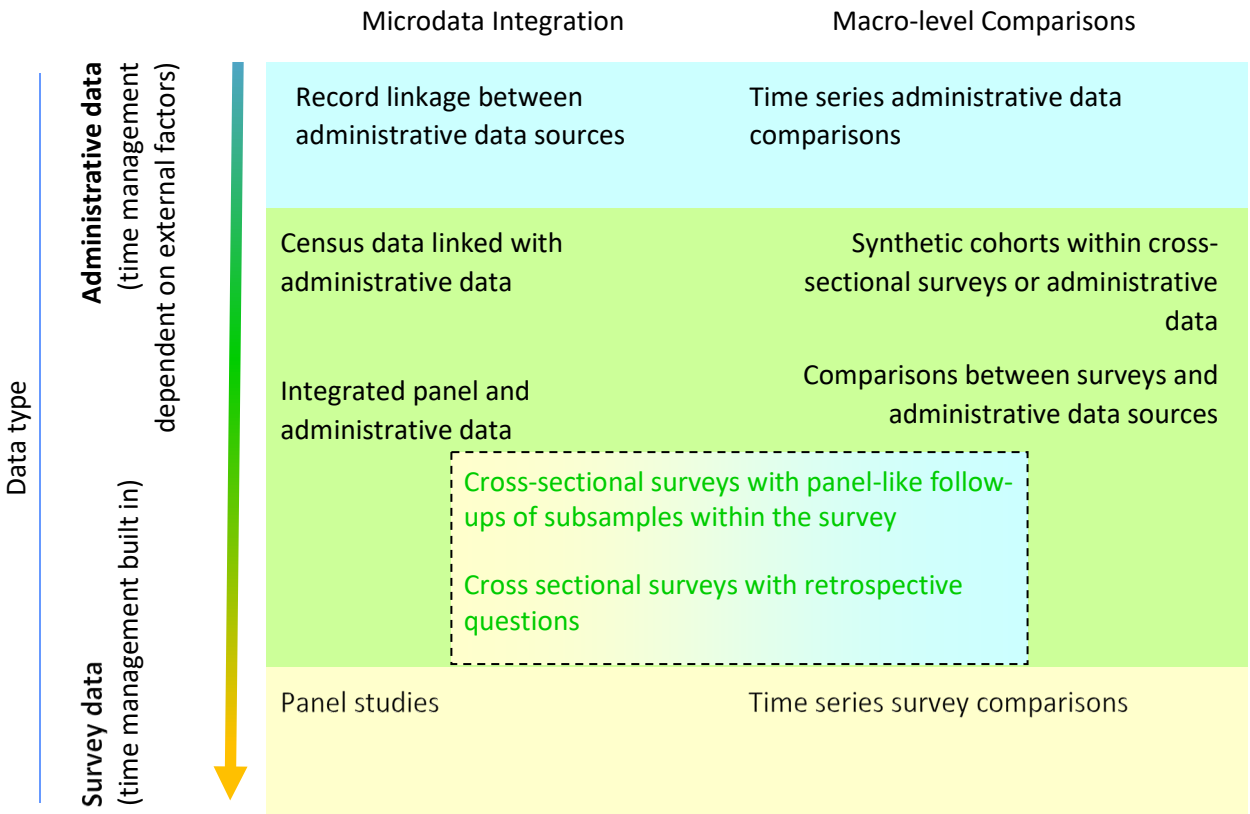
19. Examples would be panel designs embedded within cross-sectional surveys. A further example is retrospective questions to yield longitudinal evidence, within a cross-sectional survey design. Synthetic cohorts in successive censuses or surveys mimic longitudinal studies, in some ways, but do not follow the same cases over time.

20. Finally, as many countries rely on population registers to calculate migration figures, the features of these registers that make them well or less suited to assess the longitudinal aspects of migration are also studied in some detail.

2.2 Types of Data

21. The diagram below provides a conceptual typology for longitudinal information for international migration statistics.

Figure 1 Typology of longitudinal data and perspectives for migration statistics



22. As can be seen from Figure 1, data can be classified by whether they link information for the same units over time, and are therefore longitudinal data, or by whether they provide some longitudinal perspective. 'Microdata integration' refers to data that are linked at record level, over time. 'Macro-level comparisons' refer to longitudinal *perspectives* gained through time series comparisons, or by using synthetic cohorts. Unlike microdata integration, macro-level comparisons do not necessitate that the same individuals are sampled for each wave of data collection.

23. Methodological approaches range along a continuum with a varying balance between the use of survey (coloured yellow in the diagram) and administrative data (blue). Methods which make use of both survey and administrative data are coloured green; these include cases of integrated survey and administrative data and of survey/administrative data hybrids. With survey data, national statistical offices typically can control not only which data are collected but also the timing of data collection (within given resources). With administrative sources, NSOs have less control over data collection and must align time referencing as best as possible, to meet the desired research or analytical objectives.

24. Later in this chapter, two dimensions of time are considered that require careful thought when using administrative data longitudinally. The first is the *endogenous* operationalisation of time in the research design, which implies that observations are made at the time that the measurement events actually take place. The second is the *exogenous* manifestation of time in the administrative dataset, which implies that registrations are done in line with administrative practices, regulation, infrastructure and the quality of the administration. Disparities between the endogenous operationalisation and the exogenous manifestation of time need to be understood, and are a

source of potential bias in the data. The deregistration issue is an explicit example of this. This is where emigration is not recorded in administrative data, and many sources, population registers in particular, have this problem.

25. New prospective panel studies involve the selection of a sample (of individuals or households but could also include for example businesses or events). The sample is surveyed and then followed up, at regular intervals over time. Longitudinal data from panel surveys could also be in the form of new retrospective data. In this type of data, respondents are surveyed and asked about events that happened in the past. A prospective design with a retrospective benchmark combines the two types, with a retrospective benchmark study at the beginning, followed up with prospective follow-up over time.

26. National statistical offices are increasingly using administrative data sources for statistical analysis. Analysis of migrant outcomes could be based entirely on administrative records, through linkage of the records to create longitudinal datasets and time series comparisons of administrative data.

27. A combined approach can take a number of forms. It can take survey information at the beginning and update this with administrative records. Table 1, below, describes the creation of longitudinal data from the linkage of census and administrative records, in the UK Longitudinal Studies. A further alternative is the combination of cross-sectional and longitudinal designs, as seen in the EU Labour Force Survey. Retrospective questions within cross-sectional surveys also produce longitudinal insights.

28. **In the absence of longitudinal information drawn from repeated measures of the same individuals, the longitudinal perspective can be derived from synthetic cohorts.** Cohorts might be defined as 'groups of people marching together through time'; they can be defined with reference to their birth, or the period of their arrival in a country, for example.

2.2.1 Longitudinal Panel Surveys

29. A prospective longitudinal panel survey is a research design which intentionally involves observations of the same sample or of overlapping samples at different points in time. While they tend to be designed for quantitative analysis, they can include the collection of qualitative data. The observation period can be any length of time, from data collected over a number of days to studies that span decades. While longitudinal cohort studies tend to focus on an age cohort who were born in a specified month or year, panel studies typically draw their sample from the full age range. In the example from Canada in Table 1, there is a cohort of migrants defined in terms of their period of arrival in Canada, encompassing all migrants aged over 15 years.

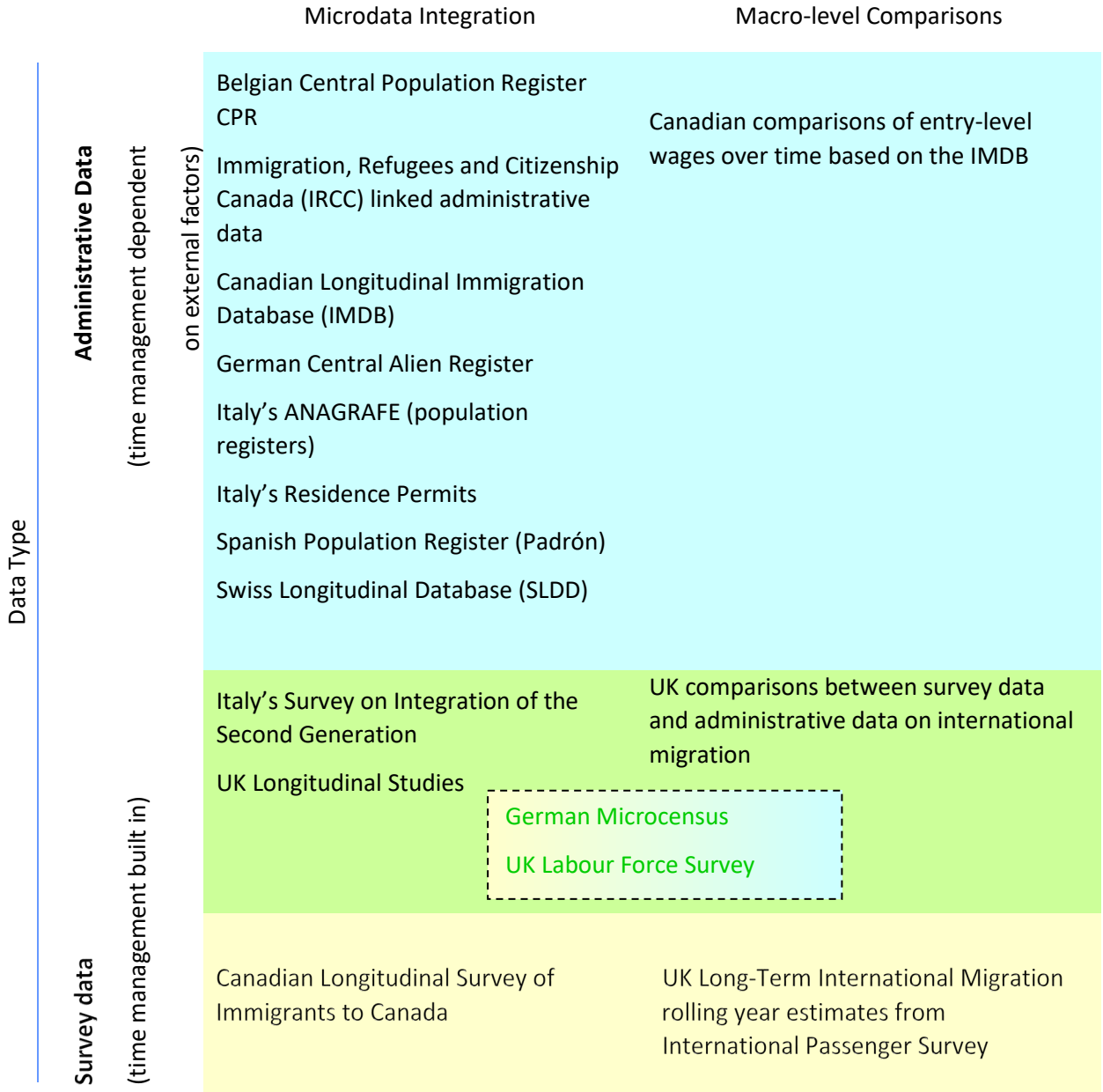
30. There are particular methodological challenges in the use of longitudinal panel surveys for analysing migrant experiences. A common concern for all longitudinal data collection is sample attrition. This occurs when sample members either fail to respond or cannot be found. Migrants tend to be more geographically mobile than non-migrants and are typically harder to contact for follow-up. Attrition can generate bias in surveys if the characteristics of those who are lost to follow-up are different to those of survey members who remain.

31. Longitudinal panel surveys that draw their sample from the general population may miss migrants, if they are in a small minority. Those who are included in the sample will be subject to sampling variation. One way to overcome this problem is to design a longitudinal panel survey specifically focussed on migrants, as seen in Canada.

2.2.2 Current Practices

32. Below is a diagram categorising current international practices collecting longitudinal data on migrants.

Figure 2 Typology of current international practices using longitudinal data and perspectives for migration statistics



33. Particular surveys and administrative data from each category are summarised in Table 1.

Table 1 Examples of current international practices in using longitudinal data for migration statistics

Name of survey/ administrative data source	Description including types of data used, coverage and sample size	Research topics, methods and examples of analyses
Longitudinal Panel Surveys		
Longitudinal Survey of Immigrants to Canada (LSIC	• Specifically designed to study the integration of new immigrants • Includes all migrants aged over 15 years who arrived in Canada between 1/10/2000 and 30/09/2001 • Data were collected at 6 months, 2 years and 4 years after migrants landed in Canada • Of 165,000 migrants within the scope of this survey, sample sizes of 12,000 (Wave 1), 9,300 (Wave 2) and 7,700 (Wave 3) were achieved • Statistics Canada used administrative sources for the sample frame (migration records) and for tracing sample members (Health registration)	• A two-stage stratified sampling design ensured that seasonal immigration patterns were reflected in the survey sample • Results are produced at national and regional levels with more granular outputs available for selected provinces and metropolitan areas (where migrants are concentrated) • Tracing migrants for follow-up is more challenging than for the general population. Almost half of migrants in the pilot study moved at least once in their first six months in Canada • Adjustment weights to account for both non-response and for untraceable cases. • To date there have been over 170 publications based on the LSIC.

Name of survey/ administrative data source	Description including types of data used, coverage and sample size	Research topics, methods and examples of analyses
Combined survey/administrative data approach		
Italian survey on Integration of the Second Generation (ISG)	• Starts with a survey and subsequently links to administrative sources • Samples foreign middle and high-school children (aged 11-19) in schools with at least 5 foreign students. • Captures 1,437 schools, with 31,700 foreign students. Includes the same number of Italian classmates as a 'control group • Data first collected in 2015. • Teachers and heads at the schools were also surveyed, allowing multi-level analysis at the individual and school levels. Data were collected through self-completion electronic questionnaires • The 2015 survey data is linked annually with data from the students register and the population register	• The 2015 survey included questions on citizenship, country of birth, parents' citizenship, school performance and linguistic ability, school wellbeing, friendships in and out of school, family, living conditions and sense of belonging and citizenship • Region of the school at NUTS2 (Nomenclature of Territorial Units for Statistics level 2) level • Supports analysis of characteristics over time comparison of 2015 aspirations and actual outcomes

Name of survey/ administrative data source	Description including types of data used, coverage and sample size	Research topics, methods and examples of analyses
Combined survey/administrative data approach		
Longitudinal Study for England and Wales (LS)	• The LS links census and administrative data • It includes over 500,000 people usually resident in England and Wales at any one time, 1.1% of the population of England and Wales • It links 1971, 1981, 1991, 2001 and 2011 censuses to vital event information (births, deaths, live and still births to sample mothers, deaths of babies born to sample mothers, cancer registrations, immigration and emigration, re-entries following emigration and armed forces enlistments) since 1971 • Sample selection is through 4 dates of birth in the calendar • Includes both migrants and non-migrants, including migrants who arrived in the UK and subsequently left, and returning migrants • Parallel studies are run in Scotland and in Northern Ireland.	• Record linkage is high quality, involving automatic and clerical matching • More granular, local area analyses are only supported where migrants are geographically concentrated • The range of census topics covered supports variety of migration research studies. The temporal scope supports research into the dynamics of socio-demographic and health status changes over time, place and inter-generationally, for migrants and their children. • Topics researched include time since immigration and language spoken, the impact of international migration on London's housing, the structural assimilation of migrants, ethnic identifiers among international immigrants and their descendants, estimating fertility of migrants, mortality of migrants and their descendants.

Name of survey/ administrative data source	Description including types of data used, coverage and sample size	Research topics, methods and examples of analyses
Cross-sectional survey with panel-like elements		
German Microcensus (MC)	• Germany's MC is a combined person and household survey, an enlarged version of the European LFS • Conducted annually since 1957, it contains a representative sample of 1% of German households, with approximately 380,000 households and 820,000 household members including migrants and non-migrants • The survey is conducted using area sampling • A two-stage area cluster sample design includes 102 survey regions (Stage 1) and population by gender, age, family status and nationality (Stage 2) • Each household remains in the sampling frame for 4 consecutive survey-rounds. Thus 25 percent of the total sampling frame is rotated annually, and there is an overlap of 75% in any two consecutive years • The sampling frame is based on a continually updated register of addresses, "proportional to space", i.e. addresses are randomly chosen within survey regions • Migrants and refugees are included if they live in private households. Refugees, who are required to live in institutions while their asylum application process, are part of the "institutionalized population" and have in the past been under-represented	• Most questions are obligatory, including family and household compositions, employment, income, education and vocational training. Optional-response questions include health, health insurance, housing situation and retirement provisions. • The large sample size supports analysis of subpopulations such as migrants • Area sampling supports regional analysis • Participants are not followed if they leave the area being sampled • A bias in the coverage of migrants is unlikely to occur as long as the likelihood of being included into the micro-census sampling frame does not vary systematically by region (assumed not to be the case) • Survey weights are the ratio of the respective population categories in the survey and the external population

Name of survey/ administrative data source	Description including types of data used, coverage and sample size	Research topics, methods and examples of analyses
UK Labour Force Survey (LFS)	• The UK's LFS is a cross-sectional survey with panel-like elements • The LFS tracks changes in members' economic status between waves • The sample covers approximately 0.2% of the population, chosen from a postcode directory. A question identifies migrants • Panels are rotated every quarter so subjects remain in the sample over 15 months • Sample attrition for migrants is high as migrants as they can be highly mobile. By the end of 5 waves, loss to follow-up is high for migrants, which will bias research findings if attriters are different to remainers	• Longitudinal analysis is carried out over 2 quarters (comparing the current quarter to the previous one) and 5 quarter periods (comparing the current quarter to the same quarter the previous year) • The LFS is supports some migration-specific outputs, including: estimates of employment, unemployment, economic inactivity and other employment-related statistics for the UK, including estimates of labour market activity by nationality and country of birth • The UK Annual Population Survey (APS) which is comprised of LFS data supplemented by sample boosts in England, Wales and Scotland (to ensure small areas are sufficiently sampled) produces annual estimates of the population of the UK by country of birth and nationality

2.2.3 Population registers, record linkage and other data integration

34. In addition to longitudinal panel surveys, administrative data sources may be used to better understand longitudinal migration. Population registers in particular are suited for this, as many countries use them already as their main source for cross-sectional migration figures.

35. Population registers are used in several ways to study longitudinal migration. At a macro-level, time series of cross-sectional indicators of migration (e.g. the number and type of migrants) are usually part of the main set of migration statistics that national statistical offices produce annually.

36. Microdata integration, involving the linkage and the follow-up of individuals, is more challenging. Linkage of individual records over time, whether between updates of a population register or between the population register and other administrative data, often presents challenges, even when a personal identification number exists. These situations are examined in more detail in chapter 3.

37. Furthermore, as explained earlier in this chapter, disparities may exist between the endogenous operationalisation of time in the research design, when observations are done, and the exogenous manifestation of time in the population register, the timing and pace of registrations. Several factors determine the latter and almost all of these are out of control of national statistical offices.

38. Extrinsic (macro level) temporal factors relate to the organization and management of an administrative source, and changes that might occur in this over time. They have an impact on the source as a whole. Typical examples are changes in legislation and definitions, and changes in how the administrative register is managed. Extrinsic factors may have an impact on both macro-level comparisons (break in time-series) and micro-data integration (follow-up of individual records).

39. Intrinsic (micro-level) temporal factors refer to issues that impact more on individual records rather than on the register as a whole. They are often structural limitations associated with how registrations are managed, making it difficult in some cases to follow individuals over time. The relationship between the underlying time reference of registrations and the time reference of the statistics produced is an issue. Emigrants who do not deregister are an example. Although these factors make micro-data integration harder, they can also impair macro-level comparisons, especially if they occur with some frequency or if they result in non-random bias.

40. A number of these issues affecting temporal measurement, both extrinsic and intrinsic, are covered in some detail below.

2.2.3.1 Extrinsic temporal factors

41. In many countries, municipalities or local administrations are the primary responsible administrations for maintaining population registers. Some countries also have a central population register, but the degree of centralization, technical features and legal matters regarding access vary greatly between countries. There are many facets to the centralization of local population registers that have a direct or an indirect impact on how and to what extent population registers can be used to produce high quality statistics.

42. On the most basic operational level, the existence of harmonized rules and guidelines on how to register persons is crucially important. In most countries there is legislation in place to govern this, however that is not a guarantee that registrations are done the same way everywhere. The

degree to which administrations comply with these rules, which may vary between or within countries, determines the quality of the registers and subsequently the statistics that are based on them. Linguistic barriers, foreign names, and cultural bias in the interpretation of concepts (e.g. the definition of a household) present further challenges.

43. Larger administrations may be able to accommodate influxes of migrants, while smaller towns and municipalities often lack the means to adapt quickly. For example, the large increase in refugees in 2015 in Europe illustrates how the quality of registration may vary over time, potentially affecting migration figures down the line, as some administrations were overwhelmed by the sheer number of migrants they had to register (Eberle, 2018).

44. In addition to operational issues, the way in which registration of the population is organized from an institutional point of view has an impact on the suitability of the register to produce demographic statistics. As mentioned, local authorities are nearly always responsible for maintaining population registers, but the way in which higher-order authorities are involved is organized differently across countries.

45. In many countries the ministry of interior is responsible for the coordination and centralization of the local registers (Poulain, e.a., 2006), as well as other departments, whether or not affiliated with the ministry, which may take up responsibility for specific facets of the work. Many countries have separate procedures, or maintain specific registers, for subpopulations like foreigners or asylum seekers. Differences in procedures, management and storage between administrations complicate the integration of information, making good communication between authorities essential.

46. Further to this, the institutional position of the NSO itself may have consequences. How ‘smoothly’ an NSO has access to the data may depend on the institutional relation between the NSO and the authorities managing the population registers. Also, in some countries the NSO is legally mandated to calculate a reference population that serves as input for a number of financial parameters (notably in financing municipalities). On the one hand this facilitates access, but it may on the other hand put a strain on the NSO to prioritize its non-statistical duties. Even the constitutional organization of a state may play a role. Not every country is hierarchically organized, and compelling non-hierarchical entities to comply with standards of data provision (or collection) can be notoriously difficult.

47. Even between countries with a centralized population register, differences can exist in the access to variables needed to construct longitudinal indicators such as personal identification numbers, names and exact dates. Finally, the periodicity of the available data is decisive. For longitudinal indicators, reliable provision of data from the population registers to the NSO on at least an annual basis is a minimal requirement.

48. Another important factor is the amount of treatment the NSO needs to apply to the data once received. This includes work like imposing quality controls, additional linkages and the centralization of the register as such.

49. In some countries, the register is very rich in information, fully centralized and of high quality. In other countries, however, the NSO must find other means to come to a stable register. This might mean doing relatively simple work, like quality controls or cleaning variables, or a more complex undertaking like integrating additional information from other databases. In some countries, there simply isn’t a ‘centralizing’ authority other than the NSO which consequently has to take on the responsibility to integrate all population registers.

50. Generally, there is almost always a need for statisticians to work the 'raw' data. From a conceptual point of view, this work implies transforming the register from an administrative database to a statistical register. As explained above, the weight of this phase may differ considerably between countries, but the result should always be quite similar: a stable consolidated database meeting a minimum number of basic prerequisites. One example of this is the need for a consistent record identifier over time which is covered in detail in section 3.6.

2.2.3.2 Intrinsic temporal factors

51. Intrinsic factors are issues that make it harder to follow up individual cases or records over time. Many of these are not only related to 'how' cases are dealt with (e.g. erroneous registrations), but also to 'when' they are dealt with and to what moment they refer to.

52. A registration might refer to the event itself, or to the time of the registration. In practice, big differences might exist between the two. But even the time of an event as such may refer to different moments associated with a single event.

53. In the first place, in the case of an immigration, there is often a difference between the moment that the person enters the country and the date of his or her registration in the population register. In most countries strict rules exist concerning this, and registration is compulsory within a given period of time, often a few weeks or months as a maximum. In the case of measuring long-term migration these few weeks or months do not make a huge difference. However, in the case of measuring migration on a yearly basis, and turning these yearly measurements into a series, this may well have an impact.

54. The most explicit example of this was once again the large increase in refugees in 2015 in Europe. The number of refugees peaked in late summer and early autumn 2015. However, many of these refugees were registered only weeks or even months after entering Europe, with those arriving later in the year potentially being registered in 2016. As a consequence, the peak in registrations followed several weeks or months after the peak in the phenomenon itself, and in some countries 2016 will historically turn up as a year in which migration was soaring, while in fact it was already over its peak, if not rapidly decreasing.

55. Furthermore, the extent to which this is the case not only depends on how these registrations were done, whether retrospective registration is administratively possible or allowed, for example.

56. The extent to which this is the case also depends on how these registrations were done and whether retrospective registration is possible or allowed. Also the degree to which statisticians have access to these dates is decisive. In Belgium, for example, both the date of arrival and the date they apply is registered for asylum seekers, but Statistics Belgium only has access to the latter.

57. Even merely the definition of what a migrant is, may influence how the moment of onset of migration is measured. Some migrants, for example, initially enter as students, short-term labourers, or even as tourists. Some do so several times before finally deciding to settle. The question then is when to start counting them as migrants – from their first registration onwards or when their legal status changed. This is an issue discussed further in chapter 4 when determining how to measure indicators related to time.

58. Furthermore, in defining a migration, there is almost always an explicit or implicit suggestion about the duration of stay. Sometimes a minimum duration of stay is explicitly mentioned in the definition, or even the intention of staying for a minimal period is supposed. In other definitions,

however, no duration criterion applies and every foreigner who registers is considered to be an immigrant.

59. The longitudinal dimension of migration and the associated assessment of time may come in to play in defining migration as such. However, the definition of migration may have an impact on how time needs to be registered and how to operationalize it in analysis. If, for example, short spells of stay in the country need to be counted as migrations, these short periods should also be registered in a reliable way. If, on the other hand, a 12-month criterion is used, it should be possible to follow persons over a period of at least a year, whether or not these 12 months are interrupted for short periods abroad. For this, it should be possible to follow the same persons over time.

60. Finally, people do not always carefully respect administrative rules, like time limits to register, even if these are legally compelled. This might be less of an issue in the case of first-time registrations because for new immigrants, motivation to comply is usually higher. However, in the case of registration after a period of absence from the country, this motivation may be lower since administrative or legal consequences are less far-reaching. This is not only the case for foreigners but also for nationals who return, possibly even more so as the incentives to register immediately after returning might be even smaller. As a consequence, registration may happen a considerable amount of time after returning and only because the first administrative consequences of not being registered begin to appear. Time spent abroad will be biased then.

61. All of the above, however, are only minor problems compared to the frequent situation where a person emigrating from a country simply leaves without further notice. This obviously has an important impact on migration figures. It biases emigration figures downwards and net migration upwards, and consequently should be addressed directly or indirectly.

62. The most straightforward way to do that is ‘simply’ identify these cases and delete them from the population register. There are different ways to perform this identification and practices may differ depending on whether the administration, the NSO, or both intervene.

63. For example, in Belgium, Italy, the Netherlands and Spain, practices, directives or legislation allow municipal administrations to deregister a person once there is proof that he or she no longer resides in the country. This will usually be triggered by administrative signals, such as new renters registering in the emigrant’s former residence or non-compliance with certain administrative duties. Usually, municipal administrations neither have the know-how nor the means to integrate different data sources, so normally they will mainly act on information in the population register itself or in other municipal registers.

64. For most NSOs, however, integrating information from different sources is simply their core business. As a consequence, linking the population register to the tax registers or to border control information may greatly optimize the search for emigrations that went unnoticed. How much an NSO is able to, or is legally allowed to, share insights with administrations will differ from country to country. As confidentiality is one of the main pillars of public statistics, some countries may have legislation in place that prohibits this.

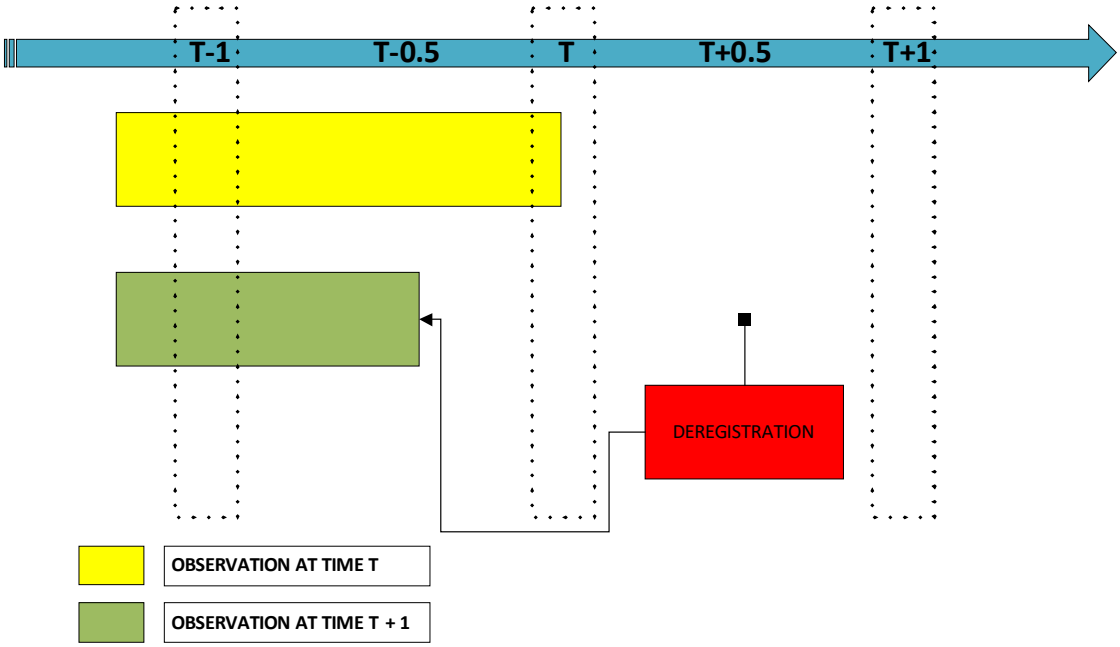
65. Figure 3 presents an example of a typical case of the consequences of emigrations that are not registered in a simple longitudinal research design with 3 measure points T-1, T and T+1. For the sake of clarity, these are considered as being annual population counts, but the issues raised may also occur when working with other administrative data.

66. The case concerns a person who entered the country a short time before T-1. At time T, the person is observed as still present in the register or administrative data, so his or her observed

length of stay at that point in time corresponds with the yellow rectangle. However, between T and T+1 it becomes clear that this person is no longer present and that he or she left the country sometime between T-1 and T. So the true status and duration of stay of this person at time T was not corresponding with the yellow rectangle but with the green.

67. The consequences of this failure for the observation at time T are evident: the population will be overestimated and the emigration figure between T-1 and T will be underestimated. However, a problem also occurs at time T+1 as the person will no longer be added to the population count without an emigration to balance this.

Figure 3 Observation of an individual's presence in a population



68. The example in Figure 3 also explains the difference between longitudinal survey designs and administrative data regarding the management and the conceptualization of time in these research designs. In a longitudinal survey, the above situation and its adverse consequences at time T would simply not occur as the person normally would no longer be eligible for follow-up due to non-contact.

69. Surveys are primary sources of statistical data. They are designed by the researchers for statistical use only and, as such, time is managed intrinsically as part of their design. This usually implies that the observations coincide with the measurement points of the research throughout time, as is the case, for example, for waves in panel surveys.

70. Administrative data primarily serves other goals than statistics. Therefore, the manifestation of time in the administrative database will coincide with its administrative purpose and this will mainly depend amongst other things on regulation, practices, infrastructure and the quality of the administration. These are of course factors that are exogenous to the research design and imply that almost always observations will no longer coincide completely with the measurement points of the research design, which is the case in Figure 3.

71. The consequence of this is that contrary to longitudinal surveys, when using administrative data for longitudinal research, two dimensions of time are relevant: (1) the endogenous operationalization of time in the research design, when the indicators need to be produced; and

(2) the exogenous manifestation of time in the administrative database, when migrations get registered. Moreover, as explained throughout the text, bias is often the result of disparities between these two. This is of course a drawback to using administrative data for longitudinal research and it needs to be countered, or its impact needs at least to be alleviated.

72. For macro-level comparisons, either modelling or correcting are ways to deal with this problem. All of these measures are not mutually exclusive and many NSOs use several to complement each other.

73. The most straightforward way **to model** out the impact of the error presented in Figure 3 is to estimate the number of similar cases for the period T-1 to T. This estimated number can then be subtracted from the population count at time T and added to the number of emigrations between T-1 and T. Another way to model this problem is to simply consider this case as an emigration at the time of deregistration, and consequently adding it to the emigration figure between T and T+1. This is also a kind of modelling because the underlying hypothesis is that what is recovered from earlier periods compensates for what is missed within the period under study. The efficiency of both ways of modelling rests on the stability of the underlying phenomena.

74. The most straightforward way **to correct** the situation in Figure 3, or at least in theory, is to correct retrospectively at time T+1 the population estimate at time T and the emigration figure for the period T-1 to T. This ensures that figures will be coherent over the whole period of observation. Needless to say, from a practical point of view this is not a straightforward solution. A second way to 'correct' is simply do nothing and ignore the case. The case will then add up to the statistical adjustment at time T+1.

75. In the case of microdata integration, dealing with these time-related issues is much more challenging. While detection and correction of errors of the kind presented in Figure 3 is straightforward, dealing with their consequences is complex. Technically, a correction of the case in Figure 3 leads, at T+1, to a new version of the database compared to time T. As a consequence, all dependencies that were valid at time T (e.g. linkages to other databases) become invalid.

76. Versioning of the database may be considered as a first order solution for these issues. The problem is, however, that in principle every new update of the data source might lead to a new version of the research database. Some NSOs receive updates from the population register on a weekly or even daily basis, making this a somewhat unpractical solution, if implemented meticulously.

77. Depending on the extent of the problems that occur between updates, it might be acceptable to produce an annual version of the research database, for example. Information for a single record or person, referring to a particular point in time, may differ from subsequent versions of the database, but changes are at least identifiable, can be analysed, and should be at least partially documented.

78. Another measure that can be taken, is imposing the time reference of the statistics on the implicit time reference of the source. This would imply that, for the example in Figure 3, the case only gets deregistered when the population register or NSO detects the error (at time T+1). Conceptually this means that the time at event is set equal to the time at registration. This introduces bias, of course, but it hugely facilitates the management of a longitudinal database.

79. Finally, more advanced data management techniques can be put to work, to optimize the practice of versioning of a database. Decomposing a database can be very helpful, for example. Some characteristics may change quite often in a person's life (address for example), others rarely or

never (citizenship, sex, date of birth). By storing every piece of information separately in smaller tables, the update pace of the characteristic that changes the most does not determine how many times the complete database is updated. While a change leads then to a new version of the corresponding sub-table, it doesn't imply producing a new version of the complete database.

80. In addition to this, making databases bi-temporal (Johnston & Weiss, 2010; Johnston, 2014) is a practice that helps controlling for their temporal dimension. A fairly straightforward way to introduce a basic form of bi-temporality is adding, in addition to the date variable referring to the information in the record as such, a second date variable referring to the validity of the record.

81. In a bi-temporal database, the case in Figure 3 would appear at least two times in the database. One time with record A, reflecting (erroneously) a stay up until time T, and a second time with a record B, reflecting the true situation (a stay up until time T-0.5).

82. Rather than giving information on the actual presence or absence of the person, the second time or date variable would give then information on the validity of these two records. The first record (A) is valid up until at least time T (and in fact time T+1 because there is no measurement between T and T+1), the second record is valid from time T+1 onwards. The figure below represents an extract of how the table might look when applying basic bi-temporality to the case in figure 1.

Table 2 Example of a basic bi-temporal table reflecting the case in Figure 3

ID	PRESENCE	DT_STRT	DT_STOP	DT_STRT_VLDT	DT_STOP_VLDT
A	1	T-1.5	T+1	T-1	T+1
A	1	T-1.5	T-0.5	T+1	31Dec9999

83. The advantage of introducing bi-temporality is that it allows to reproduce any earlier version of the database by selection on the second date variable, reflecting the validity of the records. In principle, this makes the storing of slightly different versions of the same database no longer necessary. It also simplifies management of time-related dependencies between characteristics, many of which have an update cycle of their own

84. Belgium has a more elaborate example of a (semi-)bi-temporal way of managing the inflow of a population register and how to use it to produce population frames.[1]

2.3 Complementarity and comparability of longitudinal and cross-sectional measurement

2.3.1 Retrospective content in cross-sectional surveys

85. Cross-sectional data sources that contain sociodemographic and socioeconomic information tend to be more comprehensively available than longitudinal sources. These can be used to gain longitudinal insights and perspectives over time where truly longitudinal information with repeated measures for the same respondents or cases are not available. This is achieved either through the

[1] Available at https://ec.europa.eu/eurostat/cros/content/maintenance-rich-frame-using-data-platform-example-statistics-belgium-0_en

same questions repeated over different cycles of the cross-sectional measure or through retrospective questions within a single cycle.

86. Some cross-sectional surveys ask respondents retrospective questions about how and when a characteristic changed through time. For example, when the respondent first or more recently arrived in the country, the job held before migration, or the time it took for migrants to obtain their first job after migration. This information can be analysed to provide information on the dynamic characteristics of migrants' adaptation processes.

87. Retrospective, time-related variables allow us to analyze the contemporary information in the cross-sectional data source, in the context of historic migration characteristics. Information on year of arrival may differ from year of most recent arrival, though both can be used to provide information on how socioeconomic characteristics such as being in employment vary for migrants with different lengths of stay in the receiving county. Data on country of birth together with information on parents' country of birth would support separate comparisons between first, second, third or higher generations of migrant. Caution is needed in interpreting country of birth information, given changes in international borders over time.

88. Cohorts can be defined according to any common characteristic and analysing both cross-sectional and longitudinal data on migrant groups should account for both age since birth and duration since arrival. Both have time-varying effects on socioeconomic outcomes and interact together to define migrant trajectories.

89. With variables that identify migrants (by their foreign-born status, or that of their parents) alongside socioeconomic conditions observed across various cross-sectional sources, one could analyze trends which would show the change in the prevalence of a given characteristic over time. This indicator represents the net change in that characteristic for the migrant population represented in each cross-sectional source. One limitation of this approach stems from the fact that migrant populations are typically dynamic ones and the composition of the migrant population is likely to be different at each cross-sectional snap shot. Untangling socioeconomic transitions from compositional change in the population is not easy in the absence of linked microdata.

2.3.2 Synthetic cohorts

90. This type of analysis identifies a given cohort in different cross-sectional sources and compares the socioeconomic conditions of that group over time. These are not real cohorts as they are likely to comprise different individuals in the different sources. There is a risk that compositional differences between the cohorts being compared introduce bias into the comparison. In the absence of longitudinal data on migrants, this is the approach being pursued by Germany using the German Microcensus.

2.3.3 Borrowing strength between cross-sectional and longitudinal data

91. It may be possible to 'borrow strength' between cross-sectional and longitudinal sources. Some specific applications of the term 'borrowing strength' include empirical Bayesian estimation (see for example Burke et al, 2016) and small area estimation (see Chambers et al, 2009 for borrowing strength across space). Jackson et al (2017) seek to define borrowing of strength mathematically. The use of a Generalised Structure Preserving Estimator (GSPREE) methodology has 'borrowed strength' over time, supporting the production of census-based estimates outside of

census years through the combination of administrative and survey data with the census (see Correa-Onel et al, 2016).

92. A more basic use of the term is suggested, which involves combining information from disparate sources, on the assumption that the data being combined share related and relevant characteristics. In longitudinally-linked data it may be beneficial to derive ecological data, such as area, industry or occupational characteristics, from relevant other sources with adequate sample size, such as the census, and apply these to a longitudinal survey or an administrative dataset.

2.3.4 Small-scale or longitudinal case studies

93. National statistical offices may wish to consider small-scale or longitudinal case studies to support the use of longitudinally linked administrative records. This could be considered to enrich or to check the validity of the linked administrative data, for example return migration, migrants' employment status or to tell us more about the quality of the linked data. Often, using administrative data sources to address specific research questions will involve some form of compromise if there is a gap between the ideal information you are seeking and the information that has been collected in administrative data. In this scenario, the data being analysed may provide indicators of the underlying processes or outcomes, or be used to check the validity of the linked administrative data through a parallel small-scale study. This would allow comparison between target concepts and those that have been captured by administrative data sources. Chapter 3 describes this potential disconnect in more detail.

2.4 Summary

94. Longitudinal data allow us to study and understand life events and transitions through time and, crucially, across generations. This can be particularly useful for the study of international migration and migrant integration, as longitudinal data can shed light on the geographic and socio-economic outcomes of migration. This chapter provided an overview of different longitudinal data sources and gave country-specific examples including the use of longitudinal panel surveys, administrative data including population registers, and a combination of surveys and administrative data.

95. Not all NSOs will have access to administrative data or funding to run longitudinal studies on migrants, so a more pragmatic approach may need to be taken. For example, cross-sectional surveys can provide longitudinal insights and perspectives over time and retrospective questions can be added on existing surveys. In the absence of longitudinal data, it is also possible to consider following synthetic cohorts over time, all of which is described in some detail in this chapter.

96. Where longitudinal data are available, the challenges and limitations associated with these data are recognized, and the fact that these will vary by each data source. Traditional methods such as longitudinal panel surveys have advantages in the control offered to statistical offices such as timing, coverage and measurement. However, as migrants tend to be more geographically mobile than non-migrants and are typically harder to contact for follow-up, these data sources will tend to have smaller sample sizes, suffer from attrition and may lack frequency of data collection. All of which can hinder their use for studying short- or long-term outcomes for migrant populations.

97. There are also limitations associated with the use of administrative data such as population registers or with the process of integrating administrative data to create a longitudinal dataset.

Microdata integration, involving the linkage and the follow-up of individuals, is challenging. Linkage of individual records over time, whether between updates of a population register or between the population register and other administrative data, often presents challenges, even when a personal identification number exists. These issues are described in detail in this chapter, but a key limiting factor is that these data have not been collected for statistical purposes. Therefore, control is lost over timing, population coverage and alignment to migration concepts and definitions. But administrative sources can usually produce larger sample sizes than traditional longitudinal surveys with more regular follow-up and little to no attrition due to non-response. As a result, they are, generally, well suited to study small populations of migrants longitudinally.

98. These challenges and limitations are picked up in chapter 3 which describes best practice on how to develop a longitudinal dataset for migration statistics and seeks to address limitations of different data sources.

Chapter 3: How to develop a longitudinal data set for migration statistics using integrated data

3.1 Introduction

99. This chapter provides guidance on how to develop a longitudinal data set for migration statistics based on integrated data, considering the Generic Statistical Business Process Model (GSBPM) (UNECE, 2019). The chapter focuses on elements specific to data integration, longitudinal data, and migration statistics. The UNECE Guidance on data integration for measuring migration (UNECE, 2019) should be used to supplement – especially for references to data integration.

100. Each section of this chapter will cover a phase in the process providing an explanation of how to undertake each step using concrete examples from various national statistical offices. Particular issues will be identified and solutions will be suggested.

101. The phases of developing a longitudinal data set for migration statistics based on integrated data include:

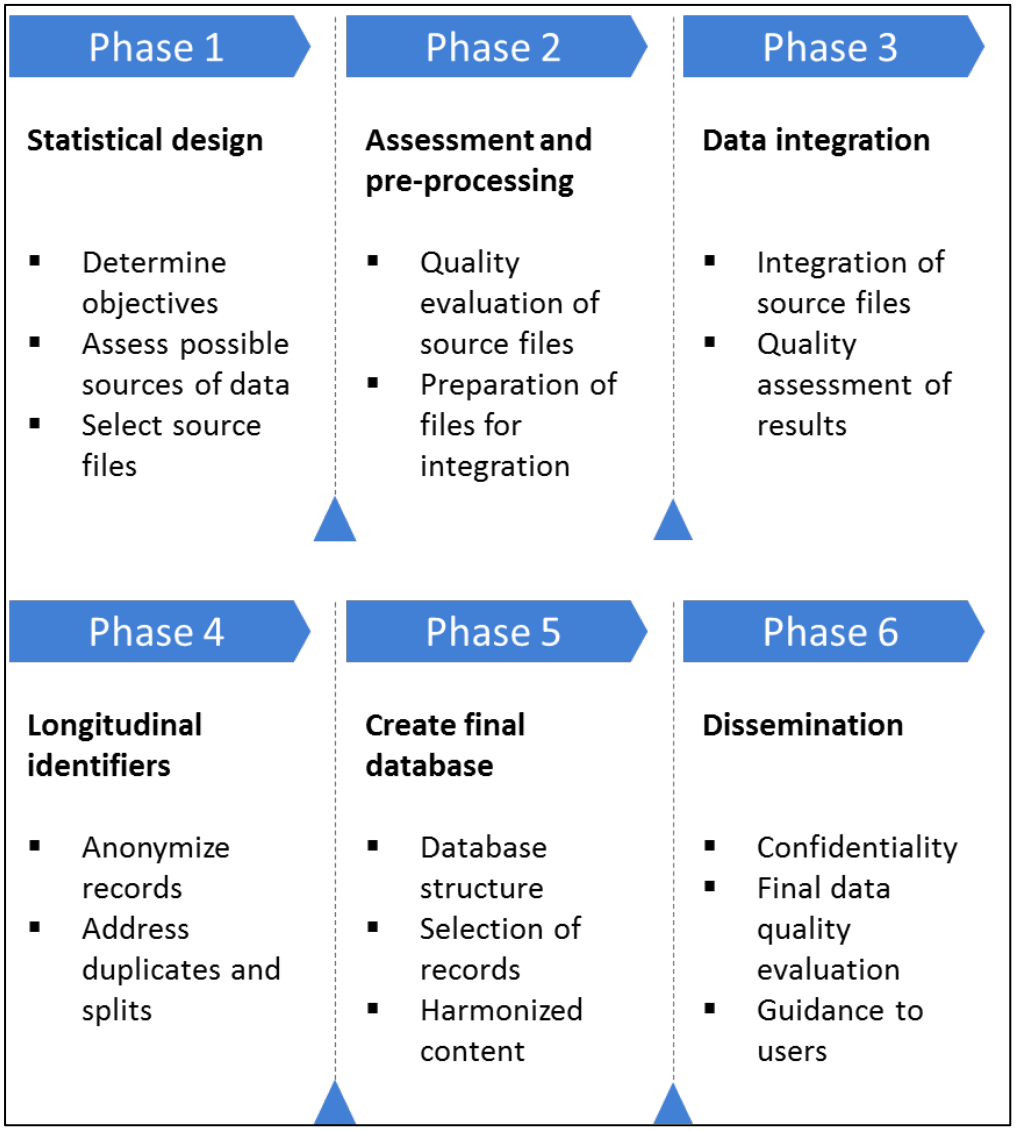
1) Statistical design
2) Assessment and pre-processing of source files
3) Data integration for longitudinal data
4) Assignment of longitudinal individual identifiers
5) Create final database
6) Disseminate results

102. This chapter provides guidance on each of these phases. They must be considered both in the context of the development of a new database as well as in the context of updating an existing one (e.g. to add new outcomes via data integration).

103. The order of the phases is important. However, iteration between these phases may be necessary depending on outcomes of later phases. For example, the assessment of the source files (phase 2) may require a reconsideration of the statistical design (phase 1). It is also possible that data quality issues associated with the integration process could only be identified during the final database evaluation (phase 5). This may require a new data integration exercise (phase 3) and the repetition of subsequent phases.

104. While all stages have to be considered, action is not required in all cases. For example, an update to an existing project may not require a reconsideration of the statistical design phase, and in the case of producing longitudinal data from a system of registers, the data integration phase may be straightforward.

Figure 4 Phases of developing a longitudinal data set



105. Throughout this chapter, rather than provide singular case studies, examples are provided from different countries for specific elements. If users are interested in how specific data sources have been developed from start-to-finish, they are encouraged to refer to available documentation on existing data sources. For example, in **Canada**, the Longitudinal Immigration Database (IMDB) is outlined by its Technical Report[2].

3.2 Error framework for longitudinal data from integrated data sources

106. Data quality is a recurring item of consideration through all phases. Source files should be selected to best address the aims of the statistical design and the quality of the potential source files is critical to this decision. The quality of the source files is further assessed in the second phase and, where possible, addressed using various methods. The data integration itself requires an assessment of integration errors and the assignment of longitudinal individual identifiers requires some reconciliation of the results of the data integration. When a final database is created, methods are

[2] Available at https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2018019-eng.htm

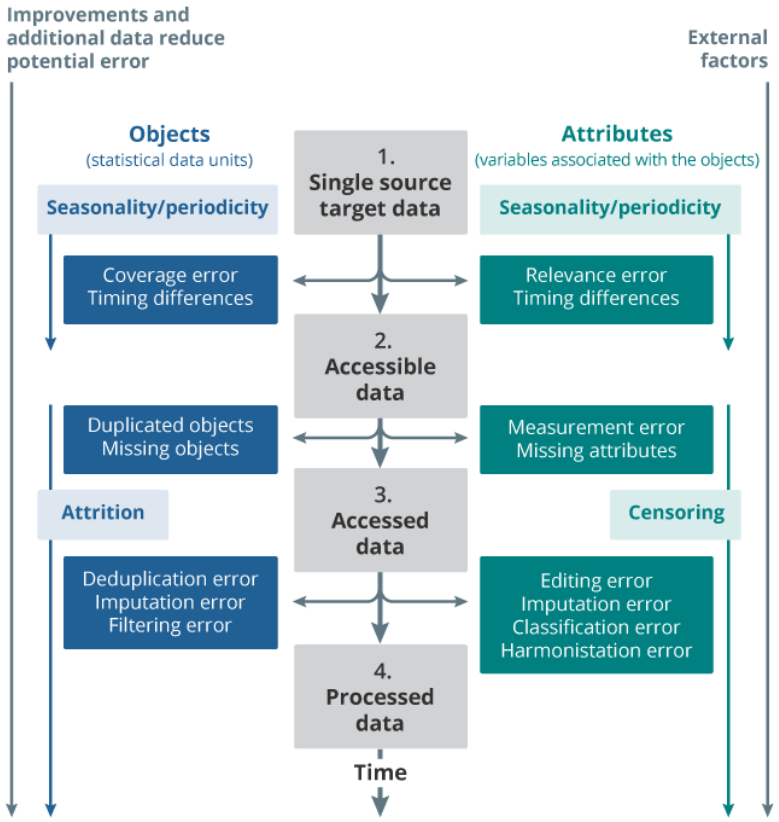
used to harmonize concepts over time and prepare the database for analysis. Finally, all data quality considerations need to be documented and appropriate guidance provided to users before disseminating the results.

107. It is useful to draw on the Total Survey Error Framework which has been developed to support statistical design (Groves, 2004), (Zhang, 2012). This provides a useful taxonomy of the different potential sources of error, relating mainly to administrative data. Not all types of error are measurable. But they need to be borne in mind when designing a longitudinally-integrated dataset. The following is an extension of Zhang's (2012) framework to discuss more fully the sources of error involved in linking administrative sources to produce longitudinal datasets.

108. The framework, developed by the United Kingdom (Blackwell & Rogers, 2019), proposes a staged approach to understanding data quality, looking first at single datasets before assessing error in the production of datasets created through the integration of multiple sources. In common with the framework proposed by Zhang (2012) and developed further by Statistics New Zealand (Statistics New Zealand 2016), the errors associated with both data objects (records) and their attributes (variables) are considered. However, in administrative data these errors are *given* by the source datasets; attributes are not designed as in a survey and what is already collected is re-purposed. At the heart of the framework are datasets which have not been designed, they are what they are.

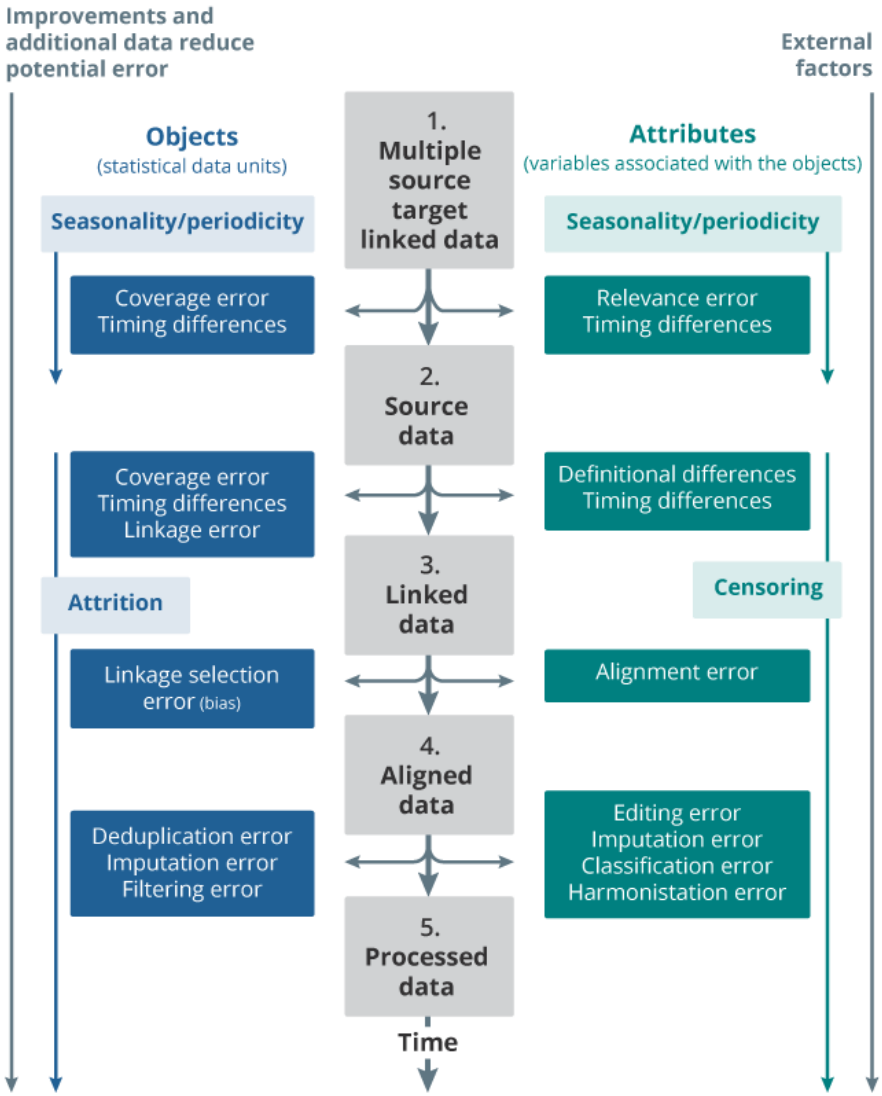
109. Figure 5 and Figure 6 describe the error frameworks for single and multiple sources. Table 3, Table 4 and Table 5 define the errors at each stage. The single source framework identifies four different stages of the data journey – target data, accessible data, accessed data and processed data. Errors are represented as a conceptual difference between data at each of these stages. Target data are conceptual – the ideal data to be collected and so errors between target data and accessible data are conceptual.

Figure 5 Single source error framework



Notes: Time is presented vertically. Target data represents the ideal data to be collected. Objects (statistical data units) refer to rows in the data source, and what those rows of data represent. Attributes (variables associated with the objects) refer to columns in the data source, and what those columns represent

Figure 6 Multiple source error framework



Notes: Time is presented vertically. Target linked data represents the ideal linked data to be produced. Objects (statistical data objects) refer to rows in the data source, and what those rows of data represent. Attributes (variables associated with the objects) refer to columns in the data source, and what those columns represent.

110. Errors are split between objects and attributes: errors occurring to objects relate to the entity the data are for, be that people, events or businesses. Errors relating to attributes relate to what you are measuring for the objects. Errors for both objects and attributes can affect each other (represented by a double arrow between the two). Not all types of errors will be applicable to each source of data and there may be other sources of error identified, particularly for processed data. In the context of this report, objects tend to refer to individuals defined by Cohort files (or cohorts defined by Outcomes files) while attributes can refer to outcomes defined by Outcomes files.

111. Error propagates through the framework, though it does not necessarily build at each stage. This framework has been used to help researchers at the Office for National Statistics in the UK understand longitudinally-linked administrative data, including Exit Checks data from the Home Office. The intention is to assess its usage for statistics on international migration. The production of the Exit Checks data begins with the supply to the Home Office of passenger data, supplied by commercial carriers at UK ports of entry/exit (United Kingdom Home Office, 2019). The Home Office

processes and manages the quality of these data, linking them with other operational data for visa nationals.

112. This application has demonstrated that in linked administrative data, error is not necessarily cumulative. This is because experts at the Home Office seek to address error in the data as they pass through processes. While error can accumulate over time, there is also the possibility that compensating errors may complicate matters. Compensating errors (e.g. imputation for missing values) may not be visible in cross-sectional distributions of the data, but can compound longitudinally. This is explained in the context of the multiple source framework (Figure 6).

113. The framework also incorporates longitudinal error created by the collection of data over time – seasonality/periodicity, attrition and censoring (Table 4).

Table 3 Single source errors

Objects[1]	**Attributes**[2]
Frame error:	*Relevance error:*
Coverage error	*Validity error*
Assessing objects that are not in the target data, or not being able to access objects that are in the target data.	The difference between ideal measurement of attributes sought about an object and the operational measure used to collect it.
Timing differences	*Timing differences*
Objects in the ideal target data that are not accessible because of a discrepancy in the time window for obtaining observations.	A conceptual discrepancy in the timing of the measurement of attributes between the ideal target data and accessible data.
Selection error:	*Measurement error:*
Duplicated objects	*Measurement error*
Objects that are represented more than once in the accessed data.	Errors arising from attributes that are not recorded accurately.
Missing objects	*Missing attributes*
Objects that *in theory* are accessible but are not in the accessed data.	Attributes that are missing from the accessed data (could be for specific objects or all of the objects).
Processing error:	*Processing error:*
Deduplication error	*Editing error*
Errors arising from deduplication of objects in the accessed dataset. This could include both deduplicating objects that are actually different (false positive error) and failing to deduplicate objects that are the same (false negative error).	Errors arising from editing the value of an attribute. This could include editing as a result of validation or quality assurance checks.
Imputation error	*Imputation error*
Errors arising from the imputation of missing objects.	Errors arising from the imputation of missing attribute values.
Filtering error	*Classification error:*
	Errors arising from classification of values into groups or derivation of new attributes.

Objects[1]	Attributes[2]
Errors arising from the selection or de-selection of accessed objects to an ideal target set.	*Harmonisation error:* Errors arising from the harmonisation of values of attributes to an ideal or target concept

[1]This refers to data units and could be events, transactions, persons, households, firms or other entries in an administrative dataset.
[2] This refers to the measures or variables that have been collected that relate to the data objects/ units

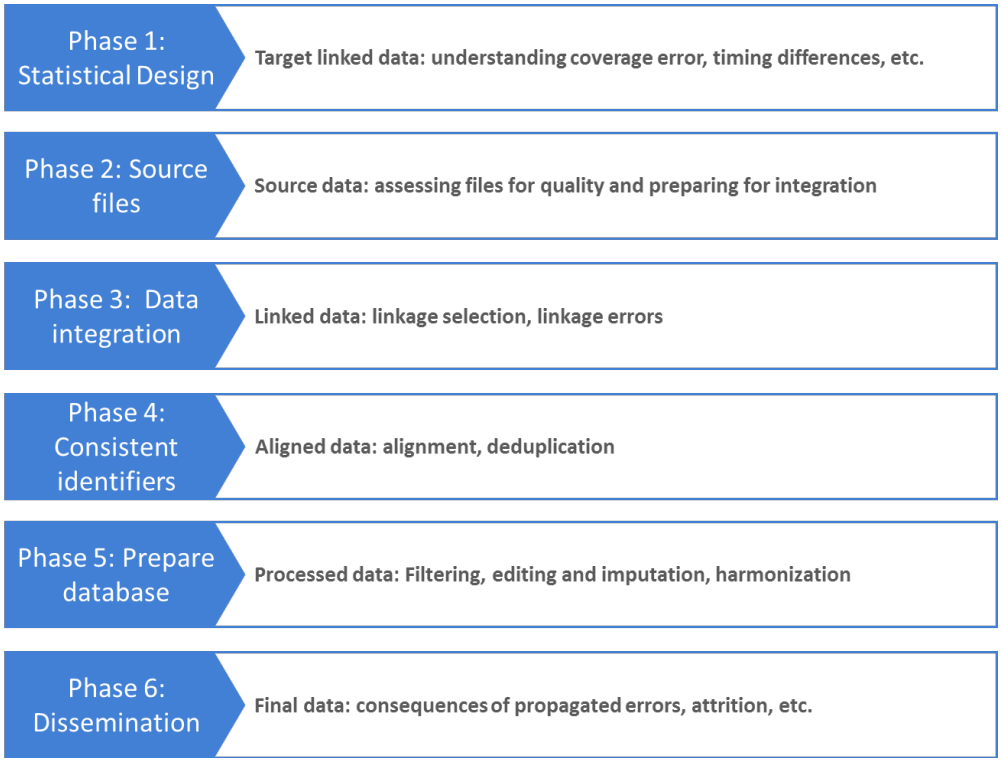
Table 4 Longitudinal error – applies to both single source events and multisource longitudinal datasets

Objects[1]	Attributes[2]
Attrition The loss of research objects or units over time. Occurs naturally, through death (or an unobserved migration). Also occurs through failure of follow-up, a refusal to take part, in the case of survey data, or through missing information or linkage failure, in administrative sources.	*Censoring* Where the value of a measurement or observation is only partially known. Right censoring is when the research object drops out of the data before the end of the observation window or does not experience the event of interest during the observation window. Left censoring is when the event of interest has already occurred, before the observation window begins.
Periodicity/seasonality error Objects are not observed because the data capture is not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).	*Periodicity/seasonality error* Measurement of attributes over time are not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).

[1]This refers to data units and could be events, transactions, persons, households, firms or other entries in an administrative dataset.
[2] This refers to the measures or variables that have been collected that relate to the data objects/ units

114. The multiple source framework (Table 5) is for the integration of multiple data sources. It consists of five stages – target linked data, source data, linked data, aligned data and processed data. In the context of this report, these stages align with the phases covered by this chapter as shown in Figure 7.

Figure 7 Stages of integration of multiple data sources



115. Errors are represented as a conceptual difference between data at each of these stages. Many of the errors in the multiple source framework are conceptually similar to the single source framework. The main difference is the fact that errors are measured between the source datasets and the ideal target linked data (rather than target data) as well as errors between the source datasets to be linked. These conceptually similar errors have the same name between the single source and the multiple source frameworks. For example, timing differences, coverage error, relevance error, imputation error, selection error and processing error.

116. Target linked data are different to the target data for each individual source. The target linked data are likely to be specific to the group of objects to be measured through linkage of multiple sources. For the final stage (processed data), processing may have occurred in the single source or in the multiple source framework. For example, imputation may have already occurred at single source, or it may be done after linkage.

Table 5 Multiple source errors

Objects[1]	**Attributes**[2]
Frame error:	*Relevance error:*
Coverage error	*Relevance error*
Observing objects that are not in the target linked data, or not being able to access objects that are in the target linked data.	The differences between ideal measurement of attributes sought about an object and the operational measures used to collect it in each source dataset.
Timing differences	*Timing differences*
	A conceptual discrepancy in the timing of the measurement of attributes between the target linked data and the source data.

Objects[1]

Objects are not observed due to conceptual discrepancies in the timing of the capture between the target linked data and source data.

Coverage error:

Coverage error

Objects are not linked due to discrepancies in the coverage of objects between data sources.

Timing differences

The difference between observed objects in source datasets due to the data being captured at different times.

Linkage error

Errors arising from linking objects together incorrectly (false positive error) and failing to link objects together that should have been linked (false negative error).

Identification error:

Linkage selection error (bias)

Errors arising from the selection of linked objects (or de-selection of unlinked objects) due to biases in the linkage, or through error in the resolution of conflicting links.

Processing error:

Imputation error

Errors arising from the imputation of missing objects.

Filtering error

Errors arising from the selection or de-selection of accessed objects to an ideal target set.

Attrition

The loss of research objects or units over time. Occurs naturally, through death (or an unobserved migration). Also occurs through failure of follow-up, a refusal to take part, in the case of survey data, or through

Attributes[2]

Mapping error:

Definitional differences

The differences between how attributes are operationally measured in each of the source datasets.

Timing differences

The differences between the values of attributes for a linked object between source datasets caused by the data being captured at different times.

Comparability error:

Alignment error

Errors arising from the alignment of the conflicting values of attributes across sources.

Processing error:

Editing error

Errors arising from editing the value of an attribute. This could include editing as a result of validation or QA checks.

Imputation error

Errors arising from the imputation of missing attribute values.

Classification error:

Errors arising from classification of values into groups or derivation of new attributes.

Harmonisation error:

Errors arising from the harmonisation of values of attributes to an ideal or target concept.

Censoring

Where the value of a measurement or observation is only partially known. Right censoring is when the research object drops out of the data before the end of the observation window or does not experience the event of interest during the observation window. Left censoring is when the event of

Objects[1]	Attributes[2]
missing information or linkage failure, in administrative sources.	interest has already occurred, before the observation window begins.
Periodicity/seasonality error	*Periodicity/seasonality error*
Objects are not observed because the data capture is not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).	Measurement of attributes over time are not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).

[1]This refers to data units and could be events, transactions, persons, households, firms or other entries in an administrative dataset.
[2] This refers to the measures or variables that have been collected that relate to the data objects/ units

117. The purpose in applying the administrative data error framework is to ensure that data quality is optimized throughout all phases of the statistical project. Initially, it is essential to identify and examine sources of error to make statistical design decisions in the production of further linkage. However, it is important to reconsider the different sources of error through all of the subsequent phases as well. The lists in Tables 4 and 5 suggest a number of quality indicators. It may not be efficient for an NSO to measure each indicator, but to identify essential indicators to include in a quality report (see section 3.4.3).

118. There is an interaction between different sources of error. For example, there may be a trade-off between linkage, coverage and imputation error. Records that have poor quality data, possibly through measurement error, may also be harder to link. This could be due to the quality of the identifiers used for linkage. One option is to develop sophisticated record linkage methods to minimise false negative matches (therefore accepting more false positives) and maximise the coverage of the linked dataset. But there is a possibility that the attributes that relate to these objects are also of poor quality, and will generate either missingness in the attribute fields, or will require imputation. Imputation is often undesirable in longitudinal data, since it can introduce spurious outcomes. The avoidance of missing data and imputation may be the over-riding concern.

119. There is also an interaction between errors in objects (or cohorts) and in attributes (or outcomes), and between the single and multiple source datasets. Missing or mis-recorded attribute data can impact the ability to de-duplicate or link records in the single source phase. This in turn impacts the ability to link objects and therefore creates errors in the multi-source dataset.

3.3 Phase 1: Statistical design[3]

120. This is the first step in the process of developing any data source. The key deliverables of this phase are to determine the ideal aims and objectives to be addressed with the longitudinal data set including the concepts and variables needed, and, ultimately, to determine which source files will be integrated in order to best meet the stated objectives. As described in sections 3.1and 3.2, it is important to consider the various errors which may have impacts throughout the project.

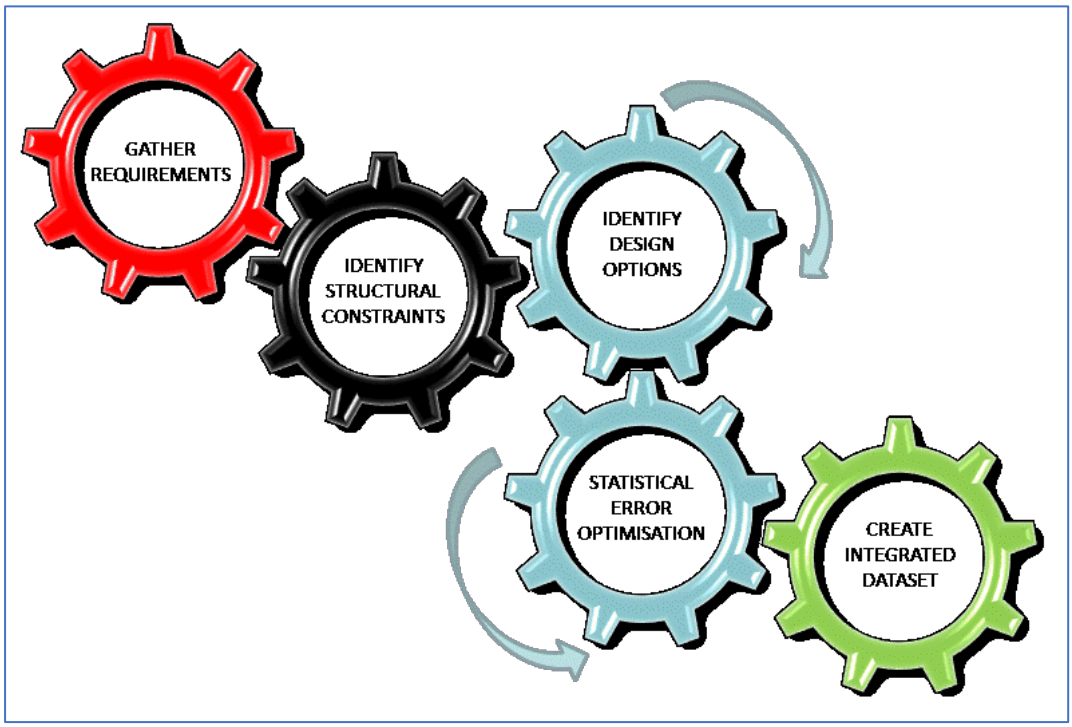
121. The design phase involves:

[3] This section builds on UNECE (2019), which is a recommended source to be used in supplement to the present Guidance

- Determining the objectives of the statistical project
- Identifying any structural constraints that may affect the statistical design
- Identifying the design options, including source file selection, with reference to statistical error

122. The goal is to choose the design, including source files, which provides the optimal balance between the different sources of error that are inherent in the production of integrated sources to address the objectives of the project.

Figure 8 Overview of the statistical design process.



123. When determining which source files are needed to meet the stated objectives, a distinction can be made between Cohort files and Outcomes files – especially in the context of data integration for longitudinal data.

124. Cohort files are source files used to identify the population(s) of interest and their characteristics while Outcomes files are the longitudinal files providing outcomes over time. In some cases, Outcomes files could serve as Cohort files as well.

125. For example, the Longitudinal Immigration Database (Canada) connects immigration administrative files (Cohort files) with longitudinal tax files (Outcomes files). The tax files could be used to identify a cohort as well, either alone or in conjunction with the immigration files. For instance, a cohort could be defined as immigrants who were admitted in a certain year (cohort defined by Cohort files only), individuals who claimed tuition amounts in a certain year (cohort defined by Outcomes files only), or immigrants who had employment income in their first year in Canada (cohort defined by Cohort and Outcomes files).

126. Ultimately, the result of this phase is the selection of the source files to best meet the stated project objectives based on the populations they cover, their temporal coverage and the measures included. Sampling and any other potential disqualifying data quality issues (e.g. Outcomes files are based on samples too small to analyze populations of interest) should be considered as well but a

more in-depth review of data quality will take place in phase 2. Any shortcomings associated with the selected source files should be well documented.

3.3.1 Determining the objectives of the statistical project

127. Before anything, it is critical to establish the objectives of the statistical project. This should be done in close collaboration with any user(s) or stakeholder(s) applicable. The objectives should be clearly determined and well documented as they will guide the rest of the phases of the project.

128. To do this, several questions need to be addressed about the aims of the project and the concepts and variables of interest:

- What are the questions to be answered with the data?
- Why is the longitudinal perspective necessary to address these questions?
- Is the intention to understand migration processes and migrant experiences? Or are we seeking to produce robust measures?
- Who is the population of interest? Can we account for this population in its entirety? Do we need a control population for comparison?
- What are the variables or concepts of interest?
- How is a 'migrant' defined? Can this be identified in the source datasets?
- Is our primary concern migration as a component of population change? If so, are we equally interested in immigration and emigration?
- Are seasonal patterns of interest?
- What about internal migration? Is the internal migration of international migrants of interest? Do we seek to know how this differs to the internal migration patterns of non-migrants?
- Are all international migrants of equal interest? Or are the migrants' intentions the over-riding concern? For example, if migrants come or leave for study, work, family, safety or other reasons?
- Do different migration policies apply to the different types of migrant?
- Do we need to identify whether migrants have temporary, short-term or permanent intentions in the host country?
- How far beyond the migration event are we interested in? For how long do we need to observe post-migration outcomes? This can include social and economic outcomes, including measures of social integration.
- Are life-course trajectories of interest? Patterns of partnering, family formation, death and cause of death.
- Are inter-generational impacts of migration of interest? Do we seek to follow up children born before and/ or after migration?
- How much granularity is required, in terms of for example characteristics and geography? Is the intention to produce national, regional or local outcomes?

129. These questions have implications on the measures needed, the populations covered and the frequency and spacing of observed outcomes.

130. Ideally, the observation window for migration itself would precede the migration event, but this would be an ambitions design, requiring international data sharing. If observation begins at the point of immigration, the follow-up period needs to extend far enough to capture the outcomes of interest.

131. Variables that could be of interest might be extensive. These could include nationality, citizenship, country of birth, age, sex, ethnicity, marital status, year of arrival, or place of residence.

132. One critical point to consider is whether the longitudinal data source will need to reproduce the same estimates as provided by the sources used in the data integration. Because integrated longitudinal data require connections across sources, duplicates (i.e. multiple identifiers for the same unit) or splits (i.e. multiple units for the same identifier) can be discovered through the integration process which would otherwise affect the source file estimates themselves. It is recommended to consider avoiding this requirement because of this complexity. Instead, consider the longitudinal data source in light of the general statistical objectives and use the source file estimates for data quality evaluation.

3.3.1.1 Understanding the statistical quality requirements

133. It is crucial to understand the requirement for statistical quality. Is the intention to produce an evidence-base for policy? For planning? For national statistics? Will it form the basis for resource allocation? If so, at what level of geography? Do stakeholders have national or local interests? Is there an international stakeholder?

134. The requirement for **accuracy** has implications for error management. For example, the design of record linkage involves a trade-off between false negative and false positive error. Is there a critical threshold of accuracy, measured in terms of statistical confidence, which the data must support? Does the project require statistical uncertainty measures? Does the accuracy requirement vary across the dataset? For example, users may tolerate a given degree of missingness for some variables but not for others.

3.3.1.2 Consistency and coherence

135. Maintaining consistent measures over time can be critical to a longitudinal study. Further, it needs to be determined whether it is necessary to have results which are comparable to other data sources. For example, is it absolutely necessary for the longitudinal data source to yield cross-sectional estimates in line with other data sources? Which variables need to be consistent over time? Will adjustment be required, for example to adjust for loss to follow-up, or attrition, from the data? Will the database use standard concepts and classifications?

3.3.1.3 Timeliness

136. Another consideration is the requirement for timely evidence. Can the project accommodate delays due to the follow-up required, for example to meet the definitional requirements? A 'long-term' migrant might be defined as one arriving and staying or leaving and staying away for at least 12 months. This implies a 1-year lag in identifying migrants. If this is unacceptable, consideration is required on estimating or imputing this measure.

3.3.1.4 Accessibility

137. The statistical design may be influenced by intentions regarding data access. If the intention is to provide the dataset in its entirety to researchers, the data preparation and processing requirements may be more onerous than if the intention is to keep the microdata in-house and supply outputs to external stakeholders. The degree of clarity and transparency required for a third party to use and understand the data may require extensive metadata, supporting documentation, guidance or even a user support service.

3.3.1.5 Relevance

138. Administrative data have their own purpose-driven operational design. In integrated datasets they are re-purposed to meet another set of requirements. This inevitably requires compromise. Consultation with administrative data providers and statistical project stakeholders may be required as an ongoing process, to ensure that data design decisions are informed by statistical objectives, and that statistical design decisions are informed by changes in the collection of administrative data. This should also take into account that different users may have divergent or varying needs.

139. If variables that are critical to users are not available in any of the source datasets, are proxies good enough? How well would they need to correspond to the desired information?

3.3.1.6 Interpretability

140. Finally, it is important to consider whether the results will be well understood by users. Proper documentation and metadata will need to be established to explain how the integrated database was developed and how to use and understand the results.

3.3.2 Structural constraints that may affect the statistical design

141. A review of the relevant sources will identify the datasets available for integration. Potential Cohort and Outcomes files should be considered on how well they address the objectives determined in section 3.3.1.

142. Before addressing content, coverage, timeliness, quality, and other statistical considerations, there may be more structural constraints that could affect the choice of files or the ability to undertake a data integration. Questions to consider when considering potential source files include:

- How will the confidentiality of the data be protected - in the data development stages and for dissemination?
- What are the costs associated with the access and use of the data?
- Are there any legal or access barriers, such as the need for respondents' informed consent? Is there a legal or institutional framework governing the access or use of this data? Are there restrictions on data integration?
- Are there any ethical concerns? Do you require ethics approval to use the data?
- Is it possible to obtain or share records with other National Statistical Offices or other foreign organizations?
- Are particular software skills required for data manipulation or management?
- What infrastructure is required to store and manipulate this data or data structure? Is the infrastructure secure? Is there enough space? How long can storage be maintained?
- Should each source be integrated to produce a linked dataset, or would it be beneficial to retain some data as a comparator, possibly for calibration or adjustment?
- Is the integration or linkage of these data sustainable over time?
- What technical capacity is required to support growth in the scale and complexity of the data over time?
- Are there any known risks to the continuity of sources? What control is there over future production? What are the risks that critical elements of the data will be changed or discontinued? What is the contingency if this happens?

143. Discontinuities in definitions and classifications between sources and over time are discussed below in the context of sources of statistical error.

3.3.3 Identifying the design options and optimizing statistical error

144. The design stage requires a series of decisions that are influenced by the project requirements, the data available and any structural constraints. They also require the optimal balance of different sources of potential error within the integrated dataset, bearing in mind the project objectives. It is recommended to refer to the framework in the introduction of this chapter for further guidance.

145. A more thorough assessment of the data quality of selected source files will follow in phase 2. However, in this phase, it is important to identify the optimal potential Cohort and Outcomes files from a statistical error perspective to meet the stated project aims, under any structural constraints. This section will include an assessment of population coverage (and sampling), measurement, temporal coverage, and frequency of collection (e.g. real-time versus periodic).

146. Potential source files may be eliminated from consideration if they fail to adequately measure either the populations or measurements of interest. This could be due to small sample sizes (after integration), missing or low-quality measures, or inadequate proxies. From a longitudinal perspective, the temporal element is critical as well – files may not adequately cover the period of interest or have frequent enough observations.

For example, in **Hungary**, a statistical project was launched in 2018 to estimate the share of circular migrants in the resident population. As a first step, definitions of circular migration and a list of variables minimally required for the exercise were established. During the project a cohort of 8 different subgroups involved in international migration processes were identified in 3 different data sources. However, after an evaluation of the input sources, only 2 subgroups were deemed to be appropriate for the exercise. The rest of the subgroups were excluded due to structural constraints (e.g. some data were not available in a longitudinal structure, while in other cases personal identifiers or other essential variables were missing). At the end of the design phase of the project, the objectives were re-defined in accordance with the results of the assessment process.

147. Questions need to be considered and weighed against the stated objectives such as:

- What time period is covered?
- Which populations are identified and covered?
- Will the sample be large enough post-integration to meet the project requirements?
- How frequently are the outcomes measured?
- What are the measurements to be included?

3.3.3.1 Differences in coverage

148. When discussing coverage, it is convenient to distinguish two different types: temporal coverage and population coverage.

149. **Temporal coverage** represents the period covered by the different data sources. This reference period may not always be ideal for the stated project aims. For example, if outcomes or cohorts are only available for periods well in the past, they may not be relevant to the current

situation. It is also possible for the quality of certain periods of collection to vary. This should be taken into consideration as well.

150. It is important to consider whether the possible Outcomes and Cohort files best address the stated objectives from this perspective. It should also be noted that the possible source files may not perfectly meet the objectives but could still yield useful results. For example, Cohort files may only include recent migrants – this may not yield results for all migrants but they could still be useful if the project prioritizes analysis of recent migrants.

151. Cross-nationally, there is wide variability in temporal coverage among the different data sources. Often this is due to when the administrative data or survey data were collected. The temporal coverage of these data sources ultimately affects how they can be used to address different statistical projects.

In **Belgium**, the central management of the Central Population Register (CPR) dates from 1985. The register contains information for records prior to 1985. However, the reliability of this information, which was kept in municipal registers, is lower.

The Belgian Labour Force Survey (LFS) has existed since 1983 but can be linked to the CPR since 2003. The Survey on Income and Living Conditions (SILC) has existed since 2004 and can be linked from 2005 onwards. Furthermore, censuses of 2001 and 2011 can also be linked to the CPR.

Immigration, Refugees and Citizenship **Canada** (IRCC) administrative records cover different periods based on how the administrative data was collected.

- Immigration records from 1952-1979 (with minimal content)
- Immigration records from 1980-present (with detailed content)
- Temporary resident records from 1980-present
- Citizenship records from 2005-present
- Settlement services records from 2013-present

These files are used as Cohort files for the Longitudinal Immigration Database (IMDB) which combines linked administrative immigration and tax data files. The database includes all immigration records and temporary resident records available and combines them with tax-based Outcomes files from 1982-present.

The Longitudinal Survey of Immigrants to Canada (LSIC) covers immigrants who landed in 2000-01 with three waves of collection at six months (2001-02), two years (2002-03), and four years (2004-05) after landing in Canada.

In **Germany**, the period of coverage of the Central Alien Register starts in 2007. The Central Alien Register contains information about all Non-Germans (included asylum applicants or refugees) currently registered in Germany (active foreigners) or previously registered in Germany.

The situation in **Italy** is more complex. Data are available since 1955. However, the use of electronic database was implemented in 1970 and the first data collection form which encodes all the foreign states was introduced in 1995. Even if the information about residence permits exists since 1995 (and made available to the Italian NSO in 2010), it's possible to use them for longitudinal analysis. Finally, a first version of the survey on integration of the second generation was carried out in 2015 and will be carried out again in 2021.

In **Spain**, the Population Register (Padrón) starts in May of 1996 and has been managed continuously since then. Padrón can be also considered as a longitudinal database, with 265M recorded movements since May 1996 (although the current population is 46.7M people). This allows for analysis of the whole history of each inhabitant in Spain throughout the last 22 years. A particularity

of the system is that if a person with a certain identifier leaves the country and later returns, the identifier assigned to him/her is different.

In **Switzerland**, the Swiss Longitudinal Demographic Database (SLDD) contains information referring to all Swiss citizens living in Switzerland from 2010 to 2013 as well as foreigners living in Switzerland from 1997 to 2013 with a regular permit.

152. **Population coverage (and sampling)** corresponds with the population covered by these sources. Sometimes, this means that only certain subpopulations are included or that some sampling has taken place. It is very important to know the degree of coverage of each prospective data source. For example: does it include all migrants or only those that migrated by regular means? The population covered by each possible source file should be checked against the stated aims of the project. If certain subpopulations are excluded, they cannot be studied using this database. If the sample size is too small, this may prevent certain types of analysis from taking place.

153. In the context of data integration, sampling is even more critical to consider. The microdata integration of two data sources will only yield results on the overlapped records. When two censuses are integrated with full coverage of the same population, at best, it can be expected to obtain integrated results for the full population covered. It can be 'at best' because integration errors may occur and, in general, an increase in false negatives is favoured over an increase in false positives (this is because false positives can generate spurious or nonsensical outcomes). However, when one of those sources becomes a sample, at best, it can be expected to obtain integrated results for the sample source only.

154. If two samples are integrated, only the sample that overlaps between the two samples can possibly have results. In the case of simple random samples, the best case scenario would be to have results for A per cent x B per cent of the population where A per cent is the sampling fraction of the first source and B per cent is the sampling fraction of the second source. For example, linking a survey with a 20 per cent sampling fraction and a survey with a 5 per cent sampling fraction would yield results for a maximum of 1 per cent of the population. If one source has a sample of analytical interest of 1,000 and is linked with another source with a 20 per cent random sampling fraction, integrated results would be available for a maximum of 200 individuals.

155. Coverage limitations exacerbate these limitations. If the temporal coverage differs between the sources, only those individuals covered by both time periods will possibly be integrated. If one file is a census of migrants and the other file is a census of hospitalization records, with no integration error, results will only be available for those migrants who had a hospitalization record. If one file is a census of migrants and the other file is a one-time sample survey, at best, results will be available for those migrants who responded to that sample survey.

156. Users will need to consider the extent to which they are prepared to work with these limitations or address them afterwards. If a source is selected with sampling or coverage limitations, the next phase will consider this in more detail as well as the possibility of addressing these limitations (e.g. using sampling weights). However, if the expected integrated sample is clearly too small for any meaningful analysis, this should lead to an adjustment in the source file selection and the statistical design more generally.

157. Similar to temporal coverage, population coverage is often dictated by how administrative data are collected or how surveys were conducted. As such, there can be a great variability in the population covered by data sources of different countries or even between different sources in the same country.

In **Belgium**, every individual legally residing in Belgium (for more than 3 months) is registered in the CPR. Asylum seekers are also included, although they are not registered by the municipalities, but centrally by the Immigration Office.

As all variables concerning migration in other sources (censuses, LFS and SILC) originate from the CPR, the same groups can be identified in these sources.

In **Canada**, there are also differences in the population covered by different available sources. The Longitudinal Immigration Database (IMDB) combines immigration and tax administrative records. While there is no sampling conducted, the population is restricted to those linked, and outcomes are only available for tax filers. This means that certain populations which are less likely to file taxes (e.g. children) are underrepresented in the Outcomes files. To address this, the IMDB recently released new outcomes files for children connecting them with their parents' economic outcomes and facilitating analysis of intergenerational economic mobility.

Immigration, Refugees and Citizenship Canada (IRCC) administrative records contain exhaustive information about immigrants who were admitted to Canada since 1952 or who became Canadian citizens since 2005. It also contains non-permanent residents who have been issued temporary resident permits since 1980. However, they would not include those who do not require permits to reside in Canada (e.g., visitors, certain dependents of temporary residents, etc.).

The Longitudinal Survey of Immigrants to Canada (LSIC) had a target population consisting of immigrants who met all of the following criteria:

- Arrived in Canada between October 1, 2000 and September 30, 2001
- Were age 15 years or older at the time of landing
- Landed from abroad, must have applied through a Canadian Mission Abroad.

Individuals who applied and landed from within Canada were excluded from the survey. These people may have been in Canada for a considerable length of time before officially "landing" and would therefore likely demonstrate quite different integration characteristics to those recently arrived in Canada. Refugees claiming asylum from within Canada were also excluded from the scope of the survey.

The LSIC population of interest was immigrants in the target population who were still living in Canada at the time of the interview. In addition, sampling was undertaken for the purposes of the survey – this was further affected by non-response. Of the approximately 165,000 immigrants in scope for this survey, about 12,000 responded to the first wave, 9,300 to wave 2, and 7,700 to wave 3.

In **Germany**, the Central Alien Register includes only foreigners, whether first / second generation, asylum seekers, refugees, etc.

In **Italy**, the population register includes all people who change their municipality of habitual residence. This administrative register excludes foreigners in an irregular situation. In the case of residence permits, the population covered are citizens who do not belong to EU-countries and stay in Italy legally. For the survey on integration of the second generation, the sampling frame was formed by 31,700 foreign students in selected secondary schools.

In **Kazakhstan**, the Statistical Population Register (SPR) contains records on resident individuals of the Republic of Kazakhstan, as well as the citizens of the Republic of Kazakhstan temporarily staying abroad. Most foreign migrants to Kazakhstan do not settle permanently, (i.e., they do not receive a residence permit), but are admitted under an employment contract and are registered temporarily.

The "Census of 2009" database contains information on resident individuals or temporary residents of Kazakhstan at the time of conducting the census.

There are two suitable sources for migration statistics. Information systems of the Ministry of Internal Affairs contain the internal and external migration data. The “Berkut” information system contains records on persons crossing the Republic of Kazakhstan border to enter/leave the country for permanent residence and gain citizenship of the country.

In **Spain**, the Population Register covers all the population resident in Spain regardless of their legal status and nationality (includes nationals and foreigners).

In **Switzerland**, the Swiss Longitudinal Demographic Database (SLDD) contains information about Swiss citizens who have lived in Switzerland from 2010 to 2013 and foreigners (with residence permits) who have lived in Switzerland between 1997 and 2013. This includes asylum seekers and nationalized immigrants, but not those people in an irregular situation.

In **Turkey**, Address Based Population Registration System (ABPRS) is a unique system from which annual population statistics are disseminated. Within this scope, annual information on international migration stocks and flows by citizenship and country of birth are produced based on the ABPRS and other additional administrative sources.

Coverage of individuals in the annual ABPRS data may vary due to specific conditions such as late registration of deaths, late registration of children under five, change of identification number after acquisition of citizenship (i.e. loss of the trace of the naturalized person) and under/over-coverage of previous/present year ABPRS.

Although registers of foreigners work in a similar way to population registers, they only cover foreigners who have legal residence in the country. In this respect, undercoverage (for those who are not in the administrative registers while present in the country) and overcoverage (for those who are in the administrative registers but not present in the country) problems come into prominence. Therefore, registration and deregistration of foreigners are crucial for determination of international migrants. Even though it is not a big problem as it was in the past, declaration is crucial for the foreigners who lived in the country for a certain length of time or who have intention of staying for at least a certain length of time and who will leave the country.

3.3.3.2 Periodicity of Outcomes

158. Further to temporal coverage, how often the data are collected or updated is also of interest. If the outcomes are only measured periodically, they may miss more frequent events (or an intensity of events) which would be better captured in more granular periods of time or even in real-time. An example would be the arrival of large numbers of foreign students at the beginning of the academic year. More detailed considerations affecting temporal measurement are covered in Chapter 2.

159. The possible Outcomes files should be considered for how often the events are being measured compared with the stated objectives of the project. It should also be noted that the possible source files may not perfectly meet the objectives, but could still yield useful results. For example, Outcomes files may only have annual results. This may not provide insights into the number of days or weeks before migrants find employment but they could still be used to understand how migrants gained employment by year.

160. It is also important to note the difference between the data being collected and the data available for a statistical project as there could be delays between these two processes.

In **Belgium**, Statistics Belgium receives a double inflow of their register (the CPR): once a year, in March, a cross-section of the complete register as it exists at that moment and every two weeks all

the registrations for the preceding two weeks. The yearly flow is used to calculate all demographic statistics, as the continuous inflow serves to update the sampling frames for social surveys.

Thus, it is possible to obtain cross-sectional outcomes continuously. Nevertheless, faulty registrations are partially corrected and removed. As a result, there may be discrepancies between two different cross-sections for the same instant made at different moments. Dealing with cases present in one cross-section but completely absent in a later version has proven difficult.

In **Canada**, there are differences in the way the information is updated in the different available sources.

In Immigration, Refugees and Citizenship Canada (IRCC) administrative records, longitudinal outcomes associated with residence permits can be observed in real-time (e.g. precise dates of permits). Most of these files are shared with Statistics Canada on a monthly basis. However, for the Longitudinal Immigration Database (IMDB), Outcomes from tax files are observed annually and are primarily comprised of information provided by individuals when they file taxes in the first few months after the end of the reference year (e.g. individuals filing their 2017 taxes would typically do so in the first few months of 2018).

Outcomes for the Longitudinal Survey of Immigrants to Canada (LSIC) were observed at specific points in time: at six months (2001-02), two years (2002-03), and four years (2004-05) after landing in Canada.

In **Germany**, the Central Alien Register contains information about all Non-Germans (included asylum applicants or refugees) currently registered in Germany (active foreigners) or previously registered in Germany. The Central Alien Register is continuously updated and statistical data refers to 31st December of each year.

In **Italy**, Anvis (Anagrafic Virtual Statistical data base) is the statistical system fully managed by Istat that, based on the information coming from the population registers, allows follow-up on an individual basis and provides outcomes on annual basis. Information about residence permits is updated annually.

Data from the Italian survey on integration of the second generation, carried out in 2015, are linked annually with the administrative data coming from the students' register.

In **Kazakhstan**, data are regularly submitted to the Statistical Population Register (SPR) by administrative sources and the 2009 Census database. This is done both systematically, in the case of an event occurrence, and at regular time intervals.

In **Spain**, the municipalities continuously update the information of registration and deregistration in the Padrón. Furthermore, the Civil Registry also updates information that deals with births and deaths. These updates are sent every month to the INE, which is responsible for updating its database. Thus, it is possible to obtain accurate data at any time of the year with updated information.

The date of registration in the Padrón is usually close to the actual date of entry into the country, which is unknown, because of the benefits that entail from being registered. On the other hand, the date of deregistration in most cases does not coincide with the date on which the citizen left the country because he/she leaves the country without any declaration.

In **Switzerland**, the Longitudinal Demographic Database (SLDD) is a linked-register based database created in two different stages. First, a demographic base is constructed by linking the Central Register of Foreign Nationals (ZAR) and the population and Household Statistics (STATPOP).

In a second step, the database is also organized to include other statistics. Up to now, the Structural Survey (2010, 2011 and 2012) and the unemployment Statistics were linked with the SLDD providing useful information on the socioeconomic characteristics of residents in Switzerland.

The longitudinal outcomes observed of the SLDD are updated annually. The database is organized in a way which allows its regular extension year after year. Such an update will increase the period under observation and therefore the capacity of the database to provide useful information on migration and integration.

3.3.3.3 Measurement differences

161. It is important to assess whether or not the measurements – either of populations or outcomes – align with the concepts stated in the objectives of the project.

162. There can be some overlap here with the assessment of the periodicity of the files as real-time outcomes. For example, becoming employed might be observed as an annual summary variable. However, there can be overlaps with population coverage as the definition of the population groups of interest may differ from what is observed in the data sources.

163. In the next phase, some differences can be addressed through pre-processing, especially for variables required for the data integration. On the other hand, differences in analytical variables may be addressed in phase 5 when content is harmonized.

164. Of particular note is how differences between measurements and concepts can pertain to migrants. Some of these differences may be due to:

- Legal status of migrants: some sources collect information about all migrants (regardless of their legal status in the new country of residence) while others only include those with a specific residence status (e.g. holding a valid residence permit).
- Minimum length of stay to be classified as a migrant: differences in the concept of migrant that is used, depending on whether a minimum stay in the country is considered or not (e.g. 12 months).

165. In the study of migrants' experiences, differences in geographic concepts may be important to note. For example, changes to boundaries over time or differences in concepts between different sources (e.g. area of residence vs. area of intended residence).

166. Differences between ideal concepts and actual measurements are explored in section 4.1 on longitudinal indicators where indicators of interest are contrasted with the limitations of how the data may be captured and collected.

For example, in **Canada**, all of the geography variables on the immigration and tax files in the Longitudinal Immigration Database (IMDB) are typically coded based on the previous census geography. Census geographies are updated every 5 years. In addition, the concept of geography of destination (found on the immigration files) differs from geography of residence. While these two are combined to understand local 'retention' of recent immigrants, it is possible that some individuals never reside in their geography of destination.

Annual tax files in the IMDB yield detailed economic outcomes of immigrants in Canada. However, because the files are annual summaries, some details are not fully measured. For example, while the annual files indicate how many immigrants had employment income in a given year (and how much), they cannot provide an employment rate for a single point in time or a reference week. Similarly, variables are available for individuals to claim post-secondary tuition amounts (a proxy for participating in post-secondary education), but details on level or major field of study or education outcomes are not available.

The tax files can also provide some of the same demographic variables found on the immigration files (such as date of birth – or age). There can be inconsistencies between the two sources of information for the same individual. These differences would need to be addressed either in pre-processing to facilitate data integration (3.2) or to harmonize concepts for dissemination (3.5).

The central alien register of **Germany** contains information about all Non-Germans (included asylum applicants or refugees) currently registered in Germany (active foreigners) or previously registered in Germany.

One of the most important challenges is to develop a work-around when dealing with changes of citizenship over time or with changes in national administrative borders.

In **Spain**, all foreigners are listed in the Population Register (Padrón) along with the Spanish-born population. While it is not possible to determine the number of refugees or foreigners in an irregular situation, it is considered that Padrón practically includes all the foreign residents because of the multiple advantages of being registered.

In the population register of Italy, information about irregular migrants is excluded, and the 12-month rule is only assumed to be true since there is no time criteria when data on migration are collected.

The residence permits data source only contains information about Non-EU citizens. In the case of the survey on integration of second generation, data pertaining to residence permits have been integrated with data on citizenship acquisitions, the Population Register, and the Social security register in order to monitor the inclusion process of non-EU citizens. This integration has been carried out to follow the students in the sample after the survey, and to compare the students' expectations and opinions expressed in the survey with their real behaviours. Data sources are subject to changes in administrative procedures and different geographic boundaries. These changes might affect the comparability of the information over time.

In the **United Kingdom**, the ONS Longitudinal Study for England & Wales (LS) considers changes to geographic boundaries over time. Initially, this stemmed from a user request to look at internal migration flows for 2011 based on old 1991 census boundaries for Local Authorities.

The smallest geographic area, for census and some population estimates, is the Output Area. This is a building block to create higher levels of geography such as Local Authority Districts (LAD). Geography experts at ONS created concordances between Output Areas (and Local Authorities) between different census years.

"Exact allocation of 1991 Local Authority District areas to individual 2011 Census records was undertaken. Each 2011 Census member record was plotted on a map, based on Output Area centroid, and assigned to the 1991 LAD digital boundary it fell within. All points were able to be allocated to an area by this method. Thus, each of the 585,771 2011 LS member records was allocated to one of the 403 local authorities."

"Please note that a decision was taken to make direct allocations in this case. This differs from the method preferred by the UK Government Statistical Service (GSS) Geography policy whereby records would have been allocated to an Output Area and then best-fit to a 1991 Local Authority. A decision was taken to directly allocate because the GSS Policy is aimed at producing tabular statistics whereas the main focus of the Longitudinal Study is to follow individuals over time."

This project was extended to allocate 1991 LS records to 2011 and 2001 boundaries.

3.4 Phase 2: Assessment and pre-processing of source files[4]

167. During phase 1, possible source files were identified and assessed for their ability to best meet the project objectives based on the populations they covered, their temporal coverage and the measures included. At this phase, data providers should be consulted and made aware of the planned integration of their data into the longitudinal data, as they will be aware of potential data quality issues or expected modifications to the content. Data providers are key partners as they have the best knowledge of the data and, over time, they can inform of changes to the data. Following this phase, the best available source files for the project have now been identified.

168. During phase 2, the quality of these source files will be assessed more thoroughly. Based on this assessment, processing (e.g. imputation of missing values) may be required to address any identified quality issues. Data quality of the source files should be assessed and documented before and after any processing. It is recommended to refer to the framework in section 3.2 for further guidance on different sources of errors.

169. If any important data quality limitations are detected in this phase, it may be necessary to return to phase 1 and consider new data sources.

170. Finally, the files will be pre-processed to prepare them for integration in phase 3. The quality of fields used to facilitate data integration will be assessed during this stage. For a longitudinal database, these phases will need to be performed every time the database is updated as changes can occur on a regular basis.

3.4.1 Quality assessment of source files

171. At this stage, the data quality of each source file should be assessed. The quality assessment should consider coverage and measurement errors as well as any errors which could affect the integration phase. The list below presents examples of errors which can be investigated along with suggested documentation and possible methods to address the errors.

172. For the purposes of this exercise, source files will continue to be defined into two types (in the context of longitudinal data from integrated source files):

- Cohort files (in the Error Framework, these are referred to as ‘objects’)
- Outcome files (in the Error Framework, these are referred to as ‘attributes’)

173. Cohort files are source files used to identify the population(s) of interest and their characteristics while Outcomes files are the longitudinal files providing outcomes over time. In some cases, Outcomes files could serve as Cohort files as well.

3.4.1.1 Issues affecting coverage

174. There are different types of issues which can affect the coverage of a longitudinal data set. These can include coverage errors, sampling error, and other sources of missing records such as unit non-response.

[4] [4] This section builds on UNECE (2019), which is a recommended source to be used in supplement to the present Guidance

3.4.1.1.1 Coverage error

175. Here, the extent to which the data source, if there were no missing records and no sampling, would be representative of the target population is considered. In effect, there would normally be two kinds of coverage error:

- Undercoverage: where units should be in-scope for the data source but are not covered
- Overcoverage: where units should be out-of-scope for the data source but are included (including duplicates of units in-scope)

176. This assessment should build upon the review conducted in phase 1 which required an evaluation of coverage limitations associated with the selected source files and how they could affect the objectives determined in the statistical design.

177. For Cohort files, coverage errors reflect the difference between the population of interest and the coverage of the Cohort file.

For example, the Longitudinal Immigration Database (**Canada**) uses immigration administrative records as the Cohort file. Migrants without administrative records (i.e. those without permanent or temporary residence permits or asylum claims) would necessarily be excluded by the Cohort file – this would include visitors and some other types of travellers. Therefore, these files would undercover the total number of individuals entering Canada.

Alternatively, these files could include multiple entries for the same individual (with different unique identifiers). This would lead to overcoverage on the Cohort file.

178. For Outcome files, coverage errors should be considered in a similar fashion. Recall that the Cohort files will be used to identify the population of interest. It will be important to consider the coverage of Outcome files in the context of both the general population as well as the population of interest.

179. Assessment of overcoverage on the Outcomes files can be limited to identifying duplicates. The integration process will restrict the final dataset to the Cohort file and thus eliminate other out-of-scope units.

For example, the **Canadian** Longitudinal Immigration Database links immigration files (Cohort) with annual tax files (Outcomes). The coverage of the tax files has elements that affect the general population (e.g. children are less likely to file taxes and are therefore undercovered) and the specific population of interest (e.g. temporary foreign workers are less likely to file taxes and are therefore undercovered).

In **Spain**, the population register (Padrón) has been continuously in operation in Spain since 1996. It reflects the place where a person is registered at a specific moment and contains some concrete variables: sex, age/date of birth, place of birth, citizenship, name and family name, level of educational attainment or official identification number.

Padrón not only stores the place where a person is at a specific time, but as a longitudinal record, it has all the places where a person has registered since 1996.

Padrón is the ideal source to obtain longitudinal information about migrants. However, it has a limitation: its starting point is in 1996, so information previous to that year is not available. To

address this, information from Padrón will be linked with the 2001 Census (which was exhaustive) and thus the information will be available for all the people.

In **Hungary**, it is well-known that administrative sources underestimate the real size of outmigration flows since migrants often do not report on their migratory events to the authorities. To reveal the migrant subgroups whose migration flows are systematically underestimated in administrative sources, administrative data was compared with survey data from the Microcensus. This comparison has shed light on the main weaknesses of administrative sources, especially when dealing with distinct forms of multiple migration.

3.4.1.1.2 Sampling error

180. In certain data integration projects, at least one of the source files may only represent a sample of the covered population. This would most often be the result of one of the source files being a sample survey, or equally an administrative data subset of the target population. There are several different situations which could affect a longitudinal data source based on integrated data.

181. A common example of **sample Cohort files with no sampling on Outcomes files** would be where a sample survey is being linked to administrative files to create a longitudinal data set. The sample survey identifies the cohort (along with certain characteristics of interest) and the administrative files provide the longitudinal outcomes.

182. In this situation, the Cohort files should have supporting information related to the sampling (including weights) as well as guidance on how the files should be used for statistical analysis. This supporting information should still be valid to account for the sampling error associated with the resulting longitudinal database.

183. An example of **sample Outcomes files with no sampling on Cohort files** would be where a longitudinal sample survey is conducted using administrative data as a frame. The administrative data defines the cohort (along with certain characteristics of interest) and the longitudinal sample panel survey provides the longitudinal outcomes.

184. In this situation, the Outcomes files should have supporting information related to the sampling (including weights) as well as guidance on how the files should be used for statistical analysis. This supporting information should still be valid to account for the sampling error associated with the resulting longitudinal database.

For example, the Longitudinal Survey of Immigrants in **Canada** (LSIC) used administrative immigration records as the survey frame.

185. The scenario of **sample Cohort files with sample Cohort files or any other combination of sampling** might be less common especially since the result of an integration of samples is a smaller sample (sample of a sample). This can be particularly problematic when studying migrants (already a fraction of the general population). However, in cases where the samples being integrated are large enough, this could still yield a useful statistical data set.

186. Advanced statistical methods would need to be used to account for more complex sampling situations and these are considered out of scope for this report. However, it is important that just as is the case for the first two types of sampling that could affect longitudinal data sources, sampling errors should be measured and addressed.

Some questions to consider when assessing coverage issues:

- Are there existing reports which document limitations affecting coverage for the Cohort or Outcomes files?
- Which migration populations or control populations are covered by the Cohort files?
- Which migration populations or control populations would be missed by the Cohort files?
- Are there duplicates on the Cohort files?
- If there are multiple Cohort files, are there duplicates across files?
- What populations are covered by the Outcomes files?
- What populations would be missed by the Outcomes files?
- Are there reasons why the Outcomes files might miss specific migration populations?
- Are there duplicates on the Outcomes files?
- Is there sampling on either the Cohort or Outcomes files? If so, what is the sample size in terms of number and sampling fraction (for the general population and the migrant population of interest)? What sampling methods were used?
- Are there records on either the Cohort or Outcomes files which are missing? What percentage are missing? Are there some populations more or less likely to be missing?
- If there are records missing from Outcomes files, are they missing intermittently, at the start only or only at the end?
- Are migrants affected disproportionately by missing records on either Cohort or Outcomes files?

3.4.1.1.3 Missing records

187. Units could be missing from either the Cohort or Outcomes source files. One common cause for this would be the result of unit non-response.

188. Records missing from the Cohort files would affect any resulting analysis as the affected units could not be matched with any outcomes.

189. Records missing from Outcomes files, on the other hand, could have varying effects. These missing records could be the result of drop-outs over time where there is complete outcomes information for earlier periods. Alternatively, these could be intermittently missing records or even records which are missing initially but become complete for later periods.

190. Migration patterns can be conflated with missing Outcomes patterns. For example, is the absence of Outcomes due to missing records *or* because the individual was not in the country for that period of time? In some cases, the migration event can be measured, but at other times it is left unknown.

191. Depending on the nature of the missing records, different methods could be employed at the analytical stage.

192. The report documenting coverage limitations could include various data quality indicators such as:

1. Distributional differences between target and covered populations (based on confrontation data[5] sources)
    a. Age and sex
    b. Migrant status
    c. Other fields of interest
2. Estimated gross and net undercoverage
3. Number and percentage of duplicates
4. Sample size (number and percentage) by migrant status on Cohort or Outcomes files
5. Number and percentage of missing records on Cohort files
6. Number and percentage of missing records on Outcomes files by period

3.4.1.2 Issues affecting measurement

193. Missing values and measurement errors on the source files also need to be assessed. Both of these issues can affect both the Cohort and Outcomes files.

194. While missing values can be fairly easy to identify, measurement errors, where the measured value is different from the true value, can be difficult to detect. In some cases, values could be invalid (e.g., values that do not exist or values which are not plausible). For example, a date of birth after the current date or, in the case of a dataset covering individuals currently alive, a date of birth hundreds of years in the past.

195. In particular, it is important to consider the quality of the fields which will be used to facilitate the integration process (for example linkable fields for record linkages such as date of birth).

[5] 'Confrontation data' refers to comparable data against which the results can be benchmarked.

Some questions to consider when assessing measurement issues:

- Are there existing reports which document limitations affecting measurement for the Cohort or Outcomes files?
- Are there missing values on the Cohort files? If so, do these missing values disproportionately affect specific migrant populations? Which variables are affected and by how much?
- Are there missing values on the Outcomes files? Are they consistently missing for the same individuals? Do the missing values disproportionately affect specific migrant populations? Which variables are affected and by how much? Are there implications for analytical outcomes?
- Are there invalid values on the Cohort or Outcomes files? If so, do these invalid values disproportionately affect specific migrant populations? Which variables are affected and by how much?
- Are there variables within the Cohort files or Outcomes files which suggest inconsistencies (e.g. individuals born after their year of migration)? Are there supplementary variables which can be used to determine which value might be more likely to be correct in cases of inconsistency?
- What are some of the risks of other measurement errors? Would these disproportionately affect migrant populations?
- How do the measures line up with other data sources?
- Are the linkage fields on either the Cohort or Outcomes files affected by missing or invalid values? To what extent is there overlap between the issues for the files to be integrated?

196. The report documenting measurement limitations could include various data quality indicators such as:

1. Frequency and percentage of missing values by variable, period, and migrant status
    a. Special attention should be paid to within-individual patterns as certain individuals may be more likely to have multiple missing or invalid values
2. Frequency and percentage of invalid values by variable, period, and migrant status
    a. Special attention should be paid to within-individual patterns as certain individuals may be more likely to have multiple missing or invalid values
3. Distributional differences for variables between comparable populations (based on confrontation data sources – i.e. alternative sources with comparable measures)
    a. Age and sex
    b. Migrant status
    c. Other fields of interest
4. Differences in variable values between linked files with comparable measures
5. Inconsistencies between variables from same source files
    a. Auxiliary variables or additional information which might support one value over the other
    b. Preference based on most commonly-occurring value

3.4.2 Pre-processing of source files

197. Now that the source files have been assessed for data quality, it is time to prepare them for data integration and address any known data quality issues where necessary.

198. The specific methods associated with pre-processing source files are not part of the present Guidance. Some elements are covered in the UNECE guidance on data integration for measuring migration (UNECE 2019b), others are well explained in general statistical literature.

3.4.2.1.1 Preparing files for data integration

199. The pre-processing of the source Cohort and Outcomes files begins with:

a. Standardizing matching variable formats and lengths (e.g. consistent format of date of birth)
b. Standardizing format of names and treatment of special characters (e.g. uppercase, removal of accents)

200. These steps are critical for the integration between Cohort and Outcomes Files. Here there are some elements of particular note for migration statistics. These include the treatment of name spellings and ensuring the ordering of given and family names is consistent between files. Common name-based issues which can affect migrants disproportionately can include:

- Changes in family name due to marriage (especially if migration is related to marriage)
- Adoption of a given name common in the host country in place of original given name
- Sub-selection of given or family names in certain data sources (e.g. possibly due to space constraints on a government form)
- Accents which may not always be captured consistently in all data sources
- Ordering of given and family names not always consistently captured (e.g. if family name is typically placed ahead of given names)
- Different spellings resulting from inconsistent name transliteration between sources (i.e. the transfer of names from one system of writing to another)

201. Date fields (e.g. date of birth) also need standardization. This includes ensuring the proper ordering of year, month, and day but also includes addressing situations where elements are either missing or appear to be erroneous. One common issue associated with administrative data is an abnormally large number of records with dates of birth falling on January 1st. This can be a default value in place of true missing values. Since migrants may not be familiar with the local calendar, this can lead to uncertainty around a precise date of birth and lead to a large number of records with only a valid year of birth.

202. Where sample selection is based on calendar date of birth, as in the UK ONS Longitudinal Study, the use of 'default' dates of birth by/ for some foreign-born migrants, in particular from Southern Asia, leads to sample bias.

203. In addition, if contact information is being used for integration, the timeliness of this information can have an impact on the results. Since migrants are, by definition, mobile, their contact information maybe more subject to change over time.

204. Besides standardizing and cleaning the source files, another best practice to consider is to use information captured at varying points in time to facilitate the linkage process. For example, if the Outcomes files are deterministically linked, they may provide variations of names and contact information over time. Taking this into account will improve the integration quality.

205. While integrating many files can lead to various integration errors such as false positive or false negative links, using longitudinal linkage files reduces the risk of integration error. If a single source is able to reliably track changes in linkage fields (e.g. name(s), addresses, etc.), its use in the

linkage process will reduce the risk of false negatives (missing links) due to the changing characteristics identified above.

In **Canada**, the Longitudinal Immigration Database (IMDB) takes advantage of a repository of linkage information captured over time. The Social Data Linkage Environment (SDLE) is used to facilitate the record linkage of the IMDB. Prior to this, the IMDB relied on a longitudinal linkage file comprised of tax information connected by Social Insurance Number (SIN). This source tracked changes in linkage fields over time.

3.4.2.1.2 Internal record linkage before integration phase to identify duplicates

206. Duplicates on Cohort or Outcomes files can be identified using an internal or within-file record linkage. This could be a useful exercise even for sources with unique record identifiers such as statistical registers. In particular, since migrants could re-enter the country multiple times, they are at non-negligible risk of obtaining more than one unique identifier over time. Undertaking this process would facilitate phase 4 on maintaining consistent identifiers over time.

In the case of **Canada**, individuals can appear on multiple Immigration, Refugees and Citizenship Canada (IRCC) administrative records. For example, individuals can have multiple temporary resident permits (e.g. international student then a temporary foreign worker) before becoming a permanent resident and then could become a citizen afterwards. Integrated, these files permit comprehensive analysis on immigration-related experience in Canada. Data are first integrated using common individual identifiers. These identifiers are issued administratively, and are intended to be singular and unique for immigrants and temporary residents in Canada.

However, while developing the Longitudinal Immigration Database (IMDB), Statistics Canada identified duplicates (cases where an individual had multiple IRCC person identifiers) through internal record linkages. While these are rare for immigrants (the working assumption is that they do not occur), they are more prevalent for temporary residents. Statistics Canada addresses these duplicates when replacing the IRCC person identifier with a new IMDB anonymized identifier. This process is explained in more depth in 3.4.

3.4.2.1.3 Addressing other data quality issues associated with source files

207. This would refer to statistical methods well established in existing literature to address data quality concerns. These could include:

- Weighting for sampling or missing records
- Correcting or removing invalid values
- Imputation for missing values

208. However, attention should be paid to how these methods could be problematic in a longitudinal integrated data source.

209. In the case of imputation for missing values, as described in section 3.2, imputation on cross-sectional files may be valid but combining these results longitudinally will lead to aberrations for every imputed value. For example, if a longitudinal database will be derived from integrating two cycles of a data source and a random 5 per cent of records were imputed on both cycles. This could lead to as much as 9.75 per cent imputation error. More cycles will lead to an increase in this error.

Imputation flags could be added to the data files to assist analysts in mitigating the impact of this risk.

210. Imputation could still be used but the risk of creating spurious outcomes should be considered carefully. In the case of imputing characteristics on the Cohort file, the risk may be smaller, for example.

211. For sampling limitations with a Cohort file, the accompanying sampling weights should be considered. Additional steps to account for integration error would require more consideration.

212. One option to address coverage limitations is to combine multiple data sources. For example, it can be the case that individual files only cover subsets of the population or only cover certain periods of migration. Combining with other sources which cover the complementary populations can address this limitation. As always, attention should be paid to any integration process and the inherent risk of integration error.

In **Spain**, the population register (Padrón) is the ideal source to obtain longitudinal information about migrants. It reflects the place where a person is registered at a specific moment and contains some concrete variables: sex, age/date of birth, place of birth, citizenship, name and family name, level of educational attainment or official identification number. Padrón not only stores the place where a person is at a specific time, but as a longitudinal record, it has all the places where a person has registered since 1996.

One limitation of the Padrón is the fact that it only has information on a few variables. On the other hand, it is worth mentioning how the last two Population Censuses in Spain have been conducted:

- 2001 Census was exhaustive/traditional and the use of administrative records was very limited (Padrón only)

- 2011 Census was combined. It was based not only on the Padrón, but also on other administrative records, and a large survey was conducted on just under 10 per cent of the population.

The plans for the 2021 Census are to try to carry out the operation only from administrative records including: Padrón, Social Security, Tax Agency, and Ministry of Education.

Within all the variables that the Census must provide, variables related to migrants include: Year of arrival in Spain/region/municipality/dwelling, previous place of residence, place of residence 1 and 10 years before.

However, Padrón has a limitation: its starting point is in 1996, so information previous to that year is not available. Therefore, the strategy or the solution that will be adopted will be to link the information of Padrón with the 2001 Census (which was exhaustive) and thus the information will be available for all the people in scope.

Data linkage between Census and Padrón has been made using different variables. First of all, using unique reference numbers. Then, for unlinked records, using auxiliary variables such as name, surname and date of birth. Finally, probabilistic imputation procedures (from other similar records) will be used to complete blank or erroneous information.

It is important to bear in mind that the variables related to migration must comply with certain rules (for example the year of arrival to Spain must be before or equal to the year of arrival to the dwelling) and if these are not fulfilled from the data sources, as previously mentioned, compliance will be forced from probabilistic imputations.

In **Canada**, the Longitudinal Immigration Database (IMDB) includes an integration of immigration files since 1980 with annual tax files. The populations not covered by this integration (largely the

non-immigrant population but also immigrants who were admitted prior to 1980) are not covered. To be able to address this limitation, the IMDB is often used in parallel with the Longitudinal Administrative Databank (LAD) a 20 per cent sample of all tax filers in Canada. The LAD is linked with the IMDB to identify those tax filers covered by the IMDB.

Ahead of the 2021 Census, Statistics Canada is investigating the possibility of replacing questions on year of immigration and immigrant status using integrated administrative records. However, the administrative records are only available since 1952. In an approach similar to that described for Spain, to address this limitation, Statistics Canada is investigating combining the administrative data with responses from previous censuses which would include years of immigration prior to 1952.

In **Hungary**, the establishment of editing rules and the imputation of missing values are crucial elements of the data cleaning process. The main data quality concerns are related to the missing or contradictory dates of migratory events (e.g. no information on immigration/return is available between multiple emigrations of the same person).

3.4.3 Report on data quality of source files

213. Finally, the quality assessment of the source files should be well documented for future users of the resulting database.

214. As a first step, if any modifications were made to the database that would affect coverage or measurement after pre-processing (section 3.4.2), the quality assessment undertaken (section 3.4.1) should be replicated.

215. Data quality indicators, such as those suggested in section 3.4.1 should be well documented before and after any data quality pre-processing. For example, if data are imputed to adjust for missing values, distributions before and after imputation should be compared with confrontation data sources (i.e. alternative data sources with similar measurements on similar populations which can be used for comparison purposes). This permits analysis of whether the processing steps improved coverage or measurement quality.

216. Rates of data quality pre-processing (e.g. imputation rates) should be documented and cross-tabulated by relevant variables (e.g. time period and variable values). For example, if sex is imputed on an annual Outcomes file, the imputation rate by year and by final sex value would be of interest.

217. Finally, the report should include guidance on the quality and usability of various fields on the Cohort and Outcomes files. This guidance will support future resulting analysis by informing users on the strengths and limitations of the source data files. The report is explained in more detail in phase 6.

3.5 Phase 3: Data integration for longitudinal data[6]

218. This phase encompasses the actual data integration process of the project. This section will outline methods of integration and provide an overview of sources of integration error. Examples are provided from different countries including approaches used to limit integration errors.

[6] [6] This section builds on UNECE (2019), which is a recommended source to be used in supplement to the present Guidance

219. This section will not cover the administrative processes which may be required before undertaking any data integration project. In order to protect confidentiality, national statistical offices have strict rules in place to govern data integration projects. These would be above and beyond the consideration to access the source files themselves including variables necessary to complete the data integration. These processes depend on the country but they should be considered before any data integration project.

220. For the purposes of this section, 'data integration' refers to the microdata combination of distinct data sets for the purposes of creating a longitudinal database. 'Record linkage' refers to the microdata process of combining the data sets and is mostly used to refer to probabilistic linkages as opposed to deterministic ones matching on a single unique identifier. However, it is possible to conduct deterministic linkage on name, date of birth, sex and geography, ahead of probabilistic linkage.

221. The ultimate result of this phase is to have a linkage key connecting the Cohort and Outcomes Files along with additional documentation on the quality of the data integration.

3.5.1 Data integration methods

"The principle of record linkage is simply to compare and bring together records from two (or more) sources that are believed to relate to the same real-world entity and whose (presumably overlapping) description can be matched in such a way that they may then be treated as a simple record" (Berti-Equille, 2007:112).

222. Data integration could be undertaken using two distinct approaches:

- Exact matching by using a common unique individual identifier (for registers or other data sources with a common identifier)
- Probabilistic matching by applying a specific algorithm and using different variables (such as name, sex, date of birth etc.)[7]

223. However, it is suggested that prior to this phase, internal linkages are undertaken on the source files to identify duplicates (see section 3.4.2).

For example, in **Canada**, the Immigration, Refugees and Citizenship Canada (IRCC) administrative records use common person identifiers which permit the different files to be connected to individuals. Regarding the Longitudinal Immigration Database (IMDB), which combines linked administrative immigration and tax data files[8], annual tax files can be connected to individuals over time through their Social Insurance Number (SIN). However, there is no common unique individual identifier between the immigration and tax records. Therefore, a probabilistic record linkage is needed. Before this, an internal linkage is undertaken as described in 3.2.2 to identify duplicates on the source files.

This record linkage process uses variables available on both data sources (e.g. names, date of birth, etc.) and currently uses the Social Data Linkage Environment (SDLE) to facilitate the linkage. Because

[7] This could also include deterministic or hierarchical deterministic matching on variables such as name, sex, date of birth, etc.

[8] The IMDB brings together immigration information from IRCC (immigration records and temporary residence records), taxation data from the Canada Revenue Agency and the date of death from Statistics Canada's Mortality Database.

migrants tend to have more changes to their characteristics over time, Statistics Canada has found it useful to rely on longitudinal linkage sources. This implies a source which can contain changes in an individual's information over time but connects these changes through an individual identifier. The use of such a file permits links to be made even if names or addresses have changed. In the case of the IMDB, historically, this longitudinal linkage source was comprised of tax data connected to a single SIN (or identified group of SINs which belong to the same individual).

In **Germany**, the Central Alien Register (AZR) – a linked-administrative record registering all non-Germans currently registered in Germany (active foreigners) or previously registered in Germany (inactive foreigners) – has a cross-period and uniform personal identifier (PI) since 2007. This PI is an anonymized AZR code.

The **Italian** National Institute of Statistics (ISTAT) has been working on administrative data for some time, applying record-linkage techniques in order to increase value derived from the information from a longitudinal point of view. It implemented a new system SIM (Integrated System of Microdata) with the aim of producing a unique statistical identification code. The data pertaining to residence permits have been integrated with data on the acquisition of citizenship, the population register, and the social security register in order to monitor the inclusion process of non-EU citizens. For the database of residence permits, while a tax code was used until now, they are being replaced by the SIM. Regarding the survey on integration of second generation, it has been linked with the Population registers and Student registers. The two databases are integrated through deterministic record linkage using the tax code.

In **Spain**, the INE (National Statistical Office) is responsible for coordinating the single national population register (Padrón), which is the combination of the register of each municipality in Spain. Every person on the register has an identity number. For most individuals (including legal foreigners in the country), this is the national identity number; for those in an irregular situation, their passport number is stored.

There are two situations when it could make sense to undertake probabilistic linkage. The first is when Padrón incorporates information from other auxiliary sources. Padrón only contains information about a few variables (date of birth, place of birth, sex and nationality). Probabilistic record linkage is thus needed to incorporate more variables from different sources. The second situation when probabilistic linkage is needed is when a Padrón member with identifier X leaves Spain and returns a few years later. On return, another Padrón identifier, for example Y, will be assigned. Consequently, to construct a longitudinal register for the whole population, record linkage can confirm that the person with identifier X is the same as the one with identifier Y.

Regarding **Switzerland**, the availability of a unique Personal Identification Number (PIN) – corresponding to the social security number – enables linkage between the Central Register of Foreign Nationals (ZAR) and the Population and Households Statistics (STATPOP) to construct the Swiss Longitudinal Demographic Database (SLDD). This 13-digit PIN (known as the NAVS13) is completely anonymous and was introduced at the end of the first decade of the 21st century. Individuals need a NAVS13 in order to work in Switzerland and it is frequently requested when dealing with authorities and administrative offices, for purposes such as getting insurance, benefits or health care. Therefore, undeclared migrants often have such a number. A number is attributed to every foreign national entering the country (either through immigration or by birth). For statistical purposes, this identifier is available to anyone who has been living in Switzerland at any time since December 31, 2010. The various registers were linked or compared using two distinct approaches: 1) exact matching by using the common identifier (PIN) and 2) matching by applying a specific algorithm. Linking both registers (ZAR & STATPOP) using a PIN was not possible for former asylum-seekers who obtained a residence permit or settlement permit prior to 2008. In those cases, a match based on the availability of either non-modifiable variables (i.e. date of birth, sex, nationality at the time of arrival in Switzerland) or rarely modifiable variables (i.e. place or residence, civil status) in

both registers was required. This algorithm takes into account two characteristics in particular: territorial changes and change in nationality of an individual (for example, a Yugoslavian takes on Serbian-Montenegrin nationality, then Serbian nationality) over time.

3.5.1.1 Linkage errors

224. Validation of the results is always an important step of data production. Administrative errors, such as duplicates, gaps, recordings of different dates and contradictory events need to be considered. Due to the nature of administrative data, variables such as name and address are captured through an administrative process and can therefore be registered with a delay. For migrants, information may change more often leading to a higher risk of linkage errors. In particular, the linkage of migrants can be disproportionately affected by various factors such as:

- Migrants taking on a given name more familiar to the new country of residence
- Migrants changing family name after marriage (marriage could be reason for migration)
- Migrants changing contact information and address after arrival in new country of residence
- Migrants not being familiar with local alphabet – leads to various spellings of names
- Migrants not being familiar with local calendar – leads to uncertainty of date of birth

225. Linkage errors arise when pairs of records are incorrectly classified. False-matches (also known as false positives) can occur when records from different individuals link erroneously. Missed-matches (false negatives) may happen when records from the same individual fail to link because of misreporting (for example typographical errors), changes over time (e.g. married women's surnames, addresses) or missing values. Linkage errors can threaten the reliability of results based on analyses of linked administrative data. Where possible, strategies have been developed to address these linkage errors. In any event, linkage errors should be estimated and well documented to inform on the fitness of use of the resulting data.

226. To assess the quality of a record linkage, best practices include:

- Qualitative assessment of linked and unlinked record pairs to estimate false negative and false positive rates
- Quantitative assessment of linkage rates. Linkage rates should be calculated from the perspective of the source files (i.e. the source files should be providing the denominators in the calculation of linkage rates).

227. For linkage rates, calculations should be done from the perspective of each source file. That is, the source file provides the denominator and the linkage rate is the percentage of that denominator which is linked. Rates then need to be interpreted through the lens of the source files. Linkage rates should only ever be expected to approach 100 per cent when everyone on the source file is in scope to be linked. Lower linkage rates could be due to linkage error but also relative differences in the presence of individuals on the other source file(s). These differences could be due to coverage or sampling error or quite simply the absence of individuals in certain linked source files (e.g. not linking individuals to hospital data could simply imply that individuals are healthy and thus not requiring visits to the hospital).

For example, in **Canada**, the Longitudinal Immigration Database (IMDB) undergoes an annual record linkage to add new cohorts and new outcomes. The linkage is assessed for quality through various steps:

1. Qualitative assessment of linked and unlinked record pairs to estimate false negative and false positive rates
2. Overall linkage rates between Outcomes and Cohort files to the Social Data Linkage Environment (SDLE) which is used to facilitate the record linkage
3. Quantitative analysis of linkage rates, calculated with each of the source files in the denominator.
4. Comparative analysis between new and previous linkage results

In particular, much time is spent on the linkage rates from the immigration files (Cohort files) perspective considering variables of key analytical interest but also variables which may be relevant to linkage errors. These include, but are not limited to, year of immigration or arrival, year of birth, sex, admission category, and country of birth.
The process and results are documented in the IMDB Technical Report.

In **Spain**, the most common errors include compound names that appear differently in each source, change of surname, birth date equal to January 1st[9], twin brother or sister with a very similar name, changes in nationality, and changes in the identification number. To deal with these issues, a probabilistic link is usually made between the sources, and links above a certain confidence threshold are accepted.

In **Germany**, the Federal Statistical Office does not deal directly with linkage errors because they are not in charge of the registers. Every year they receive a register excerpt from the register-keeping authority, namely the Federal Office for Migration and Refugees, with coded information only. This means that they receive the personal identifier (PI) of individuals as well as a few variables. These latter are both time-dependent (e.g., residence status and residence permit with issue date) and time-independent (e.g., gender, citizenship, date of birth, date of first entry, etc.). A mismatch can only happen if time-independent variables change over time. Changes in citizenship have been experienced, but this is not evidence of a mismatch. Changes in dates of birth and dates of first entry have also occurred, as a result of incomplete data (e.g., missing day and/or month) in very old data which was completed in later entries. This is also not definite evidence of a mismatch.

Mismatches are unlikely to happen as the original ID (which is derived from an anonymized transformation of a PI) is created as a time stamp on the day of first registration plus a database-generated unique counter. It is possible that a person was registered as a new case even though he/she had already been registered previously. However, it would mean that the person had a new passport, so that the passport number and previous entries for visa or residence permits were no longer visible. Under normal circumstances, such multiple registrations do not happen often as an emigrated foreigner stays as an "inactive" record in the register for an additional 12 years after his/her emigration. However, during the increase in refugees in 2015-2016, the authorities were overburdened and work was delegated to untrained personnel (e.g., armed forces) which led to many multiple registrations. The register office recently started to identify the double/multiple registrations by checking names and date of birth. They will, at a later stage, include fingerprint data for this purpose. This work is still in progress. For these reasons, the contributions and success of this process cannot yet be evaluated.

Regarding **Switzerland**, different tests were undertaken to verify the data, in particular the control of erroneous linkages, the plausibility of sequences of events and their completeness. Once these

[9] When conducting data-linkage between different sources, there is often a high number of people who were born on January 1st (especially within the foreign population). The reason is that most of those people either do not declare or have forgotten their birth date. This complicates the data-linkage process since it is possible that the same person has a different birth date than January 1st in another data source.

tests were completed, a second phase involved the identification of the sources of the errors, by consulting the Swiss Federal Statistical Office (SFSO) or by running plausibility tests with control variables. The third phase consisted of deciding how to correct the data. Duplicate records were deleted. In some cases, records concerning two different persons were linked together. The difficulty in such cases is to determine whether records are duplicated (i.e., one person recorded twice) or if the PIN number is incorrect (i.e., two persons have the same PIN number). Such cases were forwarded to the FSO which has access to complementary information (e.g., the person's name) and was able to correct the data. When there is incoherence between information from different sources, the information available in the "stock" file is prioritized.

The evaluation of the completeness of trajectories is based on different rules regarding the beginning and the end of a trajectory. The beginning of the trajectory is defined by an immigration movement or a birth. For foreigners already living in Switzerland before the beginning of observation (1998), the "stock" records as of December 31st, 1997 represent the beginning of the trajectory. The end of a trajectory is defined as departure, death or naturalisation. For people residing in Switzerland at the end of the observation period (2013), the "stock" records as of December 31st, 2013 represent the end of the trajectory. A presence at only one time without any information regarding the entry or departure is not considered as a trajectory and was therefore deleted.

Regarding **Italy**, the databases are linked by an Integrated System of Microdata (SIM) code. The SIM code is the result of an algorithm that applies a unique identification number to each individual record.

One challenge in Italy relates to maintaining good links over time. The Italian privacy law is very restrictive with respect to linking and using identifying information about individuals. Even then, identifiers are more likely to be of lower quality due to problems related to the use of different alphabets. Additionally, the tax code, used as a linking key for the residence permits, is missing for some individuals, particularly for migrants who arrived in Italy as asylum seekers.

In addition to linking the databases, the process anonymizes the contained information and removes the tax code. This process is conducted by ISTAT (the Italian National Statistical Office). ISTAT do not fully trust the result of this algorithm yet, and the result of the linkage is verified with deterministic record linkage and probabilistic record linkage. In some cases, errors are found when the linkage is made with the SIM identifier. When this happens, ISTAT informs the office that deals with SIM, who then check and record these errors by updating a transcoding table that allows the office to track changes and correct errors.

3.5.2 Documentation of integration errors

228. Documentation of the integration process should include an overview of the integration methodology as well as detailed results on the integration quality.

229. Integration errors including estimates of false positive and false negative links should be well documented. Error rates should be cross-tabulated by time-period and key variables of interest. Any subpopulations of interest which appear to be disproportionately affected by integration errors should be well documented.

230. Finally, guidance to users should be provided on the usability of the results. Particularly, guidance should be provided if certain subpopulations or time periods may have lower integration quality. For example, in Canada, immigration records from 1961 to 1972 often do not include a complete date of birth, this results in lower linkage rates for this period of immigration when linked

with other files such as the 2016 Census. These results can either lead to the exclusion of certain populations from the results or necessitate new methods to adjust for integration error.

231. The next 2 phases address outstanding issues created through the integration process. Phase 4 outlines the complexities associated with maintaining a consistent unique record identifier longitudinally (especially after repeated integrations) while Phase 5 describes the methods needed to prepare the final integrated database, including adjustments to account for integration errors.

3.6 Phase 4: Assignment of longitudinal individual identifiers

232. After phase 3, a linkage key is established between the Cohort and Outcomes files after data integration. In this phase, the key is considered in the longitudinal context. The outcome of this phase is the transformation of the linkage keys (duplicates and other longitudinal issues are addressed) into an anonymized longitudinal individual identifier ready to be used to create an analytical database. This new identifier should then be attached to the Cohort and Outcomes files (often replacing any pre-existing identifiers) to facilitate analysis of the longitudinal database. This section will examine common issues and approaches used by different countries to address the various challenges.

233. There are several benefits to statistical offices assigning a new anonymized individual identifier. This includes adding a layer of confidentiality to the resulting database. The anonymized identifier would mean nothing to users of the database and its creation eliminates the necessity of having any personal identifiers on the integrated data files. On the other hand, official identifiers generated for administrative purposes could be used to identify individuals.

234. A critical element of any longitudinal database is the capacity to connect outcomes over time to the same unit (person, family, etc.). This is facilitated through the use of unique record identifiers for the units of study.

235. When a longitudinal database is developed using integrated data sources, consistent unit identifiers need to be established. In some cases, and as is often the case for databases created from population registers, a common unique identifier already exists (i.e. the same unique identifier exists on all sources being integrated). In other cases, the individual files being integrated may have independent unique identifiers that will need to be combined.

236. An additional complexity arises for longitudinal databases, where the unique identifier is adjusted over time to account for discovered **duplicates** (i.e. multiple identifiers for the same unit) or **splits** (i.e. multiple units for the same identifier). While this situation would only arise for persons when an administrative error occurs (e.g. erroneous creation of a new identifier for someone who already has one), family units are more complex, as they can merge and split over time.

3.6.1 Data integration with common unique identifiers

237. If a longitudinal database is being developed using integrated data sources with common unique identifiers, there may be no need to create a new identifier. The common unique identifier can be reused or transformed (e.g. anonymized) to create a new unique record identifier for the purpose of the longitudinal database. The working assumption in this scenario would be that this unique record identifier exists, and is unique for every unit in scope.

238. However, the transitional nature of migrants can impact their identifiers in national databases. For example, if an individual migrates to a country two or more separate times (e.g. circular migration), they may be issued multiple identifiers. There could be an administrative process to detect such re-entries and assign one single identifier to the individual but this process may not be exhaustive. An internal record linkage prior to phase 3 could identify such cases. This was mentioned as a best practice in section 3.4.2.

239. If individuals are found to have multiple individual identifiers in the database, this identifier should be transformed to have one identifier per individual. Otherwise, there will be a disconnect between different events for the same person.

In **Switzerland**, the Swiss Longitudinal Demographic Database (SLDD) is constructed in 2 phases:

1) A demographic base is constructed, in which the Central Register of Foreign Nationals (ZAR) and the Population and Households Statistics (STATPOP) are linked
2) Other data sets, such as the Structural Survey (RS) and unemployment statistics, are incorporated to this demographic base.

These linkages are done using a unique Personal Identification Number (PIN) – corresponding to the social security number- which is available in each of the databases.

This 13-digit PIN (known as the NAVS13) is completely anonymous and was introduced in the late 2000s. NAVS13 is needed to be able to work in Switzerland, and frequently requested when dealing with authorities and administrative offices for purposes such as getting insurance, benefits, or health care. Therefore, even undeclared migrants often have such a number. A number is attributed to every individual entering the population (either through immigration or by birth). For statistical purposes, this identifier is available to anyone who has been living in Switzerland at any time since December 31, 2010.

In **Spain**, the Padrón (Population register) is created and managed by the national statistical office, INE. While there are as many registers as there are municipalities in Spain, there has been a law in force since 1996 that ensures the integration of these municipal lists into a single national database. There are also legal procedures that keep this database, and the municipal files, interconnected and updated on a monthly basis.

INE is responsible for coordinating the single national population register. It validates changes produced in every municipality to avoid duplication, and it updates it to include deaths, births, and acquisition of Spanish citizenship, which INE receives on a monthly basis from the Civil Register. Every person in the register has an identity number. For most individuals (including legal migrants in the country), this is the national identity number; for those in an irregular situation, their passport number is stored.

INE continuously receives information on registrations and de-registrations from each municipality, so that, in order to avoid duplicates, once the notification of a new registration of an individual in a municipality or consular section is received, that individual is deleted in the municipality of origin. When INE detects a duplicate, the affected municipalities are informed about the duplication of registrations detected and steps are taken to find out which is the correct record.

In order to avoid the same external identifier (national document, residence permit or passport) being used for several people, INE generates its own internal identifier for each person in Padrón. This identifier remains unchanged while the person remains in the country. If one person leaves the country because he/she is transferred abroad and then returns to Spain, a different internal identifier will be generated, even if it is the same person. An internal record linkage, as described in 3.2.2, could be used to rebuild his/her history.

In **Germany**, the Central Alien Register is also constructed using linked administrative records based on an anonymized person identifier.

In **Italy**, ISTAT has implemented a new system, SIM (Integrated System of Microdata), with the aim of producing a unique statistical identification code. For residence permits, tax codes have been used to date, but they are being replaced by the SIM. The identification code SIM is fully managed and generated by ISTAT through a specific algorithm and has been used as a substitute for the anonymization process that is required by law to link and integrate the administrative databases in one unique Microdata Integrated System (SIM).

Duplicates by SIM code are double checked since both false negatives and false positives have been identified by linking microdata with using deterministic and probabilistic methodologies. In some cases duplicates by SIM codes are in fact determined to be different individuals. These cases are referred back to the office that manages the algorithm. These communications between producers and users of the code is crucial to enhance the identification process and avoid errors in the future, errors that affect mainly specific subgroups of population such as foreigners.

In **Canada**, Immigration, Refugees, and Citizenship Canada (IRCC) data is integrated using common person identifiers, which allow the different files to be connected to individuals. These identifiers are issued administratively, and are intended to be singular and unique for immigrants and temporary residents in Canada. While developing the Longitudinal Immigration Database (IMDB), Statistics Canada identified duplicates (cases where an individual had multiple IRCC person identifiers) through internal record linkages. While these are rare for immigrants (the working assumption is that they do not occur), they are more prevalent for temporary residents. Statistics Canada addresses these duplicates, only for temporary residents as the assumption is that there are no duplicates records for immigrants, when replacing the IRCC person identifier with a new IMDB anonymized identifier.

In Canada, tax data can be connected to individuals over time through their Social Insurance Number (SIN). Canada Revenue Agency (CRA) has identified to Statistics Canada cases where individuals have multiple SINs. This is particularly common among individuals who transition from temporary to permanent resident status, as they are issued a temporary SIN prior to becoming a permanent resident. Internal record linkages have also been conducted by Statistics Canada to identify additional duplicates (or additional cases where an individual had multiple SINs). These linkages, and resulting discoveries and adjustments, are undertaken annually.

In **Kazakhstan**, common unique identifiers are also used in order to avoid duplicates. Three statistical units are produced on the basis of a legal unit: person, family and household. In the first case, a Personal Identification Number (PIN) – corresponding to the official identification of residents of the Republic of Kazakhstan – is generated by the Ministry of Justice for the Individuals' State Database, one of the main sources of information for the Statistical Population Register. This PIN is assigned to an individual in the Individuals State Database and passed to the Statistical Population Register. In order to solve some inconsistencies between different sources of information for the same individual, a system of priority between sources has been established.

In **Turkey**, the Address Based Population Registration System (ABPRS) is a unique system from which annual population statistics are disseminated. Within this scope, annual information on international migration stocks and flows by citizenship and country of birth are produced based on the ABPRS and other additional administrative sources.

The maintenance of consistent personal identifiers over time for the ABPRS has enabled the longitudinal follow up of individuals over time. However, individuals can be affected by a change of identification number after acquisition of citizenship (i.e. loss of the trace of the naturalized person).

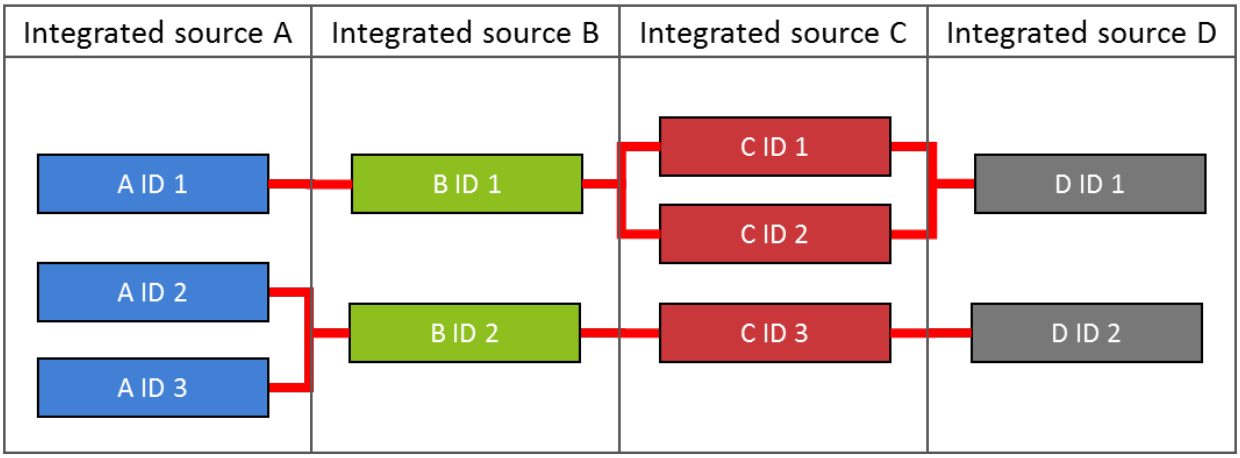
3.6.2 Data integration without common unique identifiers

240. There are countries without a population register, or without the ability to link files using a common unique identifier. For longitudinal data sets arising from integrated data sources without any common unique identifiers, a unique identifier needs to be generated. In some cases, if one of the integrated sources serves as the base of the longitudinal database (i.e. where cohorts will be defined), the unique identifier for that base could be used. Alternatively, a combination of identifiers from different sources could be used or transformed (e.g. anonymized).

241. Under these circumstances, the generated unique identifiers would need to account for **duplicates** (i.e. multiple identifiers for the same unit) or **splits** (i.e. multiple units for the same identifier), some of which could be identified in the integration process itself. In longitudinal databases that are updated over time with new linkages, additional duplicates and splits could be identified. A process to address future discoveries should be established.

242. To identify duplicates on the source files, internal record linkages prior to Phase 3 should be undertaken. This was mentioned as a best practice in 3.2.2.

Figure 9 Differences in identifiers between integrated sources



243. For example, Figure 9 illustrates an integration of 4 data sources with different record identifiers. These sources could be connected over time to create a longitudinal data source (with source A representing the earliest outcomes and source D the latest). The integration process identified discrepancies between the identifiers. For example, according to the link between source A and source B, A ID 2 and A ID 3 are the same person (B ID 2). A decision needs to be made regarding how to reconcile these discrepancies.

In **Canada**, the Longitudinal Immigration Database (IMDB) comprises the integration of tax data, with Social Insurance Numbers (SIN) as the unique identifier, and administrative immigration data, with IRCC person identifiers. The IMDB assumes that the IRCC person identifiers are singular and unique (after accounting for duplicate identifiers associated with temporary residents), and they are replaced with an anonymized IMDB person identifier. Then, as part of the record linkage process, new duplicates and splits for SINs are discovered. As a result of these discoveries, a SIN, or an identified group of SINs, is replaced with an IMDB person identifier.

In **Italy**, ISTAT has developed Anvis, an information system of demographic accounting on an individual basis with the aim of ensuring the consistency of all the individual flows that feed and/or

modify the stock of individuals in the resident population. By means of data linkage, it is possible to observe the migration history of each person, and link all the events related to the same individual.

3.6.3 Family identifiers

244. Unlike persons, families can truly merge and split over time; this adds a new level of complexity to family identifiers for the purposes of longitudinal data.

In **Canada**, the IMDB identifies families on both immigration records and tax records. On immigration records, families are identified as those who are admitted under the same permanent resident admission. In some cases (e.g. family reunification), not all family members will be under the same admission. These family identifiers do not change over time as they correspond with a single admission. On tax records, however, family units are being measured annually (based on spouses or parent(s) and children living together in the same household), and can change every year. The IMDB transforms the identifier of one of the adult immigrants in the tax family as the family identifier for all family members. This can lead to these identifiers changing every year as the composition of families evolves. In fact, even in cases where the family identifier is consistent from one year to the next, the composition of the family can change (e.g. children moving out of the household).

In **Kazakhstan**, three statistical units are produced on the basis of a legal unit: person, family and household. A unique 12-digit code is automatically assigned by the Statistical Population Register to each family and another one is provided to identify each household. In order to solve some inconsistencies between different sources of information for the same individual, a system of priority between sources has been established.

245. These complexities make it challenging to consider longitudinal analysis of families. Chapter 4.1 briefly considers longitudinal indicators related to family.

3.7 Phase 5: Create final database

246. After phase 4, a unique individual identifier is attached to the Cohort and Outcomes files. In this phase, the Cohort and Outcomes files are further processed to create a final database. This can include setting up the structure of the database, sub-selection of records relevant to the database, harmonization of content, and the standardization of variable names across files and reference periods.

247. By the end of this phase, the longitudinal database will be ready for the final phase of dissemination.

3.7.1 Confidentiality protocols

248. Following the addition of the unique individual identifier, to protect confidentiality, all personal identifiers should be removed from the Cohort and Outcomes files.

249. Further, the level of granularity in the variables should be considered at this stage. Variables should only be as detailed as is necessary for the statistical objectives. Any identifying variables such

as dates of birth, addresses, etc. should be replaced with more aggregate variables such as age or geographic area of residence. All variables should be considered in this context, yielding the appropriate analytical value while protecting the confidentiality of the individual records.

For example, the Longitudinal Immigration Database (IMDB) in **Canada**, replaces variables with dates (e.g. admission date) with years or years and months depending on the analytical need.

250. Finally, access requirements should be established to ensure that the database is only accessible by those with appropriate approval.

3.7.2 Database structure

251. There are many options for the structure of a longitudinal database.

252. If the number of outcomes is relatively small, it might be preferable to create a single flat file containing the Cohort and Outcomes variables of interest. In this scenario, the file could either be wide (where longitudinal outcomes are denoted by columns or variables specific to each outcome / reference period), or long (where each individual has a new row for each outcome / reference period).

253. For example, a basic database combining immigration details (year of immigration) with longitudinal annual employment outcomes can be presented as a wide file:

Person id	Year of immigration	Sex	Employed in 2009	Employed in 2010	Employed in 2011	Employed in 2012
1	2009	Male	0	1	1	
2	2011	Female			1	1
3	2009	Female	0	0	1	1
4	2010	Male		0	0	0

Or as a long file:

Reference year	Person id	Year of immigration	Sex	Employed
2009	1	2009	Male	0
2010	1	2009	Male	1
2011	1	2009	Male	1
2011	2	2011	Female	1
2012	2	2011	Female	1
2009	3	2009	Female	0
2010	3	2009	Female	0
2011	3	2009	Female	1
2012	3	2009	Female	1
2010	4	2010	Male	0
2011	4	2010	Male	0
2012	4	2010	Male	0

254. However, often, the number of outcomes is non-negligible and a different approach may be required. One option would be to maintain a relational database connected by the individual identifier. The Cohort and Outcomes files would be kept separately but brought together depending on the analysis of interest.

In **Canada**, the Longitudinal Immigration Database (IMDB) has a large Cohort file consisting of many variables related to immigrants' characteristics upon admission to Canada as well as some demographic content. The Outcomes files are annual tax files with hundreds of different measures spanning topics such as income, area of residence, family, and tax credits. The Outcomes files are themselves separate as there is one file per tax reference year.

This relational database format is flexible to new additions as well. Recently, temporary residence permits (itself a longitudinal dataset with individuals having multiple permits over time) was connected to the IMDB using a person-level identifier. Rather than integrate this information into the Cohort file (which would have reduced the analytical potential significantly), it was simply added as its own new relational piece of the database.

In the figure below, the current structure of the IMDB is illustrated. The Permanent resident files, citizenship files, settlement services records, and temporary resident permits are integrated to create connected files:

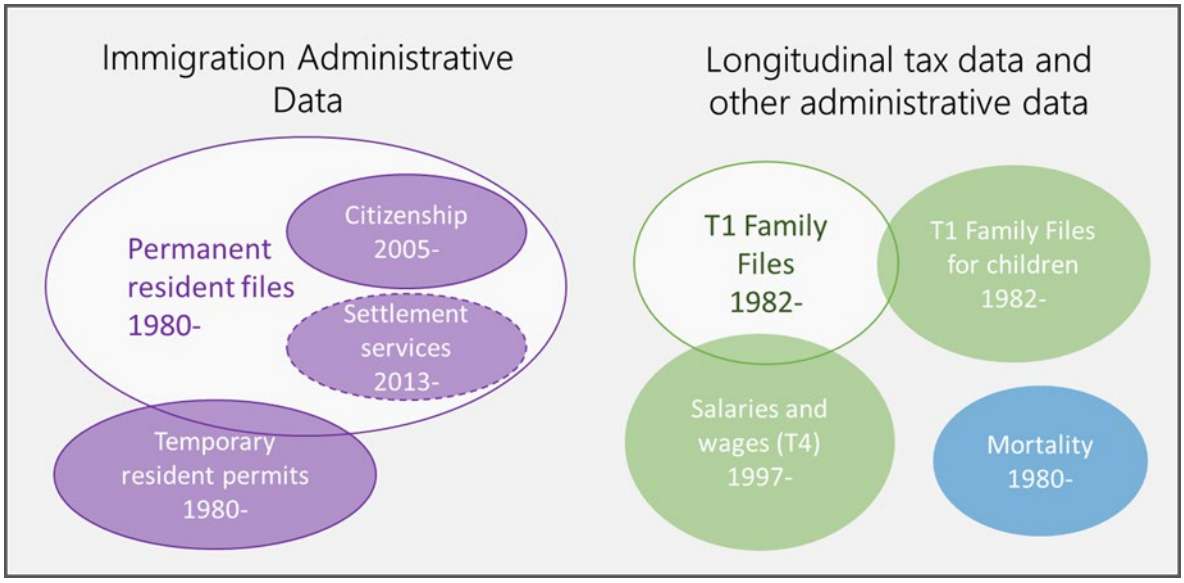
1. Individuals files (unit of analysis: person)
- Contains basic demographic information
- Contains details (if applicable) associated with conditions of becoming a permanent resident
- Contains citizenship information (if applicable)
- Contains summary variables of temporary resident experience in Canada (if applicable)
- Contains summary variables of use of settlement services in Canada (if applicable)
- Contains date of death (if applicable) from mortality outcomes file

2. Temporary resident permits (unit of analysis: permit)
- Contains details of all temporary permits held
- Dates associated with permits are shown allowing analysis in 'real-time'

3. Settlement services (unit of analysis: service)
- Contains details of settlement services received

There are separate outcomes files for each reference year since 1982. Also separate files for the main tax files (T1 family files), tax files for children, and Statement of remuneration files for salaries and wages.

Figure 10 Structure of Canada's Longitudinal Immigration Database



255. The periodicity of the Outcomes must be considered in this stage. It may be simpler to create periodic summary files instead of having real-time outcomes listed. In addition, there may be differences in periodicity between files which could be reconciled by creating such summary files.

256. However, these decisions should be weighed against the original purposes of the database outlined in phase 1.

3.7.3 Selection of records

257. Once the structure of the database is determined, it may be necessary to restrict the Cohort or Outcomes files (or both) to a subset of records relevant to the aims of the database. This could mean excluding certain populations because of their characteristics on either the Cohort or Outcomes files or restricting the database to those records which were linked. This decision should be motivated by the statistical design of the database outlined in phase 1. However, some decisions may be driven by practicality in terms of availability of other data sources compared with the technical demands of retaining a large database.

For example, the Longitudinal Immigration Database (IMDB) in **Canada**, only keeps the tax records for those linked to an immigration record. However, all immigrants and temporary residents are kept in the database, regardless of whether they were linked with a tax record. This is driven by the fact that other data sources exist for comparisons between the tax outcomes for immigrant and non-immigrant populations in Canada and the fact that retaining the additional tax records would affect the usability of the database. The IMDB is particularly supported by the Longitudinal Administrative Databank (LAD) which contains a 20 per cent sample of the longitudinal tax files in Canada with a flag identifying those linked to the IMDB.

258. Other decisions may still need to be made in the event of remaining duplicates. In particular, if an individual is linked to multiple outcome records in the same reference period, a decision needs to be made on how to address these situations.

259. Finally, this is the stage when any final weighting or calibration could be applied to the resulting database. Specific methods for weighting and calibration are outside the scope of this particular report. However, attention should be paid to how these methods will affect the intended longitudinal analysis. In particular, in the context of migration statistics, methods to address individuals 'disappearing' from the Outcomes files should consider whether this is the result of a missing outcome or an out-migration.

260. All of these decisions should be documented and outlined in a report. This should include analysis of the possible implications of the decisions taken or guidance in how to use other data sources to support comparative analysis.

3.7.4 Harmonized content and standard variable names

261. At this stage, it is important to recall some of the information gleaned in phases 1 and 2 when assessing the source files for their ability to meet the needs of the statistical design.

262. In particular, cases where concepts are not being measured the same way across files or across reference periods need to be examined. If possible, the relevant variables should be harmonized to ensure consistency. If this is not possible, differences between the concepts need to be properly documented. All variables with common (or relatively similar) concepts between files or across reference periods should be considered.

263. Harmonizing the variables over time will make longitudinal analysis much simpler as the concepts will be defined in the same way. If this is not done, differences observed over time may be due to changing classifications instead of real changes.

264. Harmonization could include the use of concordance tables to adjust for differences in classification or it could include transformation of variables across files or reference periods to reflect the same measure.

265. Another important step is to determine a naming convention for those variables which are comparable across reference periods. To make future analysis easier, keeping the exact same variable name for all reference periods is recommended. To simplify future analysis further, the reference period could be included in the variable name as well.

266. For example, in the example above of the wide file, the outcome variable could be employed YEAR where YEAR is the reference period.

267. Finally, this is the stage when any final processing should be done on the variables. This could include imputation not undertaken in Phase 2.

Statistics **Canada**, in collaboration with Immigration, Refugees, and Citizenship Canada (IRCC), harmonizes variables over time and across immigration and temporary resident files. This has included updating the databases to account for changes in the systems used by IRCC to collect the administrative data. Most of the time, these are simply changes in codesets.

To produce the IMDB, and the Longitudinal Administrative Databank (LAD) which uses the same approach, Statistics Canada first transforms the raw tax data into T1FFs (T1 family files), which include additional content of family composition, then further transform these T1FFs into longitudinal files which have harmonized variables over time.

All tax variables have a standard naming protocol (e.g. T4_I2005 provides the T4 employment income at the individual or "I" level – data is also available at the family level – and for the reference year 2005). This approach makes it easy to pull the same variable from multiple years in any statistical program. Data dictionaries include information on which reference years contain specific variables.

In **Hungary**, administrative data from the National Health Insurance Fund (NHIF) is used to produce data exclusively on the emigration and return migration of Hungarian citizens. It should be noted however that a minimum length of stay for migratory events is not required in the NHIF registration system. Therefore, in this regard, statistical adjustments are needed in order to comply with the statistical definition of migration. E.g. lengths of stay abroad (or in Hungary between return and re-migration) that are shorter than one year will be excluded from further steps of data production.

3.8 Phase 6: Disseminate results

268. After phase 5, the database itself is ready for dissemination. However, there are several other elements that should be established before any results are disseminated and before any analysis takes place.

269. In this phase, the final elements of preparing the longitudinal database for dissemination will take place. These include:

- Determining confidentiality rules for access and dissemination
- Final evaluation of the quality of the database
- Preparation of data dictionary(ies)
- Preparation or update of technical report / user guide / user support service

270. Throughout the phases of this chapter, data quality has been repeatedly examined but at various stages. In this phase, the final database will be assessed at a final time for fitness for use, but also all of the data quality considerations noted in the previous phases will make their way into formal documentation either in the data dictionary(ies) (when the data quality considerations are specific to the individual variables) or in the technical report / user guide. The technical report / user guide should also provide guidance on how the database should or should not be used based on the various data quality limitations identified.

271. After this phase, the database is truly ready for dissemination. The next chapter provides an overview of key longitudinal indicators and best practices related to the dissemination of results from a longitudinal database related to migrants.

3.8.1 Determining confidentiality rules for access and dissemination

272. It is imperative to establish strong rules to protect confidentiality of the resulting database. A database with a focus on subpopulations such as migrants is already considering small populations. The longitudinal aspect of the database yields more information than a simple cross-sectional survey. In addition, the use of administrative data warrants special consideration as well.

273. As a best practice, it is recommended to collaborate with and leave final decisions on confidentiality rules to methodologists who work in this particular domain of national statistics. The

rules should consider how the data would be accessed and the aggregate results analyzed. The risk of disclosure needs to be assessed against the utility of the data.

274. Aggregate results could have rules that consider outcomes specific to the level of geography (e.g. income dominance rule) and could imply rounding, suppression, or perturbation. Even the content of the microdata files should be scrutinized to ensure that those who access it are restricted to the information they need to undertake their analysis and nothing more. After Phase 4, all identifying variables can be removed and replaced by a randomly generated individual identifier. The level of granularity in the variables should be considered as well.

For example, in **Canada**, the Longitudinal Immigration Database (IMDB) has established rules to suppress small counts or estimates based on small populations, round other estimates, and remove outliers. These methods are outlined in the IMDB Technical Report (Evra & Prokopenko, 2018).

275. Finally, access needs to continue to be managed to restrict the database to those approved individuals who demonstrate that they require access for specified statistical purposes.

3.8.2 Final evaluation of the quality of the database

276. At this phase, the database should be evaluated for coverage after the data integration. Missing values or observations over time should be considered, as should inconsistencies between measured events (e.g. dying before migrating). If imputation has taken place, the percentage of imputed values and availability of imputation flags should be considered. Finally, the database should be assessed for fitness of use by undertaking preliminary analysis focusing on the key objectives stated in Phase 1.

277. The quality assessment should refer to the framework identified in the introduction of this chapter. How could these errors have occurred throughout the process? Were they identified and addressed?

278. To assess coverage, return to the sources used in the coverage analysis in Phase 2. Repeat this analysis but focussing only on the linked population. The temporal element added by the Outcomes files provides a means to establish that the Cohort population is in scope for those same reference periods. This makes it easier to compare against other data sources with the same reference period.

For example, in **Canada**, the Longitudinal Immigration Database (IMDB) can be compared against census estimates by restricting the comparison to the linked tax filer population for the same reference year. These comparisons can then be done by various characteristics common to both sources.

279. Patterns in the Outcomes files should be examined as well as how they line up with the Cohort files. Are there people missing in some reference periods? Are there signs of individuals disappearing? Are there outcomes occurring in a nonsensical order? Verify that outcomes begin after migrants arrive and end before they leave or die. All of these issues should be documented in the technical report / user guide. In some cases, they may warrant extra guidance for users to avoid certain types of analysis.

In **Hungary**, it is well-known that administrative sources underestimate the real size of outmigration flows since migrants often do not report on their migratory events to the authorities. However, it is assumed that administrative data on different subgroups are not biased in the same way. To reveal the migrant subgroups whose migration flows are systematically underestimated in administrative sources, administrative data were compared to survey data from Microcensus. This comparison has shed light on the main weaknesses of administrative sources, especially when dealing with distinct forms of multiple migration.

280. Any limitations identified in phases 1 or 2 should be reconsidered with the final results. Are there imputed records? Do they cause non-negligible spurious results over time? Were the coverage or measurement issues that were identified addressed?

281. If the database is updated, verify the results directly against the previous iteration. Are there notable differences in the linked population? Are there notable differences in the outcomes?

282. Finally, the database needs to be considered for the purposes identified in the Statistical Design. To undertake this exercise, begin preliminary analysis in line with the aims stated in Phase 1. If alternative data sources are available, compare the findings. Compare the outcomes for various cohorts and subsets of the population. In particular, if there are certain patterns which are expected by existing literature, verify that these patterns persist. Graphical analysis can be helpful to identify any anomalous outcomes over time.

283. Document the findings in the report and establish whether or not this database can be used to meet the aims stated in phase 1. If there are types of analysis which need to be avoided or modified, this should be stated clearly in the technical report / user guide.

3.8.3 Preparation of data dictionary(ies) and technical report / user guide

284. Now it is time to provide users with the relevant reference materials to undertake their analysis. This will include data dictionary(ies) which provide definitions and cover any data quality issues, absolute or specific to a time-period, pertaining to the variables. The technical report / user guide will provide a broad overview of the database, the various data quality considerations identified throughout this chapter, and guidance on how one should use the database for the aims identified in the Statistical Design phase.

285. As a reminder, the following were key data quality issues noted in the earlier phases which should be documented either in the data dictionary(ies) or the technical report / user guide.

Phase 1: Statistical Design

- Differences between concepts of interest and concepts measured
- Differences between populations of interest and populations measured
- Differences between temporal elements of interest and measured reference periods

Phase 2: Assessment and pre-processing of source files

- Coverage errors of source files before and after pre-processing
- Sampling error of source files before and after pre-processing
- Missing records (or non-response) on source files before and after pre-processing
- Measurement errors or missing values on source files before and after pre-processing

- Duplicates identified on source files

Phase 3: Data integration for longitudinal data

- Linkage error rates (false positives and false negatives)
- Linkage rates (carefully consider which single sources to use in the denominator)

Phase 4: Assignment of longitudinal individual identifiers

- Issues related to individual identifiers over time

Phase 5: Create final database

- Issues with harmonization of content
- Record selection issues

Phase 6: Dissemination of final results

- Coverage issues after data integration
- Missing or inconsistent observations after data integration
- Fitness for use of the database

286. The data dictionary(ies) should have an entry for each variable from the Outcomes and Cohort files. Each entry should include, at minimum:

- Variable name (following standard naming convention established in Phase 5)
- Variable definition or description
- Source of variable (along with any considerations for harmonization of content in Phase 5)
- Reference periods variable is available
- Any data quality considerations specific to the variable
- Links to relevant classifications

287. The data dictionaries do not need to delve into guidance on how to use the individual variables for analysis or go into great detail on the conceptual differences between what is ideal and what is measured. However, the variable entry should be clear about what is being measured and how.

288. The technical report/ user guide should cover much more detail. This should include sections describing the source files, methods used for pre-processing, the methodology used for the data integration, and all other steps taken in building this database. It should also cover all of the data quality elements addressed above and throughout this chapter. Finally, it should provide guidance to users on how to properly use the database. This guidance should include explanations of where the ideal concepts differ from what is being measured. These differences do not need to prevent the analysis from taking place but they can imply that the analysis needs to be adjusted accordingly.

For example, Statistics **Canada** publishes a detailed technical report on the Longitudinal Immigration Database (IMDB).

289. Descriptions of longitudinal indicators in Chapter 4 include some examples of cases where what was being measured did not align perfectly with the concepts of interest. However, the analysis can still proceed with a different approach and still yield useful results.

290. Developing a data source is always challenging but developing a longitudinal data source related to migrants through data integration includes many different obstacles not typical to simply

conducting a cross-sectional sample survey of the general population. It should be noted that while these challenges exist and while the resulting product will not be perfect, it can still be a powerful data source and yield valuable information on migrants and their conditions. The next chapter shows how to take this database further to regular dissemination.

Chapter 4: Disseminating regular migration statistics from longitudinal data sources

291. This chapter provides guidance on how regular migration statistics can be disseminated using longitudinal data sources. In particular, this chapter will include two sections:

- Key longitudinal indicators
- Best practices for dissemination of longitudinal statistics

292. The first section will outline a variety of longitudinal indicators which could be produced for migration statistics. These indicators include topics related to the migration process itself as well as socio-economic outcomes of migrants over time. Particular challenges associated with each indicator topic are presented and possible approaches to address these challenges are offered using examples from different countries. Challenges and examples focus on the use of administrative data or integrated data sources.

293. The second section will examine best practices in disseminating longitudinal migration statistics. This will include an overview of the key audiences for these statistics as well as a summary of different dissemination methods including new approaches such as interactive applications and infographics.

4.1 Key longitudinal indicators

294. Building on the UNECE publication "Measuring change in the socio-economic conditions of migrants" (UNECE 2015), this section proposes indicators on longitudinal outcomes, that is, on outcomes that occur over a span of time.

295. In the context of this report, "indicators" refer to statistics of general interest that can be replicated on a regular basis. They differ from statistics generated for the purpose of specific research. Indicators could be used to inform government policy or programme decisions but could also serve other stakeholders including community organizations and migrants themselves.

296. As a starting place, key indicator topics are listed for the primary areas of interest for longitudinal data. Within each topic, there could be multiple specific indicators proposed. While the key indicator topics may refer to general terms such as "length of time until an event", specific indicators could include the proportion of migrants who experienced that event after X years, average time until the event, or other time-based measures. The set of indicators could be defined for subsets of migrants (e.g. asylum seekers, temporary residents, etc.) and could be cross-tabulated by demographic or socio-economic status variables (e.g., age, sex, country of birth, education, conditions of admission, etc.).

297. For the purposes of this chapter, the term "migrant" will be used generally for the set of indicator topics and analyses. However, some indicators may only apply to certain subpopulations. For example, not all migrants would be in scope for indicators related to changes in legal status or citizenship. It is important to consider the populations of interest for each indicator set, as applicable. In examples, specific populations of migrants are stated clearly.

298. Each indicator will be described and issues associated with producing the indicator will be examined. Practical examples will be provided including cases where the practical indicator deviates from the desired indicator due to data source limitations. The purpose of this chapter is to provide examples of descriptive statistics which can be used as regular longitudinal indicators. More complex longitudinal analytical methods such as survival analysis or generalized estimating equations are not explored in this chapter but could be used to yield more insights on the longitudinal outcomes of migrants. The indicator sets identified in this chapter could be used as outcome variables using these methods.

299. It is important to consider the general limitations of the data sources available when considering which indicators can be estimated. Chapter 2 outlines some of the strengths and limitations of surveys, administrative data and registers while Chapter 3 goes into more detail about limitations which can exist for integrated data sources.

300. One overarching challenge is the temporal measure necessary for longitudinal indicators. If data are not observed in real-time but instead in periodic (e.g. annual) segments, time cannot be measured as precisely. Another complication could exist in determining when the clock starts – is it after the migrant arrives for the first time, after they become permanent residents, etc.? These issues need to be addressed for all of these indicators.

301. The longitudinal indicators can be classified into three major categories:

- migration patterns
- socio-economic outcomes
- family migration

302. Under each indicator, a table identifies countries who present specific examples of these statistics with their existing data, on a longitudinal or cross-sectional basis. Notably, indicator availability is not necessarily associated with regular publication of these statistics. While many countries may have available data, these indicators may not be produced regularly but as part of a standalone or a one-time publication, for example.

Table 6 Key longitudinal indicators

Category	Topic	Indicator
Migration patterns	Length of stay in country	Proportion of migrants still present in the country after X amount of time
		Average time until migrants leave the country
		(Cumulative) Length of stay since first arrival
		Length of stay since last arrival
		Average length of stay (when more than one stay per capita)
		Proportion of migrants who remain in the subnational geography after X amount of time
		Average time until migrants leave the subnational geography
	Length of stay in the subnational geography	Among those who remain in the country, proportion of migrants who remain in the subnational geography after X amount of time
		Among those who remain in the country, average time until migrants leave the subnational geography
	Time until change in residence or legal status, including citizenship	Proportion of temporary or short-term residents who become permanent or long-term residents after X amount of time
		Average time until temporary or short-term residents become permanent or long-term residents
		Proportion of asylum seekers who are admitted to settle permanently after X amount of time
		Average time until asylum seekers are admitted to settle permanently
		Proportion of foreign-born migrants who acquire national citizenship after X amount of time
		Average time until foreign-born migrants acquire national citizenship
	Circular migration indicators	Average number of stays per 'circular migrant' (to be further defined, as for censored observation, period of time, required minimum lengths of stay, etc.)
		Average length of stay per 'circular migrant'
Socio-economic outcomes	Language skill	Proportion of migrants who have ability to speak country's official languages after X amount of time
		Average time until migrants have ability to speak country's official languages
		Proportion of migrants who speak country's official language(s) at home after X amount of time
		Average time until migrants speak country's official language(s) at home
	Home ownership	Proportion of migrants who own their home after X amount of time
		Average time until migrants own their home for the first time in the new country of residence
		Housing tenure (e.g. rent vs. own) of migrants over time
		Proportion of migrants living in government subsidized or funded housing after X amount of time
		Ownership by type of residential property (e.g., apartment, semi-detached house, etc.) after X amount of time
	Employment	Proportion of migrants who are employed after X amount of time
		Average time until migrants become employed for the first time in the new country of residence
		Full-time vs. part-time employment of migrants over time

Category	Topic	Indicator
Socio-economic outcomes		Labour force status of migrants over time
	Post-secondary education	Proportion of migrants who obtained a post-secondary certificate, degree or diploma in their new country of residence after X amount of time by age
		School attendance of migrants over time by age
	Income	Average and median total income after X amount of time
		Average and median employment income after X amount of time
		Proportion of migrants with employment income after X amount of time
		Proportion of migrants with self-employment or business income after X amount of time
		Proportion of migrants with social or government assistance after X amount of time
		Time until average or median total income for migrants equals general population average or median total income
		Time until average or median employment income for migrants equals general population average or median total income
		Proportion of migrants in low income after X amount of time
	Business ownership / entrepreneurship	Proportion of migrants who are self-employed or own a business after X amount of time
		Average time until migrants are self-employed or own a business
		Number of individuals employed by migrants after X amount of time
		Similar to the indicators on labour force status, taxation records can provide a proxy for entrepreneurship by providing the proportion of migrants with self-employment or business income. Otherwise, administrative records on business ownership can be used.
		Connecting migrant entrepreneurs with information on their employees may be more complex. However, if administrative records exist connecting workers with firms and owners with firms, this analysis is also possible.
		The Canadian Employer-Employee Dynamics Database (CEEDD) is a linkage environment based on 12 administrative source-files, including the Longitudinal Immigration Database (IMDB), and includes individual, firm, and business-owner characteristics. Being a linkage environment, users must integrate the various components to fit their analytical requirements using anonymized unique identifiers on each file. The CEEDD enables analyses of firms and individuals over time, answering questions about, for instance, initial firm allocation and earnings growth, immigrant business ownership, and the trajectories of immigrant-owned firms.
	Health	Proportion of migrants who are registered with a physician within host country's health care system after X amount of time (or time to registration)
		Proportion of migrants who are hospitalized in the X years since arrival to host country of residence, overall or for various select health conditions, e.g., chronic or infectious in nature (or time to first hospitalization)
		Proportion of migrants deceased since arrival to host country of residence for various disease conditions (after X amount of time) (or time to death)

Category	Topic	Indicator
Family migration		Average time between arrival of first family member in the country and arrival of the last family member (i.e., time lag in family reunification)
		Economic outcomes of migrant families after X time in country

4.1.1 Migration patterns

303. Longitudinal indicators related to migration patterns can include indicators related to emigration, circular migration, internal migration, and changes in legal or residence status. For circular migration indicators, readers are encouraged to refer to the UNECE publication "Defining and measuring circular migration" (UNECE 2016) and the subsequent developments by Eurostat (UNECE, 2017, 2018a).

304. The following were identified as key longitudinal indicator topics of interest to UNECE member countries related to migration patterns:

- Length of stay in country
- Length of stay in the subnational geography
- Time until change in residence or legal status, including citizenship
- Circular migration indicators.

305. Table 7 shows examples provided by country on those topics.

Table 7: Key indicator topics for migration patterns with country examples

Key indicator topics	Canada	Ger-many	Italy	Spain	Switzer-land	Turkey	United King-dom
Length of stay in country	x			x	x	x	x
Length of stay in the subnational geography	x	x	x				
Time until change in residence or legal status, including citizenship	x	x	x		x		x
Circular migration indicators*							

*Please refer to UNECE (2016) and Eurostat (2017 and 2018).

4.1.1.1 Length of stay in country

306. Ultimately, international migration patterns can be complex and individuals can enter and leave a country multiple times. Tracking the many possible patterns is necessarily longitudinal. This section focuses on simple cases of singular measurements (e.g. length of time since last arrival vs. length of time since first arrival). More complex patterns are covered the existing reports prepared by UNECE and Eurostat (see section 4.1.1.4, Circular migration longitudinal statistics, for more details).

307. The following indicators are proposed for measuring length of stay in country:

- Proportion of migrants still present in the country after X amount of time
- Average time until migrants leave the country
- (Cumulative) Length of stay since first arrival
- Length of stay since last arrival
- Average length of stay (when more than one stay per capita)

308. These indicators have large implications for understanding migration patterns and their effects on host countries. They can be used to understand everything from current stocks to factors affecting the retention of migrants. Since residence can be related to different legislation and policies (e.g. qualification for citizenship), it is important for countries to consider the complex nature of these measures. Depending on the need, the indicator of most interest may be time since first or time since most recent arrival. There may also be interest in short-term but repeated stays.

309. In this context, it is also relevant to know whether migrants return to their initial country of origin or whether they move to a third country. Country of origin, which could refer to their country of birth or country of last residence, could be a variable cross-tabulated by these indicators to add such information. The countries of destination of emigrants yield valuable information on the demographic exchanges between countries. Similarly, country of origin could inform on the length of stay before emigration (e.g. arrival directly from their country of origin or via another intermediary country).

310. These are indicators which are fundamentally longitudinal in that the place of residence (inside or outside the country) of individuals needs to be observed over time. However, if changes in place of residence are not observable in real-time, there are some limits to the development of these indicators.

311. Even among countries with population registers, there can be a lag between an out-migration and its detection in the register – this can lead to an overestimate of the length of time individuals stay in the country. Moreover, in some cases, the departure might not be captured at all. For example, Statistics Netherlands (Prins, 2016) reports that: "[…] about one in three residents who leave the country to live abroad do not notify the local government of their departure."

312. For countries relying on annual administrative data, the absence of information specifically designed to measure emigration reduces the precision with which the indicator can be produced. For instance, the absence of a specific date of emigration on annual administrative data could result in annual indicators as opposed to a more precise measure of time (e.g. average number of years vs. average number of days until migrants leave the country). In many cases, administrative data rely on concepts that might not truly reflect an emigration as defined by demographers. They can also be limited by the usual flaws of administrative data such as population coverage and lags.

313. Another limitation for countries without a population register is that this measure would likely be restricted to the presence and absence of signals. In these cases, there can be some measurement error where the absence of signals is not indicative of a true absence (and in some cases the presence of signals may not reflect a true presence). Moreover, in some cases, different signals could give contradictory information. Death is a potentially confounding factor as it could also result in the absence of signals if the data cannot fully account for this event. For example, in the Canadian Longitudinal Immigration Database, a migrant who stopped filing taxes may signal that this individual has died, has emigrated, or continued to live in the country but simply didn't file their taxes. It should also be noted that technological development allows ever-increasing possibilities to be connected remotely, without necessarily being actually present in the geographic area of interest.

This may reduce the relevance of signals of administrative life as an indicator of actual presence over time.

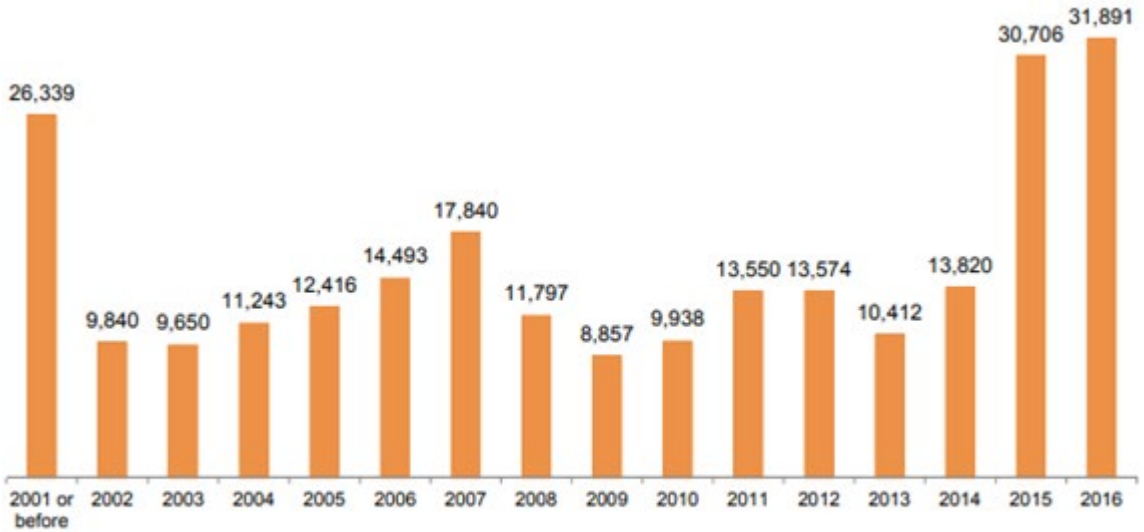
Spain publishes indicators related to this topic, including:

- Emigrants by year, year of arrival in Spain and citizenship
- Emigrants by year, year of arrival in Spain and country of birth

The main source of information pertaining to both population stocks and migration statistics in Spain is the population register, named Padrón in Spanish. Padrón is the official list of residents in each one of the 8,124 municipalities (as of January 1st, 2018) in Spain. But Padrón is also a longitudinal database, including 46.7 million people living in Spain (as of January 1st, 2018) and more than 265 million records with longitudinal information for individuals residing in Spain from 1996 to 2018. All previous places of residence (within Spain) are also captured, thus allowing longitudinal analysis and a very precise monitoring of internal migration. Although the potential of Padrón for longitudinal analysis is great, at this point the Spanish National Statistics Institute (INE) is still taking its first steps.

In **Spain**, migration statistics disseminated by INE since 2013 include information about the number of immigrants and emigrants disaggregated by age / generation, sex, nationality, country of birth, country of origin, and country of destination. For the first time in June 2019, information about the length of stay in Spain was disseminated for those who had already left the country. This information helps enrich information pertaining to emigrants. The graphic below illustrates the distribution of 2018 emigrants by year of arrival.

Figure 11 Year of arrival in Spain of emigrants in 2018



Provisional data.
Data on the year of arrival 2017 are not relevant for 2018 emigrations, because departures from the country are only considered as emigrations when the arrival in the country occurred at least 12 months before.

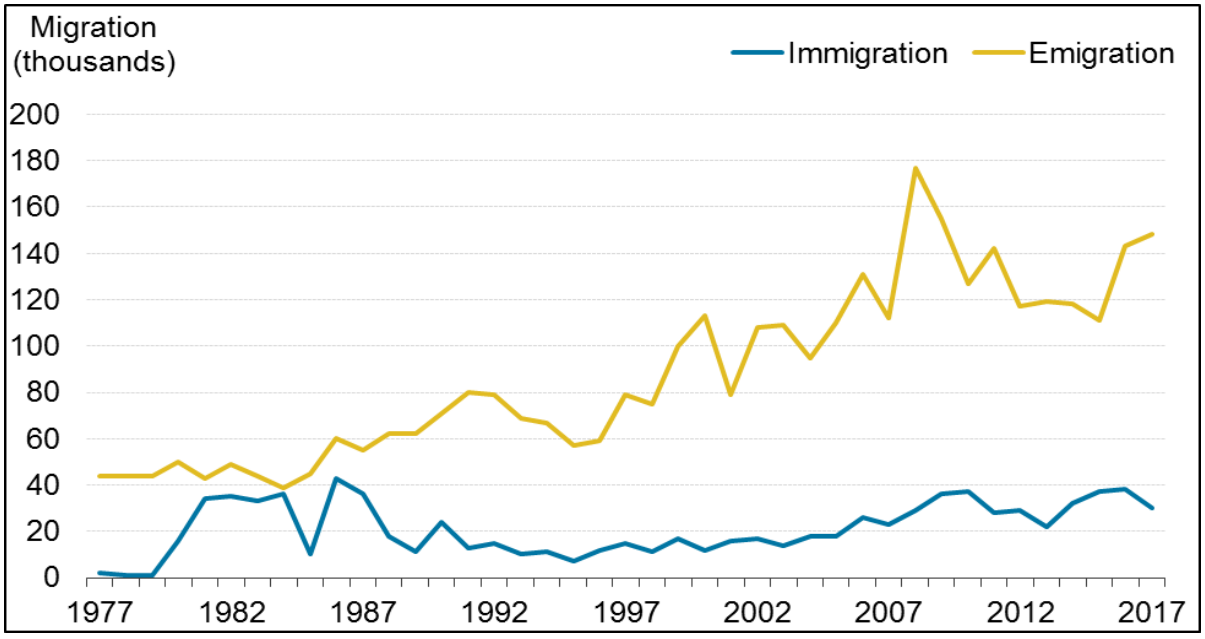
Source: Spanish National Statistics Institute (INE), Padrón, 2018.

314. More details on these indicators are available in UNECE (2018b).

Currently, the **United Kingdom** publishes data on stocks and flows of international migrants. There is no single source that can provide a measure of all movements of people into and out of the UK. The UK currently use a combination of data from different sources to calculate the official UK Long-Term International Migration (LTIM) estimates, which are published every quarter in the Office for National Statistics Migration Statistics Quarterly Report. Estimates of LTIM are about 90 per cent based on data from the International Passenger Survey (IPS), a sample survey carried out at all main UK ports. The IPS captures migration intentions and is used to identify international migrants entering and leaving the UK. To estimate LTIM, the IPS data are supplemented by Home Office and Northern Ireland Statistics and Research Agency (NISRA) data and several adjustments are made to provide a more complete estimate of migration.

Observed duration of migration for people entering or leaving the UK are also published. These estimates are based on the IPS. Actual length of stay is established by asking a retrospective question on the IPS for either year of departure from the UK (for those returning immigrants who had previously left as a long-term emigrant) or year of arrival for those leaving the UK long-term. Actual length of stay is calculated by a simple subtraction of year of arrival/departure from the year of interview (Figure 12).

Figure 12 Estimates of long-term international migration into and out of the UK, 1975-2017. Non-British, actual length of stay 1-4 years.



Source: UK Office for National Statistics, International Passenger Survey, 1975-2017(2019).

National estimates of short-term international migration are produced directly from data from the International Passenger Survey (IPS) and published annually in the Short-Term International Migration (STIM) for England and Wales bulletin. Estimates of STIM are produced for England and Wales and for local authorities. Estimates of short-term migration are available as flows (the total number of moves made over a set period) and stocks (the average number of short-term migrants in the country on an average day in a 12-month period). It should be noted that a person could migrate more than once in the same period so the data are a count of migrant moves not individuals. The UK also publish data on UK residents by country of birth and nationality as reported in the Annual

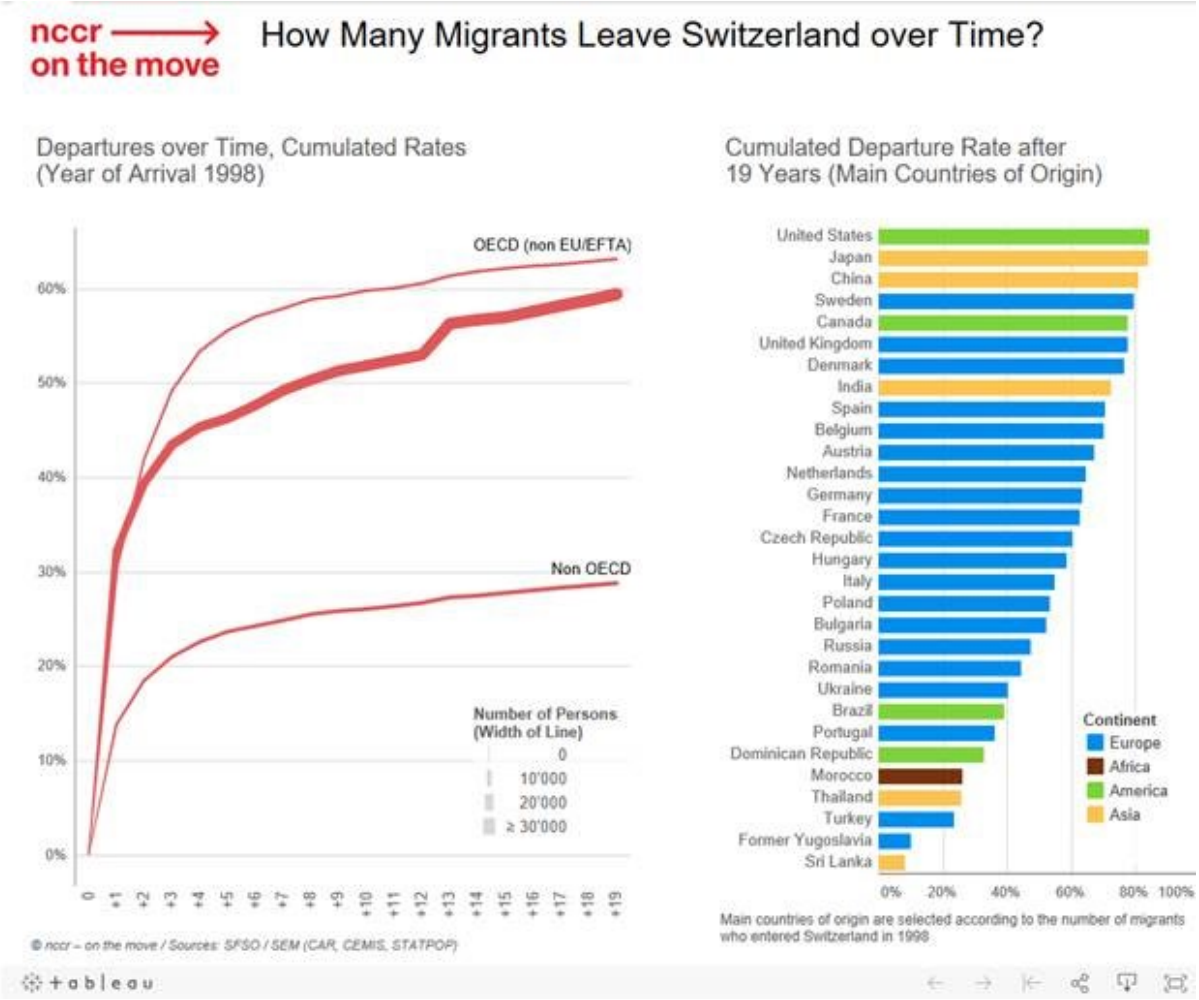
Population Survey (APS). Residents may have recently arrived or been resident in the UK for a number of years and are a count of the stock of the non-UK population.

Going forward, the UK Office for National Statistics intends to combine surveys and administrative data to form longitudinal datasets and understand the footprint migrants have while they are in the country, for instance by measuring the duration of stay and identifying return or circular migration of foreign seasonal workers.

Switzerland has an elaborate set of indicators with a focus on external migration. These are based on the Swiss Longitudinal Demographic Database (SLDD), a linked register-based database recently developed in the context of the national scientific project NCCR. On the Move, where NCCR is the National Centre of Competence in Research for migration and mobility studies. It is constructed by linking the Central Register of Foreign Nationals (ZAR) and the Population and Household Statistics (STATPOP) datasets. It covers all Swiss citizens and foreigners living in Switzerland.

The first analysis based on the SLDD emphasized the fact that the migrant population in Switzerland is quite mobile, as migrants do not usually stay in Switzerland their whole life, but only for a few years. For example, of the immigrants who arrived in Switzerland in 1998, over 80 per cent left the country within 17 years of their arrival. As the short-term migrants might be less willing to integrate into the destination country, this finding is relevant to assess in light of the current politics of integration. Switzerland publishes departure rates over time in an interactive graph through the On the Move partnership. The graph below demonstrates that length of stay, or the departure rate, varies by country of origin.

Figure 13 Cumulative departure rates for migrants who arrived in Switzerland in 1998 by years since arrival and country of origin



Source: Swiss Federal Statistical Office, Swiss Longitudinal Demographic Database (SLDD), 2017.

More information is available:

https://nccr-onthemove.ch/publications/towards-a-new-data-set-for-the-analysis-of-migration-and-integration-in-switzerland/

http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2017/mtg1/2017_UNECE_Migration_WP_07_Switzerland_WannerHeiniger_ENG.pdf

Canada's Demographic Estimates Program requires estimates of the stocks and flows of non-permanent residents (NPRs – also called temporary residents) in order to estimate the population of the country between censuses. To do so, Statistics Canada uses monthly information on administrative transactions (landing records, asylum claims, work permits, study permits, removals and applications for landing) from Immigration, Refugees and Citizenship Canada (IRCC).

Statistics Canada integrates these files and builds longitudinal profiles of NPRs. Using some basic assumptions, biographies from longitudinal data can be used to measure the administrative length of stay of each NPR in order to produce estimates of stocks and flows. It also takes into account if the person holds more than one permit in the same period.

These data has two main limitations. First, similar to the administrative immigration data, NPR data collects the intended place of destination rather than the place where the immigrant actually settles. Second, NPR data collects only the starting and ending date of the permit, not the actual arrival and departure dates of the NPR. As a result, the length of stay in the country is approximated using the duration of the permit(s) held by the NPRs. Moreover, there is a lack of information on the departure dates of NPRs as some of them leave the country before the end of the validity of their permits. Since it is not mandatory to inform the federal government before leaving the country, Statistics Canada does not currently have any additional sources to obtain this information.

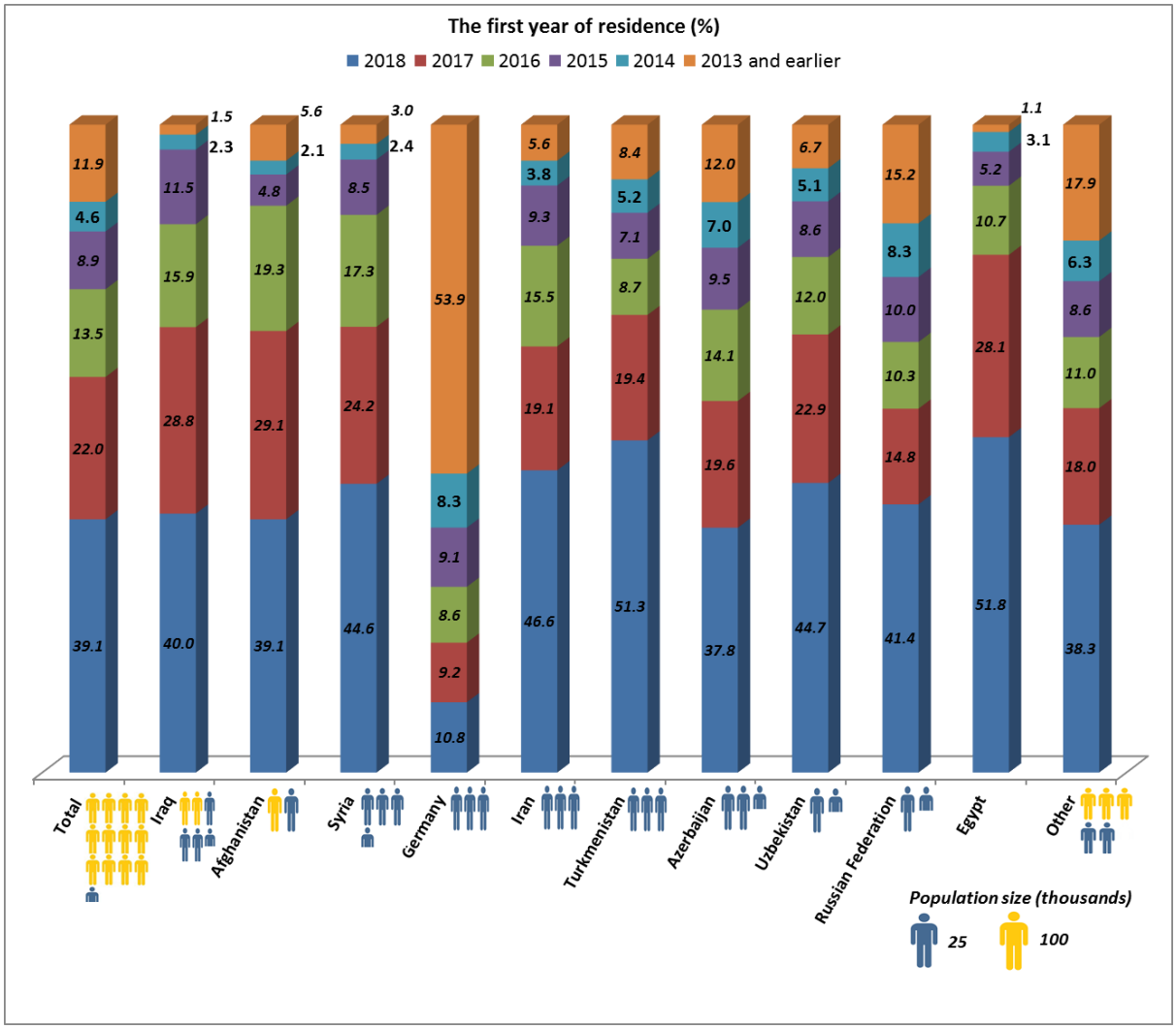
In **Turkey**, Address Based Population Registration System (ABPRS) is the unique system from which annual population statistics are disseminated. Within this scope, annual information on international migration stocks and flows by citizenship and country of birth are produced based on the ABPRS and other additional administrative sources.

In 2019, within the scope of studies on development of statistics obtained from administrative registers, "length of stay of foreign residents" study was completed and “foreign population by the first year of residence in Turkey” indicator was derived. The coverage was the foreign population registered in the ABPRS and "the first year of residence" here refers to the year that a foreign resident started his/her continuous residence until the reference date, i.e. 31 December 2018 for the first study.

The information was produced based on the retrospective follow-up and analysis of the foreign population. In the analysis, not only the last but also the previous years of ABPRS were used along with the international migration data and residence/work permit records of related years.

More details on the indicator are available on http://www.turkstat.gov.tr/PreTablo.do?alt_id=1067.

Figure 14 Foreign population by country of citizenship and the first year of residence in Turkey, as of December 31, 2018



Source: TurkStat, Address Based Population Registration System (ABPRS), 2013-2018.

4.1.1.2 Length of stay in the subnational geography

315. Similar to length of stay in the country, the indicators proposed for measuring length of stay in the subnational geography are the following:

- Proportion of migrants who remain in the subnational geography after X amount of time
- Average time until migrants leave the subnational geography
- Among those who remain in the country, proportion of migrants who remain in the subnational geography after X amount of time
- Among those who remain in the country, average time until migrants leave the subnational geography

316. Additionally, there may be some more country-specific questions, such as:

- Proportion of migrants who move from rural to urban areas after X amount of time

- Proportion of migrants who move from eastern to western regions after X amount of time

317. For indicators needing to take international moves into account, the issues raised above for "Length of stay in country" apply.

318. Otherwise, subnational migration patterns present some new (or exacerbated) limitations pertaining to geography. Geographic boundary changes pose an issue to longitudinal analysis of migration. When a boundary changes, individuals can appear as movers even if they remained in the same dwelling. This is because their official geographic area may have changed as a result of the boundary change. While this can also be an issue for international migration, internal boundary changes can be far more common.

319. Another issue could arise from combining different sources of geographic information. If the longitudinal data is the result of the integration of multiple data sources, it's possible some sources use different geographic definitions than others. City names can change from one source to the other and sources can provide geographic information that cannot be compared. For example, school or health districts could overlap with postal codes or city boundaries. A metropolitan area and a municipality could have a common name but refer to different geographical boundaries (e.g. the census metropolitan area of Toronto, Canada had a 2016 Census population count of 5,928,040 but the census subdivision (city) of the same name had a population of 2,731,571).

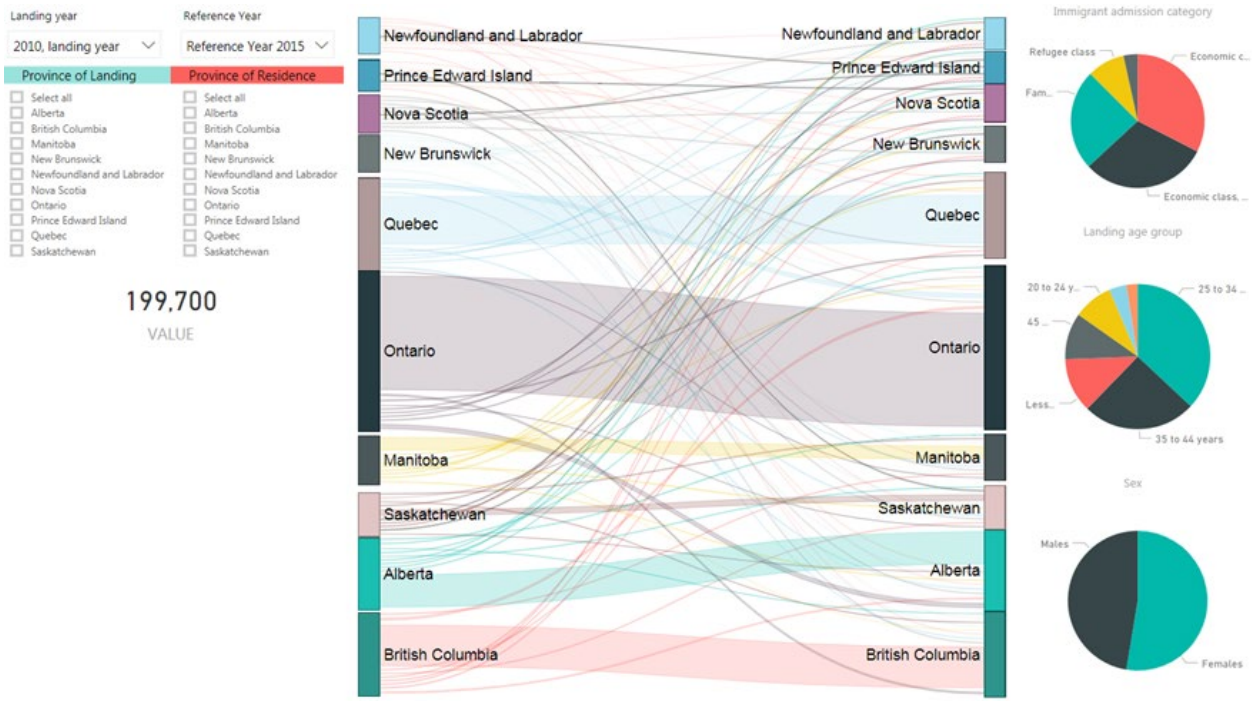
320. It can be important to consider movements within geographic boundaries as well. In some cases, especially for larger areas, migration within the geographic area could be a movement over a long distance. On the other hand, migration from one area to another could be achieved over a short distance when occurring near the areas' boundary. Longer distance moves would likely have a more profound impact on the migrants as they could relocate to communities with different labour markets, different services available, and different social contexts.

Canada regularly produces statistics on the internal migration patterns of immigrants based on the Longitudinal Immigration Database (IMDB). The IMDB is an integration of immigration administrative data and annual tax files. The annual tax files provide the current geography of immigrants by year and the immigration files provide the geography of destination when the immigrant was admitted to Canada.

One issue specific to Canada is that, in the immigration administrative data, immigrants are asked where they intend to reside. This is then compared with their actual geography of residence according to their annual tax files. Differences could be due to real movements between admission to Canada and filing taxes but they could also be due to conceptual differences (i.e. intended province of residence vs. first actual province of residence).

Despite these challenges, the IMDB is a key data source for understanding retention of immigrants by province or sub-provincial geographic area. The Sankey diagram below illustrates the shifts between intended provinces of residence for immigrant taxfilers admitted in 2010 versus their actual province of residence 5 years later, in 2015. This chart also allows users to understand the characteristics of those individuals who stayed in their intended province of residence compared with those who now live elsewhere in Canada.

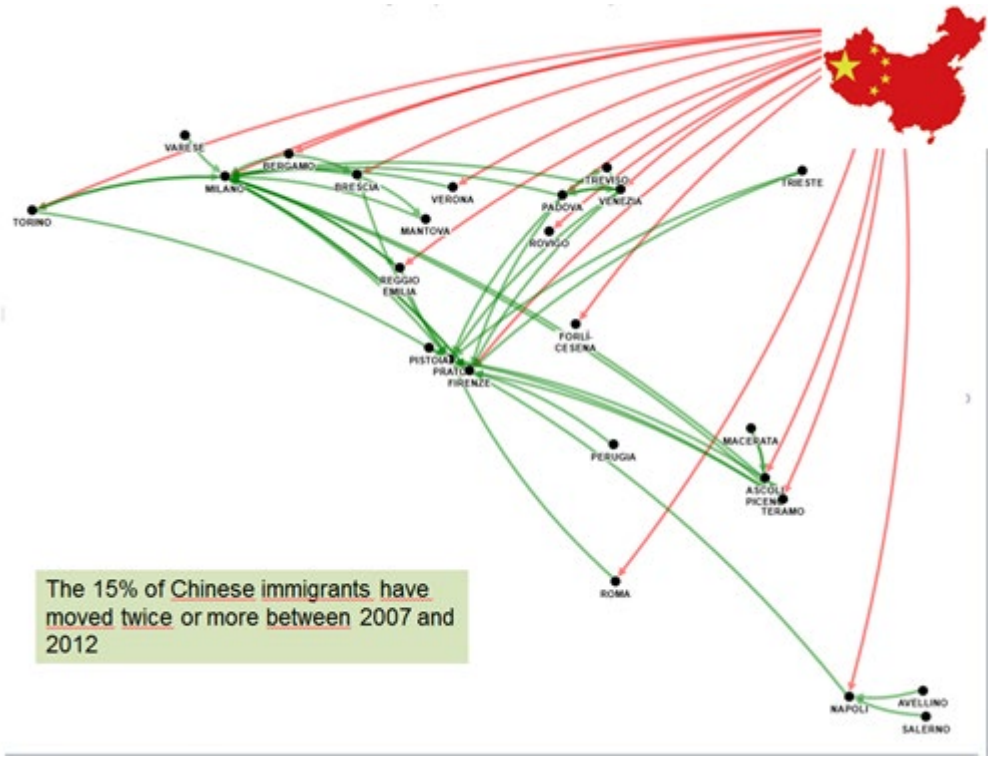
Figure 15 Shifts between intended province and province of residence 5 years after admission to Canada, immigrant taxfilers admitted in 2010



Source: Statistics Canada, Longitudinal Immigration Database, 2016.

In **Italy,** record-linkage between residence permits allows to connect international migration to internal mobility. The figure below shows an example of the Chinese migration network, both external and internal. The red lines indicate arrivals from China to Italian provinces in 2007, and the green lines represent internal mobility between 2007 and 2012.

Figure 16 Migration network for Chinese migrants who arrived in Italy in 2007 with internal mobility between 2007 and 2012



Source: ISTAT, Conti et al (2012).

Germany currently monitors annual components of change in population stocks (including internal migration). There are plans to widen the scope by looking at these factors from a longitudinal perspective. The data set will allow the monitoring of post-immigration internal mobility.

Preliminary results are based on foreigners who first migrated to Germany after 2006 or later and stayed as residents continuously until end of 2017. Post-migration internal mobility patterns were analyzed at the level of 401 counties. The following potential indicators for post-migration internal mobility were used:

- the number of moves across county borders in a defined period (by length of stay),
- the absolute regional length of stay (years spent in the same county), and the relative regional length of stay (years spent in the same county as a percentage of the overall length of stay in Germany).

Germany considers the *number of moves across county borders* as a good indicator of internal migration as long as it is applied to foreigners with a common total length of stay (in Germany). Furthermore, it considers *the relative regional length of stay* as the most flexible indicator, as it can be meaningfully interpreted even if the total length of stay differs.

The preliminary results are promising: The *relative regional length of stay* varies with the **characteristics of the foreigner** – in particular, gender, age, marital status, citizenship and residence permit status. Length of stay is longest among older married women holding the citizenship of an EU-member state and, thus, EU freedom of settlement as a permanent residence permit status.

The *relative regional length of stay* also varies by **characteristics of the region**. It is longest in regions with a high population density, a high percentage share of foreigners, and high values for GDP per

capita; it is shortest in regions with low population density, a low percentage share of foreigners and low values for GDP per capita.

4.1.1.3 Time until change in legal status

321. Typical examples of change in a migrant's legal status include acquiring the status of a permanent resident, an acceptance of an asylum seeker's claim or obtaining citizenship. The following indicators are proposed for longitudinal measurement:

- Proportion of temporary or short-term residents who become permanent or long-term residents after X amount of time
- Average time until temporary or short-term residents become permanent or long-term residents
- Proportion of asylum seekers who are admitted to settle permanently after X amount of time
- Average time until asylum seekers are admitted to settle permanently
- Proportion of foreign-born migrants who acquire national citizenship after X amount of time
- Average time until foreign-born migrants acquire national citizenship

322. If the administrative data available connects the relevant information (e.g. temporary and permanent resident records), these indicators should be fairly straightforward to produce.

323. Temporary residents can introduce some complexity for the time dimension as it may not always be evident when their residence in the country began. Another issue could be when there are international moves between an individual's first temporary residence in the country and when they become permanent or long-term residents (e.g. someone comes as a temporary student, leaves the country, returns a few years later as temporary worker, then becomes a permanent resident). As such, decisions would need to be made about when the time would start for these indicators.

324. This section also includes indicators pertaining to naturalization. For the purpose of this section, naturalization refers to the process by which a foreign-born migrant, that is someone who is either born as a foreign citizen or stateless, becomes a citizen of their new country of residence. Naturalization rates refer to the proportion of foreign-born migrants who have become citizens. The denominators for such rates may vary depending on whether they include or exclude those individuals who are not eligible for citizenship.

325. Depending on the country, not all foreign-born migrants may have the ability to become citizens. For example, in some cases, a minimum number of years of residence may be required. As a result, the calculation of these indicators can be restricted to those migrants in-scope. For example, the proportion of migrants who become citizens would be a lower proportion than the same calculation but restricted to those who were eligible to become citizens. Analysts should pay particular attention to the legal arrangements surrounding naturalization in the country of interest, to carefully identify migrants who are eligible for citizenship.

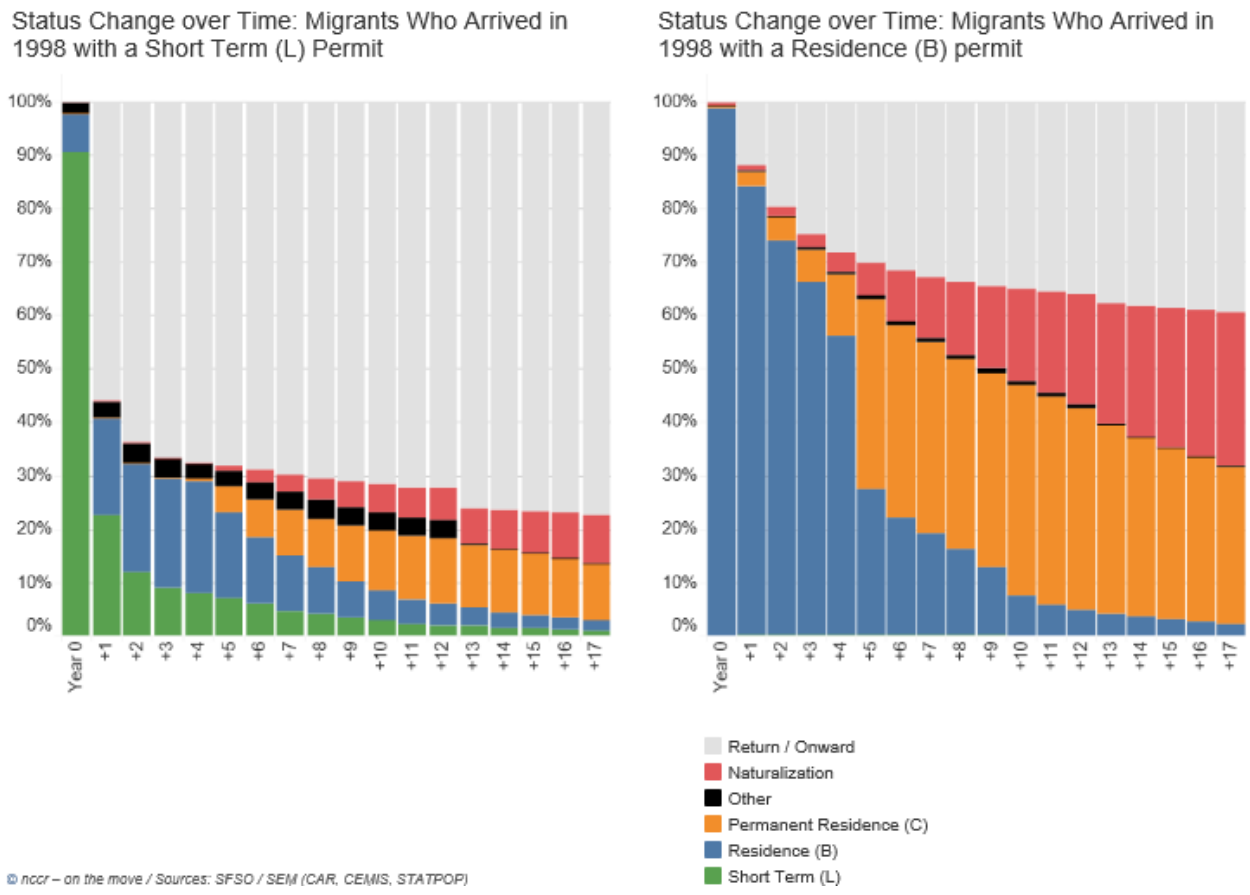
326. Similar to other indicators, if the events are not observed in real-time but instead in annual summaries, the time dimension would have to change from specific dates to years and result in less precision (e.g. number of years until temporary residents become permanent residents, number of years until permanent residents become citizens).

327. The need for a cohort approach to study naturalization has been emphasized for some time (Perrin, 2006). This perspective has recently been proposed once more by Reichel (2011), who stressed that: “to do so, it would be necessary to base the rate on the foreign population actually eligible for naturalization, or as statisticians call it, the population at risk of experiencing an event”. As pointed out by Perrin (2006), “a cohort approach towards measuring rates of citizenship acquisition would allow the calculation of the likelihood of obtaining citizenship for individual cohorts of immigrants which would provide a much better measure of the impact of policy measures on patterns of citizenship acquisition.” The same author also points out that the lack of appropriate data makes it difficult to calculate longitudinal naturalization rates, which require the availability of information on the year of immigration. This is, however, the direction in which several countries are moving, also due to the availability of statistical sources based on registers that allow the extraction of longitudinal data (Perrin, 2006).

Switzerland, through the Swiss Longitudinal Demographic Database (SLDD), has indicators related to how migrants’ legal status evolves over time connected to administrative variables pertaining to migrants’ basis for admission to Switzerland.

Further to the charts below, for more information: https://indicators.nccr-onthemove.ch/how-does-the-migrants-legal-status-evolve-over-time/

Figure 17 Evolution in legal status of migrants arriving in Switzerland in 1998 with either a short term (L) or residence (B) permit, up to 2017.



Source: Swiss Federal Statistical Office, Swiss Longitudinal Demographic Database (SLDD), 2017.

Switzerland also takes special interest in monitoring naturalization rates and has several indicators on naturalization.

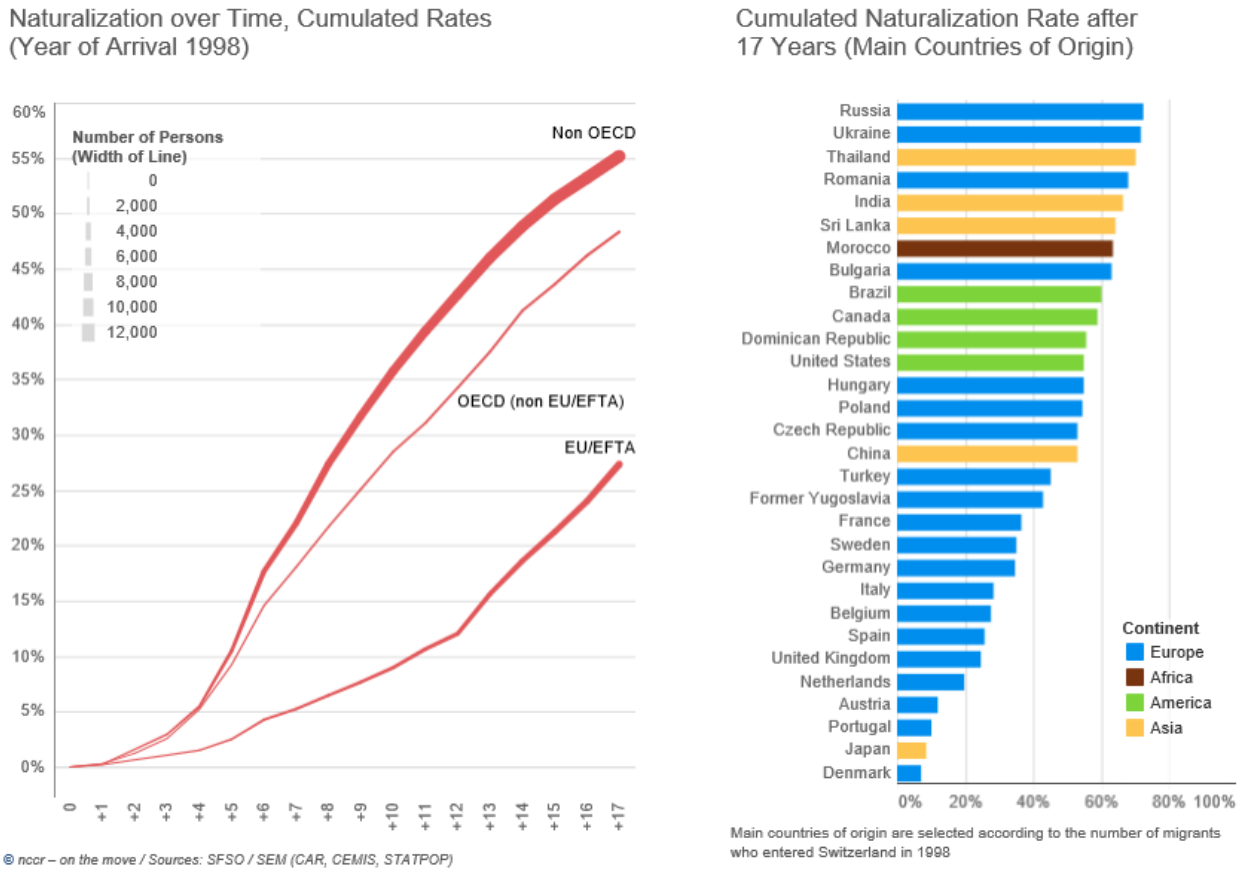
The following figures presents naturalization rates by countries of origin, canton of residence and over time, from 1998 to 2017.

For more information:

http://nccr-onthemove.ch/DataManagement/Visualization/Embed/Naturalization_Rates.html

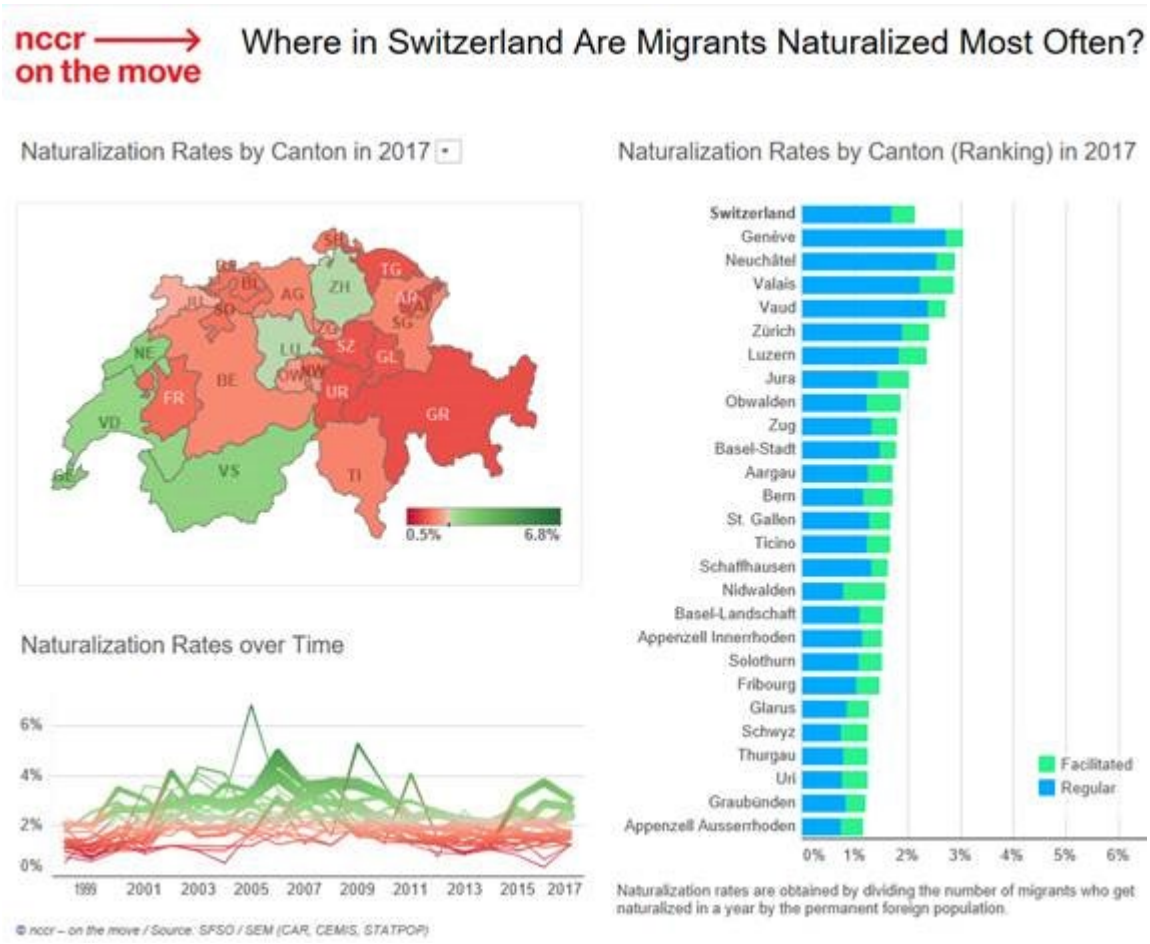
https://indicators.nccr-onthemove.ch/where-in-switzerland-are-migrants-naturalized-most-often/

Figure 18 Cumulative naturalization rates for migrants who arrived in Switzerland in 1998 by years since arrival and country of origin



Source: Swiss Federal Statistical Office, Swiss Longitudinal Demographic Database (SLDD), 2017.

Figure 19 Naturalization rates in 2017 for migrants to Switzerland by canton of residence distinguishing between regular and facilitated naturalizations



Source: Swiss Federal Statistical Office, Swiss Longitudinal Demographic Database (SLDD), 2017.

In **Italy,** the share of immigrants holding permanent or long-term residence permits is an indicator currently calculated on the basis of administrative permit-data. One method of calculation is cross-sectional. Table 8 compares two ways of calculating this indicator, by citizenship of select migrants. Column 1 divides the number of long-term permit holders by the total number of residence permits currently valid for the same period. In column 2, the denominator is restricted to the population present for more than 5 years (those eligible to apply for a long term residence permit in Italy).

Table 8 Long-term residence permits as a proportion of (1) total residence permits and (2) permits valid for more than 5 years, selected citizenships, Italy, as of January 1st, 2018.

Citizenship	1 Long-term residents as a % of total residence permits	2 Long-term residents as a % of residence permits issued before 2013
Morocco	70.3	84,1

Albania	71.6	84,6
China	56.0	63,7
Ukraine	72.3	86,1
Philippines	62.2	72,5
India	59.0	78,5
Egypt	64.6	87,2
Bangladesh	54.9	85,4
Moldova	75.6	83,8
Pakistan	49.8	83,2
Total	**61.7**	**80.0**

Table 9 illustrates the longitudinal approach from Italy. It follows two "cohorts" of new permits for non-EU citizens, issued in 2011 and 2012. For both cohort year, it shows the absolute numbers of permits issued, the percentage that were still present in Italy in January 2018, and, among those still present, the percentage who have become long-term residents. The table shows different patterns of stay and conversion to long-term residence status in Italy among migrants holding different citizenships.

Table 9 **New resident permits issued in 2011 and 2012, percentage of migrants still present in Italy at the beginning of 2018, and percentage of long-term residents among those present, by country of citizenship**

Citizenship	Year of first permit 2011			Year of first permit 2012			First permits 2011 and 2012		
	Total	% still present 1th January 2018	% long term resident of those still present	Total	% still present 1th January 2018	% long term resident of those still present	Total	% still present 1th January 2018	% long term resident of those still present
China	26,850	47.8	21.1	25,153	47.7	11.2	52,003	47.7	16.3
Morocco	30,516	48.7	52.3	21,146	45.2	36.0	51,662	47.3	45.9
Albania	24,621	50.2	48.0	18,608	48.8	32.3	43,229	49.6	41.4
India	18,371	49.1	48.1	11,666	47.0	26.2	30,037	48.3	39.8
United States	14,402	4.4	39.4	14,329	4.5	29.2	28,731	4.5	34.2
Tunisia	19,454	19.6	40.4	6,364	33.8	36.4	25,818	23.1	39.0
Moldova	16,463	38.4	51.4	8,779	35.8	40.3	25,242	37.5	47.7
Ukraine	15,595	53.1	54.3	8,596	50.8	42.3	24,191	52.3	50.2
Bangladesh	13,743	56.0	56.3	9,191	54.5	28.6	22,934	55.4	45.4
Egypt	13,082	20.6	46.5	9,834	24.7	20.9	22,916	22.3	34.4
Philippine	12,987	48.7	23.9	8,804	49.3	11.7	21,791	49.0	19.0
Pakistan	9,971	43.0	47.1	9,541	39.3	26.5	19,512	41.2	37.5
Nigeria	11,625	31.8	27.5	7,734	32.2	19.2	19,359	32.0	24.1
Sri Lanka	8,951	52.4	33.0	6,721	51.6	16.0	15,672	52.1	25.8
Senegal	8,167	47.8	48.1	6,177	41.4	28.2	14,344	45.0	40.2
Peru	8,661	43.4	40.5	5,174	41.6	24.8	13,835	42.7	34.8
Brazil	7,154	25.8	60.5	5,626	25.0	57.7	12,780	25.4	59.2
Ghana	6,375	33.4	32.8	4,528	36.7	20.7	10,903	34.8	27.5
Russian Federation	5,493	37.8	61.0	4,506	35.3	52.3	9,999	36.7	57.2

Macedonia, Republic	4,956	35.7	54.0	3,570	36.7	35.2	8,526	36.1	46.0
Other countries	*80,336*	*27.6*	*41.6*	*64,986*	*29.5*	*30.2*	*145,322*	*28.5*	*36.3*
Total	**357,773**	**37.8**	**43.4**	**261,033**	**37.4**	**27.8**	**618,806**	**37.6**	**36.9**

The indicator for the acquisition of citizenship provided to Eurostat by Italy for Regulation (EC) No. 862/2007 is currently calculated as the ratio between the number of residents who acquired citizenship in the country during a calendar year and the total number of resident foreigners in that country at the beginning of the year. The longitudinal approach may be of particular interest in this case, especially for acquisitions that are obtained by residence.

The longitudinal approach could be of particular interest in this context since not all those who acquired citizenship have remained in Italy. To calculate naturalization rates it is necessary to have the population at risk of experiencing the naturalization event. In Italy, the availability of data sources based on registers allows longitudinal data to be extracted, linking the naturalization event with the entry and exit to and from the country.

Table 10 Foreign population and Italians by acquisition of citizenship, by citizenship or citizenship of origin as of January 1st, 2018, absolute values and percentage

Citizenship/citizenship of origin	Foreign population (a.v.)	Foreign population (%)	Italians by acquisition of citizenship (a.v.)	Italians by acquisition of citizenship (%)
Romania	1,190,091	23.1	77,046	5.7
Albania	440,465	8.6	169,644	12.6
Morocco	416,531	8.1	184,333	13.7
China	290,681	5.6	12,552	0.9
Ukraine	237,047	4.6	23,096	1.7
Philippines	167,859	3.3	16,725	1.2
India	151,791	3.0	39,360	2.9
Bangladesh	131,967	2.6	22,394	1.7
Moldova	131,814	2.6	18,654	1.4
Egypt	119,513	2.3	24,125	1.8
Other countries	*1,866,681*	*36.3*	*757,332*	*56.3*
Total	5,144,440	100.0	1,345,261	100.0

To calculate naturalization rates, individual data from the 2011 census have been linked to those who acquired citizenship after the census date. For the first time in June 2019, the Italian National Statistical Office (ISTAT) disseminated data on the stock of the 'foreigner at birth' population (Table 10, above). The statistics calculate the number of those in the 'foreigner at birth' population, available at the last traditional Census (9th October 2011), with administrative flows (acquisition of citizenship and demographic flows).

Considering that in Italy non-EU Citizens can obtain citizenship for residence after a minimum of 10 years of residence, the focus is on those who were resident at the time of the Census and already in Italy in 2006.

Between 2011 and 2017, 679,470 non-EU nationals acquired Italian citizenship. The share of people who acquired Italian citizenship varies considerably by citizenship of origin: around 30 per cent for Morocco, Albania and India and below 3 per cent for China and Ukraine (Table 11). It has to be noted

that laws and regulations could be very different in the mentioned countries, especially with regard to keeping the citizenship of the country of origin, as well as differences in geo-political conditions. For example, China and Ukraine do not recognize dual citizenship, but the same is also true for India, whose citizens are more likely to acquire Italian citizenship. At the same time, Romanian citizens, being part of the European Union, may no longer be interested in acquiring citizenship, but the citizens of Poland, still an EU country, show a higher propensity.

Table 11 Retention and associated citizenship acquisition rates by non-EU citizens who arrived in Italy before 2006 and registered in the 2011 Census, by citizenship

Citizenship	Total	% still present on January 1st, 2018	% new citizens of total	% new citizens of those still present
Albania	301,606	92.7	25.2	27.2
Morocco	232,269	88.1	24.9	28.3
China	133,143	90.4	1.7	1.9
Ukraine	120,990	90.0	2.8	3.1
Philippines	92,783	91.7	4.6	5.0
Moldova	68,219	89.6	10.5	11.8
Peru	60,629	92.3	18.0	19.6
Ecuador	59,762	89.5	11.1	12.4
India	58,497	85.9	25.4	29.7
Tunisia	50,789	83.4	14.5	17.4
Macedonia	48,417	79.2	19.9	25.1
Sri Lanka	46,281	89.2	6.5	7.4
Senegal	45,820	89.1	17.7	19.9
Bangladesh	43,151	80.0	26.2	32.9
Egypt	37,827	74.1	11.4	15.4
Pakistan	34,600	84.5	27.7	32.8
Serbia	31,954	68.6	11.7	17.1
Nigeria	27,355	84.2	14.5	17.3
Ghana	25,210	80.7	31.3	38.9
Kosovo	23,624	60.9	16.5	27.2
Other countries	*255,707*	*82.6*	*13.6*	*16.5*
Total	**1,798,633**	**87.0**	**16.2**	**18.6**

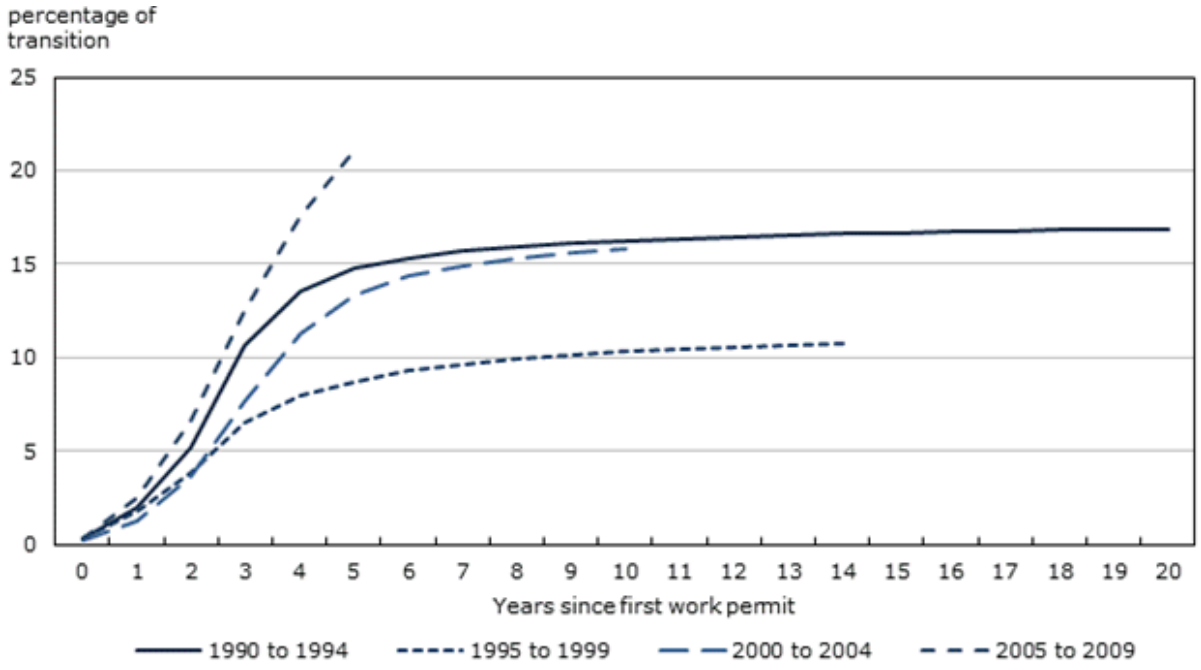
Note: citizenship acquired between 2012 and 2017

Source: ISTAT, provisional data

In **Canada**, the Longitudinal Immigration Database (IMDB), captures the dates of all temporary resident permits and the date individuals become permanent residents of Canada. This allows for an analysis of time until permanent residency. The graph below illustrates the time-varying rates at which temporary foreign workers transitioned to permanent residence based on different cohorts.

The chart shows that less than a quarter of temporary foreign workers tend to become permanent residents, but this has changed over time, with more recent workers being more likely to transition. When it does occur, this transition is more likely to take place within a few years of their first work permit in Canada.

Figure 20 Cumulative rates of transition to permanent residence among temporary foreign workers in Canada by years since first work permit and period of first work permit



Source: Statistics Canada, Longitudinal Immigration Database, 2014: Lu, Y. & Hou, F. (2017) "***Transition from Temporary Foreign Workers to Permanent Residents, 1990 to 2014***" Analytical Studies Branch Research Paper Series, Statistics Canada catalogue no. 11F0019M.

Canada recently added the date of citizenship acquisition to its Longitudinal Immigration Database (IMDB). This information is available for immigrants who have been granted citizenship since 2005. At this time, no indicators are published. This new information provides the possibility to cover several indicators including:

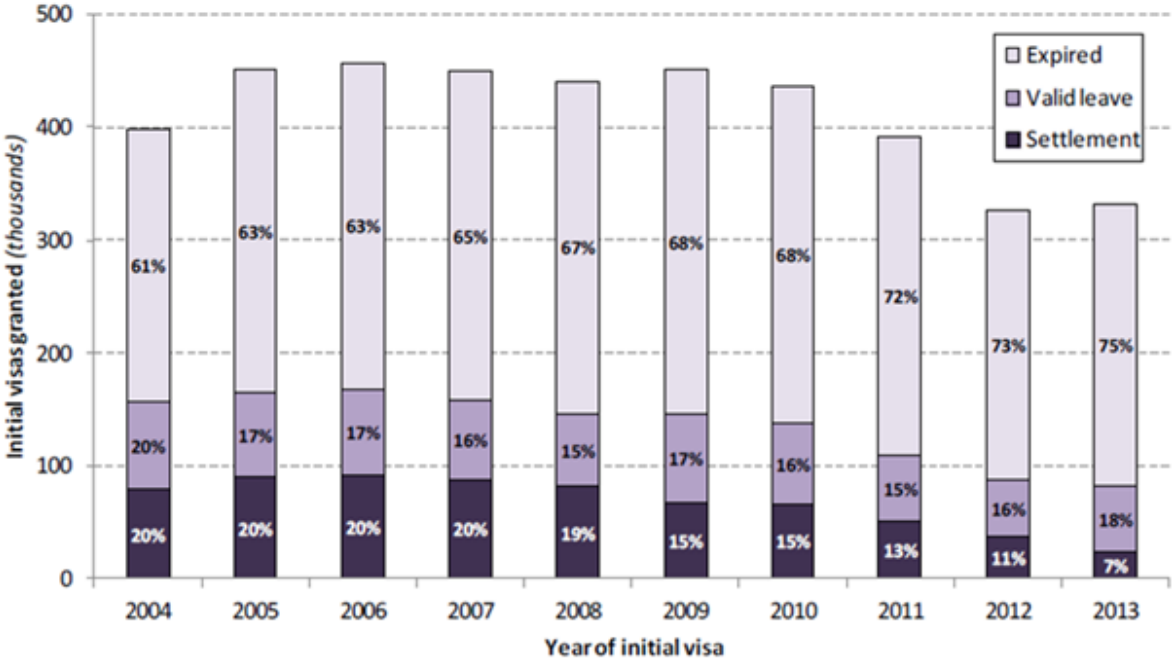
- Proportion of immigrants who become citizens by naturalization since 2005 by years since admission
- Citizenship acquisition rates since 2005 according to different key characteristics (e.g.: knowledge of country's official language(s), source country, gender)

For example, in a December 2018 article published for the release of the IMDB, wages of immigrants who acquired citizenship were compared to wages of immigrants who were admitted the same year but had not become citizens by 2016. In 2016, the median wages of male immigrants admitted in 2006, who obtained citizenship by 2016, was $40,500, while it was $31,200 for non-citizens. For female immigrants, the median wages were $28,100 if they obtained their citizenship, and $21,600 if they did not.

In the **United Kingdom**, the UK Home Office publishes an annual 'Migrant Journey' report that explores changes in migrants' visa and leave status. The report focus on journeys where there is no more than a 12-month gap between periods of leave. The latest report, covering the period up to the end of 2018, presents the data in two complementary ways. The Forward Analysis examines the immigration status of migrants issued visas to the UK between 2004 and 2016. This analysis examines the changes in immigration status by categories of entry, and excludes visitors. The latest findings discuss those issued a visa in 2013, focusing on their status after 5 years.

The backward analysis examines those whose entry to the UK ultimately resulted in them being granted settlement (permission to stay in the UK permanently) by the type of visa on which they initially came.

Figure 21 Number of migrants issued an initial visa from 2004 to 2014, by immigration status after 5 years in the United Kingdom



Source: UK Home Office (2019) Migrant Journey: 2018 Report, Statistical Bulletin 07/19.

Germany does not yet have longitudinal datasets for migrations statistics. A new longitudinal data set will allow monitoring post-immigration changes in residence status and residence permit status, transition from regular legal status to "protection status" as an asylum seeker or refugee and back as well as the exact times spent in each of these statuses.

Preliminary results are based on foreigners who first immigrated to Germany after 2006 and stayed as residents continuously until the end of 2017. Residence permit status is analyzed for citizens of EU-Member States and citizens of non-EU-Member States, separately. EU-citizens are granted EU settlement of freedom and thus, permanent residence status, from the point of immigration onward. Being granted permanent residence status is an important event for all citizens of non-EU-Member States, however. Germany measures the likelihood of being granted permanent residence status by two indicators:

- percentage of foreigners not holding permanent residence status by length of stay
- the "waiting time", that is the time span between immigration and the granting of permanent residence status (in years)

Germany considers the *waiting time* as a somewhat problematic indicator, as the value is based on those cases where permanent residence status is granted within the monitored time span. In the German case, the maximum monitoring time span was eleven years, and after these eleven years the percentage share of non-EU-citizens holding permanent residence status was only 48.1 per cent. Providing unbiased data for waiting time may well require a monitoring period of 25 years or more. The share of foreigners with permanent residence status may serve as a good substitution in earlier years.

Preliminary results of longitudinal analyses show that permanent residence status varies by age and gender. The highest rates are among men aged 45-64 years (71.6%), followed by women of the same age (69.4%). The lowest rates are among men aged 15-24 (32.3%). Men in this age group are special, however, because the majority of asylum seekers and refugees who arrived from 2015 onwards belong to this group.

Based on the issue dates of residence permits it is possible to monitor transition from regular legal status to "protection status" as an asylum seeker or refugee, and back, as well as the time spent in each status.

The preliminary results show that most, but not all asylum seekers hold protection status from the time of immigration. 8.9 per cent of all foreigners in protection status had a regular residence status before – they likely applied for asylum after their residence permit ran out to avoid deportation. The German authorities needed on average 0.8 years to process an asylum application, 0.9 years if asylum was denied. This time varies by gender, age, marital status and citizenship. On average, foreigners holding protection status in any years between 2007 and 2017 spent 3.7 years out of the 4.4 years of residence in protection status (84.1 per cent). Of these 3.7 years they spend 0.8 years waiting for a decision on their application, 2.4 years with protection granted and 0.5 years with protection denied.

4.1.1.4 Circular migration longitudinal statistics

- Average number of stays per 'circular migrant' (to be further defined, as for censored observation, period of time, required minimum lengths of stay, etc.) *
- Average length of stay per 'circular migrant'*

*Please refer to UNECE (2016) and Eurostat (2017 and 2018).

328. The measures above only refer to persons in circular migration processes, who are a subset of the entire set of migrants. For a correct interpretation of those statistics, disaggregation by country of birth and citizenship is recommended.

4.1.2 Socio-economic outcomes

329. Longitudinal indicators related to socio-economic outcomes can include a wide array of indicators related to the conditions of migrants after arrival. This can include topics such as knowledge of official language(s), employment, housing, income, and health. More information about indicators related to the socio-economic conditions of migrants can be found in the UNECE publication "measuring change in the socio-economic conditions of migrants" (2015).

330. Unlike the migration-related indicators, some indicators related to socio-economic outcomes may not be adequately measured using administrative data. In these cases, surveys or administrative proxies may be needed to measure these outcomes.

331. The following were identified as key longitudinal indicator topics of interest to UNECE member countries related to socio-economic outcomes:

- Time until migrants have ability to speak the host country's official language(s)
- Time until migrants own their home
- Migrants' labour force status over time
- Time to attain post-secondary education
- Migrants' income over time

- Business ownership / entrepreneurship for migrants over time
- Migrant's health over time.

Table 12 Key indicator topics for socio-economic outcomes by country examples

Key indicator topics	Canada	Germany	Italy	Norway	Switzer-land	United Kingdom
Time until migrants have ability to speak country's official language(s)	x	x			x	x
Time until migrants own their home						
Migrants' labour force status over time	x			x		x
Time to attain post-secondary education	x					
Migrants' income over time	x					
Business ownership / entrepreneurship for migrants over time	x					
Migrants' health over time	x					x

4.1.2.1 Time until migrants have ability to speak the host country's official language(s)

332. This is a topic which can apply to all migrants regardless of age or other factors. Some migrants will arrive with (or were indeed selected because of) knowledge of the country's official language(s). However, for others, not being able to speak the country's official language(s) can be a barrier in many respects such as obtaining a job, attending school, or even making social connections. As such, this can be an important indicator topic related to the settlement or integration of migrants in a new country.

333. Some migrants will attend language training, while others might learn the language(s) through immersion or benefit from speaking with their children who learn the language(s) at school. Overall, there is an expectation that migrants will become increasingly likely to be able to speak the official language(s) the longer they are in the country. Therefore, this is a topic which lends itself well to longitudinal statistics.

334. The following longitudinal indicators are proposed for the measurement of language skill:

- Proportion of migrants who have ability to speak country's official languages after X amount of time
- Average time until migrants have ability to speak country's official languages
- Proportion of migrants who speak country's official language(s) at home after X amount of time
- Average time until migrants speak country's official language(s) at home

335. These indicators are longitudinal in that the ability to speak the country's official language(s) can change over time. Care should be taken to exclude migrants from countries with the same official language as the destination in these calculations, assuming they are already fluent at arrival.

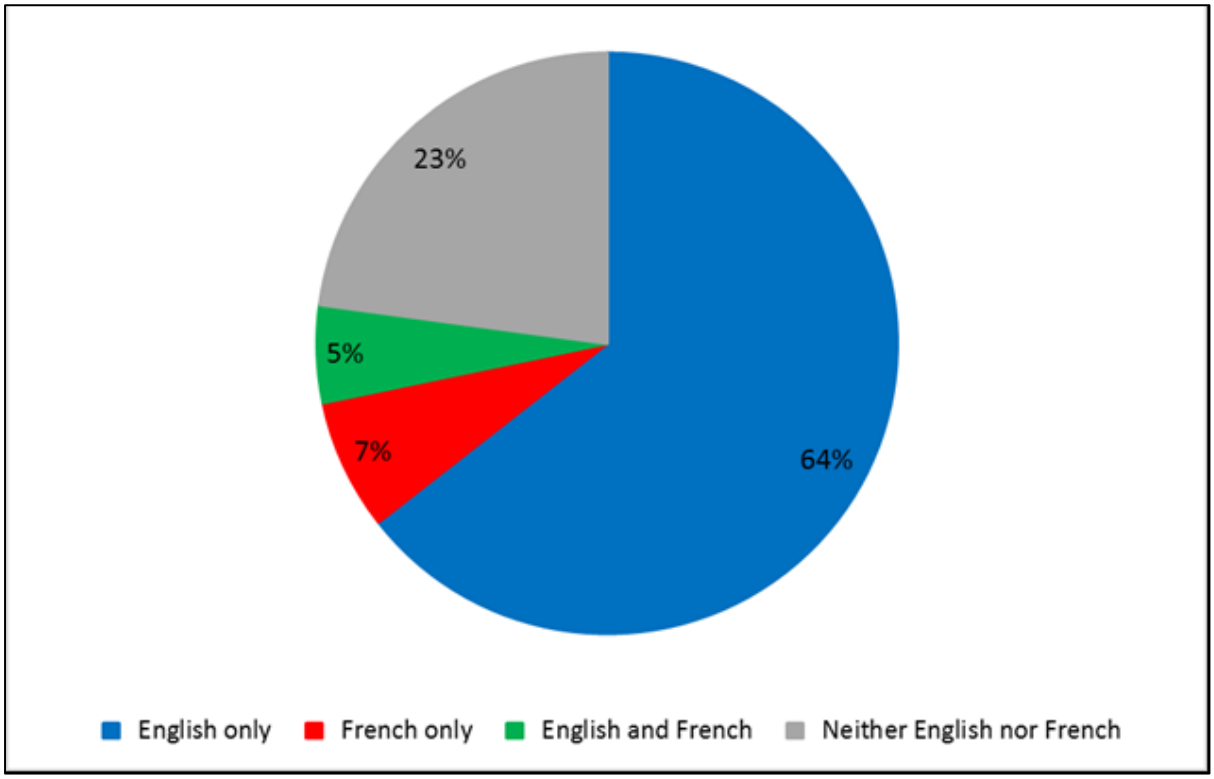
336. Unlike some of the other indicators, however, these might be difficult to measure longitudinally. A survey can ask questions about a respondent's ability to speak, write, or understand a language or whether or not they use a language at home, at school, at work, or elsewhere. However, administrative records, which would be more practical for longitudinal analysis, may be less precise.

For example, in **Canada**, detailed language questions are included on the Census of Population questionnaire. However, administrative records are often limited to the language of service (e.g. language of the administrative form).

Another limitation with administrative data is whether responses are affected by the administrative collection process itself. Immigration, Refugees, and Citizenship Canada (IRCC) collects 'knowledge of official languages' from prospective immigrants on an application form. While these responses are not used in any selection process, it's possible that individuals feel that positive answers will improve their chances at admission. On the other hand, IRCC also collects results from language standardized tests for certain prospective immigrants. These results might provide a more objective evaluation of knowledge of official languages than self-reporting on a survey.

While these results are collected by IRCC when immigrants are admitted to Canada, it is challenging to compare these measures against any follow-up results from surveys or other administrative sources. However, these comparisons can still provide valuable insights on language acquisition.

Figure 22 Knowledge of official languages in 2011 for recent (2006-2011) immigrants to Canada with no knowledge of English or French at the time of admission

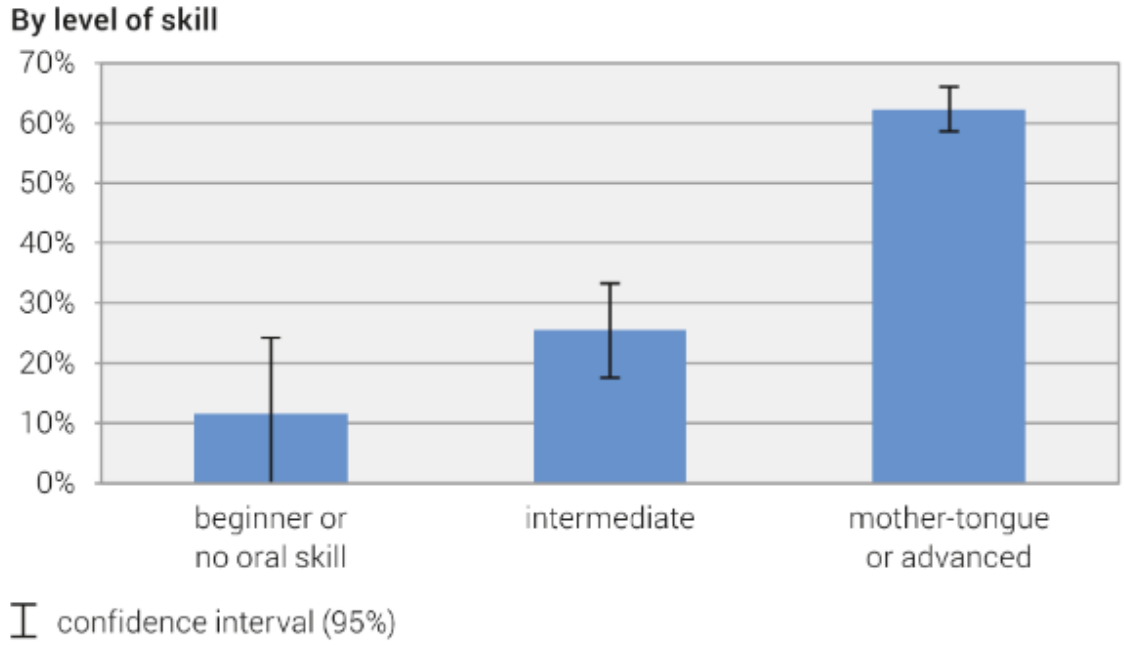


Source: Statistics Canada, 2013 Longitudinal Immigration Database integrated with 2011 National Household Survey.

Figure 22 yields valuable information about the acquisition of English or French among recent immigrants to Canada. For example, among those immigrants admitted between 2006 and 2011 who could not speak either English or French at admission, over 75 per cent reported they could speak at least one official language by May 10th, 2011 according to the 2011 National Household Survey. Further analysis could investigate the characteristic differences between those who acquired knowledge of English and French compared with those who did not.

Regional specificity may be important to consider when evaluating linguistic integration. For instance in Switzerland, cross-sectional results are available to provide a snapshot of migrants' ability to speak the official language of their canton of residence.

Figure 23 Oral language skills of immigrants to Switzerland in one of the official languages in the canton of residence, 2014



Source: Swiss Federal Statistical Office, Swiss Labour Force Survey (SLFS), migration module, 2014.

In the **United Kingdom**, whilst not set up specifically to study migration, the Longitudinal Study for England and Wales (with equivalents for Scotland and Northern Ireland) are key longitudinal data sources from which socio-economic outcomes overtime can be observed. The Longitudinal Study for England and Wales links census and administrative data over time. Two questions on main language and English proficiency were included on the 2011 census for the first time. This has enabled analysis of the effect of language skills on socio-economic outcomes such as education, health and fertility.

As an example, the England and Wales Longitudinal Study has been used to identify the causal effect of English language skills on education, health and fertility outcomes of immigrants in England and Wales (Aoki and Santiago, 2015). To begin, the research explores proficiency in English among people aged 20 and over who were childhood immigrants. Those migrating from English-speaking countries were generally proficient in English, having been exposed to the language prior to their arrival. For those migrating from non-English-speaking countries, those who migrated before the age of eight tended to be as proficient in English as those migrating from English-speaking countries.

However, proficiency rates diverge at nine years old; the later they migrated, the poorer their English language proficiency, on average. This is illustrated in Figure 24.

Using regression analysis, this study also found that better English language skills lower the probability of adults having no qualifications and raises the probability of obtaining academic degrees. The impact of language on fertility is considerable. Better English skills significantly delays age at first birth for women, lowers their likelihood of being a teenage mother and decreases fertility. There was no relationship found between English proficiency and either child health or self-reported adult health.

Figure 24 **Age at arrival and adult English proficiency in 2011 for child migrants to England and Wales**

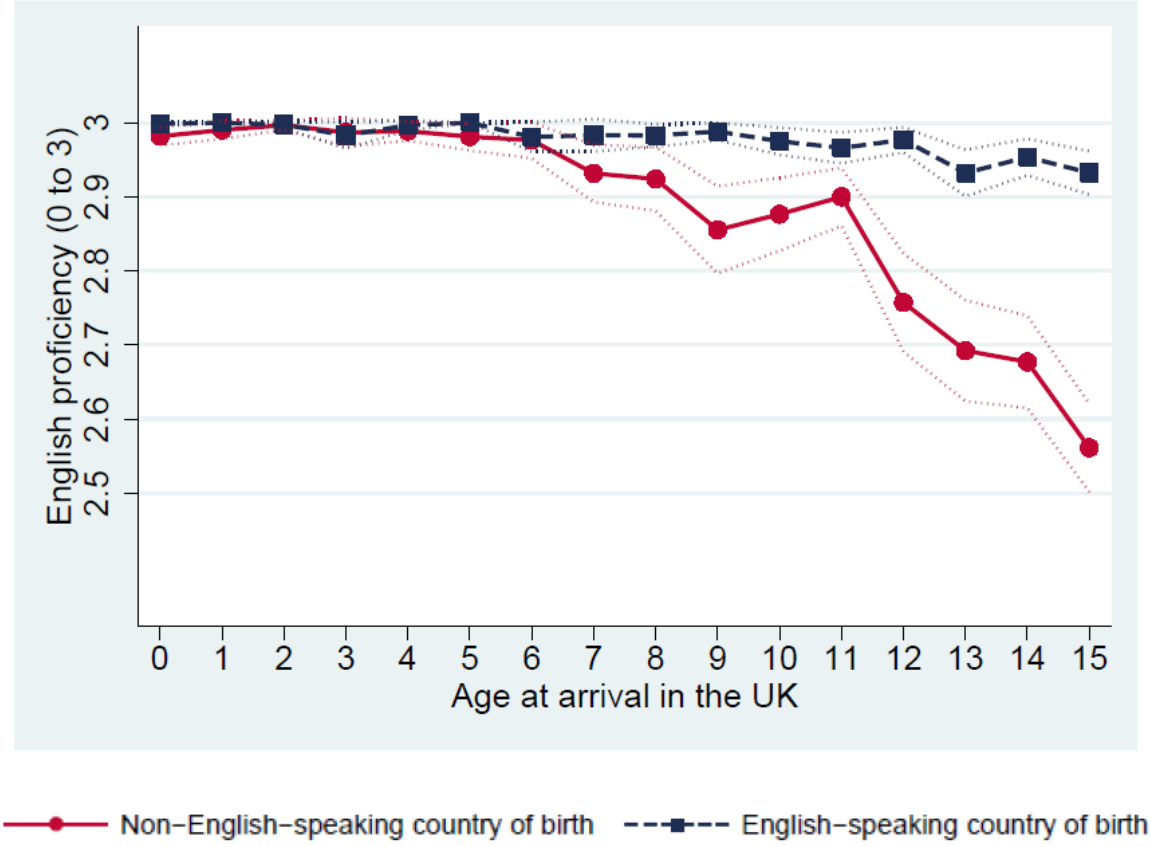


Figure 1: Age at Arrival and English Proficiency

Notes: Figure plots the average ordinal measure of English proficiency, where 3, 2, 1, and 0 correspond to speaks "very well", "well", "not well", and "not at all", respectively. English proficiency is regression adjusted for age. Two sets of outer lines correspond to 95 per cent confident intervals. The sample corresponds to childhood immigrants aged 20 to 60 at the time of Census 2011.

Source: Aoki & Santiago (2015) Education, Health and Fertility of UK migrants: The Role of English Language Skills, IZA Discussion Paper Series No.498.

In **Germany,** there are no longitudinal data sources from which socio-economic outcomes may be derived. Retrospective data in population surveys are used for this purpose instead. For immigrants, the micro-census and the German version of the European Labour Force Survey provides data on the country of birth and the year of arrival. Educational attainment, labour force participation, income earned and poverty risk are regularly analyzed for migrants by country of birth, age at first arrival and length of stay.

Germany provides indicators relating to the above topic areas for

- first and second generation migrants, and
- with and without Germany citizenship, respectively.

Time series data on these indicators are available for the years 2005 through 2018.

https://www.destatis.de/EN/Themes/Society-Environment/Population/Migration-Integration/_node.html#sprg265538

4.1.2.2 Time until migrants own their home

337. Becoming a home owner can be an indication of an individual settling in their new country or area of residence. There are certainly individuals and families who generally choose not to own their homes and prefer to rent regardless of migrant status. However, home ownership is a sign that the individuals or families have the economic means to purchase real estate and that they are willing to invest in a specific place to live.

338. The following longitudinal indicators are proposed to measure this topic:

- Proportion of migrants who own their home after X amount of time
- Average time until migrants own their home for the first time in the new country of residence
- Housing tenure (e.g. rent vs. own) of migrants over time
- Proportion of migrants living in government subsidized or funded housing after X amount of time
- Ownership by type of residential property (e.g., apartment, semi-detached house, etc.) after X amount of time

339. Home-ownership is a concept which may be measured administratively or through a population register, in addition to through a survey. Administrative records would need to be clear on the ownership of the dwelling. Alternatively, administrative records on beneficiaries of government-subsidized or funded housing or tenants would be important for certain indicators.

340. One element which can be complex with the use of administrative records is that the title of the property may only be listed under one individual. Live-in family members would need to be accounted for in this metric as well. This last component is complex longitudinally as the composition of the family may evolve over time (see section 4.2.3).

4.1.2.3 Migrant's labour force status over time

341. Finding gainful employment in a new country of residence is considered an important sign of economic integration with ties to other indicator topics listed in this chapter. In order to get a job, individuals often must meet certain requirements. For example, many jobs would require the worker to be able to communicate effectively in the official language(s) of the country. Recognized credentials, such as post-secondary attainment, may also have been required. Entering the

workforce opens new opportunities and the migrant's social network would expand to co-workers, clients and other people.

342. Labour force participation of migrants can be considered an important indicator from a national economic perspective as well. If migrants are admitted specifically to fill certain gaps in the labour market, it would be important to consider whether (and in which occupation or industry) those migrants are in fact employed after arrival.

343. Many migrants arrive in their new country of residence with employment already established. Indeed, employment in a certain occupation may have been the motivating factor behind the migration (from either the migrant's or the country's perspective). On the other hand, many migrants, including refugees, may have to source employment after arrival.

344. The following longitudinal indicators are proposed for the measurement of labour force status:

- Proportion of migrants who are employed after X amount of time
- Average time until migrants become employed for the first time in the new country of residence
- Full-time vs. part-time employment of migrants over time
- Labour force status of migrants over time

345. Similar to the set of indicators related to languages, the measure of labour force status can be difficult to estimate using administrative data.

346. In particular, while administrative data could inform about an individual's incidence of employment, distinguishing between unemployed and those not in the labour force is more complex.

347. From a longitudinal perspective, surveys tend to use a pre-determined short reference period, like a week, to determine someone's labour force status. On the other hand, administrative data may have to rely on other reference periods. If these reference periods cover larger periods of time, it may not be possible to provide a precise value for 'time until an individual obtained employment.

348. One best practice may be to ignore values where the reference period overlaps with the migrant's entry into the country. Otherwise, values for this period will be negatively biased and would be affected by when certain subgroups of migrants arrived. For example, if the reference period is annual, migrants who arrive in January would have a higher probability of being employed in that year than those who arrive in December independent of other factors.

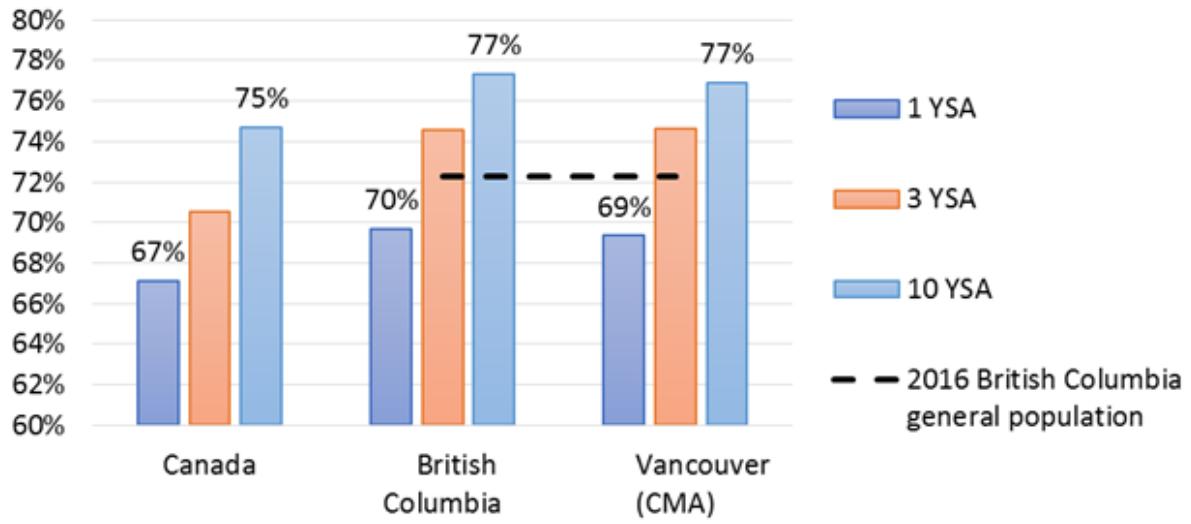
349. Another factor to consider for these indicators is the denominator as it should only include those individuals in scope to be employed or unemployed (e.g. minimum age).

In **Canada**, administrative tax files can provide a proxy for employment. In particular, the Longitudinal Immigration Database (IMDB) has annual variables for wages and salaries, industry(ies) of employment, etc. However, these variables are annual summaries. They can be used to determine how many immigrant tax filers had some employment in a given reference year (e.g. salaries and wages > 0 for the reference year), but whether that employment was in January or December is not specified. Details on hours or weeks worked are also not available. Other variables such as incidence of employment insurance are weak proxies for unemployment and capture other events such as parental leave.

Still, the IMDB does provide indicators relating to incidence of employment in years since admission to Canada for immigrants despite these limitations.

The following chart, based on data from the Longitudinal Immigration Database (IMDB) provides annual incidence of employment for refugees admitted to Canada in 2006 at 1 year, 3 years, and 10 years after admission. The figures are compared for Canada, the province of British Columbia, and the metropolitan area of Vancouver. The incidence figures are compared against the equivalent calculation for the general population living in British Columbia in 2016 (equivalent to the 2006 refugee figures 10 years after admission). Analysis could be streamlined to account for other factors such as age and sex.

Figure 25 Incidence of employment for refugee tax filers admitted in Canada in 2006, by selected place of residence and years since admission



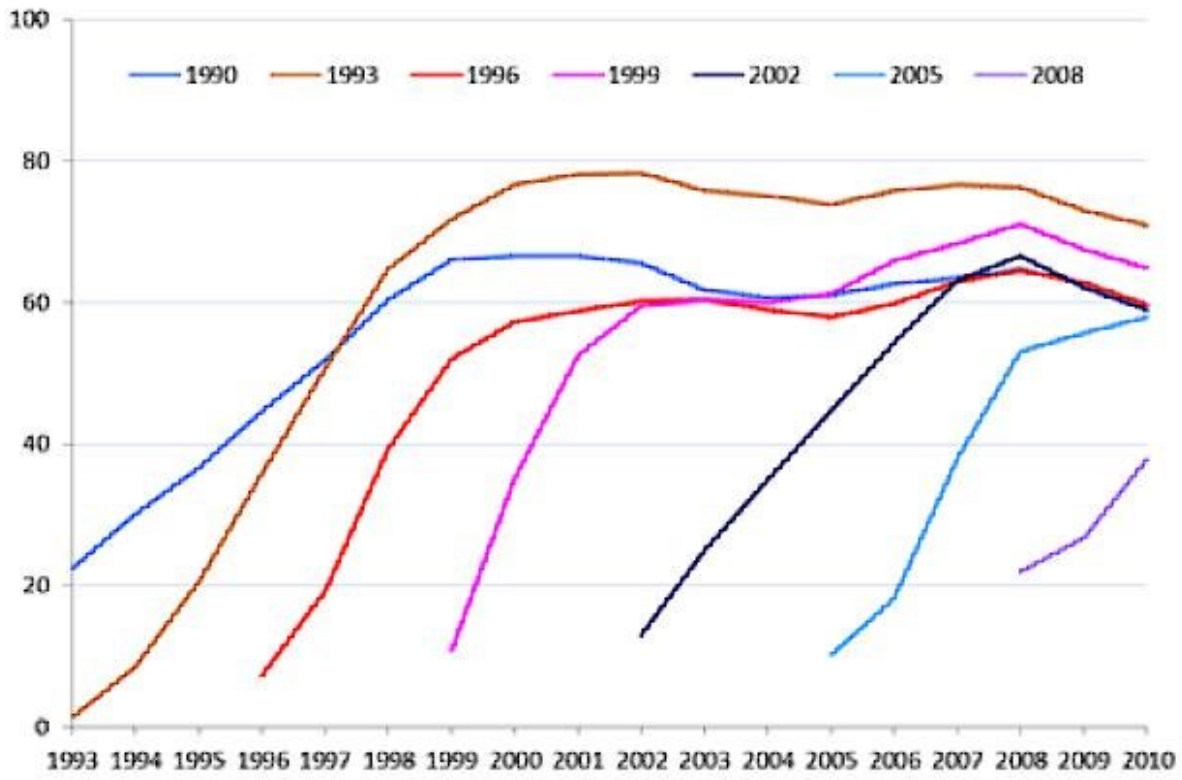
Source: Statistics Canada, 2016 Longitudinal Immigration Database, table 43-10-0014; 2017 Canadian Income Survey, table 11-10-0239-04

In the **United Kingdom**, whilst not specifically set up to study migration, the Office for National Statistics Longitudinal Study has been used to investigate unemployment gaps between first and second generation migrants to England and Wales. The authors (Zuccotti and Platt, 2016) note that research has found that the descendants of migrants show marked improvements over their first generation parents in terms of labour market disadvantage and inequality. They use data that links the results of five successive censuses (1971 to 2011) to measure individuals' family resources as children, living with immigrant parents in England and Wales, and link these with their employment outcomes as adults. By also linking information on where families lived, they are able to construct area deprivation measures to explore whether the neighbourhood that second generation children grew up in might impact their social mobility.

Results show that Pakistani and Caribbean men are around 6 per cent more likely to be unemployed than British men. When they take into account social background and origin neighbourhood deprivation, this 'penalty' is halved to 30 per cent. They also show that having a university degree considerably reduces the penalty for both male and female groups.

Norway has also published longitudinal analyses of migrants' employment. The below chart uses a cohort approach to compare employment trajectories between 7 cohorts of refugees aged 17-36.

Figure 26 Share of refugees, aged 17 to 36, who are employed, by cohort and income year, Norway, per cent



Source: Statistics Norway, Population and Income Statistics

4.1.2.4 Time to attain post-secondary education

350. For migrants (especially children and young adults), obtaining post-secondary education in the new country of residence is an important process as it gives them the opportunity to learn in the official language(s) of the country, establish a social network, and, ultimately, provides them with recognized credentials to find employment after graduation. Time to complete education can be affected by various factors such as disruptions to their previous education related to the migration event, affordability of seeking higher education versus finding employment, ability to learn in the official language(s) of the country, etc. As a result, some migrants may take longer than others to complete their education.

351. It's important for this topic to consider 'time' from an age perspective as well as from a time since arrival perspective.

352. Post-secondary education in the new country of residence can also be important for migrants with previous education. In particular, if their prior education is not recognized by employers or if they need to obtain re-training to open new labour market opportunities. In some cases, knowledge of the applicant's admission category (the programme on the basis of which the migrant was admitted) can inform the minimum credentials possessed by the individual if no other measure exists on administrative migration files.

353. Some migrants will already have established plans to attend university or higher education before arrival in the country. Similar to the indicator topic on labour force participation, attending university or higher education may be the motivating factor behind the migration event itself.

354. The indicators proposed for the measurement of time to attain post-secondary education are the following:

- Proportion of migrants who obtained a post-secondary certificate, degree or diploma in their new country of residence after X amount of time by age
- School/ university/ higher education attendance of migrants over time by age

355. If enrolment data from all post-secondary institutions in the country is collected centrally, these indicators might be relatively straight-forward to produce from a statistical register or from integrated administrative records. Alternatively, if subsets of institutions are not covered by government administrative data (e.g. private institutions), this would provide a limitation to consider for the relevant indicators.

356. This is a more complex topic since it can include at least two temporal variables and combining both in the same indicator might be excessive. There are a few possible approaches to consider to simplify the indicators. For young migrants (e.g. those who arrive aged 18 or younger), age might be a more important factor than years since arrival. So indicators could be simplified to:

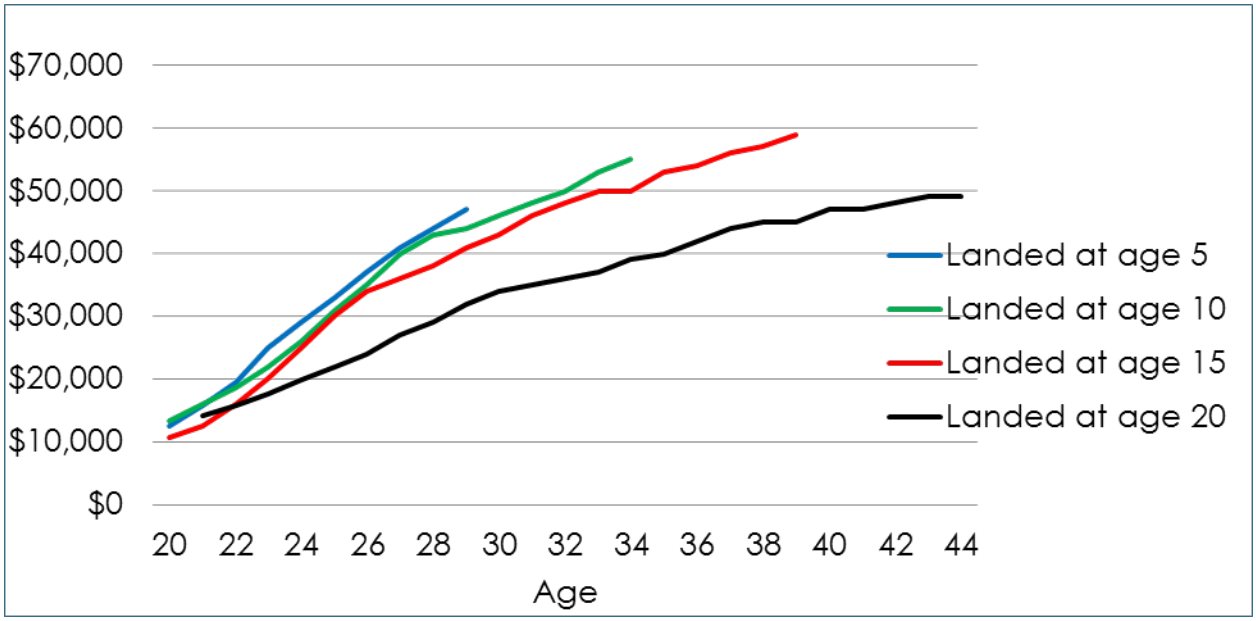
- Post-secondary school attendance by single year of age
- Post-secondary attainment by single year of age

357. In this case, it is still important to control for year of arrival. However, instead of it being part of the indicator itself, it could be a filter variable (e.g. calculate the indicators for those who arrived in a given year). The above indicators can also be reproduced for the general population allowing comparisons of post-secondary education between migrant and non-migrant youth.

358. Another temporal variable of interest, particularly when studying the outcomes of younger migrants, is age at arrival. The younger migrants were at arrival, the higher the proportion of schooling they would have received in the receiving country. This may make it easier to gain entry to post-secondary institutions in the same country.

359. For example, the chart below shows income trajectories for immigrants who landed (i.e. where admitted) in Canada in 1990 by selected ages at admission. The chart shows that, within the same cohort, immigrant children who arrive at younger ages tend to earn more when they grow up compared with immigrant children who arrive at ages closer to adulthood.

Figure 27 Average salaries and wages by age and age at admission to Canada for immigrant tax filers who landed in 1990 (2014 dollars)



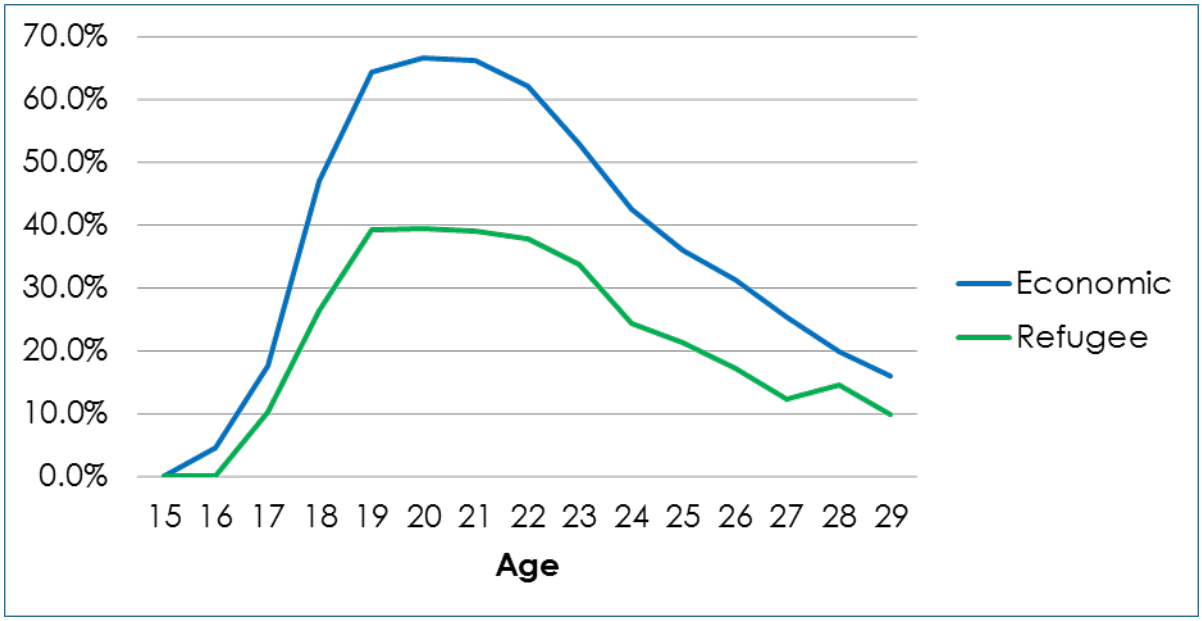
Source: Statistics Canada, Longitudinal Immigration Database, 2014

For adult migrants, years since arrival may be of more interest. In this case, it might be worth restricting the analysis to those who were aged at least 25 (or another age after post-secondary education would typically be completed) when they arrived. The purpose of this restriction is to mitigate the age effect on the corresponding analysis.

In **Canada**, the Longitudinal Immigration Database (IMDB) does not include any administrative data from post-secondary institutions directly but tax filers declare the tuition amounts for post-secondary schooling. This can then be used as a proxy for post-secondary participation. However, it does not specify type of post-secondary education (e.g. Doctoral studies vs. trades certification) nor does it identify those who complete their training.

The following chart, using the IMDB shows post-secondary participation by single year of age for immigrants admitted to Canada in 1999 at the age of 14. The results are separated by admission category show that the children of refugees are less likely to participate in post-secondary than children of economic immigrants.

Figure 28 Percentage claiming tuition by age and admission category for immigrant tax filers who were admitted to Canada in 1999 at age 14

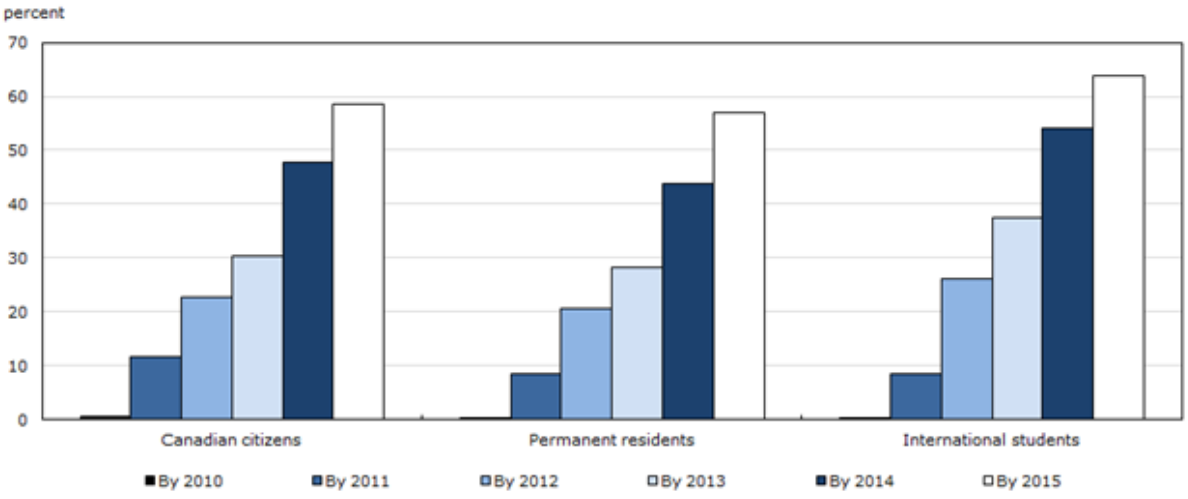


Source: Statistics Canada, Longitudinal Immigration Database, 2014.

Looking at adult immigrants, the study, Refugees and Canadian Post-Secondary Education: Characteristics and Economic Outcomes in Comparison, used the same proxy to study the characteristics and economic outcomes of those who participated in post-secondary in Canada after admission versus those who didn't.

Another Canadian analysis (Frenette, Lu and Chan 2019) used administrative postsecondary schooling data based on the Postsecondary Student Information System (PSIS). Immigrant status was captured administratively by institutions during enrolment. The chart below shows graduation rates by immigrant status after a certain number of years since entering a programme.

Figure 29 Graduation rates by immigration status of 2010 new programme entrants in Canada



Source: Statistics Canada, Postsecondary Student Information System, 2015.

4.1.2.5 Migrant's income by source over time

360. Income is a measure of economic outcomes. When broken down by key source of income, it can be understood which individuals are employed, self-employed, or receiving social or government assistance. While on its own, it cannot identify the type of job an individual has, the overall magnitude of their employment income gives a broader indication combining hourly wages with total hours worked.

361. Many migrants may face barriers to the labour market as discussed above. Overall, it is not unexpected to observe migrants earning less than the general population when they first arrive. But does this persist with more years in the country? Are there some groups which are more likely to surpass the average or median income of the general population?

362. There is interest as well in understanding the fiscal impact of immigration on receiving countries. This can be studied in various ways. For instance, one could consider overall taxes paid by immigrants compared with receipt of government assistance.

363. The following longitudinal indicators are proposed for measuring income:

- Average and median total income after X amount of time
- Average and median employment income after X amount of time
- Proportion of migrants with employment income after X amount of time
- Proportion of migrants with self-employment or business income after X amount of time
- Proportion of migrants with social or government assistance after X amount of time
- Time until average or median total income for migrants equals general population average or median total income
- Time until average or median employment income for migrants equals general population average or median total income
- Proportion of migrants in low income after X amount of time

364. Through taxation, these indicators should be straight-forward to produce using either registers or integrated administrative records. Of course, any income not declared (e.g. possibly foreign sources of revenue), may not be included.

365. From a longitudinal perspective, the reference period is important to consider for these indicators. The income values will pertain to the reference period as a whole and could be affected if income was only received in a fraction of the reference period. For example, a small employment income value over a reference period could reflect a low-paying job or it could reflect a high-paying job that was only active for a subset of the reference period. This is particularly important to consider when studying migrants since they may not have even been present in the country for the entirety of the reference period.

366. One best practice may be to ignore values where the reference period overlaps with the migrant's entry into the country. Otherwise, values for this period will be negatively biased and would be affected by when certain subgroups of migrants arrived. For example, if the reference period is annual, migrants who arrive in January would have higher incomes on average in that year than those who arrive in December, even if they held lower paying jobs.

367. Similar to the labour market indicators, it is important to consider the population in-scope for these indicators. Employment or self-employment indicators should be restricted to the population who could work. Other indicators could be considered more at the family-level. For example, low

income concepts tend to be defined at the household level – that would then be a characteristic shared by all members of the household. Social or government assistance may be considered a household or family-level concept as well. However, it is important to distinguish between support for low income families and individual-based support (e.g. temporary unemployment assistance).

368. Another important consideration, especially for longitudinal analysis, is the effect of changes in cost-of-living over time. It can be a best practice to adjust monetary values so that income values are based on a fixed currency value.

In **Canada**, the Longitudinal Immigration Database (IMDB) produces several regular longitudinal indicators related to income of immigrants. These include percent with, average and median income by source of income, years since admission, and other characteristics. As a best practice, IMDB outputs exclude outcomes in year 0 since admission. The results are available by single cohort year of immigrants to Canada with cohorts starting in 1980 and outcomes starting in 1982. Income sources selected for standard tables are consistently defined over that period of time.

Of particular interest are the measures:

Average and median total income

Average and median wages, salaries, and commissions

Average and median self-employment income

Proportion receiving social welfare benefits

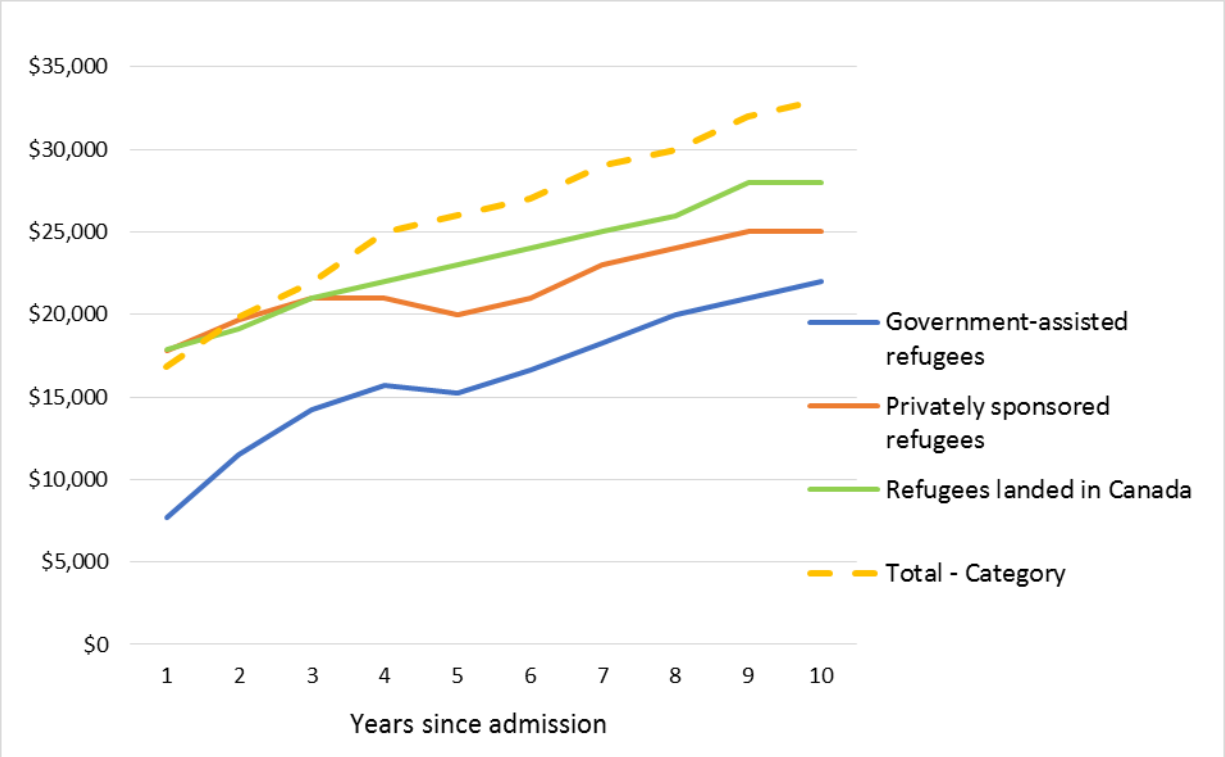
Proportion receiving wages, salaries, and commissions

These are available by year of admission, years since admission, and other characteristics. Comparisons can be made against the general population using comparable data sources such as the Longitudinal Administrative Databank.

The following are standard tables published on Statistics Canada's website and updated annually:

Immigrant Income by admission year and years since admission, Canada and provinces
Immigrant Income by admission year and immigrant admission category, Canada and provinces
Below is a chart showing trajectories of average salaries and wages for immigrants admitted to Canada in 2004 by selected admission categories.

Figure 30 Average salaries and wages of immigrants admitted to Canada in 2004, by selected admission categories and years since admission (2014 dollars)



Source: Statistics Canada, Longitudinal Immigration Database, 2014.

4.1.2.6 Business ownership / entrepreneurship for migrants over time

369. For some migrants, challenges entering the labour market can lead to self-employment; while for others, entrepreneurship or business ownership may be connected with the motivating factor behind the migration event itself.

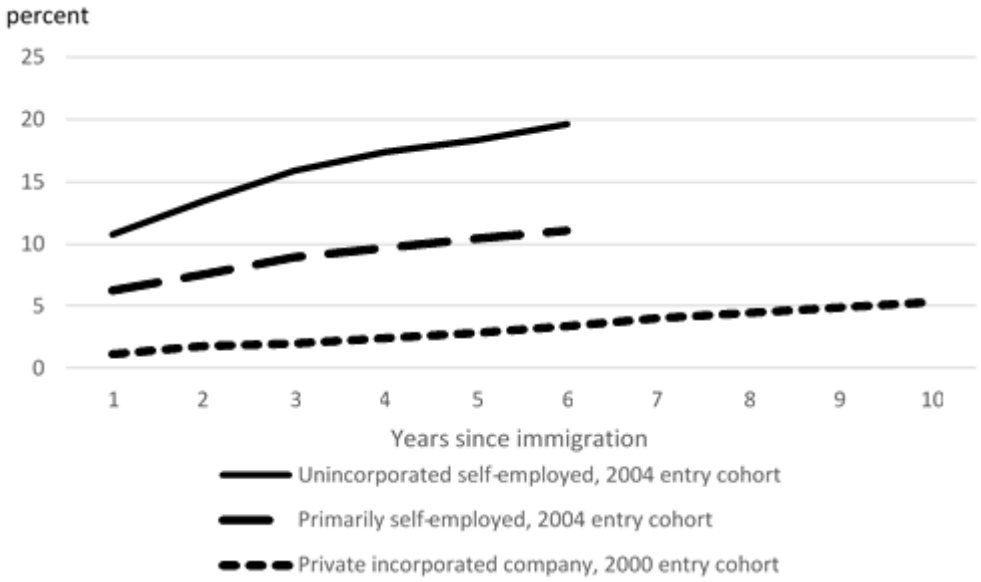
370. It's certainly true that not all individuals, let alone migrants, aspire to be self-employed or own a business. However, self-employment or entrepreneurship can be an important economic outcome for migrants. Not only does it potentially bypass some barriers to the labour market, it can create jobs for others as well.

371. The following longitudinal indicators are proposed under this topic:

- Proportion of migrants who are self-employed or own a business after X amount of time
- Average time until migrants are self-employed or own a business
- Number of individuals employed by migrants after X amount of time
- Similar to the indicators on labour force status, taxation records can provide a proxy for entrepreneurship by providing the proportion of migrants with self-employment or business income. Otherwise, administrative records on business ownership can be used.
- Connecting migrant entrepreneurs with information on their employees may be more complex. However, if administrative records exist connecting workers with firms and owners with firms, this analysis is also possible.

- The Canadian Employer-Employee Dynamics Database (CEEDD) is a linkage environment based on 12 administrative source-files, including the Longitudinal Immigration Database (IMDB), and includes individual, firm, and business-owner characteristics. Being a linkage environment, users must integrate the various components to fit their analytical requirements using anonymized unique identifiers on each file. The CEEDD enables analyses of firms and individuals over time, answering questions about, for instance, initial firm allocation and earnings growth, immigrant business ownership, and the trajectories of immigrant-owned firms.

Figure 31 Percentage of immigrant taxfilers who are business owners in Canada, by years since immigration, 2000 and 2004 entry cohorts



Source: Statistics Canada, Canadian Employer-Employee Dynamics Database, 2014; Green et al (2016) Business Ownership and Employment in Immigrant-owned Firms in Canada. Economic Insights Catalogue no. 11-626-X – No. 057.

4.1.2.7 Migrant's health over time

372. Maintenance and improvement of health among recent migrants in a new country is an important sign of successful integration, with ties to other indicator topics listed in this chapter, such as labour force participation. Without good health, it might be difficult to participate in the community, to get a job, or be able to hold on to one. At the societal level, if migrants arrive with or quickly develop serious health problems, this can become a drain on the healthcare system. As well, migrants that grow up in vastly different health environments may bring in contagious and communicable diseases that can pose a big public health risk. These are some of the reasons why some countries implement medical screening and evaluations of migrants before allowing them to enter or to settle in the host country. This may also help to explain the healthy immigrant effect that can be observed in many migrant receiving countries, reflecting the observation that most migrants, perhaps with the exception of refugees, tend to be healthy at time of arrival.

373. Health is influenced not only by an individual's biology, but also by a host of social and environmental factors identified as the social determinants of health, which include lifestyle, environmental and health care organization and access factors. These are especially important for

migrants who might have health risk factors, as well as lifestyles quite different from others in the host country, making the adjustment and integration difficult. Migrants also might not know how to navigate new health care system, as they typically arrive from countries with different healthcare environments and language requirements. Migrants who have difficulty finding quality employment or may experience stress and mental health challenges. All these factors can contribute to a loss of the initial health advantage.

374. The following longitudinal indicators are proposed under this topic:

- Proportion of migrants who are registered with a physician within host country's health care system after X amount of time (or time to deregistration)
- Proportion of migrants who are hospitalized in the X years since arrival to host country of residence, overall or for various select health conditions, e.g., chronic or infectious in nature (or time to first hospitalization)
- Proportion of migrants deceased since arrival to host country of residence for various disease conditions (after X amount of time) (or time to death)

375. In the past, migration outcomes research has relied on surveys, with indicators such as self-reported health status of migrants, self-reported unmet needs, self-reported access to medical practitioner, etc. However, true longitudinal indicators would require us to follow migrants over time in order to observe changes in health. More recently, with the increasing difficulty to obtain survey responses, together with the advancement of data linkage methodology and the decline in the cost of computing, linking administrative and survey data provides researchers with an alternative source of information.

In **Belgium**, researchers used individually linked data from the Census, the national population register and death certificates from the periods 1991-1997 and 2001-2008 to study cause-specific mortality differences between the migrant and Belgian-born population over time (Vanthomme and Vandenheede, 2019). Linking the 1991 and 2001 Belgian censuses with register data from 1991-1997 and 2001-2008, allowing an observation of all emigration and mortality in these periods. Subsequently, individual linkage with death certificates provided cause-specific mortality.

In **Canada**, certain provinces require individuals to enrol under the practice of general medical practitioners. One limitation of the provincially administered health data is that data are available only in certain jurisdictions, not centrally at the national level. The enrolment database in Ontario, the province with the largest number of immigrants, has been linked to the Longitudinal Immigration Database (IMDB) and has been used by researchers to understand differential enrolment. This is especially important in situations where there is a shortage of general medical practitioners, such as in Ontario (Globerman et al, 2018). To be able to enrol under the care of a general practitioner can be an important first step to successful health integration.

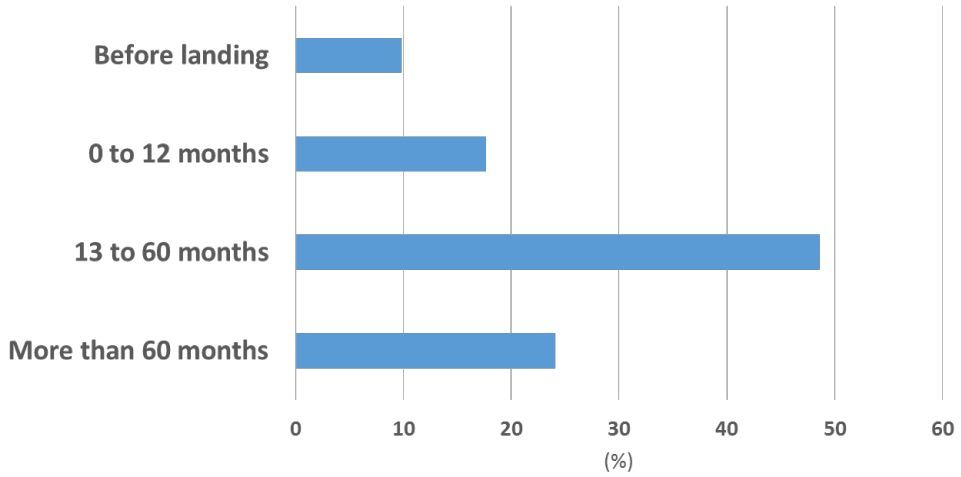
In Canada, hospital discharge records collected by the provincial and territorial health ministries have been transferred and centralized at the Canadian Institute for Health Information in a database called Discharge Abstract Database (DAD). These are subsequently shared with Statistics Canada for record linkage and research purposes. This hospital file has been linked to the IMDB to examine patterns and rates of hospitalization, either overall or for specific to chronic or infectious diseases. These data also allow for the assessment of the change in hospitalization patterns for all major causes, by gender, source regions, and immigrant admission category. Admittedly, one has to be in a serious health situation to be admitted to the hospital; thus, hospitalization is an indicator of relatively serious health situations.

For another example, the IMDB-DAD has been used to examine distribution of first hospitalizations due to Tuberculosis (TB) by time since arrival in Canada. TB is a highly contagious disease with

serious public health consequences. With the decline of TB rates among the local-born, TB has become more prevalent among immigrants but uncommon among most local-born, except for Indigenous peoples (Vachon et al, 2018).

This IMDB-DAD linked data analysis found some 10 per cent of these TB-related hospitalizations occurred prior to individual's immigration landing year, meaning that these immigrants were hospitalized when they were in Canada as temporary residents, such as foreign students or temporary foreign workers (Ng et al, 2018).

Figure 32 Percentage of Tuberculosis-related hospitalizations in Canada by time since admission to Canada, immigrants admitted 2000-2013



Source: Statistics Canada, 2000-2013 IMDB-DAD linked database (Quebec and the Territories excluded)

Limitations for this data also lie with the differential availability by jurisdiction. While hospitalization datasets may not be available for all provinces and territories, the dataset can still be used for analyses pertaining to the rest of the country. While information from physician visits or laboratory test results could provide an indication of health deterioration some time before having to be hospitalized, in Canada, such datasets reside at the provincial level only.

In Canada, the IMDB has also been linked to the Canadian Vital Statistics Death Database (CVSD), which captures all deaths occurring in Canada. As such, another relevant indicator of immigrant health, other than hospitalization, would be time to death. These data allow for the examination of the evolution over time in mortality patterns for all major causes of death, by gender, source region, and other migrant-related dimensions. One can also assess the change over time by comparing the migrant mortality differentials in the 2000s with the 2010s.

One challenge with this analysis is that since more immigrants arrived in the host country with relatively good health, time to death may require a long period of observation until outcome. In this sense, time to first hospitalization could be complementary indicator to detect health deterioration over time.

A common factor to consider for these indicators is that when a rate is to be derived, the denominator should only include those individuals in scope, that is the population at risk of experiencing the outcome. This is the issue of emigration among migrants, who may decide to

return their home country, or to move on to another country. This numerator-denominator bias challenge can be difficult to resolve, especially in countries that have no exit records. In **Canada**, administrative tax files are integrated into the Longitudinal Immigration Database, and the tax filing patterns can be used as proxy of residency of the immigrants in Canada.

Also, data analysis based on linked data involving the IMDB will obviously not have the corresponding information from the local-born reference group. One can use other data linkages to provide a comparison, if the dataset exists (Ng et al, 2016).

In the **United Kingdom**, the Office for National Statistics Longitudinal Study has been used to explore the mortality experience of first and second generation migrants to England and Wales. Migrants are often found to have low mortality compared to host populations in Western countries. The 'healthy migrant effect' is believed to wear off across generations. Wallace (2016) analyses the mortality experiences of Longitudinal Study members aged 20 to 65 years in 1991, through the study period 1991-2012. After some adjustments for missing information, this yields a sample of 555,111 people with 47,907 deaths. Migrants and their descendants are defined through a combination of ethnicity and country of birth information, which is recorded in the census. Event history analysis is used to find that migrants of all ethnicities have lower mortality, but there is variation in the mortality experiences of the descendants of migrants.

The descendants of Black Caribbean, Pakistani and Bangladeshi migrants have high mortality, which cannot be accounted for by their social background. Descendants of Indians, Black Africans and Other have low mortality which is initially masked by their social background. The author suggests that the loss of the selection effect for migration may help to explain the increased mortality among descendants. Generational differences in cultural norms, attitudes and behaviours in lifestyle and diet (including smoking) are also considered, alongside the early life development of migrants and descendants, in terms of disease profiles and social position. The authors conclude that the healthy migrant effect persists in some ethnic groups after the first generation of migration, but not in others. Social position is found to have a much greater influence on the mortality of descendants,

4.1.3 Family Migration

376. Family migration indicators were discussed but determined to be beyond the scope of this particular report.

377. The topics to be considered in relation to family migration include:

- Average time between arrival of first family member in the country and arrival of the last family member (i.e., time lag in family reunification)
- Economic outcomes of migrant families after X time in country

378. Family-level analysis in the longitudinal context is additionally complex. Individual persons are a constant unit of analysis, which cannot be divided or combined. Families, on the other hand, are not constant units of analysis as membership in a family can change over time. This makes defining a family longitudinally challenging and not without limitations. For example, a migrant family could be defined based on one member of the family. But this approach loses information whenever other members leave the family or new members join.

379. This changing-nature of families has an effect on any longitudinal indicators focusing on families as a unit of analysis. For example, the income of a family over time is directly related to the income of the members of the family. If an income-earner joins the family, the family income rises as a result of this addition.

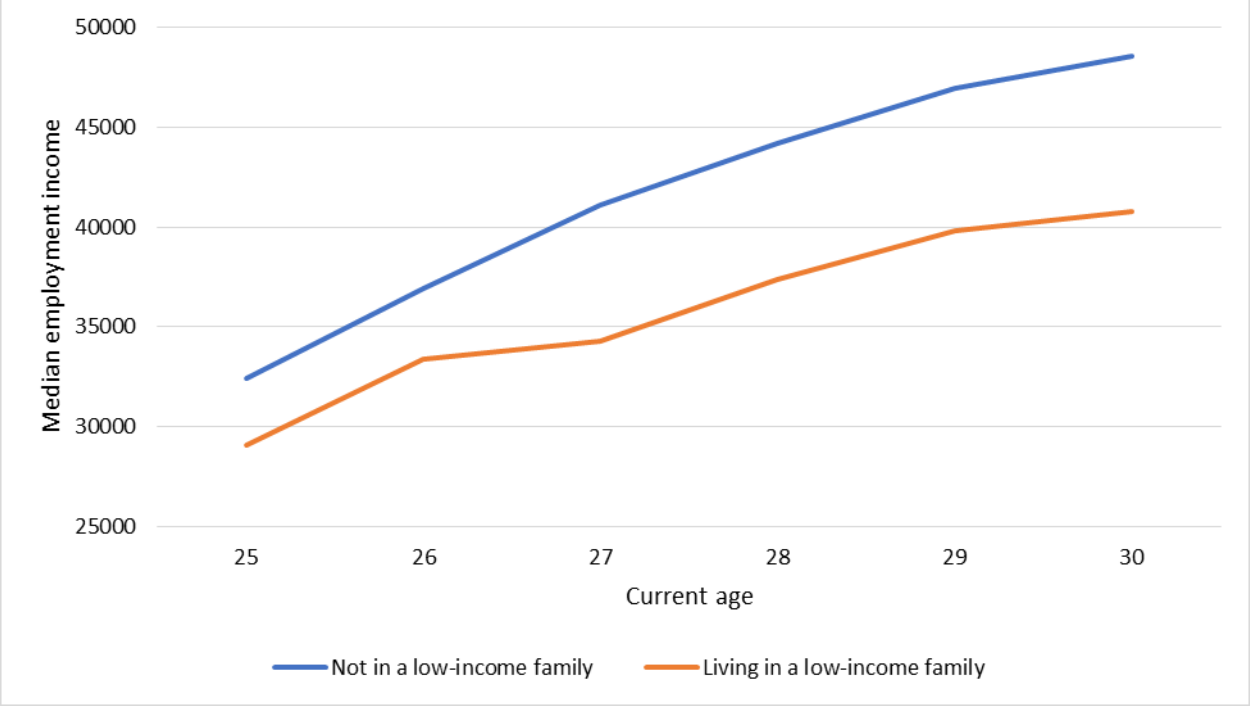
380. Defining families can be complex cross-sectionally and when referring to the general population but it can be even more so for migrants. If families are defined by those living in the same household, how are families in the process of reuniting considered? Do they only become a family once reunited?

Of the task force members, **Canada** does produce longitudinal data on migrants with family-based variables as part of the Longitudinal Immigration Database (IMDB). The IMDB includes individual- and family-level identifiers from both immigration administrative data and annual tax files. This database allows analysis of family composition over time with connections to individual immigrants and their outcomes in Canada. In 2019, Statistics Canada added a children's module to the IMDB allowing for analysis of the family economic outcomes of immigrant children. This module is then connected to the adulthood economic outcomes of these immigrants allowing analysis on intergenerational economic mobility.

The following analysis was produced using the IMDB Children module. One parent was selected as the principal parent. If the principal applicant was a filer, it was selected as the main parent, if not the spouse filer or the non-immigrant parent was. Analysis shows that the economic outcomes of immigrant children vary depending on their socioeconomic situation during childhood.

Children admitted before 5 years old and living in a low-income family two years after their admission had lower incomes than those who were not (Figure 33). The income gap increased as the immigrant children aged.

Figure 33 Median employment income of immigrants admitted to Canada before the age of 5 between 1980 and 1984 by current age and low-income status of family 2 years after admission (2016 dollars)



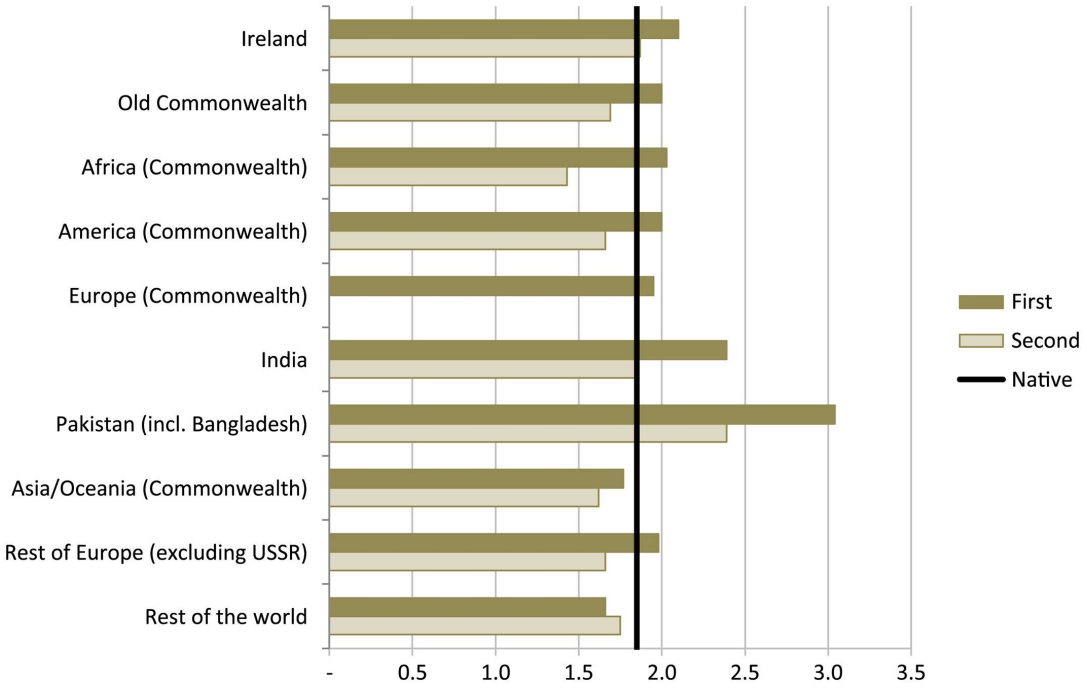
Low-income status: Identifies low-income individuals and families according to the low income measure (LIM). The LIM is one half of the adjusted median family income (taking into account the family size).

Source: Statistics Canada, Longitudinal Immigration Database (IMDB), 2016

In the **United Kingdom**, the Office for National Statistics Longitudinal Study has been used to explore family formation patterns for international migrants (Wilson and Kuha, 2017). This research used linked information on children's community population composition to assess its relationship with the completed fertility of different migrant groups. The study focusses on foreign-born women who arrived in England and Wales before the age of 16 and on the second-generation women who were born in England and Wales but who have at least one foreign-born parent. The analysis uses multilevel modelling to account for community characteristics, with specific community-level and individual-level variables as control variables. Using a multilevel model implies a comparison between immigrants and their descendants, on the one hand, and ancestral natives on the other.

The results show higher mean completed fertility for women who were child migrants compared to ancestral natives, though there is considerable variation by country of origin. The Pakistani and Bangladeshi group have a completed fertility level around 50 per cent higher than natives. This remains at 30 per cent higher for the second generation, while completed fertility for the second generation from other groups is lower than 1.85 (for ancestral natives). The multilevel model shows that child migrants are less likely to have the same fertility as natives if they are less likely to be exposed to native culture. The authors suggest that residential segregation explains some of the completed fertility for second-generation women from Pakistan and Bangladesh, which may suggest one reason why fertility of some South Asians in England and Wales may remain 'culturally entrenched'.

Figure 34 The completed fertility of different ancestry and generation migrant groups relative to ancestral natives.



Note: The figure shows the mean completed fertility for migrants (by generation and ancestry) relative to the average cumulative number of births for natives (equal to 1.85).

Source: Wilson & Kuha (2017), 'Residential segregation and the fertility of immigrants and their descendants', *Popul Space Place*. 2018; 24:e2098.

Family-level analysis of migrants is a challenging and complex topic. A more thorough review would be necessary to fully account for the different challenges and approaches which could be used to study migrant families over time.

4.2 Best practices for dissemination of longitudinal statistics

4.2.1 Different users of statistics

381. There is a wide range of users that rely on statistics. Their needs are often divergent, as are their skills to use the diverse available statistics (PARIS21 & Statistics Norway, 2009). Therefore, in order to ensure efficient communication, the target audience needs to be determined (Eurostat, 2014 & 2017). A broad distinction can be made between two major user groups:

- The specialists, who possess expert-level statistical knowledge and are able to process and interpret detailed data sets (e.g. statisticians, academia, specialized journalists and policy analysts);
- The non-specialist, i.e. people without or with limited statistical knowledge (e.g. citizens or the general public)

382. These two broad user groups have different needs and behavioural patterns as regards the use of statistics, which should be carefully considered in order to select appropriate communication channels (Eurostat, 2017). However, there is common agreement on the fact that, regardless of the target audience, the web should become the main vector of communication and diffusion. Many users are more skilled or at ease with computers or the Internet than with printed publications. Information published on the Internet are generally the most recent. The Internet is the also the quickest means to obtain information, in particular information from different sources. Users can easily and quickly download data or statistics and use them directly on their personal computer. This communication channel also allows national statistical offices to extend the dissemination of their data without generating high additional costs.

383. On the one hand, the primary interest of specialists is to receive precise and detailed information, including exact methodological definitions, presentation of the statistical trends, harmonized time series, detailed metadata, etc. (Eurostat, 2017). In this regard and given their familiarity with the topic and own ability to analyze related data, their needs are mainly for information, rather than for the explanatory techniques needed for the decoding of the statistical message by non-specialists.

384. Conversely, citizens normally do not have specialist knowledge in statistics. Their main need is thus high-level accessibility to the content of the indicators. Statistics are often perceived as a highly complex domain, out of reach for non-specialists. Therefore, in order to give statistics an appropriate role in the public debate, it is essential to find appropriate tools and channels to convey the relevant statistical information to the broader public of non-specialists, also taking into account its heterogeneity. Non-professional users are diverse. They consequently represent the group requiring the most varied communication channels. Within the general public, user groups such as university students, pensioners, families, representatives of the civil society, and generalist journalists can be considered. The media also have an important role in the sense that they distribute the information. Some general criteria should be met regarding the channels of indicator-based communication aimed at the general public. In these channels, the use of technical definition

should be limited and/or replaced by descriptions based on easy-to-understand language. In order to reach citizens, information should be presented in a way which has a direct relevance to people's everyday lives. Non-professional users need varied, but not excessively detailed statistical information.

385. Users can be national citizens, but they can also be international citizens. Both users might have different needs and interests.

4.2.2 Dissemination of longitudinal migration statistics

386. A dataset is longitudinal if it tracks the same type of information on the same subjects at multiple points in time. Conversely, cross-sectional data are data that are collected from respondents at one point in time. Therefore, while time is a major dimension in the first case, it is not in the second one. The primary advantage of longitudinal databases is thus that they can measure persistence or changes within individuals. Longitudinal data are also able to show how actions and events can affect outcomes later in life. This type of data are consequently particularly suitable when studying phenomena that evolve over time, such as migration. It can also allow researchers to examine how specific phenomena, such as migration, can affect later life outcomes, such as employment, living conditions, health, citizenship, etc. Presentations of longitudinal data should therefore focus on the most appropriate display for exploring evolutions.

387. Below are a few examples of how some countries have disseminated results based on migration longitudinal data that are the result of linked data.

4.2.2.1 Printed or online reports and/or publications

388. This channel of communication often contains a detailed description of the theoretical background, indicators, and assessment methods (Eurostat, 2014). The target audience is interested in an in-depth analysis of the respective issue and devote more time for it (Eurostat, 2017). They usually have background information on the subject and use information from analytical publications for reaching conclusion in their own research or for taking evidence-based policy decisions. They may or may not have an in-depth statistical knowledge. Most NSOs load their publications on their website, from which the publications can be downloaded free of charge. Publications, which are aimed for non-professional users, should not only contain tables. They should also have explanatory texts and graphs to guide them in their understanding of the data.

389. As an example, Canada has published various articles on diverse areas such as residence permits and earnings of migrants based on the Canadian Longitudinal Immigration Database (IMDB). Articles have ranged from short descriptive products (e.g. "Just the Facts: Asylum claimants": https://www150.statcan.gc.ca/n1/pub/89-28-0001/2018001/article/00013-eng.htm), in-depth analytical products (Bonikowska & Hou, 2015; Hou & Bonikowska, 2015; Prokopenko & Hou, 2016; Ci & Morissette, 2017; Lu & Hou, 2017), and reference materials. In 2018, a technical report was also published discussing 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 (Evra & Prokopenko, 2018). This range of products seeks to meet the needs of a broad range of stakeholders from those seeking basic stories to expert researchers making use of the data directly.

390. Italy has also published a working paper on residence permits and the acquisition of citizenship (Conti et al, 2014).

391. Switzerland has published a working paper describing the process used to create the Swiss Longitudinal Demographic Database (SLDD) (Steiner & Wanner, 2015). The paper concludes with two examples of possible applications: migration patterns and labour market integration. Wanner and Heiniger have published a second working paper in 2017 on the legal and technical developments in Switzerland that allowed the creation of the aforementioned database. It presents some results of two indicators that were constructed and calculated using this database: naturalisation and departure rates of migrant cohorts.

4.2.2.2 Data tables

392. Data tables' main function is to facilitate the comparison of different numbers (PARIS21 & Statistic Norway, 2009). Despite the increasing use of graphs, maps and other means of visualization, tables will certainly continue to be an important tool for presenting and disseminating statistics in the future. While the focus used to be on static tables, countries are now replacing them with more dynamic tools, where users can construct their own table views.

393. Alternatively, if these online tables are insufficient for user needs, there are also options to create custom tabulations based on user-specifications or provide direct access in secure facilities to anonymized data for more advanced users. Researchers need data for their work. Therefore, they have to be able to access easily online data that are well-documented and of high quality in order to improve the quality of their research (European Communities, 2005).

394. As an illustration, for the Longitudinal Immigration Database, Statistics **Canada** disseminates dynamic tables online, offers the option of customized tables and model output, and makes anonymized results available to researchers in secure Research Data Centres[10].

395. Online indicators – interactive aggregated dashboards

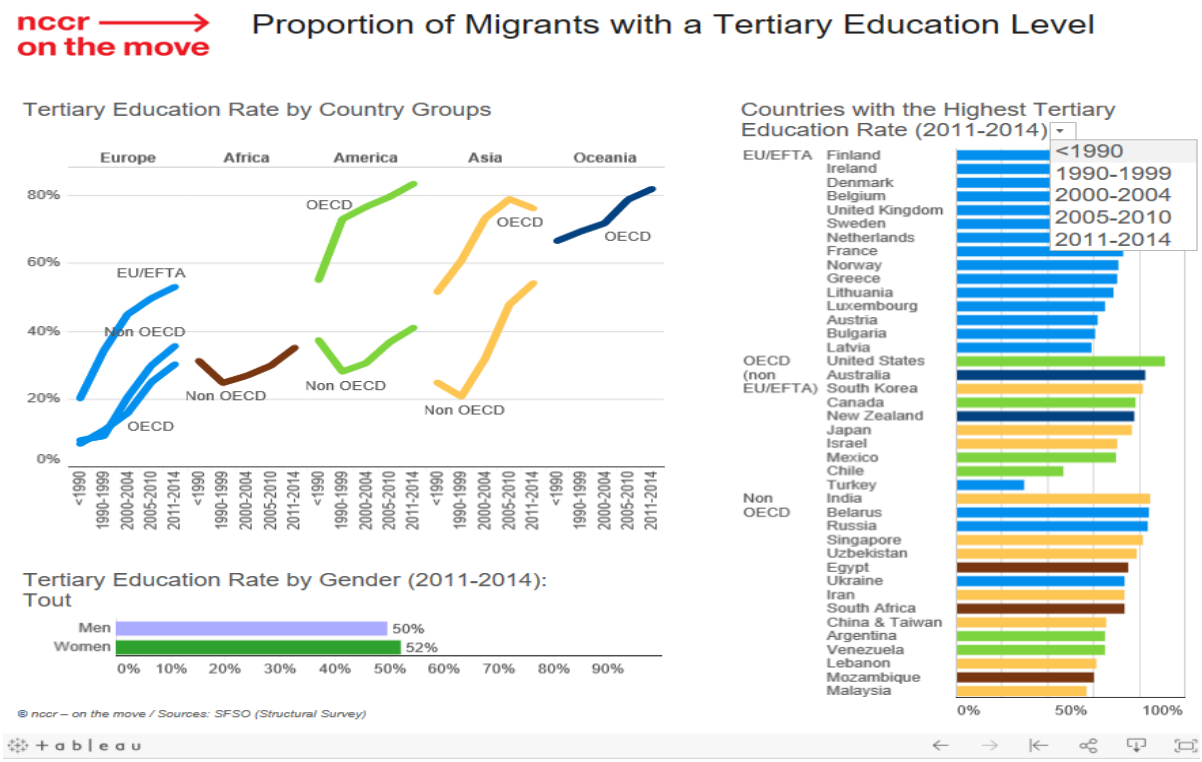
396. Information is provided in a layered manner. The use of a customizable interactive aggregated dashboard allows the user to get an overview of the situation (aggregated information) while at the same time providing access to detailed information (at the indicators' individual levels) (Eurostat, 2014). Also references to additional information concerning the theoretical background, assessment methods etc. must be provided. The target audience is interested in getting an overview of the situation (progress) and in having access to the underlying data. They are often early adopters of new technologies who prefer synthesized text-and-visual information rather than traditional publications (Eurostat, 2017). They usually have some background information on the subject which allows them to make us of the interactivity of technological products by setting their own content preferences.

397. In **Switzerland**, the National Centre of Competence in Research (NCCR) On the Move has developed ten longitudinal indicators based on the SLDD in diverse areas such as migration patterns, residence permits, naturalisation, (administrative) reasons for migration, qualification, asylum, etc. These longitudinal indicators can be consulted interactively on their website allowing personal parametrization[11]. They are therefore a tool to better understand migration and integration processes of migrant groups over a longer period of time. For example, regarding the interactive graph on the qualification of migrants in Switzerland[12], it is possible to select the countries with the highest tertiary education rate and the tertiary education rate by gender for different periods of time. Another interactive graph focusing on the administrative motives of entry in Switzerland

[10] https://www150.statcan.gc.ca/n1/en/surveys/5057
[11] https://indicators.nccr-onthemove.ch/
[12] https://indicators.nccr-onthemove.ch/how-qualified-are-migrants-in-switzerland/

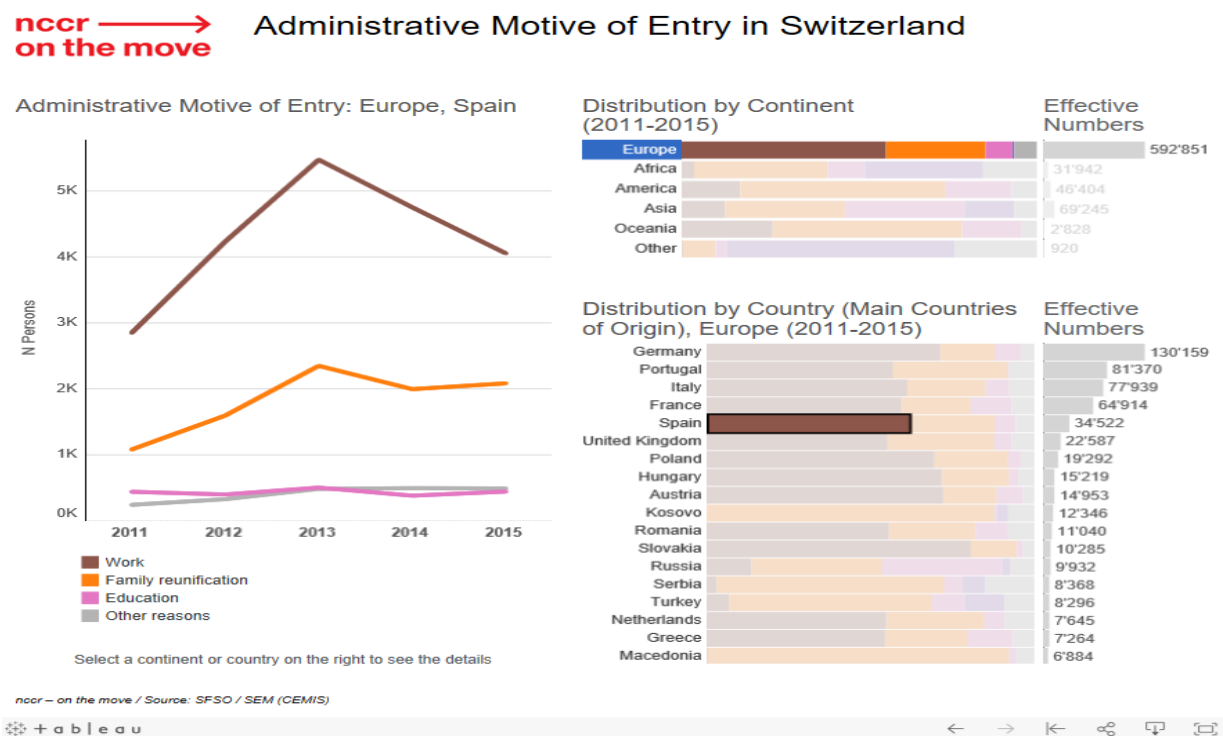
enables users to examine the evolution of the administrative motive of entry on a national level, but it is also possible to select a continent and a country[13]. Under the interactive graphs of each indicator, a small description of the data is always available. There is also some information regarding the source, the methodology, the definitions, etc.

Figure 35 Migrants' qualification in Switzerland



[13] https://indicators.nccr-onthemove.ch/for-what-administrative-reason-are-migrants-admitted-in-switzerland/

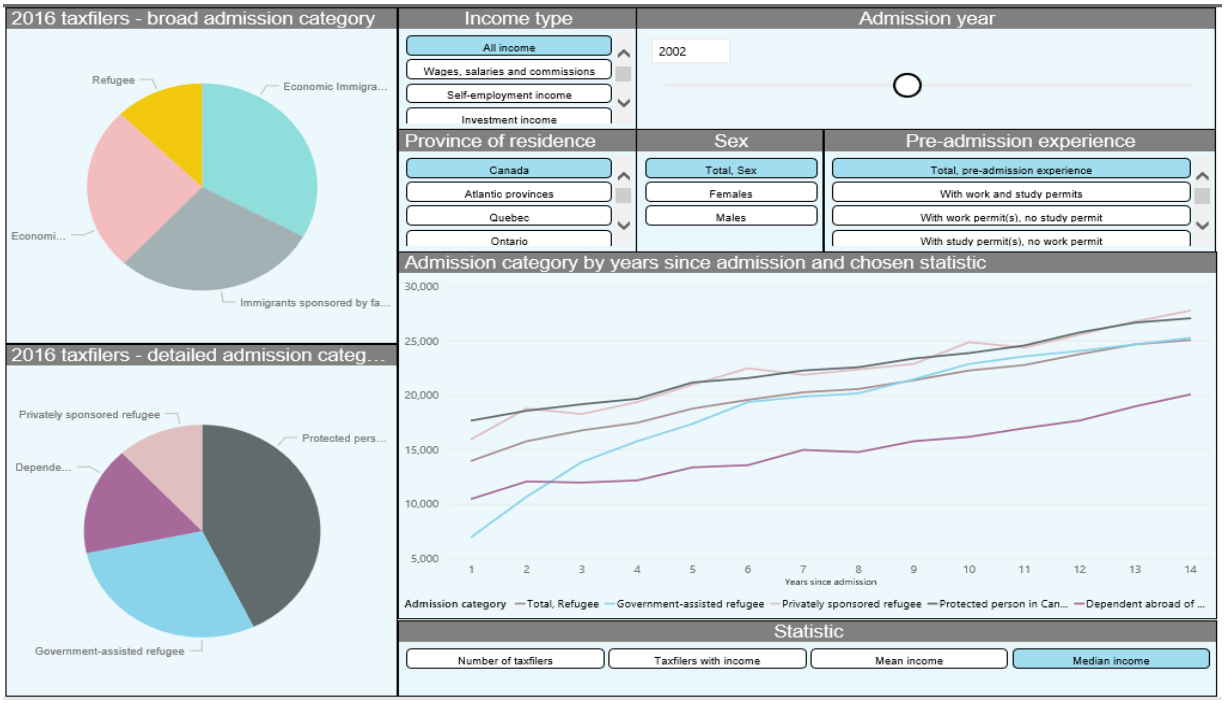
Figure 36 Administrative motive of entry in Switzerland



398. Statistics **Canada** publishes an interactive app using results from the Longitudinal Immigration Database (IMDB).[14] This interactive app connects directly to data cubes available online and displays the composition (by admission category) and income trends (by years since admission and by admission category) for immigrants based on a set of filters (year of admission, source of income, province of residence, sex, and pre-landing experience). It also allows users to easily look at differences in income trajectories between admission categories of immigrants controlling for different characteristics.

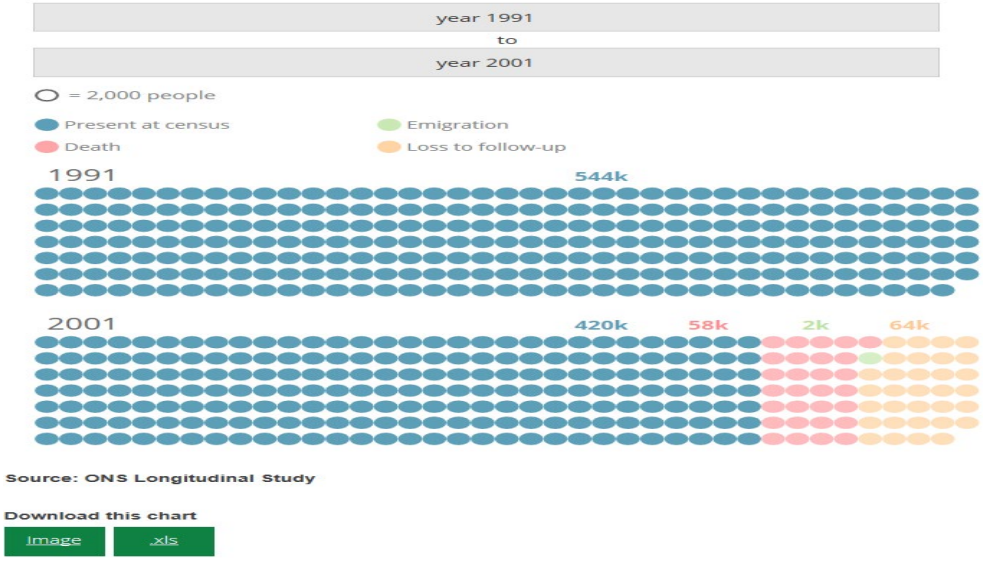
[14] https://www150.statcan.gc.ca/n1/pub/11-627-m/index-eng.htm

Figure 37 Interactive graph on compositions and income trends among immigrants in Canada



399. The **UK** does not use interactive dashboards to present results, but it uses an online cohort interactive tool to show what happens to their Longitudinal Study (LS) members over time.[15] A beginning date and an end date can be selected. Exits from the LS occur by two means: death or emigration. The quality of the death data within the LS is very high as death registration is required by law. However, emigration data is of less good quality as it relies on LS members informing the National Health Service (NHS) of their departure, causing gaps in the coverage of the data. This largely explains the number of individuals categorised as lost to follow-up.

Figure 38 Cohort interactive tool on LS members



[15] https://www.ons.gov.uk/aboutus/whatwedo/paidservices/longitudinalstudyls

4.2.2.3 Audio-visual productions

400. Videos target users who prefer video communication, have no time for long readings and/or need to be informed about the latest trends and developments or basic features of a respective issue (Eurostat, 2017). They usually use the videos as a general source of information with no specific purposes of using the information in their own analysis or research.

401. For example, the NCCR On the Move in **Switzerland** created short YouTube videos and posted them on their website[16]. At time of writing, five videos have been posted based on longitudinal data with the aim of addressing a few preconceptions about migrants such as:

- migrants come to take advantage of Switzerland;
- migrants come from poor countries;
- with that many arrivals, soon there will be more migrants than Swiss people;
- migrants only care about their country of origin, not Switzerland; and
- migrants all want a Swiss passport.

4.2.2.4 Infographics

402. According to the Oxford English Dictionary, an infographic is a "visual representation of information or data". An infographic is a collection of imagery, charts, and minimal text meant to capture attention and give an easy-to-understand overview of a topic. Infographics are great for making complex information easy to digest. They can be used to share statistics and census data. They are more focused on numbers, charts and data. They tend to contain much less text than informational infographics and have less of a narrative flow. Instead, they make statement with big numbers and standalone facts which can easily be communicated through new modes such as social media. The target audience of infographics is very wide; going from students, professionals to the general public. No country has yet published infographics using longitudinal migration data. **Switzerland** has published a few infographics on migration and integration using transversal data. [17]

[16] https://indicators.nccr-onthemove.ch/overcoming-preconceptions/
[17] https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/infographics.html?dyn_prodima=900999&dyn_publishingyearend=2019

Chapter 5: Conclusions and recommendations

5.1 Conclusions

403. To study migration is to study change: change in residence, change in legal status, and change in socio-economic outcomes. Furthermore, the related topics of integration and settlement are processes – not states. Outcomes for migrants can be short-term or long-term.

404. In order to produce statistics on migration and related topics, data measuring change must be leveraged. Longitudinal data provides the means to understand change at the individual level. Beyond simply observing changes in residence or outcomes, following individuals over time permits a better understanding of factors which can affect change. These factors may operate on the macro level, such as policy changes and economic shifts, or on the micro level, such as obtaining education or training and acquiring citizenship.

405. Researchers and official statisticians are increasingly using methods based on longitudinal approaches to better understand migratory flows and the dynamics of integration in host countries. Longitudinal approaches are the most suitable for measuring how the gap between immigrants and native population changes with time elapsed since immigration, for instance in terms of wages or unemployment.

406. With the increase in availability of administrative data, countries can now better leverage techniques such as data integration to be able to develop detailed longitudinal data for migration statistics. Analyses of integrated administrative data can answer questions that require large sample sizes with rich and detailed data on hard-to-reach populations and generate evidence with a high level of external validity and applicability for policy making. This approach allows the longitudinal observation of migration, integration or settlement outcomes without increasing the burden of data collection through surveys.

407. This Guidance shows how longitudinal data could be developed and used for international migration statistics. Chapter 2 provides an overview of types of data ranging from panel surveys to population registers. Chapter 3 guides how to develop longitudinal data sources from integrated data. Finally, Chapter 4 proposes key longitudinal indicators related to the study of international migration and best practices for dissemination.

408. While longitudinal data provides a critical and unique opportunity to study migration, integration and settlement as processes, the present Guidance recognizes the challenges and limitations associated with such data. These challenges vary by type of data source as described in chapter 2.

409. Traditional methods such as longitudinal panel surveys have advantages in the control offered to statistical offices with respect to coverage and measurement. However, due to operational considerations, these sources tend to have smaller sample sizes, suffer from attrition, and lack frequency in collection. These limitations can hinder their use for studying short- or long-term outcomes for small populations (e.g. subsets of migrants).

410. On the other hand, there are also limitations associated with the use of integrated administrative data. In particular, these data have not been collected for statistical purposes and so control is lost over coverage and measurement. Additionally, the integration process can introduce errors as discussed in chapter 3. However, these methods tend to yield large sample sizes with more

regular follow-up and little to no attrition due to non-response. As a result, they are, generally, well positioned to be used to study small populations of migrants longitudinally.

411. There are methods available to mitigate, address, or work-around the limitations of the data sources so that what is disseminated still serves the purposes outlined in the statistical design. Chapter 3 provides various best practices throughout the development of the data sources while chapter 4 illustrates how valuable indicators can be produced despite limitations with the resulting database.

412. It is important not to disregard the challenges but it is also important not to be obstructed by them. It is critical is to understand whether any particular statistics are fit for purpose. Integrated data for longitudinal migration statistics will always come with its challenges but it also opens the door to unprecedented analytical opportunities shedding light on detailed longitudinal outcomes of migrants and their impact on host countries. These data sources may not always yield perfect cross-sectional point estimates but they offer something richer – a story.

413. The power of longitudinal data for migration statistics is clearly illustrated through the examples of chapter 4. In the end, the benefits of longitudinal data for migration statistics suggest that the challenges must be overcome.

5.2 Recommendations

414. For a better understanding of migration and of the processes of integration and settlement in particular, it is recommended to increase the use of longitudinal data for migration statistics.

415. In developing a longitudinal dataset, it is recommended to follow each of the six phases described in chapter 3 in the presented order. Depending on outcomes, it may be necessary to return to one of the earlier phases, in which case the subsequent phases would need to be repeated. The phases are the following:

1) Statistical design
2) Assessment and pre-processing of source files
3) Data integration for longitudinal data
4) Assignment of longitudinal individual identifiers
5) Create final database
6) Disseminate results

416. It is recommended that countries regularly produce the longitudinal indicators for migration statistics described in Chapter 4.

5.3 Areas of future work

417. This report covered a wide range of migration-related topics which can benefit from longitudinal data. Some topics were addressed in more detail, especially in chapter 4 through the proposed indicators. However, others were deemed to be outside the scope of this Guidance.

418. In particular, further investigation would be necessary on how to measure **family-related migration**. Due to its complexity, this topic may require a more focused examination. This could include a study of subtopics such as:

- Migration patterns of families
- Migrant families as units of analysis (e.g. settlement and integration outcomes of migrant families over time)
- Family reunification and family members abroad
- Living arrangements of migrants in host country (e.g. multi-family households, multi-generational households)

419. Another topic that warrants further attention is that of **subnational settlement patterns** discussed in section 4.1.1.2 Length of stay in the subnational geography. However, increasingly countries are interested in considering the composition of neighbourhoods and encouraging migrants to settle in non-traditional locations (e.g. for economic reasons). Understanding the low-level geographic patterns can be important for stakeholders such as service providers and schools. However, as discussed through this Guidance, particularly in chapter 4, there are unique challenges associated with this type of study. In particular, boundaries fluctuate over time and sometimes a small distance movement could be considered a change in geography while a long distance movement could remain in the same geographic area.

420. Another emerging area is the ways statistical offices **disseminate results**. Some innovative approaches are presented in chapter 4. With new dissemination tools available, and with more complex data to disseminate, statistical offices are being challenged to reconsider how best to present findings to various audiences. For migration statistics, this is a particularly important challenge. With the recent increase in international migration, there is a broad ongoing public discourse across countries which demands evidence on migration patterns and the impact of migration on individuals, families, and societies. Because migration statistics are complex by nature, there is also a risk of statistical results being misinterpreted. More guidance on how to present findings to reach various audiences while respecting and acknowledging the data quality limitations could assist national statistical offices with this emerging challenge.

References

Aoki, Y. & Santiago, L. (2015). *Education, Health and Fertility of UK migrants: The Role of English Language Skills*. IZA Discussion Paper Series No.498 (ONS LS), available at: https://calls.ac.uk/wp-content/uploads/dp9498.pdf

Berti-Equille, Laure. (2007). *Measuring and Modelling Data Quality for Quality-Awareness in Data Mining*. 10.1007/978-3-540-44918-8_5.

Blackwell, L. & Rogers, N. (2019). *Placing administrative data at the heart of the UK population statistics system; how thte total survey error framework is helping to inform the use and design of integrated data solutions.* Retrieved from International Total Survey Error Workshop 2019: https://www.niss.org/news/itsew-2019-international-total-survey-error-workshop.

Bonikowska, A. & Hou, F. (2015). *Which human capital characteristics best predict the earnings of economic immigrants?*. Analytical Studies Branch Research Paper Series*,* 368, 1-30.

Burke, D. L., Bujkiewicz, S. and Riley, R. D. (2016). *Bayesian bivariate meta-analysis of correlated effects: Impact of the prior distributions on the between-study correlation, borrowing of strength, and joint inferences*. Statistical Methods in Medical Research, 0(0) 1–27.

Chambers, R., Tzavidis, N., and Salvati, N. (2009). *Borrowing strength over space in small area estimation: Comparing parametric, semi-parametric and non-parametric random effects and M-quantile small area models*. Retrieved from University of Wollongong Research Online: https://pdfs.semanticscholar.org/eede/9fa86149d5cc312da377dd24d63ab723aa4e.pdf

Ci, W., & Morissette, R. (2017). *Acquisition of permanent residence by temporary foreign workers in Canada: A panel study of labour market outcomes before and after the status transition*. Analytical Studies Branch Research Paper Series*,* 396, 1-31.

Conti, C., Quattrociocchi, L., and Rottino, F. M. (2014). *New statistics on residence permits and acquisition of citizenship: experiences of integration in Italy*. Economic Commission for Europe, Group of Experts on Migration Statistics, Work Session on Migration Statistics, Working paper 14, Geneva, 10-12 September 2014.

Correa-Onel, S., Whitworth, A. and Piller, K. (2016). *Assessing the Generalised Structure Preserving Estimator (GSPREE) for Local Authority Population Estimates by Ethnic Group in England*. GSS Methodology Series No 42, Office for National Statistics: Newport.

Eberle, J., (2018). *Foreigners seeking humanitarian protection. Measuring the stock of foreigners seeking humanitarian protection in Germany*. Note by German Federal Statistical Office, United Nations Economic Commission for Europe, Conference of European Statisticians, Work Session on Migration Statistics. Geneva, Switzerland 24-26 October 2018.

European Communities (2005). *The challenge of communicating statistics*. Proceedings of the 91st DGINS Conference, Copenhagen. Retrieved from: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/KS-EE-05-001.

EUROSTAT (2014). *Getting messages across using indicators. A handbook based on experiences from assessing Sustainable Development Indicators*. Luxembourg: Publications Office of the European Union.

EUROSTAT (2017). *Towards a harmonized methodology for statistical indicators. Part 2 – communication through indicators*. Luxembourg: Publications Office of the European Union.

Evra, R., & Prokopenko, E. (2018). *Longitudinal Immigration Database (IMDB) Technical Report*. Analytical Studies: Methods and References, 11, 1-58.

Frenette, M., Lu, Y., and Chan, W. (2019). *The Postsecondary Experience and Early Labour Market Outcomes of International Study Permit Holders*. Analytical Studies Branch Research Paper Series, Statistics Canada.

Globerman, S., Barua, B., and Hasan, S. (2018). *The Supply of Physicians in Canada: Projections and Assessment*. Fraser Institute.

Green, D., Liu, H., Ostrovsky, Y., and Picot, G. (2016). *Business Ownership and Employment in Immigrant-owned Firms in Canada*. Economic Insights Catalogue no. 11-626-X – No. 057.

Groves, R., Fowler, F., Couper, M., Singer, E., and Tourangeau, R. (2004). Survey Methodology , New York Wiley.

Hou, F., & Bonikowska, A. (2015). *The earnings advantage of landed immigrants who were previously temporary residents in Canada*. Analytical Studies Branch Research Paper Series, 370, 1-36.

Jackson, D., White, I. R., Price, M., Copas, J., and Riley, R. D. (2017). *Borrowing of strength and study weights in multivariate and network meta-analysis*. Statistical Methods in Medical Research, 26(6), 2853–2868.

Johnston, T., & Weiss, R. (2010). *Managing Time in Relational Databases*. Elsevier

Johnston, T. (2014). *Bitemporal Data: Theory and Practice*. Elsevier

Lu, Y. & Hou F. (2017): *Transition from temporary foreign workers to permanent residents, 1990 to 2014*. Analytical Studies Branch Research Paper Series, 389, 1-32.

Ng, E., Sanmartin, C., Elien-Massenat, D., Manuel, D. (2016). *Vaccine-preventable disease-related hospitalization among immigrants and refugees to Canada: use of two population-based Immigrant linked cohort studies*. Vaccine 34 (2016):4433-4442.

Ng, E., Elien-Massenat, D., Giovanazzo, G., Ponka, D., and Sanmartin, C. (2018). *Tuberculous related hospitalization among recent immigrants; a linkage study*. Health Reports, Vol 29 no 7, 14-28.

PARIS21 & Statistics Norway (2009). *User-friendly presentation of statistics. Guide to creating a dissemination strategy and dissemination guidelines for developing and transition countries*. Oslo: Statistics Norway.

Perrin, N. (2006). *A Cohort Approach to Acquisition of Citizenship Statistics*. In: Poulain, M., Perrin, N., and Singleton, A. (Ed.): THESIM: Towards Harmonised European Statistics on International Migration, Presses Universitaires de Louvain.

Prokopenko, E., & Hou, F. (2016). *How temporary were Canada's temporary foreign workers?*. Analytical Studies Branch Research Paper Series, 402, 1-31.

Reichel, David (2011). *Do Legal Regulations Hinder Naturalisation? Citizenship policies and naturalisation rates in Europe*. EUDO Citizenship Observatory.

Prins, K. (2016). *Population register data, basis for the Netherlands' population statistics*. Statistics Netherlands, Bevolkingstrends Januari 2016.

Statistics New Zealand (2016). *Guide to reporting on administrative data errors.* Statistics New Zealand, Wellington. Retrieved from: http://archive.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality/sources-of-error.aspx#

Steiner, I., & Wanner, P. (2015). *Towards a new data set for the analysis of migration and integration in Switzerland*. Working Papers – nccr on the move, 1, 1-22.

UK Home Office (2019). *Migrant Journey: 2018 Report.* Statistical Bulletin 07/19. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/803754/migrant-journey-report2018.pdf.

UK Office for National Statistics (2019). *International Passenger Survey 3.15*. Retrieved from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/datasets/internationalpassengersurveyactuallengthofstaybycitizenshiptable315.

UNECE (2015): *Measuring change in the socio-economic conditions of migrants*. New York and Geneva: United Nations. Retrieved from UNECE: http://www.unece.org/index.php?id=40542.

UNECE (2016): *Defining and measuring circular migration*. New York and Geneva: United Nations. Retrieved from UNECE: http://www.unece.org/index.php?id=44717.

UNECE (2017). *A set of tables for circular migration*. UNECE Work Session on Migration Statistics, Geneva, 30-31 October 2017.

UNECE (2018a). *Circular migration: new migration topics and revised tables*. UNECE Work Session on Migration Statistics, Geneva, 24-26 October 2018.

UNECE (2018b). *Duration of migratory episodes in Spain: Procedure of calculation and results*. UNECE Work Session on Migration Statistics, Geneva, 24-26 October 2018.

UNECE (2019a). *Generic Statistical Business Process Model, version 5.1.* Retrieved from UNECE: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

UNECE (2019b). *Guidance on data integration for measuring migration.* New York and Geneva: United Nations.

United Kingdom Home Office (2019). *User guide to the Home Office statistics on exit checks.* Retrieved from UK Government Publishing Service: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826370/user-guide-to-statistics-on-exit-checks.pdf

Vachon, J., Gallant, V., and Siu, W. (2018). *Tuberculosis in Canada, 2016*. Canada Communicable Disease Report 2018; 44(3/4): p. 75 – 81.

Vanthomme, K., and Vandenheede, H. (2019). *Trends in Belgian cause-specific mortality by migrant origin between the 1990s and the 2000s*. BMC Public Health 19:410.

Wallace, M. (2016). *Adult mortality among the descendants of immigrants in England and Wales: does a migrant mortality advantage persist beyond the first generation?*. Journal of Ethnic and Migration Studies, 42:9, 1558-1577, DOI: 10.1080/1369183X.2015.1131973. Available at: https://www.tandfonline.com/doi/full/10.1080/1369183X.2015.1131973.

Wanner, P., & Heiniger, M. (2017). *Migration and structural integration in Switzerland. A longitudinal perspective*. Economic Commission for Europe, Group of Experts on Migration Statistics, Work Session on Migration Statistics, Working paper 7, Geneva, 5-6 October 2017.

Wilson, B., & Kuha, J. (2017). *Residential segregation and the fertility of immigrants and their descendants*. Popul Space Place. 2018; 24:e2098. Available at: https://doi.org/10.1002/psp.2098

Zhang, L.-C. (2012). *Topics of Statistical Theory for Register-Based Statistics and Data Integration*. Statistica Neerlandica 66: 41 – 63.

Zuccotti, C. V., & Platt, L. (2016). *A reconsideration of ethnic penalties in inactivity and unemployment: a study of second generation men and women in England and Wales*. ISA RC28 Summer Meeting, University of Bern, Switzerland, 29 – 31 August 2016 [ONS LS].