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

1 Purpose

The Life Journeys team has produced a proof-of-concept (PoC) for the Longitudinal Population Dataset (LPD). 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 follows last year's paper to MARP (EAP214, Nov 2024), which described the scope of the LPD PoC. The current paper:

- Recaps the rationale and purpose of LPD.
- Outlines the revised scope of the LPD PoC.
- Discusses how the LPD PoC was built (guiding principles, key design decisions, methods and results).
- Assesses the quality of input data and LPD outputs.
- Outlines the cross-validation work comparing the LPD against the Refugee Information Outcomes (RIO) data and Longitudinal Study (LS) aggregate data and provides some early results.
- Highlights the value of the LPD PoC and its continuation for 2031 Census planning, implementation, and assurance.
- Seeks MARP endorsement of the recommendations made in this paper.

Further development of LPD is paused for financial year 2025/26. Further cross-validation of the LPD PoC will continue as part of the Longitudinal Study (LS) and Refugee Integration Outcomes (RIO) projects. In the context of this project, cross-validation refers to the programme of work undertaken to cross-reference the population of one longitudinal dataset to another to ensure each is picking up the correct people and linking the correct life events to the correct person.

2 Ask for MARP

MARP are invited to:

i. Validate and comment on the recommendations made throughout the paper, and (where applicable) provide additional recommendations.
ii. Advise on our outcomes and future plans for cross validating the LPD against the LS and RIO study.
iii. Advise on the usefulness of the asset for Census 2031 planning, operations, estimation and quality assurance.

3 Background

3.1 User needs for a 100% population longitudinal dataset

The ONS Longitudinal Study (LS) demonstrates the research value of longitudinal data based on linked census and administrative data. The LS has been used extensively by academic and government researchers, providing evidence for a broad range or epidemiological and social analyses (such as reviews into health inequalities, like the Black Report, and life expectancy by social class) and providing a sound basis for high-profile policy interventions (such as pension age reviews). However, the 1% LS sample size restricts the scope of what can be researched and the level of granularity available for meaningful results.

During the COVID-19 pandemic, ONS used linked Census and events data in the Public Health Data Asset (PHDA) to create critical and timely insights into the spread of the disease, and the demographic and social correlates of covid-related mortality. A criticism from users was that by 2019 the 2011 Census base for the PHDA was no longer representative of the 2019 population at risk from the pandemic. For example, health workers who had immigrated since the last census were not reflected in data and the ethnic composition of the PHDA in 2019 was misleading. Model-based estimates to address these known shortfalls relied on assumptions rather than hard data.

The LPD aims to address these limitations by rolling forward the 2021 Census population for England and Wales with population entry (birth and immigration) and exit (death and emigration) events, as well as accounting for 2021 Census undercount (ONS, Nov 2022) with admin-based proxies for Census non-respondents. Benchmarking LPD population stocks and flows against official mid-year population estimates will quantify drift. Comparing this to the Admin-Based Population Estimates (ABPEs) benchmarked against the official estimates will provide important information on the quality of the LPD compared to the ABPEs.

Recommendation 1: Where applicable, harmonise methods and governance across all longitudinal assets, starting with adapting clerical matching methodologies from RIO and user support functions from the LS.

Further, given the recent 2031 Census decision, the LPD could create an expected population to be captured at the Census. The undercoverage work can highlight non-responding addresses with admin data footprints and allow for better targeted field work. The inclusion of births and deaths data to the population spine can create a data quality measure for the Census undercoverage of young children and those near death and provide a first attempt at using admin data alongside the Census to reduce and account for census undercoverage.

Finally, the LPD person and address indices would support detailed census data quality assurance. LPD cross-validation against other longitudinal sources is providing insights into the quality of commonly used and highly rated administrative data sources such as the NHS Personal Demographic Service (PDS).

3.2 Scope as determined by ethical considerations

Advice from the National Statisticians Data Ethics Advisory Committee (NS-DEC) was to limit the variables in the LPD to only those required to link data and assess the representativeness of the LPD for the PoC. Extra variables would be linked in production. Other recommendations included (July 2022):

- Consult the Information Commissioner's Office on individuals' opt-out
- Complete and submit a Data Protection Impact Assessment
- Use lessons learnt from the Heath Data Asset and Refugee Integration Outcomes (RIO)
- Focus on public engagement
- Justify the whole-population approach
- Only necessary variables for onward linkage to be linked
- Implement strong governance controls
- Articulate the difference between the LPD asset and a population register
- Engage with the Centre for Applied Data Ethics throughout LPD development

Recommendations from NS-DEC have informed the guiding principles of the LPD design. Analysis of data from the LPD will be via a satellite-based system. The Census 2021 Data Asset provides a record level representation of the England & Wales population, from which satellite cohorts based on samples can be drawn.

3.3 Guiding principles for the design and implementation of the LPD

The guiding principals for the LPD design and implementation are:

1. Only variables necessary to make, confirm or quality assure linkage to the LPD are included.
2. The quality of data inputs and data outputs is monitored to understand how data error propagates through the LPD design and processing.
3. False positive linkage is to be avoided because it corrupts the longitudinal integrity of the data and can create misleading analysis.
4. To support (3) above, use of 'loose' matchkeys is avoided in automatic matching. Clerical matching is used to improve linkage rates and to verify automatic matches on a sample basis.
5. Linkage quality is monitored to inform data use and metadata.
6. Aggregate distributions from LPD are compared against high-quality benchmarks for data validation and to inform onward use of the LPD.
7. The 2021 Census was of high quality and is used as the base population for the LPD. Use of administrative data to account for non-respondents is primarily confined to known addresses where enumerators believed there was a non-responding household.
8. The exception to (7) is births data used to address under reported infants present in responding addresses.
9. The cohort component method is subsequently used to account for entries (births and immigrants) and exits (deaths and emigrations).

10. Monitoring cohort size in this way aims to protect against numerator/ denominator bias in longitudinal data analysis. Linkage error or an inflated base population could corrupt results.

This PoC iteration of the LPD had neither the time nor resource for clerical checking so outputs are interim, subject to further linkage and validation.

Recommendation 2: The next iteration of the LPD requires a multi-disciplinary team to develop it. The blueprint for this should bring together ONS colleagues from data engineering, in house technical support, data scientists, and social researchers to improve the design of the LPD and avoid a bottlenecks/single point of failures.

4 Design and Implementation

4.1 Data sources and variables for the LPD PoC

Table 1 outlines data sources the LPD PoC used and why they were used. Appendix A lists data sources that were excluded from the LPD PoC in comparison to the NS-DEC approval. These were either not yet available or not necessary to meet the scope of the PoC.

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

Dataset included in LPD POC	Reason for inclusion
Census 2021 and Census Coverage Survey	Population base.
NHS Digital Personal Demographic Service (PDS) – stock	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 registrations	Adding births since Census 2021 and newborns missed by Census 2021.
England and Wales death registrations	Flagging deaths since Census 2021 and adding non-respondents who died after Census 2021.
English and Welsh School Censuses, 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.
Home Office Borders and Immigration (HOBI) data, previously Exit Checks	Adding immigrants and flagging emigrations since Census Day for EEA and non-EEA nationals. Flagging the population with pre-settled or settled status

Appendix B lists variables used to produce the LPD proof-of-concept. There have been no substantive changes since our last paper.

4.2 Data architecture, linkage and indexing

The LPD datastore was built following best practice from the Generic Statistical Information Model (GSIM, UNECE, June 2024). For the LPD PoC the Census base was boosted with records found from the undercoverage investigation.

Any person or address records, either from the Census base, undercoverage investigation, or linked administrative data sources, were stored in a bespoke LPD Demographic Index (LPD-DI) or LPD Address Index (LPD-AI) respectively (collectively the LPD indices). These are separate from the ONS Demographic Index and ONS Address Index and were created to provide a population spine that had longitudinal integrity; one that could track units within the population across time periods. This also allowed the team to set their own rules to create the index to better meet the needs of the LPD and enable more control.

For illustration purposes, the processing and indexing of an administrative dataset included in the LPD can be explained using different layers.

1. Layer 1: Raw data – Data are loaded with formatting bespoke to each new dataset
2. Layer 2: Indexed data – Matching of records to the existing LPD indices. Three steps:
    a. First step is direct matches on reference numbers (e.g. NHS Number)
    b. Any residuals are matched on Personally Identifiable Information (PII) using different match keys. Match Keys are outlined in Appendix C.
    c. Any remaining unmatched records are added as new index entries
3. Layer 3: Standardised data – Data are harmonised (where possible/necessary) and standardised to follow the same coding/naming conventions so that common variables across the datasets are consistent
4. Layer 4: Analytic dataset – The data are pivoted to be wide format with a single row for each person and inclusion rules which determine the cohort at different time points. Where a person has multiple values for a variable from different datasets, for analysis purposes, a single value is chosen via an editable hierarchy.

4.3 Base population adjustments

The 2021 Census has an estimated 3% undercoverage which needed addressing via admin data to ensure the Census base was representative of the usually resident population in England and Wales. Research using the Census Intelligence Datastore (CID) identified potential addresses that didn't respond to Census 2021. CID identified three types of potential non-responding addresses: vacant properties, second residences and actual non-responding addresses.

Some UPRNs that did not respond to Census but matched with the LPD-AI had associated LPD-DI entries. These were Census non-responding addresses, listed in any other admin data source by an individual. Information about the occupant can be drawn from the admin data record, and the subject included in the LPD-DI despite the subject potentially not responding to Census. However, CID was designed to minimise Census non-response at address-level. Failure to link at person-level and limited quality checking risks adding a new record for people who had already answered Census at another address.

Recommendation 3: Carry out validation checks on the admin records found at non-responding census addresses and expand the undercoverage work to find new ways of finding Census non-responders.

4.4 Rolling forward the population

The boosted Census 2021 population spine was rolled forward to subsequent time periods. Records for new births and long-term immigrants were added and records for the deceased and long-term emigrants were flagged. For international migration, the LPD approach aligns with existing UN definitions of long-term international migrants (UN 1998). This definition requires migrants to have remained for 12-months to be considered long term. As the data used for the PoC is from more than 12-months ago, the outcome for all migrants is confirmed.

Recommendation 4: For migrants whose migration outcome is unknown, two methods should be explored and compared. A machine learning approach based on recent successful work to predict migrant outcomes in Australia and New Zealand, or waiting for the data to mature and introducing a minimum 12-month time lag.

For death and emigration, all records from the deaths data, and records flagged as long-term international emigrants in the HOBI data were matched to the LPD-DI. Flags were added to matched records showing the period within which the date of the death/emigration event occurred. Inclusion rules ensure that records only appear in analytic datasets in the period of usual residence. Currently the full dates of events are removed from the final analytic dataset for Statistical Disclosure Control. However, full date of event is required to do accurate and informative survival analysis.

Recommendation 5: Full date of event should be included in a separate secure file to allow for specific research purposes whilst maintaining Statistical Disclosure Control.

All new person records or address information from births data were added to the LPD indices. Records flagged as long-term international immigrants in the HOBI data were either linked to previous LPD-DI entries or added to the LPD-DI as new population members. Flags again indicated the period during which the event date occurred.

Recommendation 6: Compare the quality of the LPD using the current migration method against new ONS international migration estimates method which includes micro data for all migrants (European Economic Area (EEA), non-EEA, and British Nationals).

5 Quality assessment

5.1 Data inputs

Each dataset indexed on the LPD has a data dictionary with variable names, types and counts. Counts of records from each data source remain consistent throughout the indexing process. The records then combine to produce a single set of data points for each entry in the LPD-DI.

5.2 LPD Address Index and Demographic Index

Around 10% of indexed addresses have multiple entries in the LPD-AI that share a Unique Property Reference Number (UPRN) but have different street address information. In some

cases, this is legitimate where the same UPRN is given to the same accommodation unit, that has multiple versions of the street address. Other times, GeoPlace (suppliers of UPRNs) give a single UPRN to multiple accommodation units such as caravan sites or halls of residence, which can lead to infeasibly large 'households'. Future work should check household size distribution and make this consistent with the total population.

In addition, some addresses occur thousands of times in the input datasets. This can be legitimate, for example Census workplace addresses. However, some are generic addresses or addresses used as defaults when the data subject's address is unknown. For example, if someone gives no address while visiting A&E, the person entering the information will enter a generic postcode into the admin data record. This can incorrectly inflate populations at certain locations and is most common in PDS data where there are only around 25 million unique addresses from 81.5 million records. One default address occurs 580,000 times (discussed further in section 5.5).

Visual inspection of outputs showed some records on the LPD-DI were incorrectly linked together. This is because they were listed as being at the same property via a common UPRN. Coupled with the above problem, this means that generic addresses encourage false positive linkage and require further investigation. This issue is compounded further when similar names and common/generic dates of birth are added into the equation.

Many LPD-DI entries were correctly linked despite slight differences in PII. All of these different records were retained to facilitate future matching where these slight differences could appear again in admin data sources.

4.6 million record pairs match exactly on first name, last name and date of birth, but aren't determined to be a match by the matching process. This prevalence of False Negatives is down to missing or unmatched address information. All match keys required matching address information, so when address information was missing, clear matches were missed. Better address information, tweaking of the matching algorithm, and further rolling forward of the LPD can avoid this and reduce the burden on any manual matching system or clerical review.

Recommendation 7: Review the matching algorithm for suitability of the Match Keys. Alter the weight placed on address information to reduce False Positives.

5.3 LPD analytic dataset

The analytic dataset for the LPD cohort is a wide-format dataset with a single row for each person currently flagged via the input datasets as a usual resident of England and Wales at any time point from Census Day 2021 to mid-year 2022. It contains variables for characteristics such as sex, age (at key reference points), ethnicity, nationality, country of birth and location (Local Authority). It also contains flags to indicate when and how the person became a member, and/or ceased to be a member of the usual resident cohort.

Table 2 and Table 3 show quality check results for some key variables in the analytic dataset. Checks include missing and invalid rates, and consistency checks. In Table 3, the inconsistency of high-level and more detailed ethnic groups is due to a coding error which has been identified, and a solution has been found which needs implementation.

Table 2: Counts and rates of missing data (as percentage of the 59,880,685 records included in the LPD usual resident cohort at any point in time).

Variable	Missing count	Missing rate	Notes
Sex	167,430	0.3%	
Age/Date of Birth	118,590	0.2%	Analytical dataset incudes age calculated at three time points (Census Day, mid-year 2021, mid-year 2022) from the date of birth variable. Missing count is where all three are missing.
Country of birth	1,871,220	3%	
Nationality	10,251,465	17%	Nationality available only from Census, HESA and LTIM data. Census value derived from passports held and missing where no passport was indicated.
Ethnic group	2,549,815	4%	Ethnic group is available from Census, HESA, ESC and WSC.
Local authority	64,900	0.1%	Coded as K04000001 for England and Wales but LA unknown.

Table 3: Counts of invalid entries and the reason why they are believed to be invalid.

Variable	Invalid/Implausible/Inconsistent	Count records
Local authority	Outside of England and Wales	690
Age	Over 115 (at Census 2021)	505
Ethnic group	Records with inconsistent high-level vs more detailed ethnic group	4,680
Year of arrival	Records where year of arrival before birth	1,810

Appendix D outlines the counts of people included in the analytic dataset split out by their method of entry and exit.

5.4 National comparison with Census and MYEs

LPD population counts for Census Day and mid-year 2021 were compared against official population estimates published by the ONS for the same time periods, via the official Census estimate (including all aggregate level adjustments, such as coverage adjustments) and the 2021 Mid-Year Estimate (also containing aggregate level adjustments) respectively. Table 4 shows the results. LPD estimates are less than 1% below the official estimates in both cases.

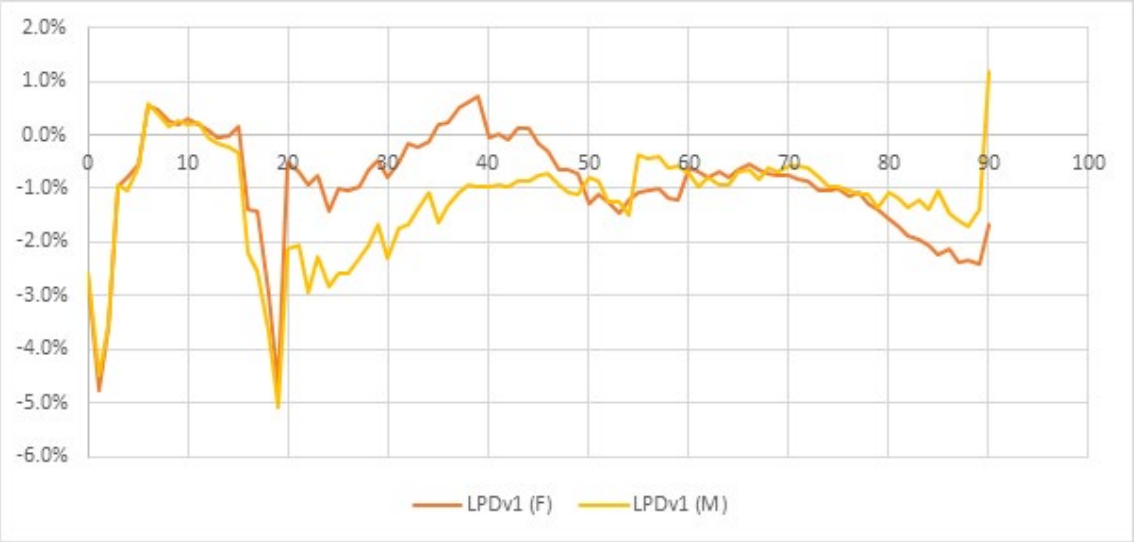
When compared to Census 2021 responses without adjustments, LPD was shown to address some of the 3% undercoverage between Census responses and Census official estimates. This can be seen in Table 4 as the LPD undercoverage is less than 1%.

Table 4: LPD estimates compared to the official population estimates of England and Wales. LPD estimates are rounded to the nearest 5 for Statistical Disclosure Control (SDC).

	Official population estimate	LPD aggregate count	% difference
Census Day 2021	59,597,529	59,233,680	-0.6%
Mid-year 2021	59,660,524	59,355,415	-0.5%

Figure 1 shows the LPD population counts for babies and young children are around 3% lower than Census estimates. This group was around 9% undercovered in Census responses. The LPD has accounted for some of this, particularly for age 0, by inclusion of all birth registrations in 2021. This identified around 28,000 records born before Census Day 2021 but with no Census record. However, a gap remains. Linking pre-2021 births may further improve coverage for young children.

Figure 1: LPD percentage difference from Census by age and sex



The dip in population coverage at age 19 mirrors Census responses and reflects the difficulties of enumerating student populations. The LPD corrects for around half of the Census response undercount for this age group. For 20-50 year olds the LPD female population is closer to Census estimates than the male population. This reflects greater Census undercoverage of males and the difference is widened by generally better coverage of females in administrative data.

Around 58,000 deaths were flagged after Census Day with no associated Census record. Census undercoverage for 90+-year-olds was reduced from around 4% to around 1% by the LPD. However, the LPD has overcoverage for the oldest males compared to Census estimates. A high proportion is due to the LPD undercoverage adjustment at non-responding addresses and includes some at implausibly old ages (130 years and above). The subjects of these records are either deceased, or the result of incorrect data entry into the admin source. Regardless, they occur because of quality issues in the admin data sources. Future datasets

being added should help to avoid this as the different datasets shouldn't contain the same errors and so variable consistency between data sources can be used to check the quality.

Recommendation 8: Use LPD findings to inform Census 2031 planning and quality work. Especially in addressing undercoverage of the very young and very old via administrative data, and in planning for investigating non-responding addresses.

Recommendation 9: Sense check certain records and include filtering of the analytic dataset to remove impossible records.

5.5 Sub-national comparisons

LPD estimates were also compared to the official estimates at Local Authority (LA) level. All official estimates include their aggregate level coverage adjustments.

Overall, the LPD performs well at both Census Day and Mid-Year 2021. At Census Day, only 12 LAs have more than 3% undercoverage, and no LAs have more than 3% overcoverage. At Mid-Year 2021, only 14 LAs have more than 3% undercoverage, and no LAs have more than 3% overcoverage. A total of 39 local authorities have an LPD estimate higher than the official estimates. LAs with the greatest differences between LPD estimates and official estimates are seen in Table 5 and Table 6.

Table 5: LA populations with greatest differences between LPD and official Census 2021 estimates (more than 1% overcoverage, or more than 3% undercoverage). Census 2021 is drawn from published figures, and LPD values are rounded to the nearest 5 for SDC.

Local authority	LPD	Census 2021	LPD % difference from Census 2021 estimates
Boston	71,555	70,509	1.48%
Wolverhampton	267,335	263,725	1.37%
Hackney	262,570	259,143	1.32%
Hammersmith and Fulham	185,435	183,153	1.25%
Camden	212,435	210,131	1.10%
Halton	129,790	128,470	1.03%
Shropshire	313,895	323,619	-3.00%
Richmondshire	48,170	49,770	-3.21%
Bridgend	140,795	145,500	-3.23%
Brent	328,820	339,816	-3.24%
Eden	52,850	54,738	-3.45%
Lincoln	100,230	103,812	-3.45%
Westminster	197,070	204,233	-3.51%
Harborough	94,140	97,619	-3.56%
Rutland	39,225	41,051	-4.45%
City of London	8,180	8,584	-4.71%
Oxford	153,670	162,032	-5.16%
Cambridge	135,630	145,681	-6.90%

Table 5 shows that the LPD has differences from official Census estimates in LAs where the Census had coverage challenges such as City of London; in LAs with significant special populations such as Rutland; and in LAs with universities. Table 6 shows that these differences persist in Mid-Year 2021 comparisons.

Table 6: LA populations with greatest differences between LPD and official mid-year estimates (MYE) for 2021 (more than 1% overcoverage, or more than 3% undercoverage). 2021 MYEs are from published figures, and LPD values are rounded to the nearest 5 for statistical disclosure control.

Local authority	LPD	MYE 2021	LPD % difference from mid-year estimates for 2021
Kingston upon Hull, City of	272,110	266,516	2.10%
Hammersmith and Fulham	185,845	183,310	1.38%
Hackney	263,585	260,082	1.35%
Wolverhampton	267,815	264,260	1.35%
Coventry	348,140	344,151	1.16%
Boston	71,600	70,815	1.11%
Halton	129,955	128,570	1.08%
Tower Hamlets	303,030	312,715	-3.10%
Kensington and Chelsea	139,700	144,266	-3.16%
Shropshire	313,820	324,669	-3.34%
Bridgend	140,855	145,738	-3.35%
East Devon	146,895	152,065	-3.40%
West Devon	55,440	57,480	-3.55%
Eden	52,800	54,951	-3.91%
Westminster	197,560	205,759	-3.98%
Oxford	153,945	160,379	-4.01%
Harborough	94,170	98,254	-4.16%
Richmondshire	48,155	50,313	-4.29%
Rutland	39,230	41,342	-5.11%
City of London	8,195	8,689	-5.69%
Cambridge	135,930	145,022	-6.27%

The emergence of Kingston upon Hull at the top of Table 6 was investigated and the default address used in PDS when a patient's address was unknown is in this LA. Significant numbers of new migrant records used in the cohort maintenance were indexed and then linked to health records that used the default address. This caused inflated in migration to Kingston upon Hull, and inflated the mid-year population of the LA.

Recommendation 10: Generic UPRNs relating to 'default' addresses should be removed from the matching process to avoid high false positive rates

Internal migration (moves within England and Wales) is not yet included in the cohort maintenance, and this will introduce some differences to the mid-year population estimates

at the local authority level. Sub-national comparisons beyond Mid-Year 2021 would include out-dated address information.

Recommendation 11: Future LPD work should prioritise the development of an internal migration element to cohort maintenance. This should be compared against new official methods that are in development.

5.6 Population components of change

The LPD components of change (births, deaths and international migration) between Census Day and Mid-Year 2021 were compared to corresponding official estimates. Births and deaths agree at Local Authority, age and sex levels.

However, LPD international immigration is 57% of the official value, and for emigration it is only 18%. The LPD international migration levels are low compared to the official estimates because:

- This period covers the transition period for Brexit, and those with EU Settled Status were moving without needing a visa.
- 16% of records from Home Office data were linked to multiple LPD-DI entries and need clerical review to determine the correct match. Without time for clerical review, these were removed from the final figures.
- No British National migration included in the LPD.
- EEA record level data was very low compared to the official estimates because the official estimates used RAPID which is aggregated and not record level.

Representation of international migration flows at the LA level is poor as address data are not included in the Home Office migration data. Only a small proportion of the migration data matched to other LPD address information. Therefore, 71% of immigrant and 68% of emigrant flows have no known LA.

The immigration flow to Kingston-upon-Hull is 6,050 between Census Day 2021 and Mid-Year 2021; almost 14 times higher than official estimates. As discussed above this is a result of quality issues in health data addresses.

Recommendation 12: Any quality issues in source data should be passed on to relevant data owners for further investigation regardless of LPD continuing.

6 Cross-validation

6.1 Overview

To carry out quality assurance of the LPD methods, processes, and data, cross-validation against other longitudinal assets has been carried out. The other assets are the LS (as described in section 3.1) and the Refugees Integration Outcomes (RIO) study (tracking resettled refugees and asylum seekers over time and determining outcomes via a range of administrative data).

Cross-validation against the LS has so far been restricted to checking aggregates (population total, sex breakdowns, mid-year populations etc.) of both datasets due to difficulties in getting both datasets onto the same system. However, the checks show promising results.

Cross-validation for LPD and RIO has been able to proceed at a record level via linking of the two datasets. This has allowed for checking whether records appear on both datasets where expected. It also allowed for more in-depth checks of the migration, deaths, and personal information data that the LPD has assigned to records against a 'gold standard'. The results are promising, generally have high agreement, and are set out in the following sections.

6.2 Cohort Membership

LPD and RIO were linked on the unique identifiers NHS Number and Census ID. There were 76,905 records linked on both identifiers that were for the same person. However, it is the converse of this that is most interesting. 1,360 records linked on both identifiers but gave different people via their PII. These are either False Positive (FP) links in RIO where the wrong person was linked to the wrong identifiers, or False Negatives (FN) in LPD where real links weren't being correctly identified by the matching algorithm. FN links account for around 90% of these cases.

Some records linked using NHS Number joined LPD records with a Census ID to RIO records with no associated Census ID. Some of these were Census visitors in the LPD and RIO does not include visitors. Others require further investigation.

The RIO dataset was also indexed using the LPD PII match keys. Where this method made matches, the same matches are made as when unique identifiers were used. Limited investigation into the unmatched residuals shows linkage failure due to matchkey differences in the two processes. Further analysis would be helpful.

There were around 2,600 records on RIO that arrived pre-Census, have no linked death or emigration event, but have no Census ID or NHS Number. Essentially, they have no admin footprint beyond the Home Office data so are not in the LPD. This is an important avenue for further investigation to understand why this occurs.

6.3 PII Agreement

The linking of records across LPD and RIO has allowed for the consistency of the PII of both records to be checked. In general, records on RIO that are linked to records on LPD have good agreement on sex and date of birth.

There are 625 instances where sex doesn't agree between the two datasets. For example, LPD males linked to RIO females. Other PII agree, suggesting that both records refer to the same person, but sex could be incorrect on one dataset. Further investigation is required.

6.4 Death Flags

In most cases, LPD picks up death flags consistent with the RIO death flags for the years of data available.

75 RIO records with a death flag from pre-2021 link LPD-DI/LPD data with no death flag. Some of these may be drawn in to the LPD-DI from PDS and so won't have a linked death event because the pre-2021 deaths data was not available. Future iterations of the LPD should include pre-2021 life events data to better capture these individuals.

For 2021 deaths, 20 people that died on RIO don't have a death flag on LPD, and 10 people that have a death flag on LPD, don't have one on RIO. Further investigation is required.

6.5 Emigration

The linking of records on Census ID has allowed the identification of RIO records who have been listed as emigrating via HOBI data but are found on LPD; many have a Census ID for 2021. This suggests some RIO cohort members who are believed to have left the country are still producing an admin data footprint and have been linked to Census 2021. This needs further investigation by the RIO project.

Recommendation 13: The initial success of the cross-validation work highlights its importance and should be continued and expanded to all longitudinal assets, regardless of the continuation of the LPD. This should help to answer some of the questions this work has produced and continue to find quality issues in source datasets.

7 Conclusion

The LPD PoC demonstrates the feasibility of producing and maintaining a high quality 100% longitudinal asset, despite the use of limited data sources. There are methodological challenges to address. The asset would provide unique insights in health and migration statistics and support Census 2031 planning, operations and quality assurance.

Recommendation 14: Datasets to prioritise for inclusion are:

- **Life events from more years, including pre-2021 to capture more Census undercoverage;**
- **Marriage data to pick up on name changes;**
- **Individualised Learner Records and Lifelong Learner Record for Wales to find students not linked to Census;**
- **Citizenship data to flag naturalised citizens;**
- **Births to cohort mothers to provide information on family units, and check the inclusion of mothers (particularly migrant mothers) in the LPD.**

Appendices

Appendix A

Table 7: Data sources not in the LPD proof-of-concept and reason for exclusion

Dataset	Reason for exclusion from PoC
NHS Digital Personal Demographic Service (PDS) – monthly update files	PDS stocks have been included to identify addresses and individuals missed by Census 2021. Internal migration was taken out of scope and new patient registrations from abroad, and emigrations up to mid-year 2021 are covered by the PDS stocks included.
England and Wales birth notifications	Birth registrations have been included to add births since Census 2021. Births notifications would allow us to add registered mothers and fathers which could further reduce undercoverage and provide relationship information.
England and Wales marriages and civil partnerships, and divorces and civil partnership dissolutions	Full data not yet available to ONS and linkage pilot indicated data quality challenges.
Home Office Citizenship	Data not yet available.
Electoral registers for England and Wales	No need to confirm residence as of March 2021 in England and Wales as internal migration was taken out of scope. Impact of missing British emigrants not identified in PDS data through overseas voter status is likely to be minimal between Census and mid-year 2021.
Individualised Learner Record (ILR), and Lifelong Learning Wales Records (LLWR)	School Censuses and Higher Education Statistics Agency (HESA) have been included to validate census or PDS records not linked to another source where children or students are present. Exclusion of ILR and LLWR has likely minimal impact.

Appendix B

Table 8: Variables used to produce the LPD proof-of-concept

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.	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.	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/departure dates, and UK visa start/expiry dates	Used to filter the data for linkage purposes, and to report on linkage quality.	HOBI
Alternative addresses 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 new registration or removal	Used in cohort maintenance and to report on linkage quality.	PDS
Date of death	Used in cohort maintenance and to derive age at death.	Death registration

Appendix C

Table 9: Match key number and the variables used in the matching

Matchkey	Variables included in matchkey	Inconsistency resolved by matchkey
1	SEX | DATE OF BIRTH (DOB) | FULL NAME | LPD ADDRESS ID	Strictest matchkey, allows for missing unique id.
2	SEX | DOB | FIRST NAME | LAST NAME | LPD ADDRESS ID	Incorrectly reported middle name.
3	SEX | YEAR OF BIRTH (YOB) | MONTH OF BIRTH (MOB) | FIRST NAME | LAST NAME | LPD ADDRESS ID	Picks up small errors in reported day of birth
4	SEX | YOB | DAY OF BIRTH | FIRST NAME | LAST NAME | LPD ADDRESS ID	Picks up small errors in reported month of birth
5	SEX | DOB | AGED OVER 30 | FULL NAME (Levenshtein distance < 3) | LPD ADDRESS ID	Allows for slight errors in names. Full names must be more than 3 characters long
6	SEX | DOB | AGED OVER 30 | FIRST NAME (Jaro-Winkler similarity > 0.7) | LAST NAME (Jaro-Winkler similarity > 0.7) | LPD ADDRESS ID	Allows for slight errors in first and last name
7	SEX | DOB | AGED OVER 30 | FIRST NAME (Levenshtein edit distance > 0.6) | LAST NAME (Levenshtein edit distance > 0.6) | LPD ADDRESS ID	Allows for slight errors in first and last name
8	SEX | DOB | AGED OVER 30 | FIRST NAME = LAST NAME (Jaro-Winkler similarity > 0.7) | LAST NAME = FIRST NAME (Jaro-Winkler similarity > 0.7) | LPD ADDRESS ID	Transposed first and last name with some slight errors allowed when comparing transposed names
9	SEX | DOB | AGED OVER 30 | FIRST NAME = LAST NAME (Levenshtein edit distance > 0.6) | LAST NAME = FIRST NAME (Levenshtein edit distance > 0.6) | LPD ADDRESS ID	Transposed first and last name with some slight errors allowed when comparing transposed names
10	SEX | DOB | FIRST NAME (Jaro-Winkler similarity > 0.9) | LAST NAME (Jaro-Winkler similarity > 0.9) | LPD ADDRESS ID	Allows under 30s and allows for larger difference in first and last names
11	SEX | DOB | FIRST NAME (Levenshtein edit distance > 0.9) | LAST NAME (Levenshtein edit distance > 0.9) | LPD ADDRESS ID	Allows under 30s and allows for larger difference in first and last names
12	SEX (matches exact or missing) | DOB | AGED OVER 30 | FIRST NAME (Jaro-Winkler similarity > 0.85) | LAST NAME (Jaro-Winkler similarity > 0.85) | LPD ADDRESS ID	Allows for missing sex but slightly tightens rules on differences between names
13	SEX (matches exact or missing) | DOB | AGED OVER 30 | FIRST NAME (Levenshtein edit	Allows for missing sex but slightly tightens rules on differences between names

	distance > 0.85) | LAST NAME (Levenshtein edit distance > 0.85) | LPD ADDRESS ID	
14	SEX | DOB (Levenshtein distance < 2) | AGED OVER 30 | FIRST NAME (Levenshtein edit distance > 0.85) | LAST NAME (Levenshtein edit distance > 0.85) | LPD ADDRESS ID	Allows for slight differences in the date of birth
15	SEX | YOB | FIRST NAME | LAST NAME | DAY OF BIRTH=MOB | MOB=DAY OF BIRTH | LPD ADDRESS ID	Allows for transposed day and month of birth but year still has to match exactly.
16	SEX | YOB | MOB | FIRST NAME (Levenshtein edit distance > 0.9) | LAST NAME (Levenshtein edit distance > 0.9) | LPD ADDRESS ID	Allows for error in the day of birth and allowed for errors in the first and last names
17	SEX | YOB | DAY OF BIRTH | FIRST NAME (Levenshtein edit distance > 0.9) | LAST NAME (Levenshtein edit distance > 0.9) | LPD ADDRESS ID	Allows for error in the month of birth and allowed for errors in the first and last names

Appendix D

Table 170: Counts of LPD cohort members by their entry and exit route to the cohort. Counts have been rounded to the nearest 5, and any low counts less than 10 have been suppressed for SDC reasons

Cohort start flag	Cohort end flag	Count
Census 2021 response	NULL	57,416,940
Census 2021 response	Died before 21 Mar 2021 (Census Day) OC	4,900
Census 2021 response	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	107,910
Census 2021 response	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	218,350
Census 2021 response	Emigration to EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	30
Census 2021 response	Emigration to non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	3,200
CCS 2021 response	NULL	73,415
CCS 2021 response	Died before 21 Mar 2021 (Census Day) OC	#
CCS 2021 response	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	25
CCS 2021 response	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	70
CCS 2021 response	Emigration to non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	#
Born before 21 Mar 2021 (Census Day) UC	NULL	27,820
Born before 21 Mar 2021 (Census Day) UC	Died before 21 Mar 2021 (Census Day) OC	355
Born before 21 Mar 2021 (Census Day) UC	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	30
Born before 21 Mar 2021 (Census Day) UC	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	15
Undercoverage from CID path 3 (non-responding addresses)	NULL	1,378,320
Undercoverage from CID path 3 (non-responding addresses)	Died before 21 Mar 2021 (Census Day) OC	7,405
Undercoverage from CID path 3 (non-responding addresses)	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	3,460
Undercoverage from CID path 3 (non-responding addresses)	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	3,985
Undercoverage from CID path 3 (non-responding addresses)	Emigration to non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	110
Born between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	NULL	324,370

Born between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	800
Born between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	NULL	170,725
Born between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	455
Born between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	85
Immigration from EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	NULL	3,215
Immigration from EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Died before 21 Mar 2021 (Census Day) OC	#
Immigration from non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	NULL	62,470
Immigration from non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Died before 21 Mar 2021 (Census Day) OC	#
Immigration from non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	#
Immigration from non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	15
Immigration from non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	Emigration to non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	#
NULL	Died between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	20,175
NULL	Died between 1 Jul 2021 and 30 Jun 2022 (MY 2022)	37,945
NULL	Emigration to EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	55
NULL	Emigration to non-EEA between 22 Mar 2021 and 30 Jun 2021 (MY 2021)	14,025

= suppressed count below 10