Integrated statistical design for the transformed population and social statistics system; overview and longitudinal design principles for a 2021 Census Cohort Study

Authors: Nicky Rogers and Louisa Blackwell

Post meeting note

Since presenting our proposal for a 2021 Census Cohort Study to MARP, we have sought ethical approval for a Proof of Concept from the National Statistician's Data Ethics Advisory Committee (NSDEC). On NSDEC advice we have reviewed our approach to the design of the study which has led us to renaming the 2021 Census Cohort Study the "Census 2021 Data Asset".

The Census 2021 Data Asset (CDA) will form the core population for the whole of England and Wales. It will be statistically controlled (through the use if weights) rolled-forward record-level representation of the usually resident population of England and Wales, produced on an annual basis. Whilst the CDA provides the core population, we see satellite cohorts based on samples where the cohort membership comes from the replenished rolled-forward population (CDA). On NSDEC advice we propose a parsimonious approach to admin data linkage for the satellite cohort studies, where we only link the necessary variables to the cohort. The NSDEC application will be made available to view here under Minute and Agenda – July 2022.

Structure:

1. Purpose of paper

This paper provides an overview of the ONS longitudinal data landscape and puts this in context of the proposed new population and social statistics system. Development of the longitudinal components is at an early stage, and we welcome MARP's views on the design principles that are taking shape for a 2021 Census Cohort Study.

This paper describes:

i. Rationale for and purpose of a 2021 Census Cohort Study

ii. How this and other cohort studies fit into the integrated statistical design for the new population and social statistics system, focusing on population

iii. Coherence and benchmarking; the virtuous statistical triangle, borrowing statistical strength across systems, estimates and data

iv. Design principles for the 2021 Census Cohort Study

v. Next steps

This research will provide evidence to support the 2023 Recommendation to the National Statistician on the future of traditional censuses and support the ambition to accelerate the production of more frequent population statistics.

2. Ask for MARP

We ask MARP members to:

- Comment on the typology we have drawn up for longitudinal perspectives (Figure 1).
- Comment on the overall design of the 2021 Census Cohort Study, including any omissions
- Advise on quality assurance plans and design principles for the longitudinal component
- Advise on any further ideas or sources for replenishment of the 2021 Census Cohort Study

3. Background

We have previously presented a paper to MARP (December 2021) outlining plans to develop a demographic accounting model for the population of England and Wales. This research is led by the Social Statistics Transformation, Analysis and Research (SSTAR) Directorate, in collaboration with ONS Methodology, the Universities of Southampton and Warwick, and with John Bryant of Bayesian Demography Limited in New Zealand.

Demographic accounts will provide our best, timely estimates of population stocks and the components of population change. The intention is that they will also

provide benchmarks and control totals for other elements in the transformed statistical system. This includes a rolled-forward 2021 Census Cohort Study.

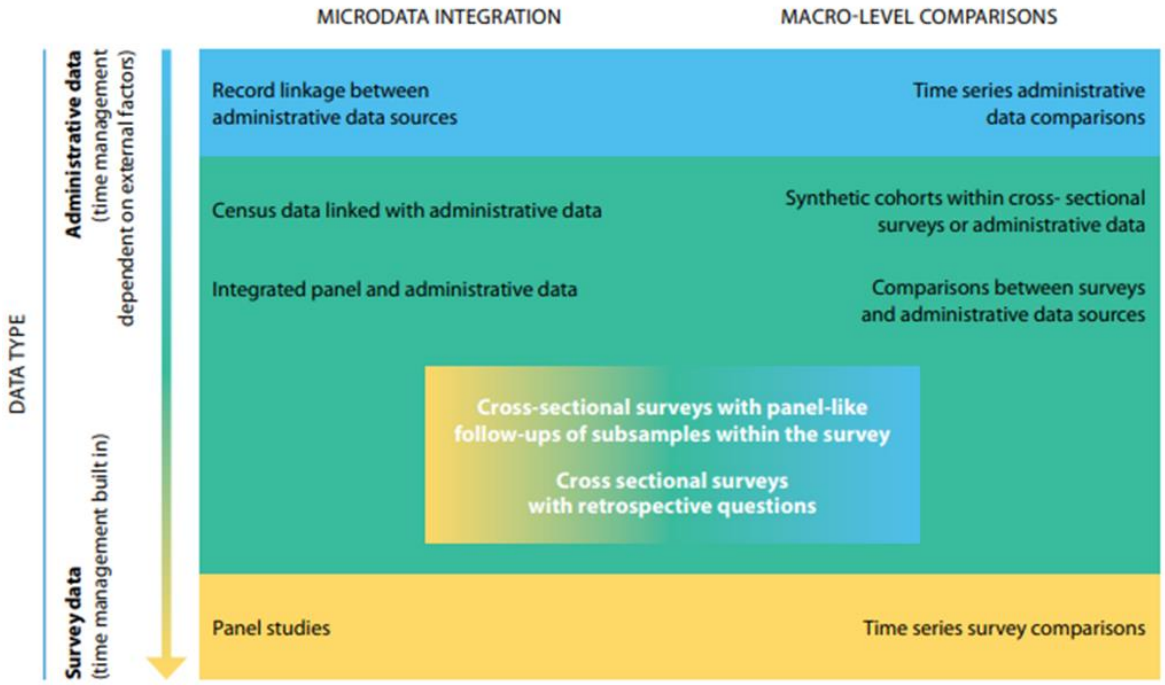
The ONS Health Data Asset demonstrated its analytical value in response to the Covid-19 pandemic, allowing analysis of Covid-related mortality in terms of 2011 Census characteristics. The intention is to build on that experience, adding not just mortality but also other administrative and survey data to a cohort that is defined by presence at the 2021 Census. Unlike the Health Data Asset, for the 2021 Census Cohort Study we aim to maintain the representativeness of this longitudinal asset through cohort replenishment: adding births and immigrants and flagging deaths and embarkation. Longitudinally linked microdata, in combination with study weights to account for cohort replenishment, 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.

We are working closely with the ONS Longitudinal Scientific Advisory Panel (LSAP) and the Census Research Assurance Group (CRAG) to develop our thinking on the design and development of the 2021 Census Cohort. We have incorporated feedback to date from both these advisory panels. We are in the process of engaging with the National Statistician's Data Ethics Advisory Committee (NSDEC) on our plans for such a cohort study.

4. ONS Longitudinal Landscape and Longitudinal Components

Longitudinal data refers to information which is collected from the same units of analysis, such as individuals or households, over time. Different types of longitudinal data are described in Figure 1. For example, cross sectional data are included to show how they can be used to provide a longitudinal perspective on population and their experiences. Complementarity and comparability between longitudinal and cross-sectional measures are also depicted, as well as designs that combine both approaches. (UNECE 2021, Blackwell & Rogers 2021).

Figure 1: Typology of longitudinal data and perspectives



Source: UK Office for National Statistics.

We were consciously inclusive in the development of this typology, recognising that there is international variation in access to and investment in longitudinal datasets. The typology includes 'longitudinal perspectives', gained from what are essentially a series of cross sectional data. Internationally, the longitudinal landscape ranges from Scandinavian countries, with fully integrated population and business registers, to others who rely on censuses and surveys but are committed to increased administrative data use.

Figure 1 aims to summarise the range of administrative and survey-based data balances that exist and is a useful framework for thinking about where we are in terms of a longitudinal landscape within ONS.

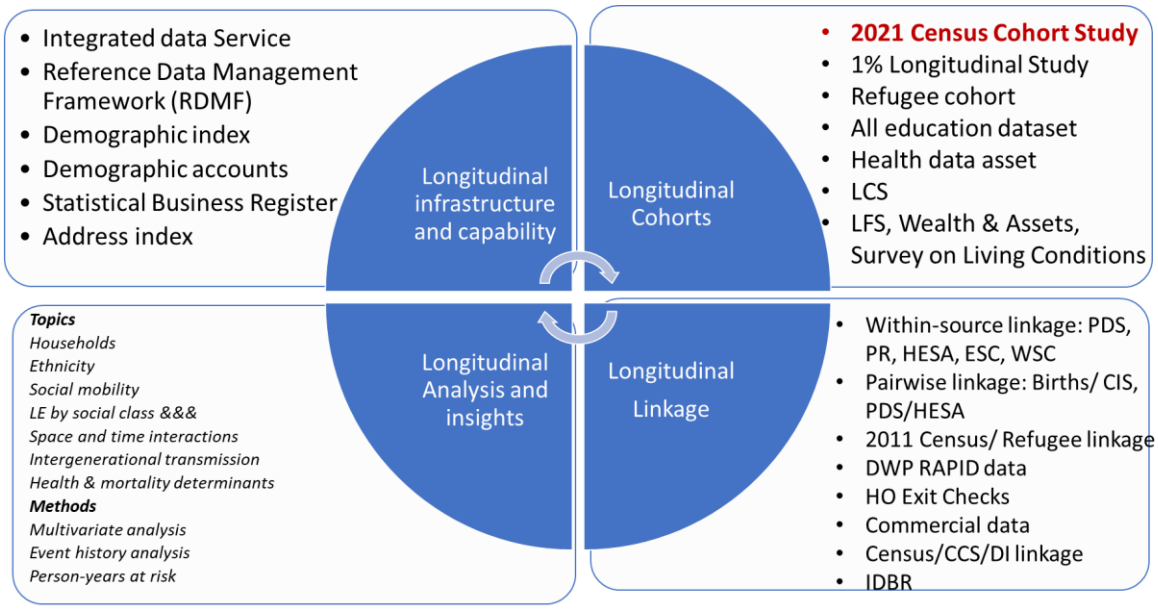
Briefly, in Figure 1, microdata linkage products, which can be considered to be truly longitudinal, follow the same individuals over time and are shown on the left hand side of the figure. Quasi-longitudinal perspectives, for example time series comparisons and synthetic cohorts are shown on the right. Running from top to bottom we capture the administrative (blue) and survey (yellow) data mix. Hybrid longitudinal perspectives are shown in the middle. These hybrid perspectives may include cross-sectional survey with panel follow-up or cross-sectional surveys with retrospective questions. We are not reliant or restricted to one approach. However, an important distinction is that we have less control over design and timing of administrative data collection than for surveys. This typology has relevance broader than international migration and we have applied this to draw up the longitudinal landscape shown in Figure 2.

Figure 2 summarises the current ONS longitudinal landscape and what this may be in the future. Further detail on selected longitudinal components is highlighted in

Appendix 2. We show truly longitudinal cohort studies (the ONS Longitudinal Study (LS) and the Refugee Cohort Study, alongside longitudinal infrastructure and examples of longitudinal linkage. We summarise the topics that longitudinal analysis and insights may deliver, as well as the methods used for analysis.

We have deliberately included Demographic accounts as part of an enabling longitudinal infrastructure. Demographic accounts provide a longitudinal perspective, using the cohort component model to capture our best estimate of the population at risk at any given point in time (monthly estimates will be available).

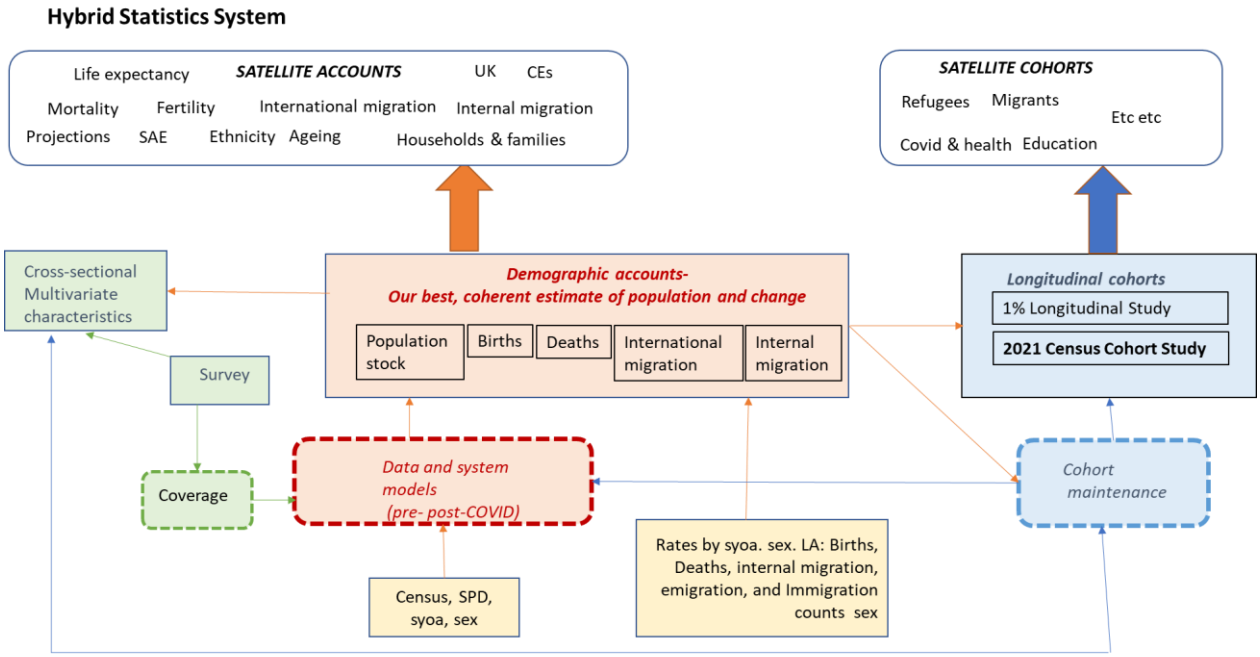
Figure 2: ONS Longitudinal Landscape



5. Integrated statistical design for the new population and social statistics system

At the heart of the new system will be a demographic accounting system. New requirements for very timely population estimates at subnational level have emerged over the past 18 months, in response to the COVID-19 Pandemic. Annual population estimates, lagged by more than 12 months and rebased decennially, are not timely enough. Aggregate-level, model-based approaches are required to respond to and report population dynamics that are not in a 'steady state'. Figure 3 shows how longitudinal data are integral to the proposed new population estimation system.

Figure 3: Overview of the proposed population and social statistics system



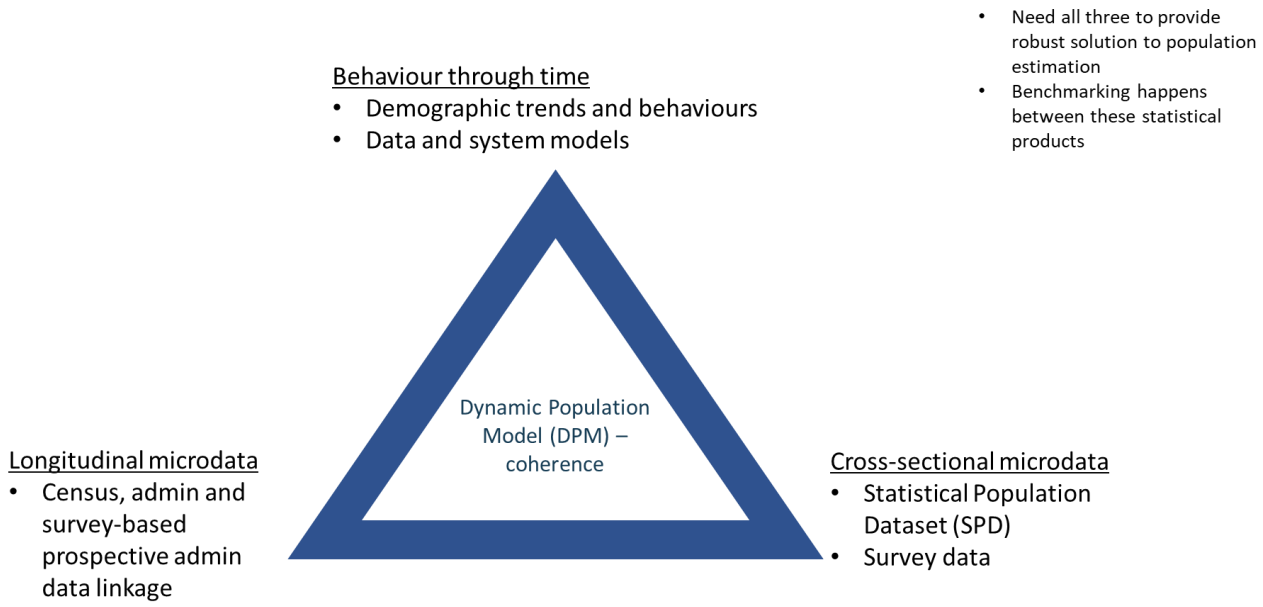
Coherence and benchmarking; the virtuous statistical triangle

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 4.

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) will feed into Demographic Accounts, alongside coverage insights from the rolling survey. Entries to and exits from the 2021 Census Cohort 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 2021 Cohort Study will be investigated and understood.

Figure 4: Benchmarking between statistical products



6. 2021 Census Cohort: Rationale and purpose

The 2021 Census Cohort Study would be a statistically controlled Study covering the population of England and Wales. It is not intended to be a population register, but a study to be used only for research and statistical purposes by approved researchers for approved research projects. The Study is not concerned with individuals; identifying information is being used for linking only and then being removed before analysis takes place, so there will be no impact on individuals.

Whilst we aim for full coverage of the England & Wales population, this is not always possible through under- and over-coverage in the Census and administrative data. Therefore the Study will use statistical methods to address representativeness and will include some element of estimation to account for under-coverage. For example, entries to and exits from the 2021 Census Cohort due to international migration will be constrained, first to provisional and then to final estimates of immigrants and emigrants from the Demographic Accounts.

The full cohort study would have a governance structure to ensure that proposed research projects meet ethical approval and would not harm any member of the cohort. Proposed research projects would be expected to go through an ethics self-assessment that is agreed with the ONS Data Ethics team. Any outputs produced by approved researchers would be subject to disclosure control measures.

Our proposal for a 2021 Census Cohort Study will demonstrate how a replenished cohort study will improve representativeness and provide a flexible and responsive analytical asset to better meet the need for:

- A longitudinal record level asset to support longitudinal and cross-sectional analysis, incorporating the true population at risk, therefore improving on the ONS Health Data Asset. For example, analysis of life expectancy and healthy

life expectancy estimates by characteristics and occupations not possible through the ONS Longitudinal Study (LS) which is a 1% sample.

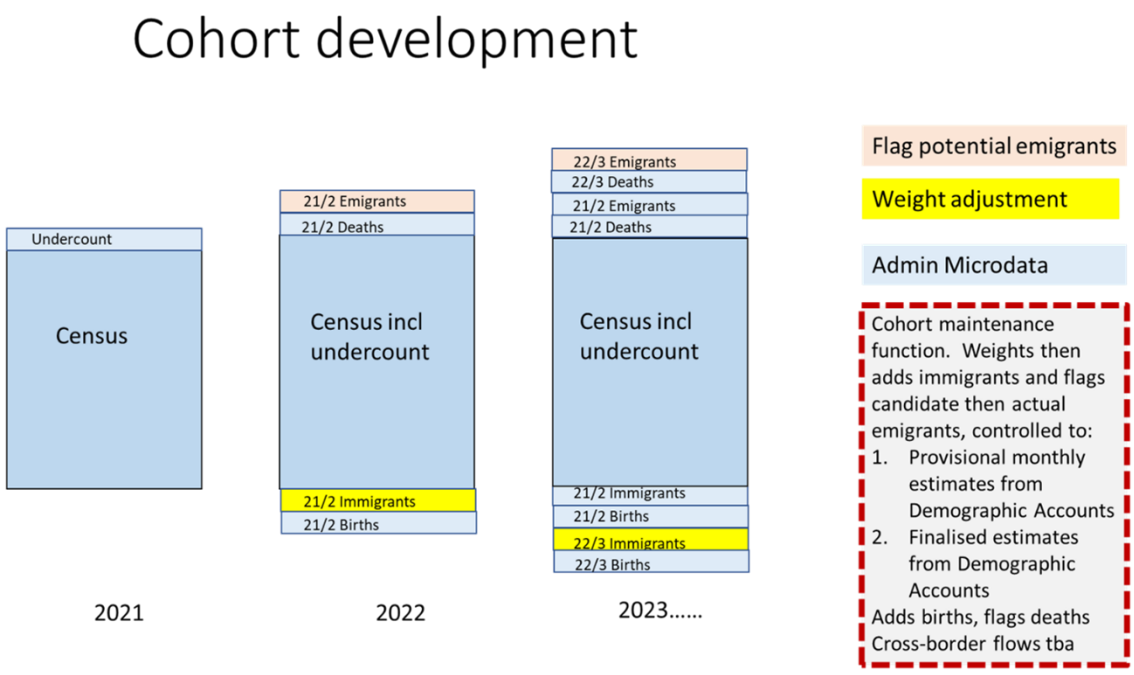
- More granular analysis of the whole England & Wales Population than is currently possible through surveys and the LS to support decision making and investments to improve people's lives. LS numbers quickly fall to single figures when reporting longitudinal outcomes by age and ethnic group sub-nationally. For example, ability to identify changes in morbidity and mortality at a local level across population groups, such as ethnicity, homeless, children in care.
- Additional attributes not currently covered by the LS e.g. educational attainment, health outcomes, providing new insights to inform decision making and target services.
- Inclusivity in our statistics for vulnerable or currently under-represented population groups, supporting decision making to ensure the needs of the whole population can be met. Incorporation of studies such as the Refugee Cohort will strengthen the data inclusivity of the Study.
- Multivariate analysis of different characteristics to improve understanding and monitoring of inequalities and inform policy development, decision making and service provision to improve people's lives, for example outcomes for children in care.
- Ability to understand the outcomes people experience in life, set against their characteristics, lived experiences and behaviours at a local level and therefore identify associated impacts on society, businesses, local economies and the environment.
- Addressing the intercensal gap that currently exists by rolling forward the population and linking data. Therefore building capability to produce fully inclusive statistics on the whole population, not just private households, but people living in communal establishments, travellers, circular migrants, homeless etc.
- Providing insight into where local and national services and infrastructure (transport and construction) are needed; and where policy and resources can be targeted to support economic growth (and "levelling up") and improvements to the environment.

The 2021 Census Cohort Study is aligned to the ONS strategic objective of inclusivity and recommendations made by the National Statistician's Inclusive Data Task Force (IDTF), to reflect the experiences of everyone in our society so that everyone counts, and is counted, and no one is forgotten. This will ultimately help inform local authority, government, charities and other organisations with resource allocation for these vulnerable populations as well as the potential to increase public awareness of societal issues.

The proposal is to repeat the success of the Health Data Asset with a 2021-based cohort (ONS 2021a), but this time including cohort replenishment, with the inclusion

of births and immigrations, alongside flagged deaths and emigration departures. Figure 5 illustrates how this would work.

Figure 5: 2021 Census Cohort replenishment over time



We plan to understand public acceptability of a full cohort study through an engagement strategy with the public and groups representing sectors of society, building on engagement strategies for the 2021 Census ('It's all about us') and the Refugee Cohort Study. We would ensure that are transparent about the role and purpose of the Study and how personal information is being managed and used.

Through these communications we will address public acceptability of a full 2021 Census Cohort Study. We will communicate the need for such a study to our key audiences, and ensure they are aware of the benefits of the study for the public good and that it will not be used for harm i.e. analysis of an individual's outcomes. We plan respond to feedback transparently by publishing outcomes from any engagement.

Initially we propose a proof of concept to test the feasibility of a replenished population ahead of deciding to go ahead with a full cohort study. The proof of concept will produce the following outputs:

- Evidence based recommendation for a 2021 Census Cohort Study design that will meet needs for representative statistics on our society.
- Adaptive linkage methods to optimise linkage for those people who may be excluded due to linkage failure. We plan to build on methods developed through the Refugee Cohort Study

- Use of weights from Demographic Accounts to produce a provisional representation of the population and then later confirmed as final.
- Research findings and recommendations to feed into a consultative report for users of our statistics and the general public

6.1. Quality assurance to measure success

We are starting with a proof of concept project to test the feasibility of producing a rolled-forward true representation of the England and Wales population at mid-year 2021 and a *provisional* mid-year 2022 population. If successful we aim to then full out the full cohort study. Success will be determined by how well our proposed maintenance schedule preserves the cohort (see sections 6.4 and 7). We welcome feedback from the Panel on other criteria:

1. Linkage rates and precision metrics by key characteristics (e.g. age, sex, ethnicity), with a focus on harder to link populations. This will include comparisons with ONS Longitudinal Study (LS) linkages as a 'gold standard' benchmark.

2. Loss to follow up metrics by specific population characteristics (e.g. age, sex, ethnicity, in a communal establishment, migratory status).

2. Coverage of hard to reach populations and their characteristics in administrative data.

3. Consistency of stocks and flows for the rolled forward 2021 mid-year population with estimates produced by the Demographic Accounts.

4. Proportion of immigrants and emigrants correctly flagged and then updated with an administrative data record.

5. Monitor signals for identifying potential under-coverage or linkage failure e.g. unlinked deaths or unlinked births to sample mothers.

6. Assessment of how well the cohort study : (1) maintains households (as defined by a census), (2) updates characteristics captured in censuses e.g. ethnicity, occupation, geographical location and, (3) can produce alternative measures for social class and socio-economic group.

6.2. Ethical and legal considerations

A comprehensive strategy of legal, ethical and statistical assurance will inform the development of the 2021 Census Cohort Study.

Legal advice on the linkage of additional data will be source-specific, related to the conditions of data supply to ONS.

In parallel with this proposal the Integrated Statistical Design Team is developing proposals for promoting best practice in the linkage of data for marginalised groups, drawing lessons from the Refugee Cohort Study to develop best practice for other

marginalised and statistically less- or invisible groups in society. This strategy will add statistical strength to the 1% ONS LS and 2021 Census Cohort Study.

We are aware that any divergences between population estimates produced from the Demographic accounts and cross-sectional populations implied by the 2021 Census Cohort Study may surface undocumented populations. We are seeking National Statistician's Data Ethics Advisory Committee (NSDEC) advice on our plans and their approval to develop a proof of concept study. We will return to NSDEC at each stage of implementation; we are following a similar strategy for the development of the Refugee Cohort Study, which is welcomed by NSDEC.

6.3. Data security and statistical disclosure rules for access and dissemination

We aim to protect the confidentiality of the 2021 Census Cohort data through the following protocols:

1. Data linkage will be undertaken by ONS data linkage experts within ONS secure environments.

2. A clear separation between ONS analysts who access identifying data and ONS analysts who access the resultant linked data with all personal identifiers removed.

3. Assignment of a non-disclosive study member and family member IDs to protect confidentiality.

4. A statistical disclosure control strategy to manage the risk of identification of study members. This will include 'top-coding' of variables that otherwise allow the identification of individuals through either their characteristics, living arrangements or geography.

5. Derived or more aggregated variables that yield the appropriate analytical value of the data while protecting the confidentiality of the individual records.

6. Use of 'X-Files[1]' and disclosure control practices similar to the ONS Longitudinal Study (LS) to facilitate epidemiological research that requires accurate time to event analysis.
7. Removal of all personal identifiers from the analysis database to be made available to researchers via the Integrated Data Service.

[1] X-Files contain identifiable data, for example Postcode for linkage to ecological variables, dates e.g. date of birth and will only be accessed by ONS approved researchers in the DAP environment.

8. Researchers accessing the database are Approved Researchers and research proposals are endorsed by a Project Board.

9. A user support service to monitor access, agree research proposals, provide support to users when creating analysis datasets from the Analytical Database, output clearance and quality control.

10. This study is also strictly for research purposes and therefore the statistical disclosure control strategy will ensure that it cannot be used operationally at an individual level. We will engage with ONS statistical disclosure control experts to ensure that the data and the study database will be protected to the standards of the Data Protection Act (2018) and the General Data Protection Regulation (2016). Statistical disclosure control will include organisational and technical controls separating linkage and analysis functions, pseudonymisation and suppression of identifying characteristics.

6.4. Cohort creation and maintenance

In 2021 the Census cohort would include all usual residents found in the Census. We would use signs of life in administrative data and Census/ Demographic Index matching to include microdata records for the estimated under-coverage, matching on characteristics identified by the Census coverage adjustment methodology where possible and using Census field information to identify non-responding households. There may be some over-coverage in the Census, for example duplicate records for individuals that are not removed. The process of linking the 2021 Census to the 1% ONS LS does identify duplicate records for LS members, which we will draw upon. We will also develop linkage routines that can surface possible duplicates at pre-processing stages and flag ‘primary’ records.

For 2022, 2021/2 deaths would be flagged, and new birth records added. For births we would include data available at birth notification and registration and on entry to the Personal Demographic Service, including ethnicity. Our cohort maintenance functions will aim to address orphan records, i.e. unlinked deaths or births not linked to mothers. Whilst they won’t retrospectively create a new study member, as they may be a visitor to the UK or short-term migrant, they will be flagged and added to linkage updates.

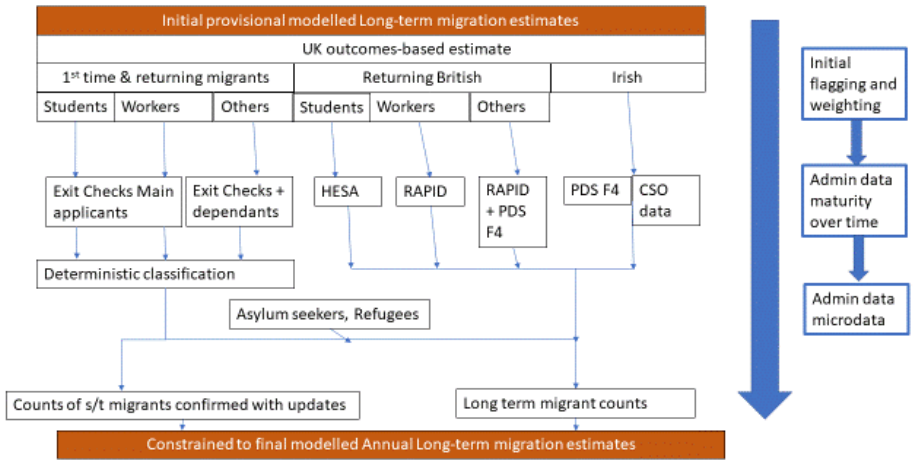
Adding and removing migrants is essential if we are to maintain the representativeness of the longitudinal cohort. Migrants are a particularly challenging group to follow through time (UNECE, 2021). We will draw them into the 2021 Census Cohort in a statistically controlled way, using our best available estimates of international migration from the Demographic Accounts. Unless we do this within a statistically controlled cohort maintenance process, we run the risk of snowballing cohort inflation as short-term migrants and visitors would be erroneously added to the cohort and emigrants remain unflagged.

In addition, non-Western names pose a range of record linkage challenges and false negative linkage could generate multiple identities. Part of this process will be to explore use of the Demographic Index (DI) and Reference Data Management Framework (RDMF) for maintaining the 2021 Census Cohort. As the DI aims to represent a complete picture of the 'ever registered admin data based population', understanding the DI linkage processes, linkage quality for marginal populations, and how best to identify usual residents is essential and high priority. Work is underway to link 2021 Census/CCS records to the DI within CCS postcode clusters.

A positive feedback loop from the longitudinal evidence back to the Demographic Accounts would signal areas for investigation. For example if administrative data implied more migrants that were not included in the Study. Since Home Office border data and NHS Flag 4 information, for example, are used in international migration estimation in the Demographic Accounts, this would suggest model failure and would require review.

For 2021/2 immigrants and emigrants we would adjust using the provisional estimates from our Demographic Accounts. For immigrants we would use weights to flag potential immigrants, for example by age, sex, nationality. In this way we would have a temporary migrant population able to accrue events. For emigration we would flag as candidate emigrants a sample of records with no recorded admin activity, selected using the characteristics determined by migration estimation (age, sex nationality, reason for migration). A cohort maintenance strategy would periodically replace 'candidate emigrants' with new admin-based activity with matching (on characteristics) records that remained dormant. Careful thought is required on the treatment of visa expiry/ visa extensions/ short trips away with returns on new visas (see ONS 2017 and 2018). Linkage of Home Office EU Settlement Scheme (EUSS) and citizenship data to the Cohort Study would identify those who subsequently achieved citizenship.

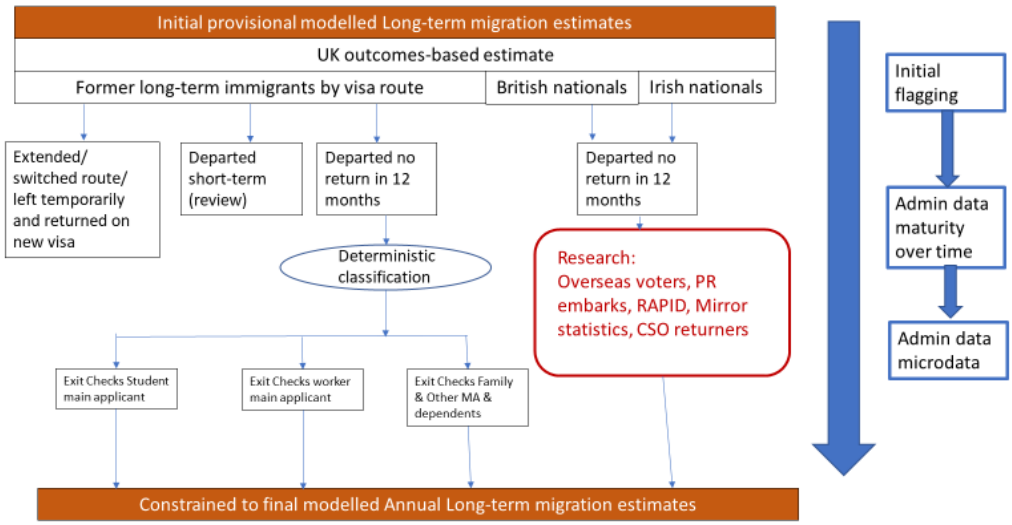
Figure 6: Estimating and adding immigrants to the 2021 Census Cohort



By 2023 we should have fuller admin-based information for both immigrants and emigrants and would aim to replace weights and candidate emigration flags with

microdata for confirmed arrivals and departures, where migrant records have been identified using administrative sources. Home Office border crossings data offer great promise with the integration of European together with non-EU migrants in the Exit Checks programme. Figure 6 above suggests administrative sources for identifying immigrants.

Figure 7: Estimating and flagging emigrants in the 2021 Census Cohort



Emigrating British migrants are the most difficult to verify. Our initial suggestions for identifying migrants in administrative microdata are provided in Figure 7. This is subject to confirmation through research. This statistically controlled approach, using known characteristics of migrant populations, should help to manage the potential error around capturing emigrants. For example, we know the nationalities and highly seasonal nature of student migration and can use this to help to identify student emigrants. Erroneous flagging of candidate emigrants would be corrected through the cohort maintenance regime if activity was detected on flagged records.

A cohort maintenance strategy will also need to consider the possibility that the 2023 Recommendation proposes that there isn't a 2031 Census. We would therefore need to establish options to deal with characteristics not captured in administrative data but rather through surveys, but also populations not captured in administrative data e.g. undocumented migrants. This can be challenging as populations not well captured in administrative data tend not to be well captured in surveys (e.g. migrants).

7. Guiding design principles for the 2021 Census Cohort

Guiding design principles for the Study:

Overarching

1. Cohort stocks and flow distributions should align with the Demographic Accounts; the design of both builds on borrowing statistical strength across the statistical properties of each statistical product. Any divergence will be urgently investigated.
2. The base population will be the population of England and Wales present on 21 March (Census Day) 2021.
3. This Cohort will form the basis for the 2021-based Health Data Asset.
4. The study design will be a hybrid of regular linkage of key sources to the study, with bespoke user-defined linkage available on a case-by-case basis.
5. Incorporation of studies such as the Refugee Cohort would strengthen the data inclusivity of the 2021 Census Cohort Study, but this should be at no cost to linkage quality for the Refugee Cohort, which would retain ring-fenced clerical resource to minimise loss to follow-up.

Retrospective data linkage:

6. Linkage to 2011 Census will provide both 2011 and 2021 Census information for those present at both, supporting longitudinal analysis.
7. Linkage to 2011 Census will provide some historic information for the admin records used to populate the 2021 undercount and help to assess the plausibility of this proposed admin microdata usage.

Prospective data linkage:

8. Only high-quality administrative sources will be allowed to create new records for immigrants, including NHS Personal Demographic Service (PDS), Home Office Border Crossings, Higher Education Statistics Agency (HESA) data, all with associated admin-based activity.

   The quality standard for adding records to maintain statistical integrity of the cohort will higher than is required for the attribution of characteristics and events. We plan to use error frameworks (Blackwell & Rogers 2021) to understand coverage and potential biases of data sources and use existing source specific knowledge from past and current research to inform use of each data source. We will validate new records by linking to other sources for example, the use of school census/PDS to validate immigrant children added via Exit Checks A list of sources is appended.

9. Emigration is the most difficult event to capture in administrative data as there is no duty on emigrants to de-register anywhere. Home Office Exit Checks data will

help us to identify emigration of visa-holders. We will research the use of overseas UK voter registration, to help identify British emigrants (we expect to identify fewer than 10% through this route) but also, perhaps most fruitfully, to help us to understand the admin data shadow that is left by those living abroad, for example in HMRC data, so that we can recognise 'candidate emigrants' through these patterns where this is necessary . The error associated with identifying the wrong candidates will be quantified and shared with users.

Cohort maintenance:

10. Cohort maintenance will be a regular processing activity (addition of weekly or monthly births and deaths, for example) ensuring that active records are not flagged as emigrants and proactively maintaining the integrity and representativeness of the cohort study. The frequency of updates is yet to be determined but will be dependent on data availability.

    We will identify quality signals that demand corrective action, for example a rise in linkage failure for deaths or births to sample mothers will alert us to under-coverage in the study.

11. Cross-referencing linkage between the ONS Longitudinal Study (LS) and the 2021 Census Cohort Study will help us to identify where linkage processes can be improved in both studies. It is envisaged that clerical input to the Study linkage will be minimal but focused on harder to link populations.

12. Additional data can efficiently be added to the LS through passing over relevant links from the 2021 Census Cohort Study cohort. This would be to add admin-based attributes with a known and acceptable linkage error, not to influence LS cohort membership.

13. We will explore the use of survey data, potentially the Labour Market Survey (LMS), to report on drift in the accuracy of intermediate-level characteristics that are not captured well or in a timely way in administrative data due to lags, including household size/ structure/ formation/ dissolution and internal migration. A longitudinal coverage survey will need to follow people through time which in itself will present challenges and therefore may not resolve the issues it set out to address in the first place. For example may suffer from loss to follow up or population coverage of difficult to reach groups (migrants).

14. We plan to monitor and address linkage failure through a maintenance strategy and provide reporting on quality metrics. We will assess the quality of our linkages to adjust and improve linkage algorithms, for example using the Longitudinal Study as a quality benchmark against which linkage can be assessed.

15. We will aim to understand reasons for loss to follow up in the Study and develop strategies for the treatment of loss to follow up in different population sub-groups and provide quality metrics to assess potential bias in the Study among these groups due to 1) linkage failure, 2) data set coverage, 3) data missingness 4)

event missingness. We are in the process of developing a loss to follow up strategy for the Refugee Cohort Study which we will adapt for other populations e.g. migrants.

Characteristics over time:

16. Administrative data attributes will be used to derive or impute Census-type characteristics with known and shared accuracy levels. However, this will need further exploration, for example, social class or socio-economic group. We aim to integrate work currently being carried out by ONS colleagues to investigate the use of administrative or survey data to replace or enhance questions in censuses.

17. Co-residents and their characteristics in households are identified through linkage to the 2021 Census. We believe the Census to be the best source even though it only gives us a point in time viewpoint. Co-residents will not be followed up. Further research is needed to understand household formation and dissolution implied through admin data sources. At best admin data provide evidence on co-residents at an address rather than household composition.

18. Differences in characteristics recorded on admin-data sources will be accepted into the study to respect that certain characteristics (gender, ethnicity, nationality) can change over time. This will allow researchers to decide how to deal with changes over time.

Governance, data security and statistical disclosure rules for access and dissemination

19. Study governance and access will draw lessons from the ONS Longitudinal Study (LS), with a dedicated user support service to support and promote use of this study.

20. All processing and analysis of the data will be carried out in such a way that Cohort members cannot be identified.

8. Next Steps

We welcome MARP's views and recommendations on our proposal for a 2021 Census Cohort Study. Specifically at this stage we are seeking comment on:

- The typology we have drawn up for longitudinal perspectives (Figure 1).
- The overall design, including any omissions
- Advice on the design principles for the longitudinal component
- Advice on any further ideas or sources for replenishment of the 2021 Census Cohort Study

We are developing a series of work packages to attend to the most pressing design considerations for the Study, including but not limited to:

i. Requirements gathering including stakeholder engagement

ii. Whether and how we can maintain population characteristics, households, families and geographies over time through administrative data linkage

iii. Ethical, legal and security considerations

iv. Systems and environments

v. Inclusion strategy for disadvantaged groups or hard to link populations e.g. homeless, veterans, non-household populations such as children in care, prisoners as they may 1) not be in the administrative data at all, 2) be in the data but not identifiable through a disadvantage variable, 3) or identifiable, but their data are of poorer quality/suboptimal.

vi. Assessment of existing data linkage strategies and use in the Study

vii. Dealing with loss to follow up and maintenance schedule

viii. Integration with the Dynamic Population Model

ix. Uncertainty in admin data, particularly around lags for use to measure international or internal migration. For example, the use of fractional counting for 1) weighting or inclusion rules in the study population, 2) geographical placement for usual residence (dealing with misplacement or lags in administrative data) or 3) assignment of characteristics e.g. ethnicity

9. References

Astin, A. W. and Boruch, R.F. (1970) A 'Link' System for Assuring Confidentiality of Research Data in Longitudinal Studies, American Educational Research Journal, Vol 7. No 4, Nov 1970, pp 615-624

Blackwell, L., & Rogers, N. (2021). A Longitudinal Error Framework to Support the Design and Use of Integrated Datasets. In Measurement Error in Longitudinal Data. : Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198859987.001.0001/oso-9780198859987-chapter-3.

Office for National Statistics (2017) International student migration research update: August 2017, Newport, Wales.

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/internationalstudentmigrationresearchupdate/august2017

Office for National Statistics (2018) Report on international migration data sources: July 2018, Newport, Wales.

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/reportonthecomplexityandqualityofinternationalmigrationstatistics/july2018

Office for National Statistics (2020a) Pilot record linkage study for Vulnerable Persons and Vulnerable Children's Resettlement Scheme data, Application to the National Statistician's Data Ethics Advisory Committee, Newport Wales.

https://uksa.statisticsauthority.gov.uk/publication/nsdec-minute-agenda-and-papers-february-2020/

Office for National Statistics (2020b), Census Refugees Matching Methodology Report, *Unpublished,* Newport Wales.

Office for National Statistics (2021a) Statistical properties of coronavirus (COVID-19) mortality data: error in longitudinally linked survey and administrative sources, Newport Wales.

https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/statisticalpropertiesofcoronaviruscovid19mortalitydataerrorinlongitudinallylinkedsurveyandadministrativesources

Office for National Statistics (2021b), Vulnerable Persons and Vulnerable Children Resettled Refugee Linkage Pilot: Methodology Report, *Unpublished* Newport Wales.

Office for National Statistics(2021c), Longitudinal cohort study for refugees. Application to the National Statistician's Data Ethics Advisory Committee, Newport Wales.

https://uksa.statisticsauthority.gov.uk/publication/nsdec-minute-and-agenda-february-2021/

UNECE (2021) Guidance on the use of longitudinal data for migration statistics, United Nations Economic Commission for Europe, February 2021, Geneva.

https://unece.org/statistics/publications/guidance-use-longitudinal-data-migration-statistics

OPCS (1973) Cohort Studies: New Developments, London: HMSO

Appendix 1: List of data sources to be used in cohort development

We plan to use the 2021 Census as the population base and link high quality administrative data to replenish the cohort to produce a 2022 population and then a 2023 population. Data will include:

a) 2021 Census (post edit and item imputation) and Coverage Survey data to form the population base. Census Non-Response Linkage Study data, Census Address Register and history files.
b) NHS Digital Personal Demographic Service (PDS) data to identify addresses and individuals missed by the Census. Monthly PDS update files will identify internal migration moves since the Census, new patient registrations from abroad and embarks.
c) England & Wales birth notification and registration data to update the cohort with births since the Census.
d) England & Wales death registration data to update the cohort with flagged deaths since the Census.
e) England & Wales Marriages, Civil Partnerships. Divorces and Civil Partnership dissolutions data to address missed links due to name changes.
f) Home Office Exit Checks data to update the cohort with new flagged immigration records since Census Day for EU and non-EU nationals subject to immigration control and then later confirmation after enough time has elapsed. To also flag potential EU and Non-EU emigrants after Census Day and then later confirmation after enough time has elapsed.
g) Home Office EU Settlement Scheme data for those with pre-settled/settled status – to use as an outcome flag for settled status.
h) Home Office Citizenship data to use as an outcome flag for British citizenship.
i) Electoral Registers for England & Wales data for overseas voters to flag British emigrants not identified in PDS data.
j) HMRC P85 data to flag possible emigrants (those who notified HMRC they are moving abroad)
k) English and Welsh School Census, Individual lifelong learning records for England and Wales, and HESA data (UK) to validate Census records not linked to PDS where children are present/validate PDS records not linked to Census where children are present.

Appendix 2:

ONS Longitudinal Study

The ONS Longitudinal Study (LS) contains linked administrative and Census records for a representative sample of 1% of the population of England and Wales. Starting in 1973, it now contains data from six Censuses (1971-2021) and supports a vast array of research projects, providing unique longitudinal insights to support policy.

Record linkage for the LS is conducted for ONS by NHS Digital, using the Master Patient Service. The linkage strategy was originally intended to provide data protection through the separation of analysis and linkage processes (OPCS, 1973 and Astin and Boruch, 1970).

In the context of the transformed population and social statistics system, the power of the LS is that it supports the analysis of longitudinal outcomes across the life course and inter-generationally. Since the high accuracy of LS linkage, including clerical search using longitudinal history, isn't feasible for a full population cohort, maintaining independence between the LS and the 2021 Census Cohort Study would support cross-referencing of linkage to identify where linkage processes can be improved in both studies.

The ONS Health Data Asset

The ONS Health Data Asset, which linked 2011 Census with deaths and health data to provide pivotal timely evidence on the Pandemic last year demonstrated the high value of linking events prospectively to Census data (ONS 2021a).

Our experience of working on the Health Data Asset emphasised the importance of study representativeness over time (i.e. representative of the population at risk of death involving coronavirus on 2 March 2020).

By design, as people emigrate and die, they are not replaced with new births and immigrants in the Health Data Asset, so the cohort becomes less representative of the contemporary population over time. Clearly this becomes a major problem where we are measuring outcomes by nationality, ethnicity or migratory status.

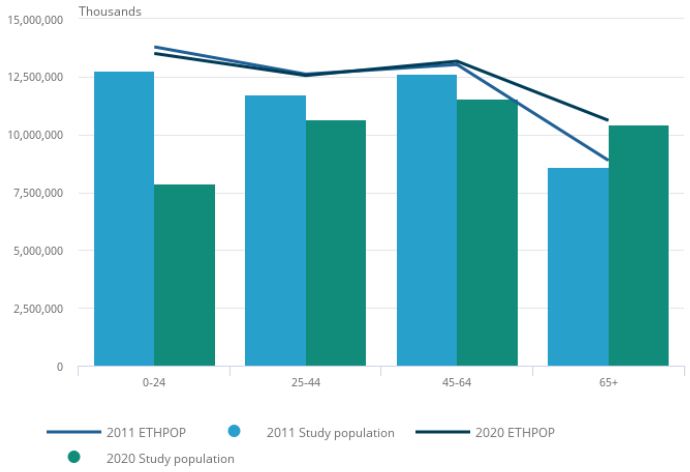
Figures 8 and 9 compare broad age, sex distributions for the White and Black Ethnic Groups observed in the ONS Health Data Asset study population in 2011 and 2020, to distributions observed in Ethnic Population Projections (ETHPOP)[2] Database for 2011 and 2020.

The 2011 study population closely mimics the 2011 ETHPOP projection by broad age, as ETHPOP uses Census as its base. By simulating the future study population (without replenishment), the absence of children born after the 2011 Census and

[2] ETHPOP projections use Census, survey, official Mid-Year estimates and Vital Statistics data for England, Wales, Scotland, and Northern Ireland. Office of National Statistics (ONS), General Register Office of Scotland (GROS) and Northern Ireland Statistics and Research Agency (NISRA) provide the data.

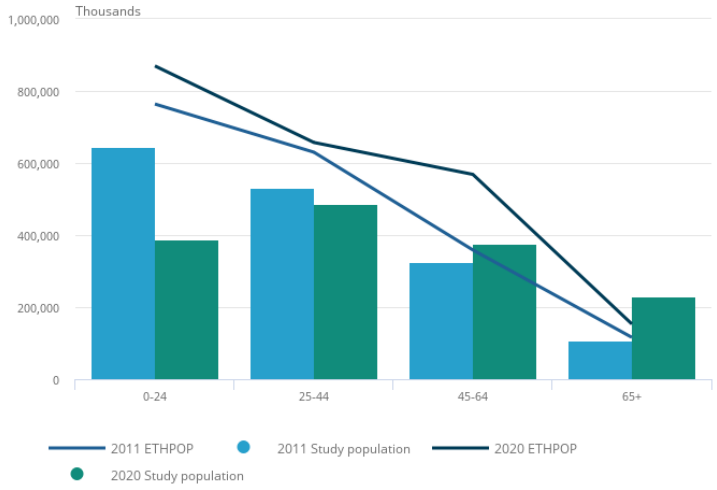
immigrants affects the representativeness of the younger population the most. The population aged 65 years and over is the least affected. The undercount of the study population at younger age groups is much more pronounced for ethnic minority groups, reflecting immigration over the period.

Figure 8: Representativeness of the study population, England and Wales, by broad age, White ethnic group, 2011 and 2020



Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data.

Figure 9: Representativeness of the study population, England and Wales, by broad age, Black ethnic group, 2011 and 2020



Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data.

Longitudinal Refugee Cohort Study

A collaboration between ONS and the Home Office (HO), the Longitudinal Refugee Cohort Study aims to create a research resource that provides unique insights into social and economic integration and COVID-19 outcomes for cohorts of approximately 103,000 asylum route refugees and 22,000 resettled refugees (via the Vulnerable Persons (VPRS) and Vulnerable Children's (VCRS) Resettlement Schemes). Cohorts are based on those who arrived in the UK between 2015 and 2020.

A pilot study (ONS, 2020a) was developed to address design considerations for a future refugee cohort study based on longitudinally linked admin data. The pilot aims included, developing linkage algorithms to deal with non-Western naming conventions, assessing linkage rates and evaluating the viability of a future refugee cohort study.

The pilot study achieved linkage rates of 96.9% with precision[3] of 99.9% (ONS 2021a) when VPRS/VCRS records were linked to Home Office Exit Checks data. The pilot also achieved linkage rates of 96.3% with precision of 100% when linked to Patient Demographic Service data (ONS 2021b).

Within admin data source longitudinal linkage

Within ONS Population and Migration Statistics Transformation (PMST), the Data Sources and Quality work package (DSQ) aims to link individual administrative data sources longitudinally (e.g. NHS Personal Demographics Service (PDS), Higher Education Statistics Agency (HESA), and the English School Census (ESC). Each longitudinally linked dataset consists of four tables:

1. A high level table with monthly activity and absence flags

2. A demographic table which includes DoB, Sex, Gender, Ethnicity etc.

3. A detailed activity and absence table that provides information on the type of activity or absence e.g. Started School or Embarked

4. A geography table containing a first address and then any subsequent changes

The intention is to use these datasets to gather detailed information on activity and absence of recorded individuals to inform and improve other PMST products, such as the Statistical Population Dataset (SPD) and estimates of internal migration.

[3] Precision is defined as the proportion of links made that are true matches and is used as a standardised measure of linkage quality. A precision rate of 99.9% means that 99.9% of the links made were true matches.