United Nations ECE/CES/2020/8

Economic and Social Council

Distr.: General
31 March 2020

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-eighth plenary session
Geneva, 22-24 June 2020
Item 4 (f) of the provisional agenda
Reports, guidelines and recommendations prepared under the umbrella of the Conference: Longitudinal data on migration

Guidance on the use of longitudinal data for migration statistics[1]

Note by the Task Force on the use of longitudinal data for migration statistics

Summary

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.

The current short version of the Guidance is prepared for translation purposes. It includes the introduction, introductory sections of all substantive chapters and the full chapter with conclusions, recommendations and further work.

The full text of the Guidance has been sent to all members of the Conference of European Statisticians (CES) for electronic consultation. It is available at the Conference webpage: http://www.unece.org/index.php?id=53381.

Subject to a positive outcome of the consultation, the CES plenary session will be invited to endorse the Guidance.

[1] This document was scheduled for publication after the standard publication date owing to circumstances beyond the submitter's control.

GE.20-04744(E)
2004744

I. 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.

II. Chapter 2: Overview of longitudinal data sources for migration statistics

A. 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.

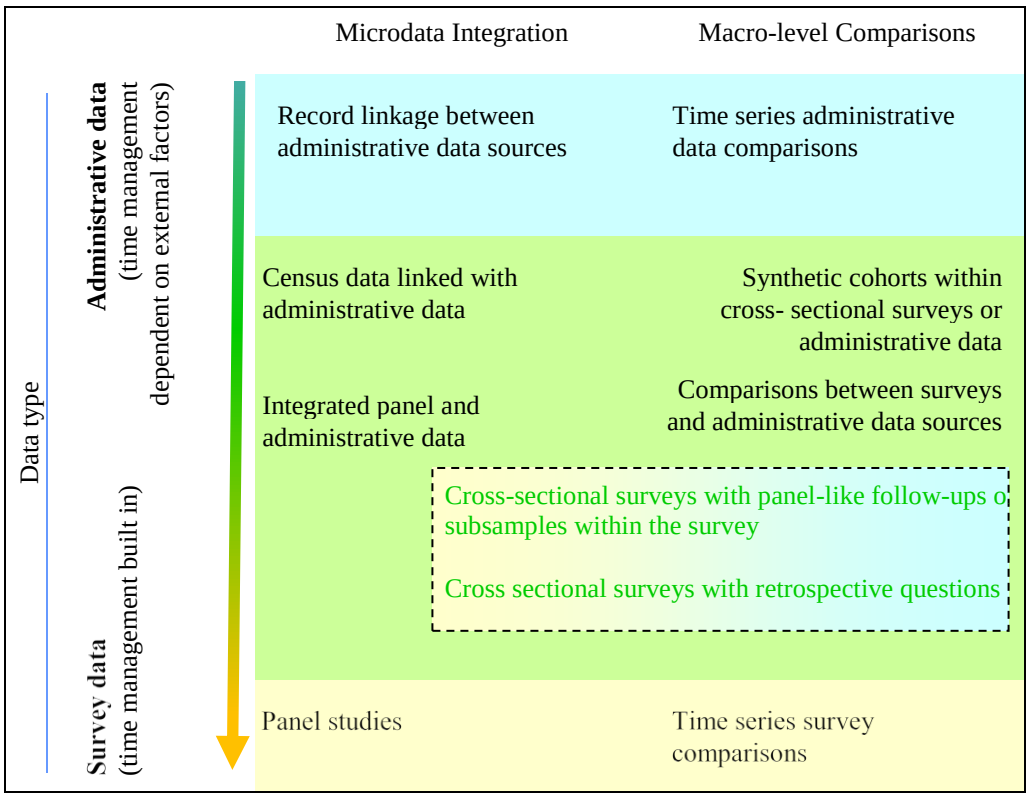
19. Examples would be panel designs embedded within cross-sectional surveys. A further example is retrospective questions to yield longitudinal evidence, within a cross-sectional survey design. Synthetic cohorts in successive censuses or surveys mimic longitudinal studies, in some ways, but do not follow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

B. 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.

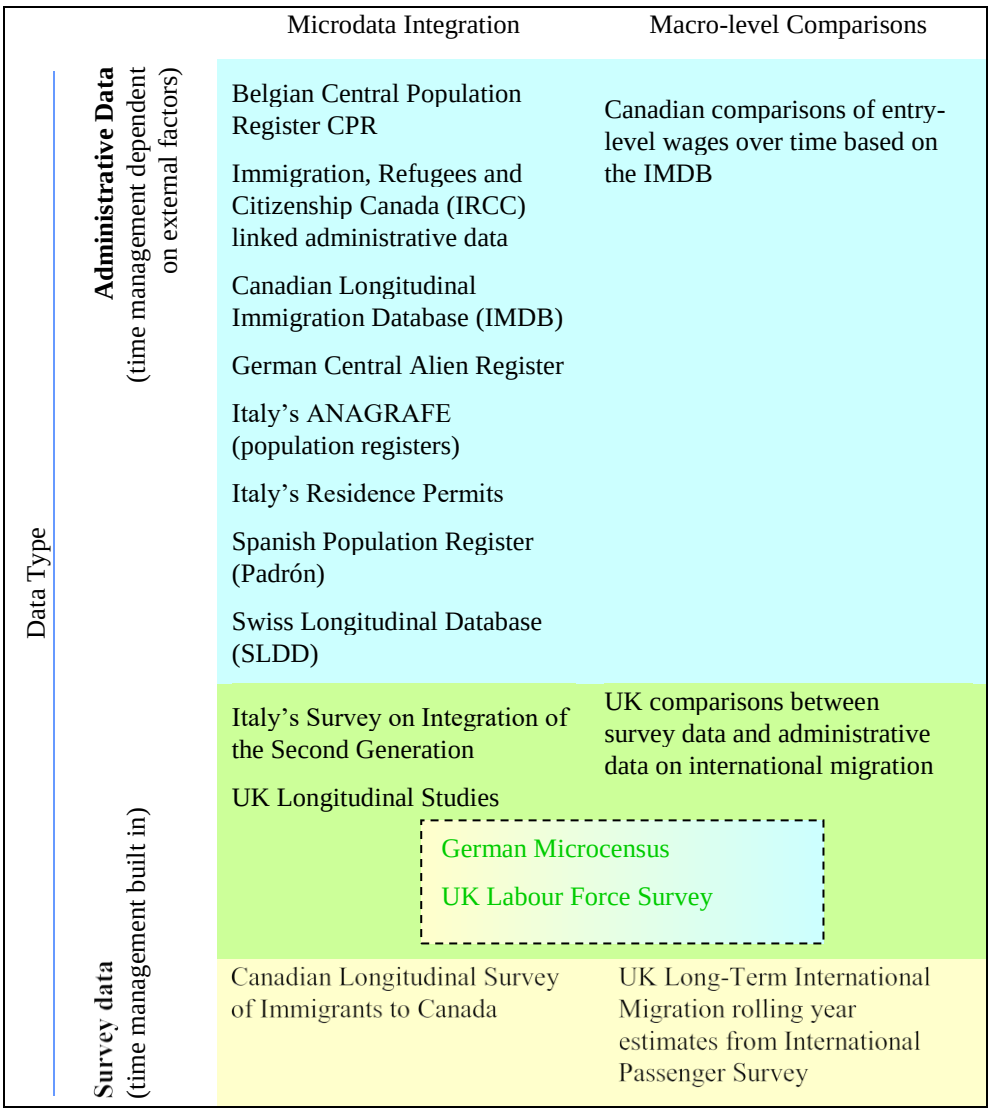
27. A combined approach can take a number of forms. It can take survey information at the beginning and update this with administrative records. Table 1, below, describes the creation of longitudinal data from the linkage of census and administrative records, in the UK Longitudinal Studies. A further 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.

Current practices

29. 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



C. Summary

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

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

A. Introduction

35. 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.

36. 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.

37. The phases of developing a longitudinal data set for migration statistics based on integrated data 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.

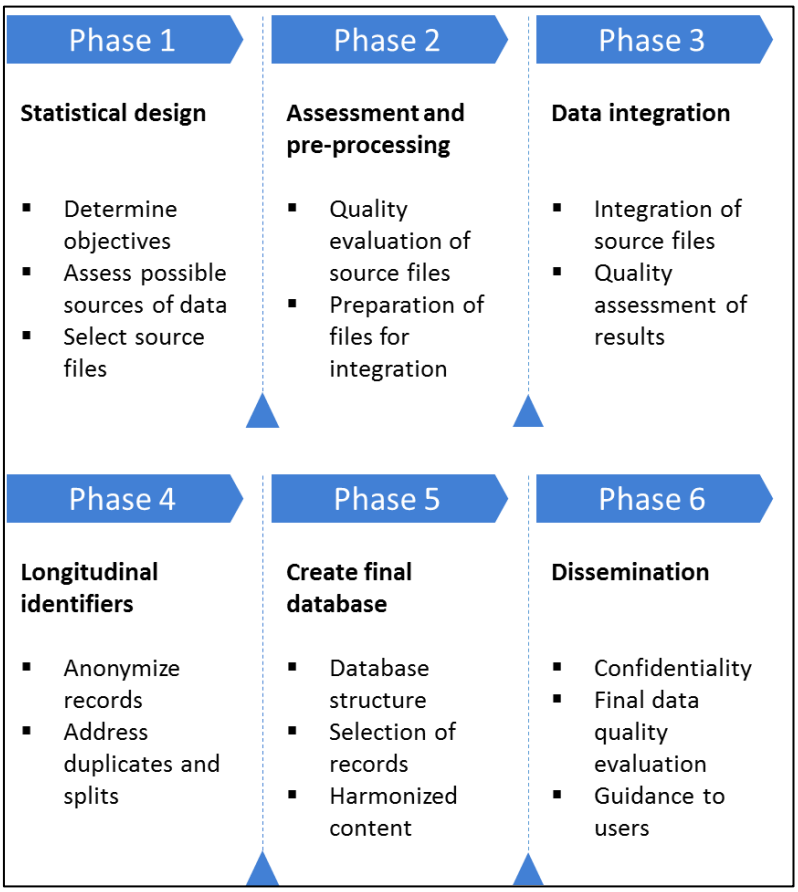
38. This chapter provides guidance on each of these phases. They must be considered both in the context of the development of a new database as well as in the context of updating an existing one (e.g. to add new outcomes via data integration).

39. 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.

40. 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 3
Phases of developing a longitudinal data set



B. Error framework for longitudinal data from integrated data sources

41. 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 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.

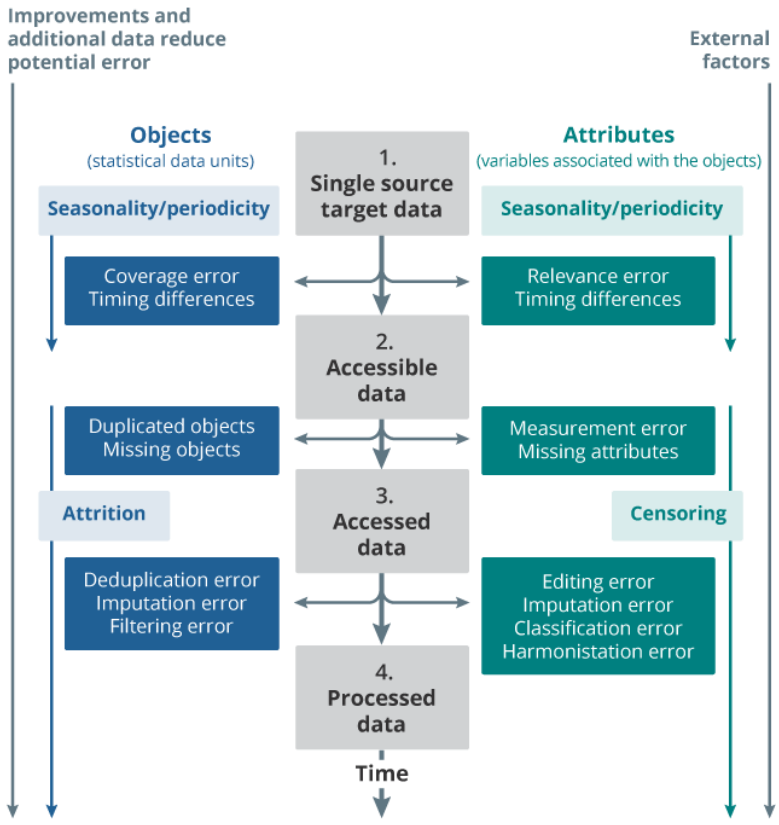
42. It is useful to draw on the Total Survey Error Framework which has been developed to support statistical design(Groves, 2004), (Zhang, 2012). This provides a useful taxonomy of the different potential sources of error, relating mainly to administrative data. Not all types of error are measurable. 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.

43. 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.

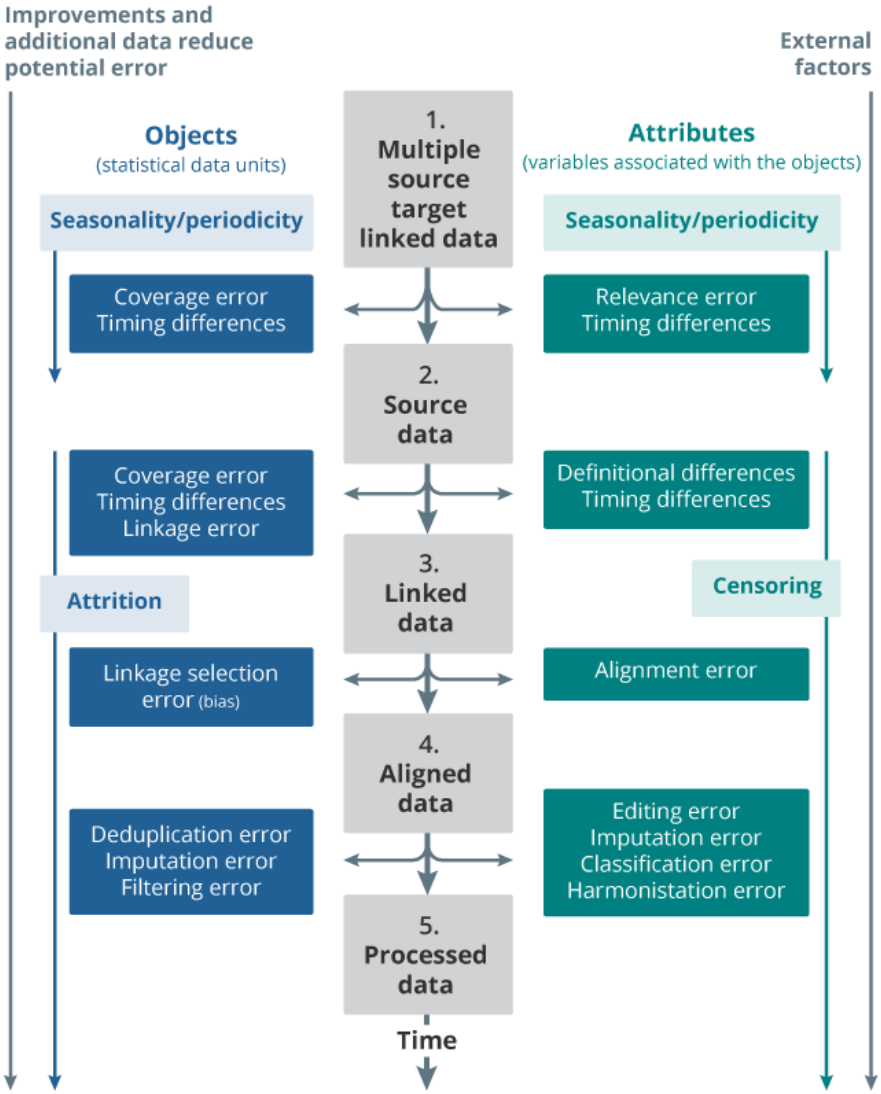
44. Figure 4 and Figure 5 describe the error frameworks for single and multiple sources. Table 3, Table 4 and Table 5 define the errors at each stage. The single source framework identifies four different stages of the data journey – target data, accessible data, accessed data and processed data. Errors 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 4
Single source error framework



Note: 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 5
Multiple source error framework



45. 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.

46. 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.

47. 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).

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

Table 1
Single source errors

Objects[1]	**Attributes**[2]
Frame error:	*Relevance error:*
Coverage error Assessing objects that are not in the target data, or not being able to access objects that are in the target data.	*Validity error* The difference between ideal measurement of attributes sought about an object and the operational measure used to collect it.
Timing differences Objects in the ideal target data that are not accessible because of a discrepancy in the time window for obtaining observations.	*Timing differences* 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 Objects that are represented more than once in the accessed data.	*Measurement error* Errors arising from attributes that are not recorded accurately.
Missing objects Objects that *in theory* are accessible but are not in the accessed data.	*Missing attributes* Attributes that are missing from the accessed data (could be for specific objects or all of the objects).
Processing error:	*Processing error:*
Deduplication 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).	*Editing 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 Errors arising from the imputation of missing objects.	*Imputation error* Errors arising from the imputation of missing attribute values.
Filtering error Errors arising from the selection or de-selection of accessed objects to an ideal target set.	*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

[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

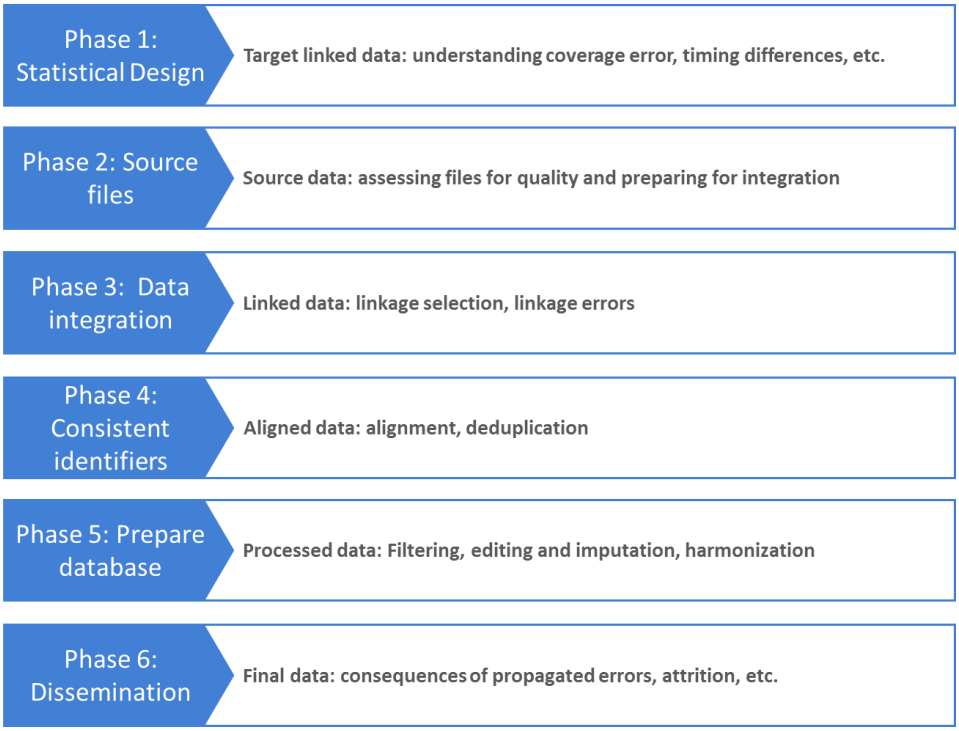
Table 2
Longitudinal error – applies to both single source events and multisource longitudinal datasets

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

49. The multiple source framework (Table 3) 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 6.

Figure 6
Stages of integration of multiple data sources



50. 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.

51. 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 3
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*
Objects are not observed due to conceptual discrepancies in the timing of the capture between the target linked data and source data.	A conceptual discrepancy in the timing of the measurement of attributes between the target linked data and the source data.
Coverage error:	***Mapping error:***
Coverage error	*Definitional differences*
Objects are not linked due to discrepancies in the coverage of objects between data sources.	The differences between how attributes are operationally measured in each of the source datasets.
Timing differences	*Timing differences*
The difference between observed objects in source datasets due to the data being captured at different times.	The differences between the values of attributes for a linked object between source datasets caused by 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:	*Comparability error:*
Linkage selection error (bias)	*Alignment error*
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.	Errors arising from the alignment of the conflicting values of attributes across sources.
Processing error:	*Processing error:*
Imputation error	*Editing error*
Errors arising from the imputation of missing objects.	Errors arising from editing the value of an attribute. This could include editing as a result of validation or QA checks.
Filtering error	*Imputation error*
Errors arising from the selection or de-selection of accessed objects to an ideal target set.	Errors arising from the imputation of missing attribute values.
	Classification error:

Objects[1]	Attributes[2]
	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.
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

52. 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 **Error! Reference source not found.** in the full version of the Guidance).

53. 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.

54. 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.

IV. Chapter 4: Disseminating regular migration statistics from longitudinal data sources

55. This chapter provides guidance on how regular migration statistics can be disseminated using longitudinal data sources. It 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.

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

57. 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.

58. 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.).

59. 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.

60. 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.

61. 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.

62. 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.

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

- Migration patterns
- Socio-economic outcomes
- Family migration.

64. 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 4
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

Category	*Topic*	*Indicator*
	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
		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
Socio-economic outcomes		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

Category	*Topic*	*Indicator*
		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)
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

V. Chapter 5: Conclusions and recommendations

A. Conclusions

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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).

72. 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.

73. 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.

74. 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.

75. 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.

B. Recommendations

76. 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.

77. 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:

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

78. It is recommended that countries regularly produce the longitudinal indicators for migration statistics described in Chapter 4.

C. Areas of future work

79. 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.

80. 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)

81. Another topic that warrants further attention is that of subnational settlement patterns discussed in section **Error! Reference source not found.** of the full version of the Guidance - **Error! Reference source not found.**. 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.

82. 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.