Proof-of-concept for the Longitudinal Population Dataset for England and Wales

Version: 2.0

Date: 18 December 2024

Authors: Stephan Tietz, Louisa Blackwell, Neus Beascoeachea-Segui, Paola Signoretta, Mike Bracher, Merilynn Pratt, Francis Ongondo and Kevin McCafferty

Revision History

Version	Date	Owner	Summary of changes
1.0	01/11/2024	Stephan Tietz	Final version for MARP panel
2.0	18/12/2024	Stephan Tietz	Final version for publication

Version 2.0 of this paper was created to incorporate our responses to feedback we received from MARP panel members before our presentation on 12 November 2024 and which we answered during our presentation. This paper therefore represents an accurate account of the work on this date.

1 Purpose

As part of the Future of Population and Migration Statistics program, ONS is currently producing a proof-of-concept for the Longitudinal Population Dataset (LPD) for England and Wales. The LPD was previously referred to as the 100% Longitudinal Cohort, 2021 Census Cohort Study (EAP178, Feb 2023) and Census Data Asset (ONS, Dec 2023).

This paper:

- recaps the rationale for and purpose of the LPD,
- outlines the scope of the LPD proof-of-concept,
- provides an update on the progress we have made towards the LPD proof-of-concept including the key design decisions taken and the methods we are exploring,
- discusses the opportunities for scale benefits by supporting the rapid production of longitudinal assets (referred to as satellite cohorts) and how to ensure coherence.

2 Ask for MARP

Members of the panel are invited to:

i. advise on the scope of the LPD proof-of-concept,
ii. comment on the design decisions taken and methods used to create the Census 2021 base for the LPD proof-of-concept,
iii. comment on the design decisions taken for linkage and indexing of the LPD proof-of-concept,
iv. comment on the design decision taken and the methods we are exploring to address census undercoverage, maintain the cohort and track internal migration, and
v. advise on any further ideas on the rapid production of longitudinal assets (referred to as satellite cohorts) and how to ensure coherence.

3 Background

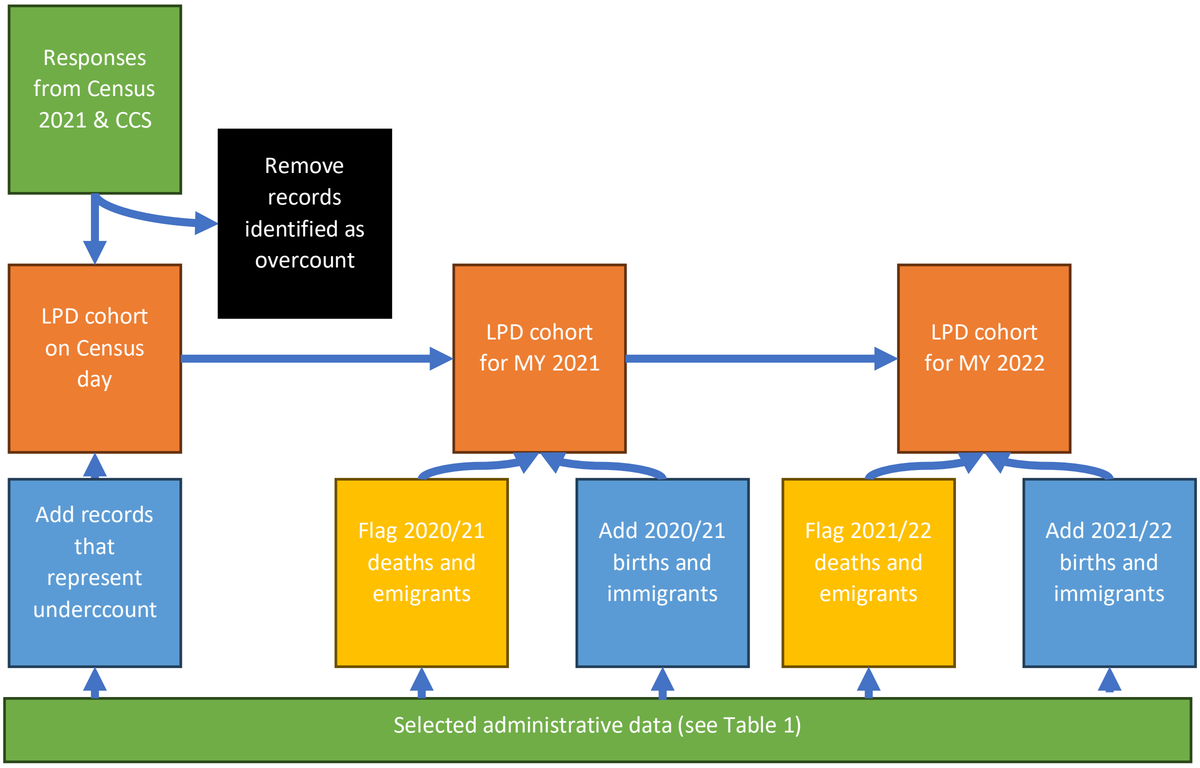
3.1 Summary

The Longitudinal Population Dataset (LPD) is an anonymised person-level data asset based on Census 2021 and selected administrative data sources. The LPD is a longitudinal dataset that combines population characteristics (e.g. age, sex, ethnicity, nationality) with information on life events (e.g. birth, death, internal and external migration events). It aims to fully represent a cohort of usual residents of England & Wales by:

- addressing the 3% Census 2021 undercount,
- replenishing (updating) the cohort to reflect the changing makeup of the population through births, deaths, and international migration and
- recording internal migration to ensure geographic representatives and enable analysis by household/address/co-residency.

By maintaining a representative person-level data asset of near to 100% of the usual resident population of England & Wales (as shown in Figure 1), the LPD will enable production of multivariate and longitudinal population statistics.

Figure 1 Cohort maintenance and replensihment for the LPD


Responses from Census 2021 & CCS
Remove records identified as overcount
LPD cohort on Census day
LPD cohort for MY 2021
LPD cohort for MY 2022
Add records that represent underccount
Flag 2020/21 deaths and emigrants
Add 2020/21 births and immigrants
Flag 2021/22 deaths and emigrants
Add 2021/22 births and immigrants
Selected administrative data (see Table 1)

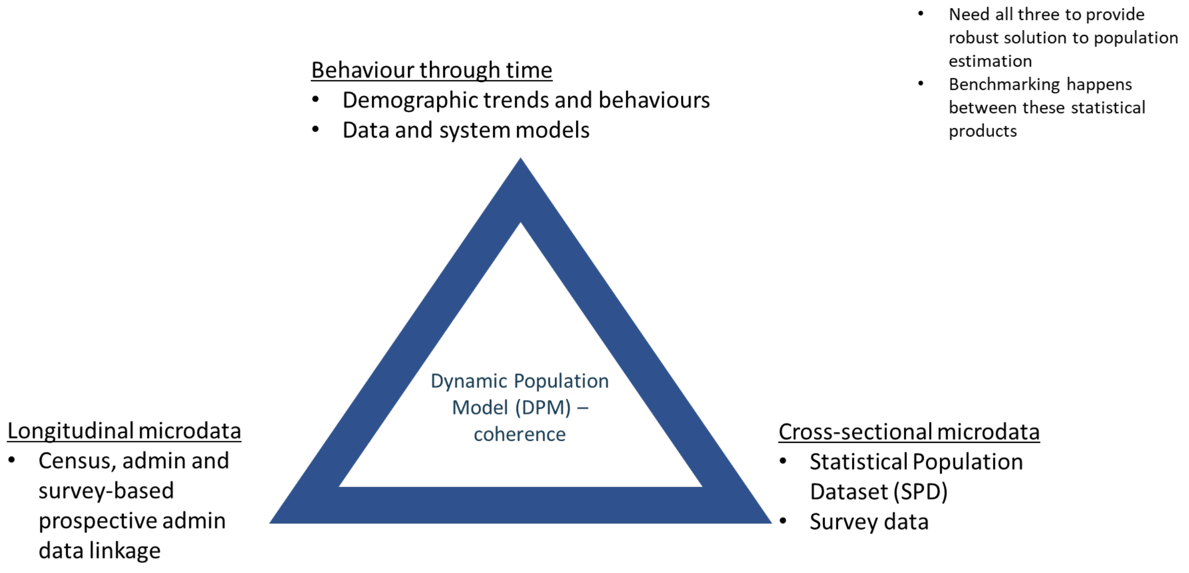
This work builds on the success of the Public Health Data Asset (ONS, Feb 2022) which demonstrated its analytical value in response to the Covid-19 pandemic, allowing analysis of Covid-related mortality in terms of 2011 Census characteristics. Longitudinally linked microdata, in combination with cohort replenishment and potentially applying study weights to ensure representativeness, will support far more granular analysis of the population characteristics associated with key longitudinal outcomes than is currently possible with the ONS 1% Longitudinal Study (ONS).

3.2 Demographic accounts and benchmarking

The heart of ONS' integrated statistical design for population and social statistics system is a demographic accounting system, because the underlying aggregate-level, model-based approaches are able to respond to and report population dynamics that are not in a 'steady state'. The proposed system design aims to make best possible use of the 2021 Census. However, we are also seeking to establish whether we can produce robust timely sub-national population estimates, including social and demographic characteristics, without a decennial census. As a minimum, we aspire to produce intercensal estimates that have less bias from intercensal drift than the current mid-year estimates and maintain a consistent level of bias over time.

To achieve this, we are building in statistical benchmarking, borrowing strength across the different properties of our administrative and survey data as suggested in Figure 2. Demographic behaviours and trends through time will be informed by the cross-sectional and longitudinal microdata and will feed into the demographic accounts. Statistical Population Datasets (SPD; ONS, Feb 2023) will feed into demographic accounts, alongside a robust coverage adjustment (methods in development). Entries to and exits from the LPD due to international migration will be constrained, first to provisional and then to final estimates of immigrants and emigrants from the Demographic Accounts. Divergences between the SPDs and cross-sectional populations implied by the LPD will be investigated and understood.

Figure 2 Benchmarking between statistical products


Behaviour through time
• Demographic trends and behaviours
• Data and system models
• Need all three to provide robust solution to population estimation
• Benchmarking happens between these statistical products
Dynamic Population Model (DPM) – coherence
Longitudinal microdata
• Census, admin and survey-based prospective admin data linkage
Cross-sectional microdata
• Statistical Population Dataset (SPD)
• Survey data

3.3 User needs

The LPD aims to be a statistically controlled microdata asset covering the population of England and Wales. It is not intended to be a population register and is only to be used for statistical analysis by approved researchers for approved research projects. Personal identifiable information (PII) is being used for linking only and then being removed before analysis takes place. We are currently developing the proof-of-concept for the LPD. The full cohort study would have a governance structure to ensure that proposed research projects meet ethical approval, are subject to disclosure control measures and would not harm any member of the cohort.

The full LPD would meet multiple user needs by providing:

- a census-like (fewer variables) micro-level utility dataset which can be used for multivariate analysis,
- high-quality satellite cohorts for longitudinal research like the Refugee Integration Outcome (RIO)[1] cohort study and Public Health Dataset Asset (PDHA); satellite cohorts provide insight into life outcomes and transitions which cross-sectional data assets cannot,
- a robust method to identify coverage drift of admin data and thereby providing a quality control for cross-sectional data assets and underpinning assumptions for the DPM, and
- data for novel analytical products like life expectancy by ethnic group.

4 LPD proof-of concept

4.1 Scope of the LPD proof-of-concept

By June 2025 we aim to complete a proof-of-concept of the LPD. We aim to produce LPD development datasets for Census day, mid-year 2021 and mid-year 2022. To achieve this, we are developing methods and data processing to address census undercoverage, cohort maintenance (births, deaths and international migration), and internal migration and co-residency. We will have undertaken a full quality assessment of the development datasets and cross-validation at record level with the Longitudinal Study (LS) and the RIO Study. Where the quality falls short of supporting user needs, we will identify the further developments and methods needed to reach the desired quality.

4.2 Data sources and variables

The proof-of-concept is using selected data sources to ensure that the representativeness of the cohort is maintained over time. Table 1 outlines which data sources will be used for the LPD proof-of-concept and reasons for inclusion.

[1] RIO was set up to provide quantitative data on refugees' long-term integration outcomes in the UK. It links Home Office Vulnerable Persons Resettlement Scheme (VPRS), Vulnerable Children's Resettlement Scheme (VCRS) and Asylum Route Refugees (ARR) data to administrative data collected by other government departments and Census 2021. (ONS, June 2023)

Table 1 Data sources included in the LPD proof-of-concept and reason for inclusion

Dataset	Reason for inclusion
Census 2021 and Census Coverage Survey	Population base.
NHS Digital Personal Demographic Service (PDS) – stock and monthly update files	Identifying addresses and individuals missed by Census 2021. Identifying internal migration moves since Census 2021, new patient registrations from abroad, and emigrations.
England and Wales birth notifications and registrations	Adding births since Census 2021.
England and Wales death registrations	Flagging deaths since Census 2021.
England and Wales marriages and civil partnerships, and divorces and civil partnership dissolutions (not yet available)	Addressing missed links because of name changes.
Home Office Border Intelligence (HOBI) data, previously Exit Checks	Adding immigrants and flagging embarkations since Census day for EEA and non-EEA nationals. This may be subject to later confirmation after enough time has elapsed. Flagging the population with pre-settled or settled status from Home Office EU Settlement Scheme (EUSS).
Home Office Citizenship (not yet available)	Flagging non-UK nationals who achieve citizenship.
Electoral registers for England and Wales	Confirming residence as of March 2021 in England and Wales. Flagging British emigrants not identified in PDS data through overseas voter status.
English and Welsh School Censuses, Individualised Learner Record for England, Lifelong Learning Wales Records and Wales, and Higher Education Statistics Agency (HESA)	Validating census records not linked to PDS where children or students are present. Validating PDS records not linked to census where children or students are present.

Table 2 outlines variables used to produce the LPD proof-of-concept, the reason for inclusion, and which of the data sources they can be found in.

Table 2 Variables used to produce the LPD proof-of-concept, reason for inclusion and source data

Variable	Reason for inclusion	Source data
Full name	Used in linkage and cohort maintenance.	All listed
Date of birth	Used in linkage and cohort maintenance. Used to derive age to report on linkage quality and analysis.	All listed
Sex	Used in linkage, cohort maintenance, and to report on linkage quality and analysis.	All listed
Address (including postcode)	Used in linkage and cohort maintenance. Used to assign local authority to report on linkage quality and analysis.	All listed
Nationality	Used in linkage, cohort maintenance, and to report on linkage quality and analysis.	Census, HOBI, HESA
Country of birth	Used to report on linkage quality and analysis.	Census, death registration
Month and year of arrival	Used to filter data for linkage purposes, and to report on linkage quality.	Census
Arrival and departure dates, and UK visa start and expiry dates	Used to filter the data for linkage purposes, and to report on linkage quality.	HOBI
Alternative addresses (including postcode), e.g. usual address one year ago, second residence, and term-time addresses	Used for linkage and to report on linkage quality.	Census
Term-time postcode or domicile address	Used in linkage and cohort maintenance.	HESA
Previous postcode	Used in linkage and cohort maintenance.	PDS
Ethnic group	Used to report on linkage quality and analysis.	Census, School Censuses, HESA, birth notifications
NHS number	Used in linkage and cohort maintenance.	PDS, death registrations
Date of NHS registration or date of patient UK entry	Used to filter the data for linkage, and for cohort maintenance.	PDS
Reason for removal flag and other flags for new registrations	Used in cohort maintenance and to report on linkage quality.	PDS

5 Progress towards the LPD proof-of-concept

5.1 Census base

Our key aim when constructing the Census 2021 base for the LPD proof-of-concept was to have a dataset that can be effectively linked to administrative data sources. Therefore we have to ensure that we have a clean dataset with minimal statistical processes applied (e.g. imputation).

Table 3 Reviewed rule-based edits and resolution applied to LPD base

Description of issue	Rule-based edit for census	Resolution for LPD base
Postcode validation failed for non-enumeration address	Lookup correct postcode If lookup fails the postcode and associated fields removed	Applied as address information unusable otherwise (and not able to assign UPRN)
Erroneous lookup during coding (e.g. ethnic group)	Reverted to accurate value	Applied
Age value changed for various statistical reason (and does no longer match date of birth)	Set date of birth to missing or recalculated	Not applied, keep original date of birth
Following discovery of unexpected sex ratios, a clerical review of 105+ y/o identified spurious records (e.g. protest responses and unrealistic data values) and incorrectly transcribed/unrealistic date of birth (e.g. 105 y/o in full-time employment or student)	Removed spurious records Set date of birth to missing	Could not restore spurious records Restored original date of birth
Data capture returned erroneous/out of range data for month and year of arrival (<10 records)	Set to missing	Applied
Electronic questionnaire allowed month and year of arrival up to May 2021 to allow students to respond that were delayed due to Covid-19. Paper questionnaire only allowed month and year of arrival up to March 2021.	Disregarded records that arrived after census unless they are student in full-time education For the latter set month of arrival to missing and keep year of arrival as 2021	Could not restore disregarded records Applied month of arrival being set to missing to keep paper and electronic questionnaire consistent

We chose to use an extract of the Census data just after resolution of multiple responses (and before edit and imputation). This means that the different data streams (paper and electronic questionnaires) have been processed to provide clean data and a (within-postcode) linkage process was applied to remove duplication from multiple responses (ONS webpage, June 2023).

There are a number of rule-based edits that are applied throughout the census processing to resolve different issues. We reviewed these and made a decision to apply/keep them or not apply/revert them. Table 3 provides an overview of the rule-based edits that were reviewed and how we decided to resolve this for the LPD base.

5.2 Data architecture, linkage and indexing

The datastore for the LPD proof-of-concept has been built following best practice from the Generic Statistical Information Model (UNECE, June 2024). At this point we have linked the Census bases with PDS, births/deaths, school census and HESA. At this point linkage relied on the linkage algorithm from the resolution of multiple responses and where unique identifiers like census response ID or NHS number are available these were used instead of personal identifiable information (PII). We also relied on the bespoke linkage of Census to PDS which was produced for the Covid Infection Survey as well as the 2021 Census to Census Coverage Survey linkage which fed into census estimation. We have already identified a few issues with the linkage, for example twins being collapsed into the same person cluster. As we improve our linkage method, we will make use of clerical resource to provide quality assurance.

We have built up our own address and demographic index, instead of relying on the RDMF Demographic Index (DI) used by the Statistical Population Dataset (SPD). We made this choice, because the RDMF DI is only composed of a limited number of core datasets (e.g. it does not use the Census 2021 and HOBI data), focuses on linkage maximisation (over minimisation of false-positives) and does not have longitudinal integrity as a design principal. This provides us a platform to fully explore how to build a demographic index that focuses on longitudinal integrity.

The design of the LPD datastore allows us to save all values of the population characteristics (see Table 2 for a list) from the selected administrative data sources as well as the census response for each person. This means both consistent and inconsistent values are retained for reference and analysis. We use the “value when becoming an LPD cohort member” for each population characteristic for initial analysis and to monitor the representativeness of the cohort. This requires us to standardise and harmonise the population characteristics variables (e.g. ethnic group) and attempt to complete missing census data from the most trusted (using a deterministic hierarchy of belief starting with Census 2021) and chronologically closest admin data.

5.3 Addressing Census undercoverage

Census undercount should not be seen as an exception but rather as the norm (United Nations, Oct 2022). The 2021 Census count was 97% of the final estimate (ONS webpage, Nov 2022). Addressing undercoverage is necessary to ensure that the LPD is representative of usual resident population including hard-to-count. It is worth noting that part of the population that is hard-to-count during Census will also be socially disengaged and might not show up on administrative data sources.

There are multiple ways we can approach undercoverage:

- Exploring non-responding addresses to Census
- Looking at particular groups (Communal Establishments, homeless people, children undercount, young males, etc.)

We are currently focusing on identifying non-responding addresses using Census Intelligence Datastore (CID). CID is an address-level dataset which combines Census, Census Coverage Survey (CCS) and administrative data sources together for England and Wales. For LPD, CID can provide insights for addresses that did not respond to the census. Census fieldwork officers collected information about non-responding address, such as refusals to answer and if an address is vacant or a second residence. Similarly, CID includes record level information from Council tax, Housing Associations and utilities data.

No source within CID is considered the gold standard or the sole source of truth. For example, Council tax provides a flagging system for empty houses and second residences, but its accuracy cannot be guaranteed due to administrative errors or the possibility of gaining financial advantage from providing inaccurate information. Similarly, Housing Association data on vacancies might differ from Council tax records and we lack information on the most up-to-date source. Therefore, no source is considered more reliable than another; all sources are treated equally and have the potential to provide key insights.

Another limitation of CID is timeliness and reference dates. While all information should reference Census day (21 March 2021), Census forms could be returned early (or late) and fieldwork didn’t finish until 9 May 2021. Similarly, CID has a Council tax annual dataset and an April update (including updates up to the 9th).

All decisions have been made taking the stand of the research question around addressing Census undercoverage to improve the LPD, not Census undercoverage itself. The rationale is that we have been thinking in terms of ‘pockets’ of people who might have failed to answer Census for different reasons and trying to locate those people. We haven’t been trying to get a high-quality reliable list of empty houses (UPRNs) or to understand better how people reported 2nd residences or how accurate that was, both of those potential very useful outputs for Census. The CID is a means to an end of finding the missing people on the LPD demographic base.

We are following 3 paths: vacant properties, second residences and non-responding addresses. For each path, we get a list of addresses (identified by UPRN) from the CID summary table and match with the corresponding sources table for validation (e.g. council tax for vacancies). Then, we take the subset of UPRNs that have not responded to the Census and match with LPD address and demographic index. This way we find personal records of people reporting those addresses in admin data who did not respond to Census. For example, people who refused to answer 2021 Census who have children, their address will appear in the School Census and PDS. Those matched records can then be included in the LPD cohort. This allows us to filter the large number of records on the administrative data that are not on the 2021 Census to people which are on our selected administrative sources at non-responding addresses.

5.4 Cohort maintenance – births and deaths

Adding births is reasonably straight forward as the data are timely and of high quality. However, there are significant time lags in death registrations for referrals to coroners. Currently, the median time to registration of a death not requiring further examination by a medical registrar or coroner is 7 days. These 'routine' deaths account for approximately three quarter of all registered deaths, with the remaining one quarter being referred to the medical registrar or coroner for further investigations.

For deaths involving referral for further investigations, the median time to registration is approximately 20 days. Approximately 5% of total deaths are referred for further investigation and will be subject to an inquest, in which cases, delays between death and registration can be substantially longer (for example, in the case of suicides, the median time to registration is approximately 200 days). The number of death registrations that have taken more than year to be registered has been increasing since 2020 and was 1.4% in 2023. For a given reference date, we will initially look up to 1 year ahead in the data to minimise this impact. We will look at lack by age and sex to test this assumption and minimise the introduction of bias.

5.5 Cohort maintenance – international migration

We are working closely with the Long-Term International Migration (LTIM) team to add immigrants to the LPD cohort and flag embarkations. Our strategy follows four key principles:

- Align with existing definitions of usual residents and long-term international migrants (12+ months, see UN 1998) and migrants' streams (non-EEA, EEA, British nationals)
- Widen existing definitions e.g. short-term international migrants (3-12 months, see UN 1998 and UNECE & ONS, Sep 2024), dependent on resource and available data.
- Compare and assess output quality against existing data sources and estimates.
- Demonstrate from the research work how the international migration part of the LPD cohort maintenance regime can produce a stable cohort.

We are primarily relying on the HOBI dataset. The LTIM team produces their estimates using Q2 and Q4 estimates each year. We will use the LTIM dataset built to produce the May 2024 ONS publication (reference period 2019 to December 2023) and Q1 2024 data for linking personal identifiable information (PII) to the LPD proof-of-concept. In addition, Q1 2024 will be used to assess the quality of the linkages between LTIM and Q1 2024. We will attempt to validate that we can trust that HOBI ID refers to the same person across extracts and further investigate findings from the RIO study which identified cases where multiple refugees (e.g. family members) were incorrectly clustered into one record.

As highlighted, the ONS' LTIM estimates use the same data and similar methodology to those that we will use for the LPD proof-of-concept. The official LTIM estimates apply a number of aggregate adjustments (based on patterns observed in previous years) to account for VISA overstayers and early leavers. The LPD proof-of-concept has the luxury of looking at HOBI data going up to 12 months after the reference data and therefore in most cases we do not need to predict peoples' future behaviour. However, we do not expect the data to be perfect and plan to explore model-based approaches to identify 'candidate' migrants to compensate for incomplete and time-lagged data. This work will build on collaboration with statisticians from Stats New Zealand (Stats NZ, Aug 2021) and the Australian Bureau of Statistics who have successfully developed machine learning approaches to forecast migration outcomes as their data matures (i.e. catches up with time lags).

5.6 Internal migration and co-residency

Change of address is the primary mechanism through which internal migration activity can be identified in datasets that will feed into the LPD. Most datasets enable one of two kinds of address change identification:

- Moves: individual moves between addresses within a given time period.
- Transitions: whether the address is the same or different at two different time points (but may not capture other moves within that time frame)

Currently, the most common method for identifying moves is through the PDS data, which uses data on GP registrations to identify address changes. Other datasets that feed into the LPD may help identify address changes, however these are often transitions (e.g. English School Census) rather than moves and might not provide full addresses (e.g. HESA has only postcode).

We are reviewing best practice from the internal migration team and the Statistical Population Dataset (SPD) team to decide if existing methods meet the LPD's needs or require further development. Currently we are not able to fully capture moves from England or Wales to Scotland or Northern Ireland.

While Census 2021 data provides information on households and families. However, administrative data does not provide robust information on households and families. Therefore we think that co-residency (i.e. people living at the same address as defined by Unique Property Reference Number) is the only concept we can derive from administrative data with reasonable consistency.

5.7 Characteristics, satellite cohorts and other use cases

The LPD wants to meet a number of different user needs. Among them, it might provide an alternative framework for an admin data census. We think there are two possible routes:

- survey-like weighting of the full LPD dataset using DPM population estimates or
- census-like approach requiring estimation, adjustment and imputation.

As part of the proof-of-concept we will explore this more and specify which additional methods and research would be required, and if the LPD can provide the required input to an estimation and imputation process.

The LPD can also provide data for novel analytical products such as life expectancy by ethnic group. This type of granular analysis has not been feasible using the 1% LS. This analysis will build on the official statistics in development, Ethnic differences in life expectancy and mortality from selected causes in England and Wales: 2011 to 2014 publication (ONS, Jul 2024), which used Census 2011 to produce life expectancy by ethnic group. There were several limitations outlined in this analysis that the LPD could help address. As part of the proof-of-concept we will fully explore the viability of life expectancy by ethnic group.

As part of the proof-of-concept we are also working with RIO, HAPI (Health Analysis and Pandemic Insights, for Public Health Data Asset) and other potential satellite cohorts (Veterans, PACE for record-level migration) to demonstrate how we produce the relevant sample from the LPD development dataset and add key characteristics from Census 2021. This will demonstrate the viability of using the LPD as sample frame for satellite cohorts.

In section 5.2, we stated that we use the “value when becoming an LPD cohort member” for each population characteristic for initial analysis and to monitor the representativeness of the cohort. We are aware that many characteristics are not fixed in time and we will work with topic experts to develop and apply suitable methods for deriving these from administrative data sources.

6 Economy of scale

The new Life Journeys area within the FPMS brings together the management of three longitudinal data assets: The LPD, the ONS Longitudinal Study and RIO. This is an opportunity to derive efficiencies from economies of scale and to promote and develop best practice in the design, development and analysis of longitudinal data. It also allows us to drive data quality improvements by comparing the three studies, given that they have been developed independently from each other.

In January 2025 we will bring RIO and the LPD and then the LS and LPD together, at microdata level, to understand the records on one source but not on the other. This will highlight any systematic biases resulting from the studies’ respective record linkage methods, and hopefully improve linkage approaches.

This record-level cross-validation is likely to lead to some convergence in linkage methods within Life Journeys. We will also seek to harmonise design and processing of these longitudinal assets, working towards best practice. For example, linkage of Home Office HOBI data to RIO and the LPD has developed insights and expertise that can be brought to bear on the LS, providing an opportunity to capture immigration of new LS members from Home Office as well as from NHS (currently PDS) data. This would also potentially allow us to use HOBI data to address the known under-capture of LS members’ embarkations.

Managing the longitudinal studies under the Life Journeys umbrella may also bring benefits at the end of the processing chain, as we are considering the pros and cons of a single user support function spanning the three cohorts.

7 Conclusion and next steps

We outlined what we want the LPD proof-of-concept to achieve and our progress to date. We are looking forward to feedback from the panel. Our plan is to provide another progress update to the panel before summer 2025 and going to the ONS Design Authority. We hope that in 2026, the LPD starts to provide an alternative framework for an admin data census (through survey-like weighting of the full LPD dataset using DPM population estimates or a census-like approach requiring estimation, adjustment and imputation) and delivers economy of scale benefits by supporting the rapid production of longitudinal assets including RIO, Longitudinal Study (LS) and Public Health Data Asset (PHDA).