Office for National Statistics

# Statistical properties of coronavirus (COVID-19) mortality data: error in longitudinally linked survey and administrative sources

Analysing the quality of the bespoke study dataset used in the published ONS articles on deaths involving the coronavirus (COVID-19) by ethnicity in England and Wales.

Contact:
Nicky Rogers
demographic.methods@ons.gov.uk
+44 (0)1329 444866

Release date:
4 June 2021

Next release:
To be announced

## Table of contents

# 1 . Introduction

This article describes the quality of the bespoke study dataset used in the published Office for National Statistics (ONS) articles on deaths involving the coronavirus (COVID-19) by ethnicity for England and Wales. This article should be read before undertaking any analysis of the bespoke dataset.

Ethnicity is not recorded on the death certificate. Deaths were linked to 2011 Census microdata to provide self-reported ethnicity and other demographic characteristics for the deceased. Deaths that occurred from Census Day on 27 March 2011 up to 28 July 2020, and registered by 24 August 2020, were linked. This linkage is continuing to support further reporting on deaths during the coronavirus pandemic. The period from 27 March 2011 to 28 July 2020 was investigated but the study dataset specifically focused on the period between 2 March 2020 to 28 July 2020. The study dataset is available to accredited researchers in the ONS Secure Research Service (SRS). For further details see the SRS website.

ONS methodologists worked alongside ONS mortality experts to understand and report on the quality of the linked census and deaths data. We applied our longitudinal error framework to these linked survey and administrative data to understand how statistical error can occur. We have assessed these sources of error and have provided quality indicators to report and address them.

# 2 . High-level description of the study dataset and linkage methods

A prospective cohort study design was used where deaths to individuals were linked to their characteristics measured at a point in time (the 2011 Census). A prospective cohort study follows a group of individuals (cohort) over time to determine how certain characteristics can affect rates of a certain outcome (in this case, mortality).

A dataset was created which included all records in the England and Wales 2011 Census microdata and linked death registration records within England and Wales from 27 March 2011 to 28 July 2020. It was created first by linking the 2011 Census to the GP Patient Register (PR) records between 2011 and 2013, and adding an NHS number (a unique personal identifier) to each linked Census record.

Death registration records were linked to the 2011 Census microdata on NHS numbers between 28 March 2011 and 28 July 2020, registered by 24 August 2020. Further deterministic methods using personal identifiers were used to link death registrations directly to the 2011 Census without using a NHS number. The linkage rate for deaths occurring between 28 March 2011 and 28 July 2020 was 90.2%.

To form the study dataset, the linked data was filtered to include deaths that occurred between 2 March 2020 and 28 July 2020. It was created to analyse mortality among ethnic groups during the coronavirus (COVID-19) pandemic. The population at risk of death involving COVID-19 were those alive on 2 March 2020. The dataset was updated weekly with linked death registration records. It contains deaths linked to usual residents and non-usual residents[1] for the reporting period 2 March to 28 July 2020. At the time of linkage, 2 March 2020 was the date of the first recognised death involving COVID-19 in England and Wales that had been registered within the study period. This will therefore exclude any deaths involving COVID-19 registered in late January and February 2020, of which there were very few.

Table 1 shows the number of linked and unlinked deaths at various stages of the dataset production. In this article our initial assessment of data quality was based on the first iteration of these data (deaths registered by 17 April 2020) and later on two further iterations (registrations by 29 May and 24 August, respectively), as well as all deaths linked to the 2011 Census for England and Wales since 28 March 2011. We highlight throughout the report which iteration this quality analysis refers to.

Table 1: Number of linked and unlinked death registration records by dataset iteration 2 March to 28 July 2020

| Iteration | Reporting period for deaths occurring between: | Total deaths occurring in period | Linked deaths in period | Unlinked deaths in period | Linkage rate (%) |
|---|---|---|---|---|---|
| 1 | 2 March 2020 and 10 April 2020, registered by 17 April 2020 | 75,905 | 68,155 | 7,750 | 89.79 |
| 2 | 2 March 2020 and 22 May 2020, registered by 29 May 2020 | 160,886 | 145,976 | 14,910 | 90.73 |
| 3 | 02 March 2020 and 28 July 2020, registered by 24 August 2020 | 253,194 | 229,983 | 23,211 | 90.83 |
| 4 | 27 March 2011 and 28 July 2020, registered by 24 August 2020 | 4,880,407 | 4,400,129 | 480,278 | 90.16 |

Source: ONS analysis of linked 2011 Census microdata and death registration records

Notes

1. 2011 Census data includes non-usual residents.

### Notes for: High-level description of the study dataset and linkage methods

1. Non-usual residents are identified through the 2011 Census “Intention to Stay” question as they would have entered the UK in the year before the 2011 Census took place. We recommend excluding non-usual residents from any analysis of these data because of their high propensity to have left the UK by March 2020.

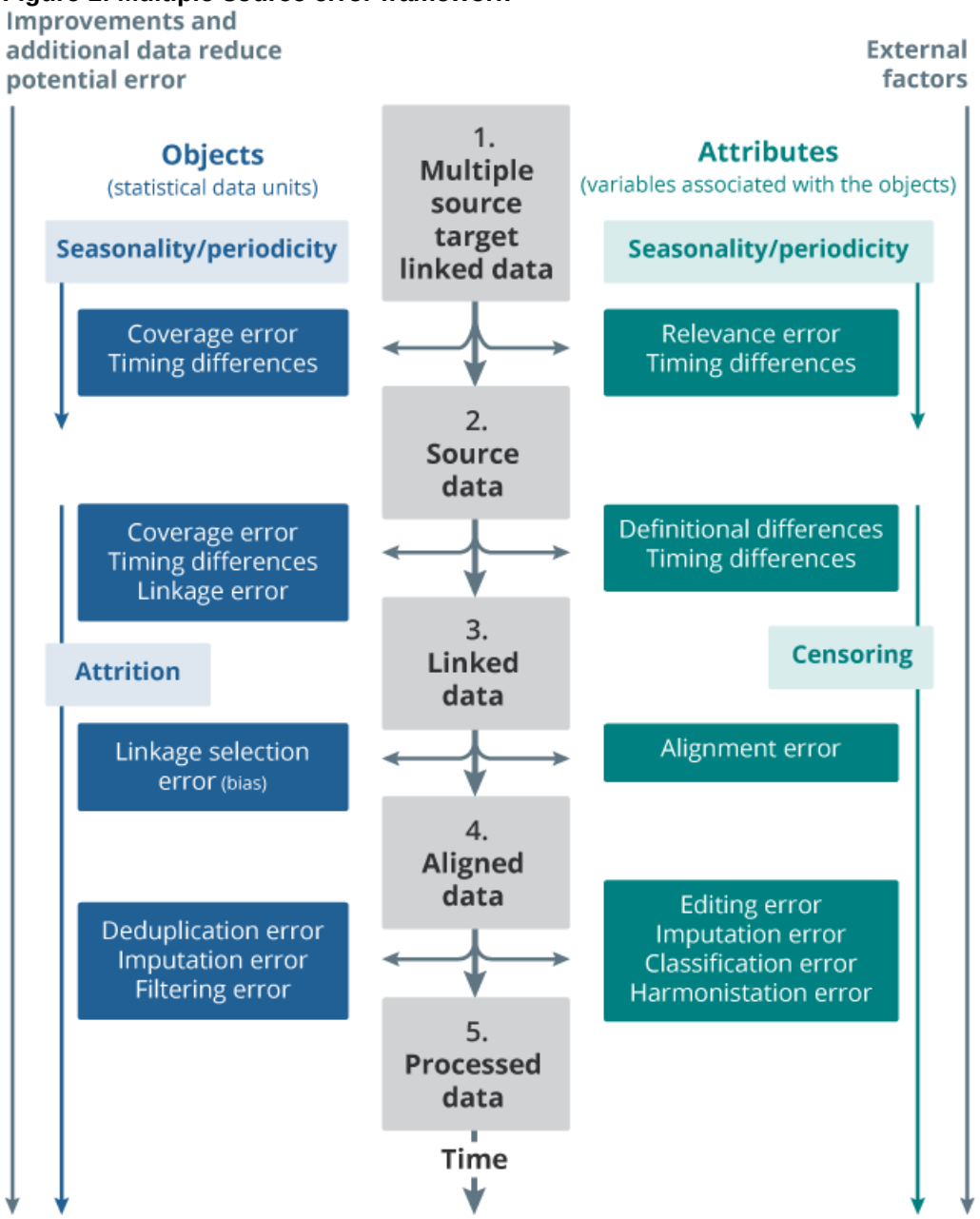
## 3 . Our longitudinal error framework

Office for National Statistics’ (ONS) methodologists have developed an error framework for longitudinally linked data. The framework can be applied to linked survey and administrative data, in this case the linked 2011 Census and death registrations data for England and Wales. It is designed to help researchers and statisticians understand the strengths and limitations of source data for linkage and the data created through the linkage of different datasets.

Building on the framework developed by Statistics New Zealand, we developed a general framework for single longitudinal data sources and multiple sources that are longitudinally linked. In this article we apply the multiple-source error framework to the linked census and death registrations data. Further information on statistical error are provided in the Appendices of our error framework report, published in February 2020.

The multi-source framework (Figure 1) addresses error arising through the processing and linking of objects (the statistical data units, for example, individuals) and their attributes (the variables associated with these individuals). Section 4 considers error introduced through dataset coverage and representativeness, linkage, and timing differences and Section 5 considers the relevance and consistency of ethnicity and household composition over time.

**Figure 1: Multiple-source error framework**




**Source: Office for National Statistics adaptation of Statistics New Zealand error framework**

**Notes:**

1. Time is presented vertically
2. Target linked data represent the ideal linked data to be produced.
3. Objects (statistical data objects) refer to rows in the data source, and what those rows of data represent.
4. Attributes (variables associated with the objects) refer to columns in the data source, and what those columns represent.
5. More information of the terms used can be found in Appendix A.

# 4 . Understanding study dataset coverage

Ideally the study population would represent the population at risk of death involving coronavirus (COVID-19) on 2 March 2020. Here we report on errors where there is a difference between our ideal study population and the study population resulting from linking the 2011 Census to death registration records. The types of error that occur include:

- coverage error – people in our ideal study population that are not in our source data; subsection Census coverage error in Section 4 covers people not enumerated in the 2011 Census
- linkage bias – bias in the representativeness of the study population; Representativeness of the analysis dataset in Section 4 analyses differential linkage rates between different groups in the linked 2011 Census and death registration records false positive linkage error[1] and false negative linkage error[2]; serror in Section 4 estimates linkage error in the linked 2011 Census and death registration records
- timing differences – errors in coverage caused by timing differences between the source data

There will be under-coverage in the study because of immigration and births since 27 March 2011. There will be also be over-coverage in the population at risk on 2 March 2020 if emigrants and deaths prior to this are not accounted for. The subsection “Using a simulated population to account for emigration in the cohort study population” in Section 4 describes the methods for developing emigration rates to adjust for emigrations between 27 March 2011 and 1 March 2020. This subsection “Representativeness of the cohort study population over time” in Section 4 uses the Ethnic Population Projections (ETHPOP) database to assess the representativeness of the non-replenished study population by age and ethnic group, after adjustment for deaths and embarkation between 27 March 2011 and 2 March 2020.

## Census coverage error

There is known non-response in the 2011 Census, causing under-coverage error in the 2011 Census microdata. We compared the 2011 Census microdata to the published 2011 Census estimates, which were adjusted for Census non-response. Non-response adjustment weights (1) were calculated by quinary age group, sex and broad ethnic group as follows:

$$\text{Adjustment weight}_{(\text{quinary age, sex, ethnic group})} = \frac{P\text{ublished 2011 Census estimates (quinary age, sex, ethnic group)}}{\text{2011 Census count (quinary age, sex, ethnic group)}}$$

Figures 2 and 3 show coverage shortfalls in the 2011 Census by sex, quinary age, and broad ethnic group. Under-coverage is particularly concentrated in the 20 to 39 years age range and is higher for males and Mixed, Chinese, Black and Other Ethnic minority groups. These weights can either be applied to the analysis dataset or used in models to adjust for this error.

**Figure 2: Coverage weights for the 2011 Census for males, by quinary age group and broad ethnic group**

| Males | All | White | Indian | Bangladeshi & Pakistani | Black | Chinese | Mixed | Other |
|---|---|---|---|---|---|---|---|---|
| 0 to 4 | 1.11 | 1.09 | 1.07 | 1.08 | 1.22 | 1.23 | 1.23 | 1.26 |
| 5 to 9 | 1.09 | 1.07 | 1.06 | 1.08 | 1.19 | 1.17 | 1.19 | 1.22 |
| 10 to 14 | 1.07 | 1.06 | 1.05 | 1.07 | 1.16 | 1.15 | 1.16 | 1.20 |
| 15 to 19 | 1.08 | 1.07 | 1.08 | 1.08 | 1.16 | 1.18 | 1.19 | 1.28 |
| 20 to 24 | 1.15 | 1.12 | 1.17 | 1.14 | 1.28 | 1.35 | 1.30 | 1.46 |
| 25 to 29 | 1.17 | 1.14 | 1.18 | 1.16 | 1.36 | 1.41 | 1.40 | 1.59 |
| 30 to 34 | 1.14 | 1.11 | 1.11 | 1.12 | 1.31 | 1.32 | 1.37 | 1.48 |
| 35 to 39 | 1.10 | 1.08 | 1.07 | 1.09 | 1.22 | 1.24 | 1.26 | 1.33 |
| 40 to 44 | 1.07 | 1.06 | 1.05 | 1.07 | 1.18 | 1.17 | 1.21 | 1.25 |
| 45 to 49 | 1.06 | 1.05 | 1.04 | 1.05 | 1.14 | 1.11 | 1.17 | 1.17 |
| 50 to 54 | 1.04 | 1.04 | 1.03 | 1.05 | 1.10 | 1.10 | 1.13 | 1.14 |
| 55 to 59 | 1.04 | 1.03 | 1.03 | 1.04 | 1.10 | 1.08 | 1.13 | 1.13 |
| 60 to 64 | 1.03 | 1.03 | 1.02 | 1.03 | 1.08 | 1.07 | 1.10 | 1.09 |
| 65 to 69 | 1.03 | 1.02 | 1.02 | 1.03 | 1.07 | 1.06 | 1.09 | 1.10 |
| 70 to 74 | 1.03 | 1.02 | 1.02 | 1.03 | 1.06 | 1.07 | 1.10 | 1.11 |
| 75 to 79 | 1.03 | 1.02 | 1.03 | 1.03 | 1.07 | 1.07 | 1.11 | 1.10 |
| 80 to 84 | 1.03 | 1.03 | 1.03 | 1.04 | 1.06 | 1.08 | 1.11 | 1.10 |
| 85+ | 1.03 | 1.03 | 1.03 | 1.04 | 1.07 | 1.09 | 1.10 | 1.12 |
| Total | 1.08 | 1.06 | 1.08 | 1.09 | 1.19 | 1.23 | 1.23 | 1.30 |

**Source: 2011 Census microdata and published 2011 Census**

**Notes:**

1. Data are for England and Wales 2. Non-usual residents are excluded

**Figure 3: Coverage weights for the 2011 for females, by quinary age group and broad ethnic group**

| Females | All | White | Indian | Bangladeshi & Pakistani | Black | Chinese | Mixed | Other |
|---|---|---|---|---|---|---|---|---|
| 0 to 4 | 1.11 | 1.09 | 1.07 | 1.09 | 1.21 | 1.21 | 1.23 | 1.25 |
| 5 to 9 | 1.09 | 1.08 | 1.06 | 1.08 | 1.18 | 1.14 | 1.19 | 1.22 |
| 10 to 14 | 1.07 | 1.06 | 1.05 | 1.07 | 1.15 | 1.13 | 1.16 | 1.18 |
| 15 to 19 | 1.08 | 1.07 | 1.07 | 1.07 | 1.16 | 1.15 | 1.18 | 1.22 |
| 20 to 24 | 1.10 | 1.09 | 1.09 | 1.08 | 1.18 | 1.23 | 1.21 | 1.26 |
| 25 to 29 | 1.08 | 1.07 | 1.06 | 1.06 | 1.15 | 1.19 | 1.18 | 1.23 |
| 30 to 34 | 1.07 | 1.06 | 1.05 | 1.06 | 1.12 | 1.15 | 1.14 | 1.19 |
| 35 to 39 | 1.05 | 1.05 | 1.04 | 1.05 | 1.11 | 1.12 | 1.13 | 1.16 |
| 40 to 44 | 1.04 | 1.04 | 1.03 | 1.04 | 1.09 | 1.09 | 1.11 | 1.13 |
| 45 to 49 | 1.04 | 1.03 | 1.03 | 1.03 | 1.08 | 1.07 | 1.09 | 1.10 |
| 50 to 54 | 1.03 | 1.03 | 1.02 | 1.03 | 1.07 | 1.06 | 1.09 | 1.09 |
| 55 to 59 | 1.02 | 1.02 | 1.02 | 1.03 | 1.07 | 1.06 | 1.08 | 1.09 |
| 60 to 64 | 1.02 | 1.02 | 1.02 | 1.03 | 1.06 | 1.06 | 1.07 | 1.09 |
| 65 to 69 | 1.02 | 1.02 | 1.03 | 1.03 | 1.06 | 1.07 | 1.09 | 1.10 |
| 70 to 74 | 1.03 | 1.02 | 1.03 | 1.04 | 1.06 | 1.07 | 1.09 | 1.10 |
| 75 to 79 | 1.03 | 1.03 | 1.03 | 1.04 | 1.06 | 1.08 | 1.08 | 1.10 |
| 80 to 84 | 1.03 | 1.03 | 1.03 | 1.04 | 1.06 | 1.10 | 1.06 | 1.10 |
| 85+ | 1.04 | 1.04 | 1.03 | 1.05 | 1.06 | 1.09 | 1.08 | 1.09 |
| Total | 1.06 | 1.05 | 1.05 | 1.06 | 1.13 | 1.14 | 1.17 | 1.17 |

**Source: 2011 Census microdata and published 2011 Census**

**Notes:**

1. Data are for England and Wales 2. Non-usual residents are excluded.

## Representativeness of the analysis dataset

Linkage rates can only ever be 100% when we expect all individuals on source A to be on source B with certainty. Lower linkage rates could be because of links that have been missed, but also other factors such as coverage error, selection error or simply the absence of individuals or a population sub-group from one of the data sources.

Death registration records will not link to the 2011 Census when:

- people were not enumerated in the 2011 Census
- children were born after 27 March 2011 (Census day)
- people immigrated to England or Wales after 27 March 2011
- people moved from Scotland or Northern Ireland after 27 March 2011 (cross-border flows)
- there were false negative linkage errors (links that were missed)

Linkage failure can introduce bias into the study dataset if the missing data are not random. Understanding the representativeness of the study dataset allows us to account for this bias in the interpretation of analytical results.

In Section 2, Table 1 provides overall linkage rates for each iteration of the analysis dataset. Here we examine linkage rates by age and sex, country of birth and ethnic group. We compare linkage of the 2011 Census and death registration records for all causes of death and for deaths involving coronavirus (COVID-19) occurring between 2 March 2020 and 22 May 2020.

Table 2 highlights the high linkage rates achieved across all deaths and deaths involving COVID-19 at around 90%. Linkage rate cannot reach 100% because some deaths were of children born after the 2011 Census, people immigrating since the Census, or deaths of 2011 Census non-responders. However, linkage rates are unlikely to be significantly impacted by post-Census migration and Census non-response as most deaths occurred at older ages, and these issues are more typical among younger age groups.

Table 2: Linkage rates for all deaths and COVID-19 deaths, 2 March and 22 May 2020.

| | Unlinked | Linked | Total | Match rate (%) |
|---|---|---|---|---|
| **All deaths** | 14,910 | 145,976 | 160,886 | 90.73 |
| **COVID-19 deaths** | 4,392 | 39,228 | 43,620 | 89.93 |

Source: ONS analysis of linked 2011 Census microdata and death registrations records

Notes

1. Deaths involving COVID-19 were identified through the International Classification of Diseases codes (ICD) “U071” (confirmed COVID-19) and “U072’” (suspected COVID-19).

2. Non-usual residents in the 2011 Census were filtered out.

3. Deaths of non-usual residents of England and Wales according to death registrations were filtered out.

Like all deaths, COVID-19 deaths were concentrated among those aged 65 years and over. Linkage rates for all-cause deaths and deaths involving COVID-19 were higher for both males and females aged 65 years and over, and lower in the younger age groups (Table 3). The linkage rates for COVID-19 deaths were only marginally lower than the rates for all-cause deaths, except for the linkage rates for 0 to 24 year olds. Although there are relatively few deaths in this age group, it is likely that the low linkage rate for all-cause deaths reflects mortality in infants who do not have a Census record.

Table 3: Linkage rates for all deaths and COVID-19 deaths by age and sex, 2 March to 22 May 2020

| | All deaths | | COVID-19 deaths | |
|---|---|---|---|---|
| | Linkage rate (%) | Linked | Linkage rate (%) | Linked |
| **Females** | | | | |
| **0-24 years** | 25 | 72 | 80 | 12 |
| **25-44 years** | 78.56 | 645 | 81.4 | 140 |
| **45-64 years** | 87.4 | 5,680 | 85.95 | 1,254 |
| **65+ years** | 92.27 | 67,129 | 91.22 | 16,154 |
| **Males** | | | | |
| **0-24 years** | 22.28 | 82 | 75 | 15 |
| **25-44 years** | 69.26 | 721 | 67.03 | 185 |
| **45-64 years** | 79.96 | 7,766 | 78.58 | 2,223 |
| **65+ years** | 92.04 | 63,881 | 91.04 | 19,245 |

Source: ONS analysis of linked 2011 Census microdata and death registrations records

Notes

1. Deaths involving COVID-19 were identified through the International Classification of Diseases codes (ICD) “U071” (confirmed COVID-19) and “U072” (suspected COVID-19).

2. Non-usual residents in the 2011 Census were filtered out.

3. Deaths to non-usual residents of England and Wales according to death registrations were filtered out.

Linkage rates were higher for deaths to people born in the UK. Therefore, deaths to people born overseas are under-represented in the linked 2011 Census and death registration records (Table 4).

Table 4: Linkage rates by country of birth category, 2 March to 22 May 2020

| | All deaths | | COVID-19 deaths | |
|---|---|---|---|---|
| | **Linkage rate (%)** | **Linked** | **Linkage rate (%)** | **Linked** |
| **United Kingdom** | 92.2 | 129,398 | 92.25 | 32,828 |
| **EU** | 82.11 | 5,085 | 80.69 | 1,433 |
| **Europe (non-EU)** | 83.11 | 792 | 82.56 | 322 |
| **Africa** | 77.37 | 2,339 | 75.66 | 1,113 |
| **Middle East/Asia** | 80.52 | 5,158 | 79.76 | 2,227 |
| **Americas and the Caribbean** | 80.52 | 2,546 | 80.79 | 1,089 |
| **Other** | 89.05 | 179 | 87.5 | 42 |

Source: ONS analysis of linked 2011 Census microdata and death registrations records

Notes

1. Deaths involving COVID-19 were identified through the International Classification of Diseases codes (ICD) “U071” (confirmed COVID-19) and “U072” (suspected COVID-19).

2. Non-usual residents in the 2011 Census were filtered out.

3. Deaths to non-usual residents of England and Wales according to death registrations were filtered out.

4. Deaths with the country of birth code of “969” have been removed from analysis, as we have been unable to find out what this code stands for.

Linkage rates are highest for the White ethnic group, and lowest for Black, Chinese, Indian and other ethnicities (Table 5). Therefore, deaths of people of non-White ethnicity will be under-represented in the linked 2011 Census and death registration records.

Table 5: Linkage rates by broad ethnic group, 2 March to 22 May 2020

| | All deaths | | COVID-19 deaths | |
|---|---|---|---|---|
| **Ethnicity** | **Linkage rate (%)** | **Linked** | **Linkage rate (%)** | **Linked** |
| **White** | 91.65 | 136,963 | 91.29 | 34,677 |
| **Mixed** | 82.18 | 772 | 82.1 | 260 |
| **Indian** | 79.38 | 2,305 | 78.76 | 1,028 |
| **Bangladeshi & Pakistani** | 83.53 | 1,698 | 84.85 | 788 |
| **Chinese** | 75.68 | 324 | 75.18 | 138 |
| **Black** | 79.94 | 3,026 | 81.97 | 1,518 |
| **Other** | 75.54 | 1,319 | 76.91 | 645 |

Source: ONS analysis of linked 2011 Census microdata and death registrations records

Notes

1. Based on deaths occurring between 2 March 2020 and 22 May 2020.
2. Deaths involving COVID-19 were identified through the International Classification of Diseases codes (ICD) “U071” (confirmed COVID-19) and “U072” (suspected COVID-19).
3. Non-usual residents in the 2011 Census were filtered out. 4. Deaths to non-usual residents of England and Wales according to death registrations were filtered out.
4. Deaths to non-usual residents of England and Wales according to death registrations were filtered out.

## Linkage error

Linkage errors occur from linking records incorrectly (false positive error) and failing to link records together that should have been linked (false negative error). The two types of linkage error trade off each other and the consequence of each type of error should be considered when linking data. False positive linkage error can affect the accuracy of analytical findings from linked data, whereas false negative linkage error may introduce bias into the analysis datasets if particular sub-groups of records are more or less likely to link.

A clerical review of a sample of linked and unlinked record pairs is typically used to estimate false negative and false positive error rates. In this section, the false positive linkage rate for the linked 2011 Census and death registration records is estimated in this way. However, the false negative linkage rate is calculated by estimating and removing other reasons for linkage failure (such as coverage error). This is also broken down by ethnicity.

## False positive linkage error in the Census deaths linked dataset

The precision of the linked 2011 Census and death registration records was assessed through a clerical review of approximately 2,000 linked records based on deaths occurring between 2 March and 22 May 2020. The false positive error rate is estimated to be 0.2% (2), which indicates that the precision of the linked data is very high.

$$\frac{\text{True positives}}{(\text{True positives} + \text{false positives})} x\,100$$

## False negative linkage error in the unlinked deaths

Death registration records may not link to the 2011 Census for England and Wales when:

- people were not enumerated at the 2011 Census
- children were born after 27 March 2011 (Census day)
- people immigrated to England or Wales after 27 March 2011
- people moved from Scotland or Northern Ireland after the 27 March 2011 (cross-border flows)
- there were false negative linkage errors (links that were missed)

For unlinked death registration records we have attempted to account for each other source of linkage failure to isolate and estimate the unlinked deaths relating to false negative linkage error. This analysis is based on the fourth iteration of the analysis dataset, that is, deaths from 27 March 2011 (Census day) to 28 July 2020, registered by 24 August 2020.

Between 27 March 2011 and 28 July 2020 there were 4,880,332 deaths in England and Wales. Of these, 4,400,129 (90.2%) were linked to the 2011 Census and 480,203 were not (9.8%). Of the unlinked deaths, 28,149 (5.9%) can be accounted for by births since 27 March 2011 (deaths to under 10 year olds[3]). Coverage weights were used to estimate the population not enumerated by Census, by age, sex, and ethnic group. Census undercount (non-response) was estimated to account for 128,479 (26.8%) of unlinked deaths. The number of deaths to post-Census migrants was estimated by applying mortality rates to IPS estimates of migration[4]. This method estimated that 17,686 (3.7%) unlinked deaths were to immigrants over the period.

It is assumed that the remaining 305,631 unlinked deaths out of a total of 4,880,332 deaths are because of false negative linkage error or cross-border flows since 27 March 2011 (Table 6). From this the overall false negative linkage rate is estimated at 6.3% (3).

$$\frac{\text{All unlinked deaths} - (\text{deaths to under 10 year olds} + \text{deaths to census non} - \text{responders} + \text{deaths to immigrants})}{(\text{All linked deaths} + \text{All unlinked deaths})}$$

Table 6: Estimates of the sources of linkage error for unlinked deaths

| | Estimate of unlinked death registration records | % of unlinked deaths |
|---|---|---|
| **Children born after 2011 Census** | 28,149 | 5.87 |
| **Census under-coverage** | 128,479 | 26.77 |
| **Immigrants arriving after 2011 Census** | 17,686 | 3.69 |
| **Remaining unlinked deaths** | 305,631 | 63.684 |
| **TOTAL** | 479,945 | 100 |

Source: Analysis of unlinked death registration records that occurred between 27 March 2011 and 28 July 2020

Notes

1. Data are for England and Wales.

## Ethnic differences in false negative linkage

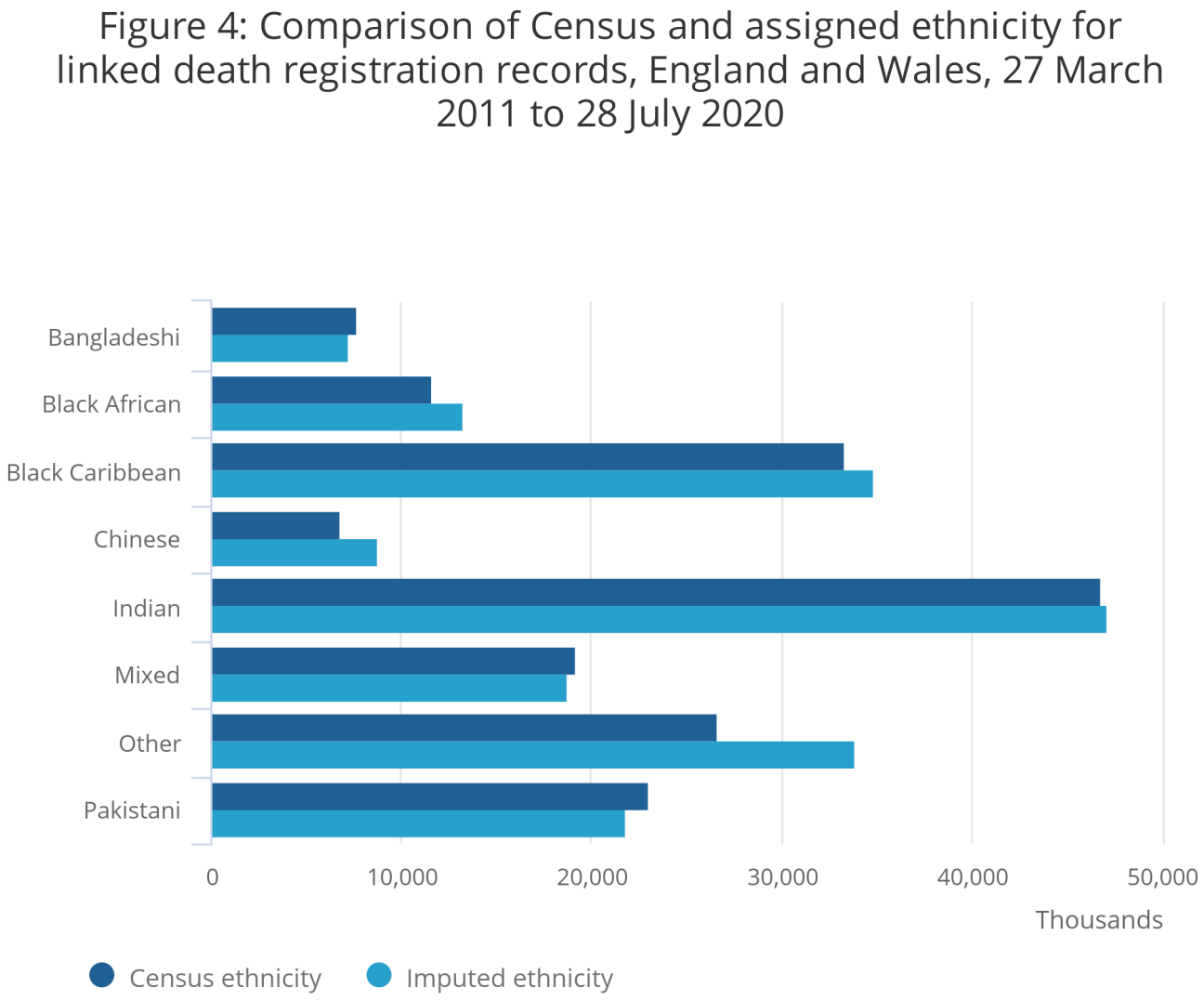
Given the known ethnic differences in COVID-19 mortality, it is important to understand whether false negative linkage error varies by ethnic group and therefore whether higher false negative linkage error for Ethnic minority groups could be masking or distorting excess COVID-19 mortality.

Since ethnicity is not recorded in the deaths data but the country of birth is, the ethnic distributions were inferred by broad age and sex using a country of birth to ethnic group lookup using the whole enumerated 2011 Census population. This lookup was used to assign ethnicity to the unlinked deaths.

Any error from assigning ethnicity using a lookup method was evaluated by comparing the distribution of assigned ethnicities for unlinked deaths against the 2011 Census ethnicities for the linked deaths. This comparison assumes that the country of birth to ethnicity relationship is the same for linked and unlinked deaths.

Figure 4 shows that the distributions of assigned and self-reported ethnicity for deaths are very similar. However, assigning ethnicity using a lookup attributed more Chinese and Other ethnicities than reported at the 2011 Census (a difference of 28.5% and 27.1% respectively). This implies that the country of birth and ethnicity lookup used to derive ethnicity for unlinked deaths, which was based on the entire enumerated 2011 Census population, is not representative of those from the Census population that subsequently died.

**Figure 4: Comparison of Census and assigned ethnicity for linked death registration records, England and Wales, 27 March 2011 to 28 July 2020**




**Source: Analysis of linked death registration records to 2011 Census microdata that occurred between 27 March 2011 and 28 July 2020**

**Notes:**

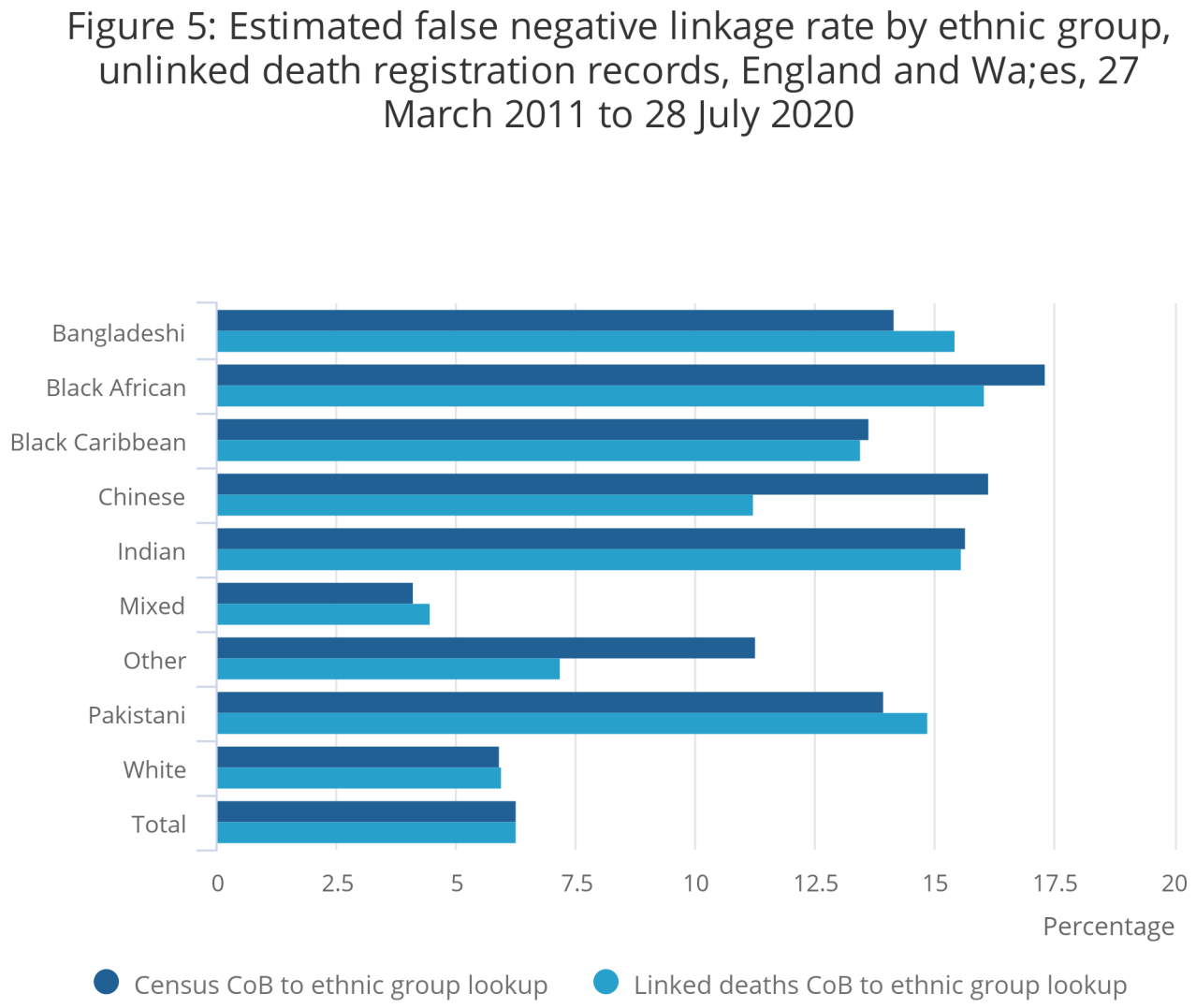
1. The White ethnic group is excluded because it dwarfs the other ethnic groups. There were 4,214,360 deaths assigned as White compared to 4,224,207 self-reported White ethnicities in the 2011 Census.

2. While there were 713 linked deaths with an unknown self-reported ethnicity from Census, this category was not used in the country of birth to ethnic group lookup.

Given the differences seen for the Chinese and Other ethnic groups in Figure 4, we created a lookup between country of birth and ethnic group using only the linked death registration records. This assumes that the relationship between country of birth and ethnic group in the linked death registration records is the same for the unlinked death registration records. While this should be more representative of the ethnic composition of the unlinked deaths, lack of data in some age, sex and ethnicity groups in the linked deaths may affect the estimated false negative linkage rate.

Figure 5 shows the false negative linkage rate by ethnic group using the country of birth to ethnic group lookup based on the linked death registration records.

**Figure 5: Estimated false negative linkage rate by ethnic group, unlinked death registration records, England and Wa;es, 27 March 2011 to 28 July 2020**




**Source: Analysis of unlinked death registration records that occurred between 27 March 2011 and 28 July 2020, registered by 24 August**

**Notes:**

1. This figure used the country of birth to ethnic group lookup based on the linked deaths between 28 March 2011 and 28 July 2020.

The results show that the false negative rate is substantially higher for all Ethnic minority groups (except for the Mixed and Other ethnic group) compared with the White ethnic group. The Black African ethnic group has the highest false negative linkage rate at 16.0%, closely followed by Indian (15.6%), Bangladeshi (15.4%) and Pakistani (14.9%). This indicates that the linkage methods are not effectively linking Ethnic minority groups and therefore Ethnic minorities are underrepresented in the study dataset. Further analysis is required to understand aspects of the linkage methodology which are biasing linkage towards people in the White ethnic group.

## Using a simulated population to account for emigration in the cohort study population

The calculation of mortality rates requires estimates of the population at risk of death. The study dataset includes people who emigrated since the 2011 Census, which inflates population denominators as the deaths of this embarked population are not captured in England and Wales deaths data. This section describes emigration rates based on survey and administrative data to account for emigration from the study population between 27 March 2011 and 1 March 2020.

The emigration rates were based on observed and unobserved embarkation information from the GP Patient Register (PR) linked to the Office for National Statistics' (ONS) Longitudinal Study (LS) and on International Passenger Survey (IPS) based measures applied to the study dataset. This is the best evidence currently available for estimating emigration between 2011 Census and the study period.

To adjust for emigration, we derived rates that can be used in weighting. These rates are available on request from demographic.methods@ons.gov.uk.

Further information on the methods used to create the simulated population and emigration rates can be found in Appendix A.

## Representativeness of the cohort study population over time

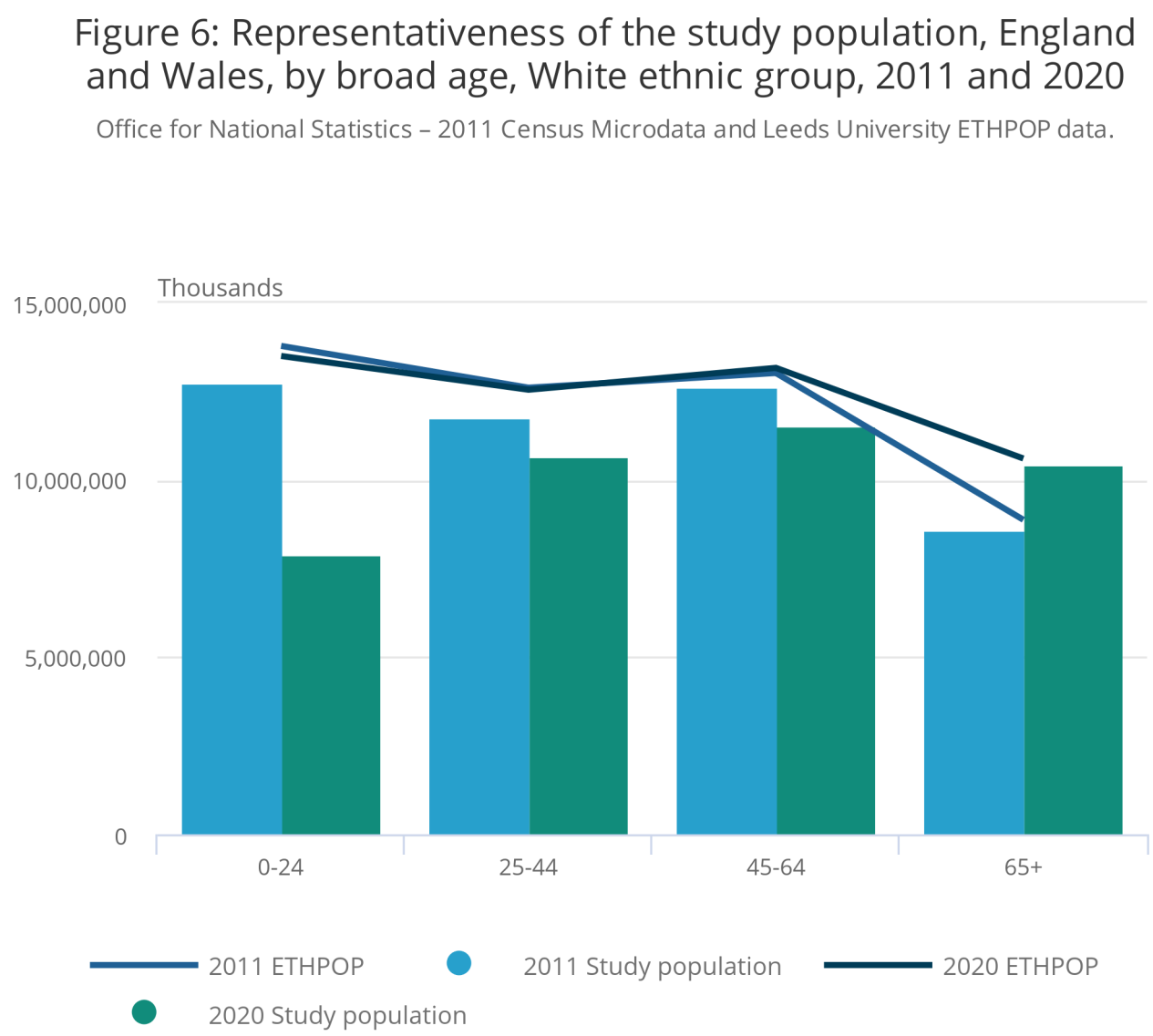
In the subsection Linkage error in Section 4. we were concerned with the representativeness of the linked study dataset by age, sex, and ethnicity. Here we consider how the representativeness of the linked cohort changes over time. By design, as people emigrate and die, they are not replaced with new births and immigrants, so the cohort becomes less representative of the contemporary population over time.

To understand the representativeness of the study population we compared broad age, sex and broad ethnicity distributions observed in the study population in 2011 and 2020 to distributions observed in Ethnic Population Projections (ETHPOP)[6] Database for 2011 and 2020. This is illustrated in Figures 8 to 13 in Appendix B. An example is shown in Figure 6 for the White ethnic group.

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 immigrants affects the representativeness of the younger population the most in the study population. 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 6: 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.**




**Notes:**

1. ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates.

2. Wohland P, Burkitt M, Norman P, Rees P, Boden P and Durham H, ETHPOP Database, ESRC Follow on Fund "Ethnic group population trends". www.ethpop.org. Date of extraction [24,04,2020].

## Using a simulated population to account for new entrants to cohort study population

We explored simulating a population that was representative of the population at March 2020 to produce new weights that could be applied to analysis of the study population. These weights adjust for Census undercoverage (using the weights created in the subsection Census coverage error in Section 4, and account for the addition of births and new migrants using International Passenger Study (IPS) data. We called these replenished population weights.

We found that the replenished simulated population over-estimated the population in 2020 when we compared to Ethnic Population Projections (ETHPOP) (we assumed ETHPOP to be accurate). We concluded that using weights derived from the simulated replenished population would:

- over-estimate the population in all ethnic groups, for both males and females aged 65 years and over and, therefore, under-estimate mortality rates at these older ages
- under-estimate the 0 to 44 years age group for Indian males and females in the population
- under-estimate the 0 to 24 years age group for Bangladeshi and Pakistani males and females in the population
- under-estimate Chinese male 0 to 24 year olds in the population

In developing the replenished simulated population, we were unable to construct equivalent hybrid emigration rates based on the IPS and the Office for National Statistics (ONS) Longitudinal Study for England and Wales (LS) because of time constraints. We needed to construct new LS embarks that accounted for immigration and births. Without these we would not emigrate enough people from the population. We recommend that further exploration of these LS rates is undertaken alongside consideration of using ETHPOP projections for the population at risk.

### Notes for: Understanding study dataset coverage

1. Erroneous links.
2. Linkage failure.
3. Deaths to children born after Census (date of birth 28 Match 2011 onwards) were removed. As deaths did not include a date of birth, age at death and year of death where used. For example, deaths to 0-year olds in 2011, deaths to 0-1-year olds in 2012, continuing up to deaths to 0-9-year olds in 2020.
4. The number of deaths to post-Census migrants was estimated by applying the mortality rates to International Passenger Survey estimates of immigration and deducting the resulting values from the remaining unlinked deaths.
5. If the denominator is adjusted for under-coverage, then these figures must be used to adjust the numerator. Therefore, making these linkage failures relevant.
6. 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.

# 5 . The relevance of 2011 Census characteristics over time

In this section we deal with relevance error introduced from using socio-demographic characteristics collected at the 2011 Census to report on deaths occurring during 2020.

Relevance error can be introduced through timing differences between the ideal measurement of attributes for deceased persons in 2020 and the measure used to capture these attributes (the 2011 Census). Timing differences are a conceptual discrepancy that are difficult to quantify. Here we consider the use of more contemporary data. For ethnicity, we compare ethnicity collected at the 2011 Census with ethnicity collected on Hospital Episode Statistics (HES) data (Stability of ethnicity over time in Section 5). The GP Patient Register (PR) is used to report on household composition (Stability of household composition over time in Section 5).

## Stability of ethnicity over time

To understand error arising from using 2011 Census ethnicity for deaths in 2020 we compared 2011 Census ethnicity to Hospital Episode Statistics (HES) data ethnicity in 2011 to 2012 and 2019 to 2020.

HES data consists of three datasets: accident and emergency (AE), outpatients (OP) and admitted patient care (APC). The information within these three datasets is at episode level (each finished period of care under a consultant). A person-level dataset was created by de-duplicating a NHS number and date of birth. Records with a missing NHS number or date of birth were removed. A primary ethnicity was chosen where the value was not consistent across episodes for a person by taking the modal ethnicity value.

To link 2011 to 2012 and 2019 to 2020 HES data to the 2011 Census microdata, the Census data were first linked to the GP Patient Register (PR) to assign a NHS number to each Census record. HES data were then linked by NHS number and date of birth. The linkage rate of 2011 to 2012 HES data to 2011 Census was 88.5%, and 2019 to 2020 HES data to 2011 Census was 83.7% (Table 7).

Table 7: Linkage rates for 2011 to 2012 HES to 2011 Census and 2019 to 2020 HES to 2011 Census

| | Linked | Total HES | Linkage rate (%) |
|---|---|---|---|
| **2011/12 HES to 2011 Census** | 21,030,406 | 23,774,266 | 88.46 |
| **2019/20 HES to 2011 Census** | 21,436,133 | 25,621,108 | 83.67 |

Source: 2011 Census microdata linked to 2011/12 HES data and 2019/20 HES data

Notes

1. Excludes people born since Census on HES. This is because the data was equivalised and therefore removed individuals under 10 years of age.

The linked Census and HES data were first compared to understand the quality of HES ethnicity in 2011 to 2012, using Census ethnicity as a gold standard. The comparison of ethnicity in HES data for 2011 to 2012 and 2019 to 20 shows how reported ethnicity changes over time and the comparison of linked 2011 Census and HES data in 2011 to 2012 and 2019 to 2020 shows whether the Census ethnicity is still relevant in 2019 to 2020.

It should be noted that HES ethnicity is based on the 2001 Census classification. There are differences between the 2001 Census classification and the 2011 Census classification for ethnicity, which may affect the interpretation of analyses. The differences are listed below:

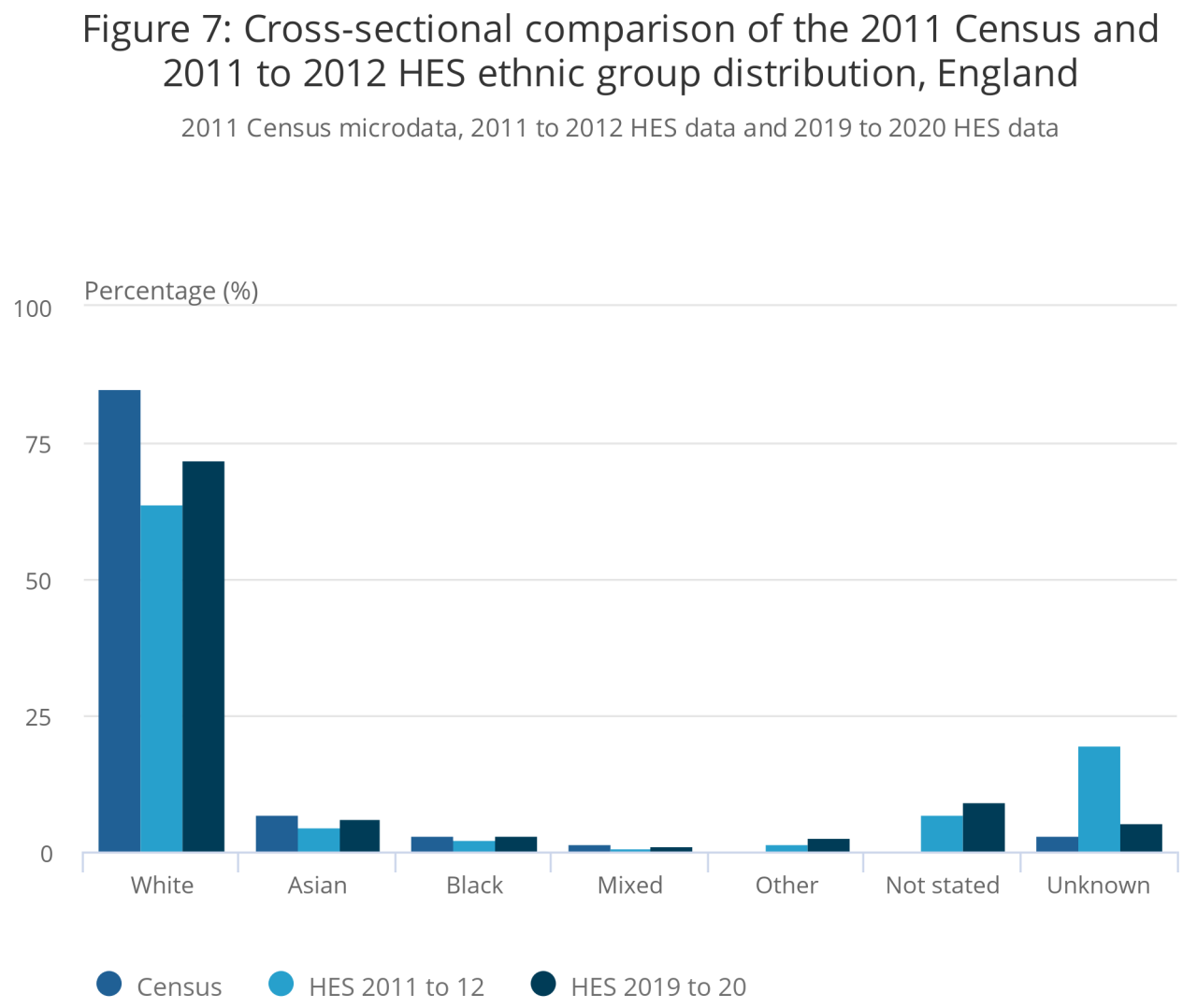
- in 2001, Gypsy or Irish traveller was categorised as Other whereas for 2011 the classification was White
- in 2001, Chinese was categorised as Other whereas for 2011 the classification was Asian
- in 2001, there was no category for Arab whereas for 2011 the classification was Other

## Comparisons of Census ethnicity to HES in 2011 to 2012 and 2019 to 2020

Hospital episode statistics (HES) data contains a higher proportion of ethnicities reported as Unknown. Not stated or Other compared with the 2011 Census, indicating that HES data are less complete. In 2011 to 2012 HES, 28.1% of ethnicities reported are in these three categories compared with 3.5% of 2011 Census (Figure 7). However, this decreased to 17.7% in 2019 to 2020 because of a lower proportion being reported as Unknown ethnicity.

**Figure 7: Cross-sectional comparison of the 2011 Census and 2011 to 2012 HES ethnic group distribution, England**

**2011 Census microdata, 2011 to 2012 HES data and 2019 to 2020 HES data**




**Source: 2011 Census microdata, 2011/12 HES data and 2019/20 HES data**

**Notes:**

1. HES ethnicity is based on the 2001 Census classification.
2. HES contain a category “Not stated” for ethnicity. There is no comparable category on the 2011 Census ethnicity classification.
3. Excludes people born since Census on HES.

A longitudinal comparison of linked 2011 Census and 2011 to 2012 HES ethnicity shows agreement of 69.3% (Table 8). This agreement increases to 80.1% for 2019 to 2020 HES (Table 9) because of improvements in quality of HES ethnicity data over the decade.

Table 8 : Distribution of recorded ethnic group in linked 2011 Census and 2011 to 2012 HES

| Census ethnicity | White | Asian | Black | Mixed | Other | Not stated | Unknown | TOTAL |
|---|---|---|---|---|---|---|---|---|
| **White** | 62.43 | 0.05 | 0.05 | 0.14 | 0.51 | 5.82 | 16.12 | 85.11 |
| **Asian** | 0.19 | 3.95 | 0.03 | 0.08 | 0.29 | 0.50 | 1.50 | 6.55 |
| **Black** | 0.10 | 0.03 | 1.73 | 0.10 | 0.09 | 0.22 | 0.62 | 2.90 |
| **Mixed** | 0.45 | 0.07 | 0.12 | 0.40 | 0.09 | 0.14 | 0.39 | 1.66 |
| **Other** | 0.09 | 0.08 | 0.01 | 0.02 | 0.08 | 0.04 | 0.10 | 0.41 |
| **Unknown** | 1.95 | 0.20 | 0.12 | 0.05 | 0.15 | 0.23 | 0.67 | 3.37 |
| **TOTAL** | 65.21 | 4.38 | 2.06 | 0.79 | 1.21 | 6.95 | 19.40 | 100.00 |

Source: 2011 Census microdata linked to 2011/12 HES data

Notes

1. HES ethnicity is based on the 2001 Census classification.
2. HES contain a category “Not stated” for ethnicity. There is no comparable category on the 2011 Census ethnicity classification.
3. Agreement is based on the sum of percentages across the diagonal in the table (excluding Not Stated and Unknown ethnic groups). Percentages within the table sum to 100%.

Table 9: Distribution of recorded ethnic group in linked 2011 Census and 2019 to 2020 HES
2011 Census microdata linked to 2019 to 2020 HES data

| Census ethnicity | White | Asian | Black | Mixed | Other | Not stated | Unknown | TOTAL |
|---|---|---|---|---|---|---|---|---|
| **White** | 72.46 | 0.07 | 0.06 | 0.25 | 0.97 | 7.08 | 4.17 | 85.05 |
| **Asian** | 0.23 | 4.88 | 0.04 | 0.11 | 0.49 | 0.71 | 0.34 | 6.80 |
| **Black** | 0.11 | 0.04 | 2.00 | 0.13 | 0.20 | 0.35 | 0.15 | 2.96 |
| **Mixed** | 0.53 | 0.08 | 0.13 | 0.49 | 0.14 | 0.19 | 0.10 | 1.66 |
| **Other** | 0.11 | 0.10 | 0.01 | 0.02 | 0.10 | 0.06 | 0.02 | 0.42 |
| **Unknown** | 2.03 | 0.24 | 0.13 | 0.06 | 0.20 | 0.29 | 0.15 | 3.11 |
| **TOTAL** | 75.47 | 5.40 | 2.36 | 1.06 | 2.10 | 8.68 | 4.93 | 100.00 |

Source: 2011 Census microdata linked to 2019/20 HES data

Notes

1. HES ethnicity is based on the 2001 Census classification.
2. HES contain a category “Not stated” for ethnicity. There is no comparable category on the 2011 Census ethnicity classification.
3. Agreement is based on the sum of percentages across the diagonal in the table (excluding Not Stated and Unknown ethnic groups). Percentages within rows and columns sum to 100%.

Table 10 and 11 show the distribution of Census ethnicities within each HES ethnicity (column percentages). Of those recorded as Mixed and Other in 2011 to 2012 HES, only 50.6% and 6.7% were also recorded as Mixed and Other on the 2011 Census. The Other category in HES has been known to be used as a “catch-all” category.

Of those recorded as White, Asian, and Black in 2019 to 2020 HES, there was similar or higher agreement on Census compared with 2011 to 2012 HES data (Table 10). There was, however, lower agreement between the sources for the Mixed and Other ethnic groups compared with the 2011 to 2012 HES data. In the 2019 to 2020 HES data, a higher proportion of the Mixed and Other ethnic groups were classed as White at the 2011 Census, potentially indicating a change in the way ethnicity is reported over time.

Table 10: 2011 to 2012 HES ethnic group distribution by 2011 Census ethnic group
Census microdata linked to 2011 to 2012 HES data

| Census ethnicity | White | Asian | Black | Mixed | Other | Not stated | Unknown |
|---|---|---|---|---|---|---|---|
| **White** | 95.74 | 1.19 | 2.45 | 17.25 | 41.68 | 83.70 | 83.09 |
| **Asian** | 0.29 | 90.18 | 1.63 | 10.65 | 23.99 | 7.26 | 7.73 |
| **Black** | 0.15 | 0.70 | 83.98 | 13.22 | 7.15 | 3.18 | 3.20 |
| **Mixed** | 0.69 | 1.53 | 5.68 | 50.64 | 7.69 | 2.01 | 2.01 |
| **Other** | 0.14 | 1.93 | 0.41 | 1.98 | 6.73 | 0.53 | 0.50 |
| **Unknown** | 2.99 | 4.47 | 5.85 | 6.25 | 12.76 | 3.32 | 3.47 |
| **TOTAL** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Source: 2011 Census microdata linked to 2011/12 HES data

Notes

1. HES ethnicity is based on the 2001 Census classification.
2. HES contain a category “Not stated” for ethnicity. There is no comparable category on the 2011 Census ethnicity classification.
3. Column percentages sum to 100%.

Table 11: 2019 to 2020 HES ethnic group distribution by 2011 Census ethnic group
Census microdata linked to 2019 to 2020 HES data

| Census ethnicity | White | Asian | Black | Mixed | Other | Not stated | Unknown |
|---|---|---|---|---|---|---|---|
| **White** | 96.01 | 1.27 | 2.38 | 23.39 | 46.13 | 81.63 | 84.62 |
| **Asian** | 0.31 | 90.26 | 1.52 | 10.68 | 23.37 | 8.21 | 6.84 |
| **Black** | 0.14 | 0.73 | 84.50 | 11.96 | 9.32 | 3.99 | 2.99 |
| **Mixed** | 0.70 | 1.48 | 5.54 | 46.12 | 6.74 | 2.19 | 1.98 |
| **Other** | 0.14 | 1.86 | 0.38 | 1.94 | 4.82 | 0.64 | 0.43 |
| **Unknown** | 2.69 | 4.41 | 5.68 | 5.91 | 9.62 | 3.33 | 3.14 |
| **TOTAL** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Source: 2011 Census microdata linked to 2019/20 HES data

Notes

1. HES ethnicity is based on the 2001 Census classification.
2. HES contain a category “Not stated” for ethnicity. There is no comparable category on the 2011 Census ethnicity classification.
3. Column percentages sum to 100%.

## Comparisons of Hospital Episode Statistics (HES) ethnicity in 2011 to 2012 and 2019 to 2020

The 2011 to 2012 Hospital Episode statistics (HES) data was longitudinally linked to the 2019 to 2020 HES data to show the stability of ethnicity over time. The person-level HES data were linked using a NHS number and date of birth (linkage rate of 44.0% of 2019 to 2020 HES and 51.6% of 2011 to 2012 HES).

A longitudinal comparison of 2011 to 2012 and 2019 to 2020 HES data shows agreement on ethnicity for 67% of linked records (Table 12). This low agreement is in part because of improvements in data completeness since 2011 to 2012.

Table 12: Longitudinal analysis of 2011 to 2012 HES to 2019 to 2020 HES
2011 to 2012 HES linked to 2019 to 2020 HES data

| | White | Asian | Black | Mixed | Other | Unknown | TOTAL |
|---|---|---|---|---|---|---|---|
| **White** | 60.05 | 0.11 | 0.09 | 0.22 | 0.47 | 4.24 | 65.19 |
| **Asian** | 0.10 | 4.19 | 0.03 | 0.05 | 0.19 | 0.39 | 4.94 |
| **Black** | 0.08 | 0.03 | 2.03 | 0.08 | 0.12 | 0.28 | 2.62 |
| **Mixed** | 0.14 | 0.05 | 0.08 | 0.51 | 0.05 | 0.09 | 0.92 |
| **Other** | 0.34 | 0.16 | 0.06 | 0.04 | 0.58 | 0.16 | 1.34 |
| **Unknown** | 17.46 | 1.52 | 0.74 | 0.31 | 0.71 | 4.25 | 24.99 |
| **TOTAL** | 78.2 | 6.1 | 3.0 | 1.2 | 2.1 | 9.4 | 100.0 |

Source: 2011/12 HES linked to 2019/20 HES data

Notes

1. Agreement is based on the sum of percentages across the diagonal in the table (excluding Unknown).
2. Unknown category also includes Not Stated.
3. Total linked records = 12,626,736.

Of those ethnicities recorded as Mixed or Other in 2019 to 2020, only 42.1% and 27.4% were recorded in these respective groups in 2011 to 2012, indicating movement between these categories (Table 13). There is a clear movement between Mixed and Other in 2019 to 2020 and White in 2011 to 2012. This movement can also be seen in the comparisons between 2011 Census and HES 2019 to 2020 (Table 11).

Table 13: Comparison of ethnicity in linked 2011 to 2012 and 2019 to 2020 linked HES data
2011 to 2012 HES linked to 2019 to 2020 HES data

| | White | Asian | Black | Mixed | Other | Unknown |
|---|---|---|---|---|---|---|
| **White** | 76.83 | 1.88 | 2.88 | 18.49 | 22.29 | 45.01 |
| **Asian** | 0.12 | 69.13 | 0.86 | 3.88 | 8.91 | 4.17 |
| **Black** | 0.10 | 0.57 | 67.15 | 6.60 | 5.46 | 3.01 |
| **Mixed** | 0.18 | 0.77 | 2.60 | 42.10 | 2.28 | 1.00 |
| **Other** | 0.43 | 2.56 | 2.05 | 3.63 | 27.44 | 1.69 |
| **Unknown** | 22.34 | 25.09 | 24.47 | 25.30 | 33.62 | 45.12 |
| **TOTAL** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Source: 2011/12 HES linked to 2019/20 HES data

Notes

1. Column percentages sum to 100%.
2. Unknown category also includes Not Stated
3. Total linked records = 12,626,736

## Stability of household composition over time

In this section we consider how reliable it is to analyse 2020 deaths using household composition recorded at the 2011 Census. Using linked 2011 Census and 2011 GP Patient Register (PR), we compared 2011 PR co-resident counts within Unique Property Reference Numbers (UPRNs) with the equivalent 2011 Census data (Table 14).

Table 14: Comparison of 2011 GP Patient Register and 2011 Census co-resident counts

| 2011 Census household siz | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 9.05 | 1.58 | 0.74 | 0.41 | 0.21 | 0.10 | 0.05 | 0.03 | 0.02 | 0.01 | 0.04 | 12.23 |
| **2** | 1.31 | 20.50 | 4.24 | 1.54 | 0.65 | 0.31 | 0.15 | 0.07 | 0.04 | 0.02 | 0.07 | 28.88 |
| **3** | 0.29 | 1.73 | 12.40 | 3.41 | 1.01 | 0.42 | 0.22 | 0.11 | 0.05 | 0.03 | 0.08 | 19.75 |
| **4** | 0.14 | 0.45 | 2.20 | 15.82 | 2.54 | 0.81 | 0.36 | 0.20 | 0.10 | 0.05 | 0.12 | 22.80 |
| **5** | 0.06 | 0.13 | 0.32 | 1.19 | 6.27 | 1.17 | 0.41 | 0.19 | 0.11 | 0.06 | 0.11 | 10.01 |
| **6** | 0.02 | 0.05 | 0.09 | 0.19 | 0.49 | 2.13 | 0.59 | 0.25 | 0.13 | 0.07 | 0.12 | 4.14 |
| **7** | 0.01 | 0.02 | 0.02 | 0.04 | 0.06 | 0.13 | 0.57 | 0.16 | 0.06 | 0.03 | 0.05 | 1.14 |
| **8** | 0.00 | 0.01 | 0.01 | 0.02 | 0.02 | 0.03 | 0.06 | 0.23 | 0.07 | 0.03 | 0.04 | 0.52 |
| **9** | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.10 | 0.03 | 0.04 | 0.24 |
| **10** | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.05 | 0.04 | 0.13 |
| **>10** | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.10 | 0.15 |
| **Total** | 10.89 | 24.48 | 20.04 | 22.63 | 11.26 | 5.11 | 2.43 | 1.29 | 0.70 | 0.39 | 0.79 | 100.00 |

Source: Linked 2011 Census microdata and 2011 GP Patient Registration data for England and Wales

Notes

1. This is a comparison of the number of co-residents within each person's address; these could refer to different addresses for people who moved house, and addresses will feature more than once in the table. Even for people with the same number of co-residents on both sources, the counts may refer to different individuals.

2. The table is indicative of consistency between the sources.

3. Co-resident counts are within Unique Property Reference Number (UPRN).

4. Agreement is based on the sum of percentages across the diagonal in the table. 5. The total number of linked 2011 Census to 2011 PR records in the analysis dataset is 47,800,385.

Co-resident counts are consistent for 67.2% (total percentage on the diagonal) of people on the linked dataset. For 9.3%, more co-residents are included on Census returns and for 23.5% there are more co-residents in the PR than in Census. Reasons for this difference could be:

- people who moved to a different property or household with a different number of co-residents
- co-residents who moved to or from a person's address

The higher percentage above the diagonal than below indicates that the PR has higher co-resident counts than the Census. This could be because of:

- Census under-coverage; the Census data used here are not adjusted for Census undercount
- list inflation in the PR; there can be time lags between someone moving and registering with a new GP, and emigrants' records are known to stay on the PR if they have not notified the NHS of their departure

Analysis of inconsistencies between 2011 PR and 2011 Census co-resident counts reveals that this is consistent between males and females but varies by age and ethnic group.

Table 15 shows that consistency between the 2011 Census and PR is highest for those aged 65 years and over (84.4%).

Table 15: Comparison between co-resident counts in the linked 2011 Census and the 2011 PR by age and sex

| Census group | Same in PR (%) | More in PR (%) | Fewer in PR (%) | N |
|---|---|---|---|---|
| **All persons** | 67.20 | 23.54 | 9.26 | 47,800,385 |
| **Males** | 67.26 | 23.73 | 9.00 | 22,975,653 |
| **Females** | 67.14 | 23.35 | 9.51 | 24,676,895 |
| **Age 0-24** | 62.24 | 26.38 | 11.38 | 14,125,761 |
| **Age 25-44** | 61.49 | 27.74 | 10.77 | 12,566,573 |
| **Age 45-64** | 67.34 | 23.43 | 9.23 | 12,820,119 |
| **Age 65+** | 84.37 | 12.30 | 3.33 | 8,167,503 |

Source: Linked 2011 Census microdata and 2011 GP Patient Registration data for England and Wales

Notes

1. This is a comparison of the number of co-residents within each person's address; these could refer to different addresses for people who moved house, and addresses will feature more than once in the table. Even for people with the same number of co-residents on both sources, the counts may refer to different individuals.

2. The table is indicative of consistency between the sources.

3. Co-resident counts are within Unique Property Reference Number (UPRN).

4. The total number of linked 2011 Census to 2011 PR records in the analysis dataset is 47,800,385.

Inconsistencies between 2011 Census and 2011 PR co-residence counts are higher in the Ethnic minority groups, and highest for the Black ethnic group, with only 46.3% of linked records having the same number of co-residents on both sources (Table 16).

Table 16: Comparison between co-resident counts in the 2011 Census and the 2011 PR by ethnic group

| Census ethnic group | Same in PR (%) | More in PR (%) | Fewer in PR (%) | N |
|---|---|---|---|---|
| **All persons** | 67.20 | 23.54 | 9.26 | 47,800,385 |
| **White** | 69.64 | 21.35 | 9.01 | 40,698,461 |
| **Mixed** | 57.63 | 31.65 | 10.72 | 830,069 |
| **Asian** | 51.12 | 38.22 | 10.66 | 3,243,709 |
| **Black** | 46.26 | 42.31 | 11.43 | 1,306,092 |
| **Other** | 51.43 | 37.82 | 10.75 | 182,384 |
| **Census ethnic group** | Same in PR (%) | More in PR (%) | Fewer in PR (%) | N |

Source: Linked 2011 Census microdata and 2011 GP Patient Registration data for England and Wales

Notes

1. This is a comparison of the number of co-residents within each person's address; these could refer to different addresses for people who moved house, and addresses will feature more than once in the table. Even for people with the same number of co-residents on both sources, the counts may refer to different individuals.

2. The table is indicative of consistency between the sources.

3. Co-resident counts are within Unique Property Reference Number (UPRN).

4. Ethnicity totals do not sum to "All persons" because of missingness in ethnicity data collected at the 2011 Census.

5. The table is indicative of consistency between the sources.

Table 17 shows that, according to the PR, just 45.7% of people had the same number of co-residents in 2019 has they had in the 2011 Census.

Table 17: Comparison of 2011 Census and 2019 Patient Register co-resident counts

| 2011 Census household size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6.22 | 1.78 | 0.89 | 0.54 | 0.26 | 0.12 | 0.06 | 0.03 | 0.02 | 0.01 | 0.32 | 10.24 |
| **2** | 3.43 | 15.87 | 3.91 | 2.45 | 0.94 | 0.42 | 0.20 | 0.10 | 0.05 | 0.03 | 0.31 | 27.69 |
| **3** | 1.07 | 4.54 | 7.90 | 4.08 | 1.57 | 0.66 | 0.31 | 0.16 | 0.08 | 0.04 | 0.15 | 20.55 |
| **4** | 0.79 | 2.51 | 5.27 | 10.54 | 2.80 | 1.09 | 0.49 | 0.25 | 0.13 | 0.07 | 0.23 | 24.17 |
| **5** | 0.32 | 0.78 | 1.34 | 2.39 | 3.60 | 1.14 | 0.48 | 0.22 | 0.12 | 0.06 | 0.17 | 10.62 |
| **6** | 0.12 | 0.27 | 0.39 | 0.61 | 0.86 | 1.09 | 0.48 | 0.24 | 0.12 | 0.07 | 0.13 | 4.40 |
| **7** | 0.03 | 0.07 | 0.09 | 0.13 | 0.16 | 0.21 | 0.25 | 0.12 | 0.06 | 0.03 | 0.05 | 1.21 |
| **8** | 0.01 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.05 | 0.03 | 0.05 | 0.56 |
| **9** | 0.01 | 0.01 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.02 | 0.04 | 0.26 |
| **10** | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.14 |
| **>10** | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.05 | 0.16 |
| **Total** | 12.00 | 25.88 | 19.86 | 20.84 | 10.29 | 4.85 | 2.40 | 1.26 | 0.69 | 0.39 | 1.53 | 100.00 |

Source: Linked 2011 Census microdata and 2019 GP Patient Registration data for England and Wales

Notes

1. This is a comparison of the number of co-residents within each study member's address; these could refer to different addresses for people who moved house, and addresses will feature more than once in the table. Even for people with the same number of co-residents on both sources, the counts may refer to different individuals.

2. The table is indicative of consistency between the sources.

3. Co-resident counts are within Unique Property Reference Number (UPRN).

4. Agreement is based on the sum of percentages across the diagonal in the table.

5. The total number of linked 2011 Census to 2019 PR records in the analysis dataset is 43,211,343.

Table 18 shows that this varied by age (but not by sex), with stability in the number of co-residents highest for those aged 65 and over (66.9% with the same number of co-residents).

Table 18: Comparison between co-resident counts in the 2011 Census and 2019 Patient Register by age and sex

| Census group | Same in PR (%) | More in PR (%) | Fewer in PR (%) | N |
|---|---|---|---|---|
| **All persons** | 44.66 | 28.24 | 26.10 | 43,211,343 |
| **Males** | 45.99 | 28.39 | 25.61 | 20,759,926 |
| **Females** | 45.35 | 28.09 | 26.56 | 22,325,074 |
| **Age 0-24** | 37.80 | 32.32 | 29.88 | 13,564,247 |
| **Age 25-44** | 38.38 | 39.68 | 21.94 | 12,158,913 |
| **Age 45-64** | 52.43 | 17.74 | 29.83 | 12,016,886 |
| **Age 65+** | 66.92 | 15.46 | 17.62 | 5,367,445 |

Source: Linked 2011 Census microdata and 2019 GP Patient Registration data for England and Wales

Notes

1. This is a comparison of the number of co-residents within each person's address; these could refer to different addresses for people who moved house, and addresses will feature more than once in the table. Even for people with the same number of co-residents on both sources, the counts may refer to different individuals.

2. The table is indicative of consistency between the sources.

3. Co-resident counts are within Unique Property Reference Number (UPRN).

4. The total number of linked 2011 Census to 2019 PR records in the analysis dataset is 43,211,343.

Inconsistencies between 2011 Census and 2019 PR co-residence counts were higher in the Ethnic minority groups, and highest for the Asian ethnic group, with only 32.7% of linked records having the same number of co-residents on both sources.

Table 19: Comparison between co-resident counts in the 2011 Census and the 2019 Patient Register by ethnic group

| Census ethnic group | Same in PR (%) | More in PR (%) | Fewer in PR (%) | N |
|---|---|---|---|---|
| **All persons** | 44.66 | 28.28 | 26.1 | 43,211,343 |
| **White** | 47.63 | 26.08 | 26.29 | 36,585,388 |
| **Mixed** | 37.58 | 38.53 | 23.88 | 786,650 |
| **Asian** | 32.71 | 42.12 | 25.17 | 3,083,561 |
| **Black** | 32.98 | 44.84 | 22.17 | 1,235,602 |
| **Other** | 33.51 | 43.05 | 23.44 | 171,323 |

Source: Linked 2011 Census microdata and 2019 GP Patient Registration data for England and Wales

Notes

1. This is a comparison of the number of co-residents within each person's address; these could refer to different addresses for people who moved house, and addresses will feature more than once in the table. Even for people with the same number of co-residents on both sources, the counts may refer to different individuals.

2. The table is indicative of consistency between the sources.

3. Co-resident counts are within Unique Property Reference Number (UPRN).

4. Ethnicity totals do not sum to "All persons" because of missingness in ethnicity data collected at the 2011 Census.

5. The total number of linked 2011 Census to 2019 PR records in the analysis dataset is 43,211,343. 5.2 Household size (Table 19).xlsx

# 6 . Conclusion

This article describes the quality of the bespoke study dataset used in the published Office for National Statistics (ONS) articles on deaths involving the coronavirus (COVID-19) by ethnicity for England and Wales. We applied our longitudinal error framework to support the use and understanding of these data and in doing so we have highlighted the types of potential errors that need to be considered when interpreting analysis based on these data. Potential errors include but are not limited to:

- coverage of the study dataset
- linkage error
- error introduced from using socio-demographic characteristics collected at the 2011 Census to report on deaths occurring during 2020

Our analysis has highlighted that high linkage rates were achieved across all deaths and deaths involving COVID-19 at around 90%. The low false positive rate of 0.2% indicates a high level of precision for the linked data. However, we identified that some Ethnic minority groups are under-represented in the study dataset. We recommend that further analysis is required to understand aspects of the linkage methodology which are biasing linkage towards people in the White ethnic group.

A longitudinal comparison of ethnicity recorded at the 2011 Census and in 2019 to 2020 Hospital Episodes Statistics (HES) data has identified close agreement between ethnicity recorded at these two time points. We also considered how reliable it is to analyse 2020 deaths using household composition recorded at the 2011 Census. Inconsistencies between the 2011 Census and 2019 Patient Register (PR) co-residence counts were observed in Ethnic minority groups. These can be attributed to several factors. For example, Census co-residence counts are lower due to Census under-coverage and non-replenishment of the study population. PR co-residence counts are higher, because of time lags between someone moving and registering with a new GP, or where emigrants' records are known to stay on the PR if they have not notified the NHS of their departure.

We are very keen to receive feedback and observations on our work, including from those who find it useful, and, those who think it needs further thought and refinement. Please contact us at demographic.methods@ons.gov.uk with any comments.

# 7 . Authors:

Nicky Rogers, Louisa Blackwell, Sarah Cummins, Eleanor Fordham, Amy Large, Sonya Ridden, Elzemiek Scott-Kortlever, Gemma Hanson

# 8 . Appendix A: Using a simulated population to account for emigration in the cohort study population

## Rates based on the Office for National Statistics (ONS) Longitudinal Study (LS) for England and Wales

We calculated emigration rates using Patient Register (PR) data linked to the LS for England and Wales for 2011 to 2016. The LS cohort is a 1% sample of those present at the 2011 Census. Embarkations were identified in the LS using PR data, where a person de-registered with a GP because they were moving abroad (observed embark) or where a patient registration was cancelled by the GP surgery (unobserved embark).

## Rates based on the International Passenger Survey (IPS)

We reviewed all available data sources to calculate comparative emigration rates and concluded that the IPS was the best available contemporaneous data source with estimates available up to year ending March 2019.

There is a break in the emigration time series for the decade, with low and flat emigration rates following the EU exit referendum in 2016. We had concerns that using an average and extrapolating the LS rates forward might obscure true patterns of emigration that have been observed in migration estimates since June 2016.

The IPS does not collect data on ethnicity but does collect data on citizenship. We therefore produced an IPS-based out-migration rate using a lookup between citizenship and Ethnic group by age and sex. The lookup was produced from published 2011 Census tables for England and Wales.

We equivalised the IPS data to mimic our study population.

This involved excluding anyone who immigrated after the 2011 Census or subsequently emigrated in the period of interest (2011 to 2020). IPS emigrants were adjusted for births in 0 to 9 year olds, reflecting non-replenishment in the cohort.

We took the 2011 Census microdata (usual residents) as our starting population in 2011 to 12. We adjusted for under-coverage using the weights produced in Census coverage error in Section 4.

To move to our new base population in 2012 to 2013 we subtracted those who emigrated in 2011 to 2012 from the base population by broad age, sex, and ethnicity. We then deducted deaths by broad age, sex, and broad ethnicity.

We aged the 2012 to 2013 base population on a year. There were no new-borns entering the population. We assumed a constant age distribution where we aged-on 1/25 of those aged 0 to 24 years and added these into ages 25 to 44 years. For ages 25 to 44 years we aged on 1/20 and added into ages 45 to 64 years and so on.

We also accounted for not adding in new-borns over time. We needed to reflect that in 2012 to 2013 the 0 to 24 years age group will become 1 to 24 year olds, 2 to 24 year olds in 2013 2014 and so on.

Our rates (numerators and denominators) now mirrored the ageing of the 2011 cohort.

Emigration rates for the cohort by broad age (0 to24 years, 25 to 44 years, 45 to 64 years and 65 years and over), sex and broad ethnicity (White, Mixed, Indian, Chinese, Pakistani and Bangladeshi, Black, Other) and year (2011 to 2012 through to 2019 to 2020) were calculated using the formula (4):

$$\mathrm{Out-Migration\ Rate}_{\text{age, sex, ethnicity}} = \frac{\mathrm{Emigrants(age, sex, ethnicity)}}{\mathrm{Base\ population(age, sex, ethnicity)}}\ x\ 1000$$

We assumed that, at the time we created the rates, out-migration rates in the period April 2019 to March 2020 were the same as the period April 2018 to March 2019.

## Extrapolation and hybrid approach

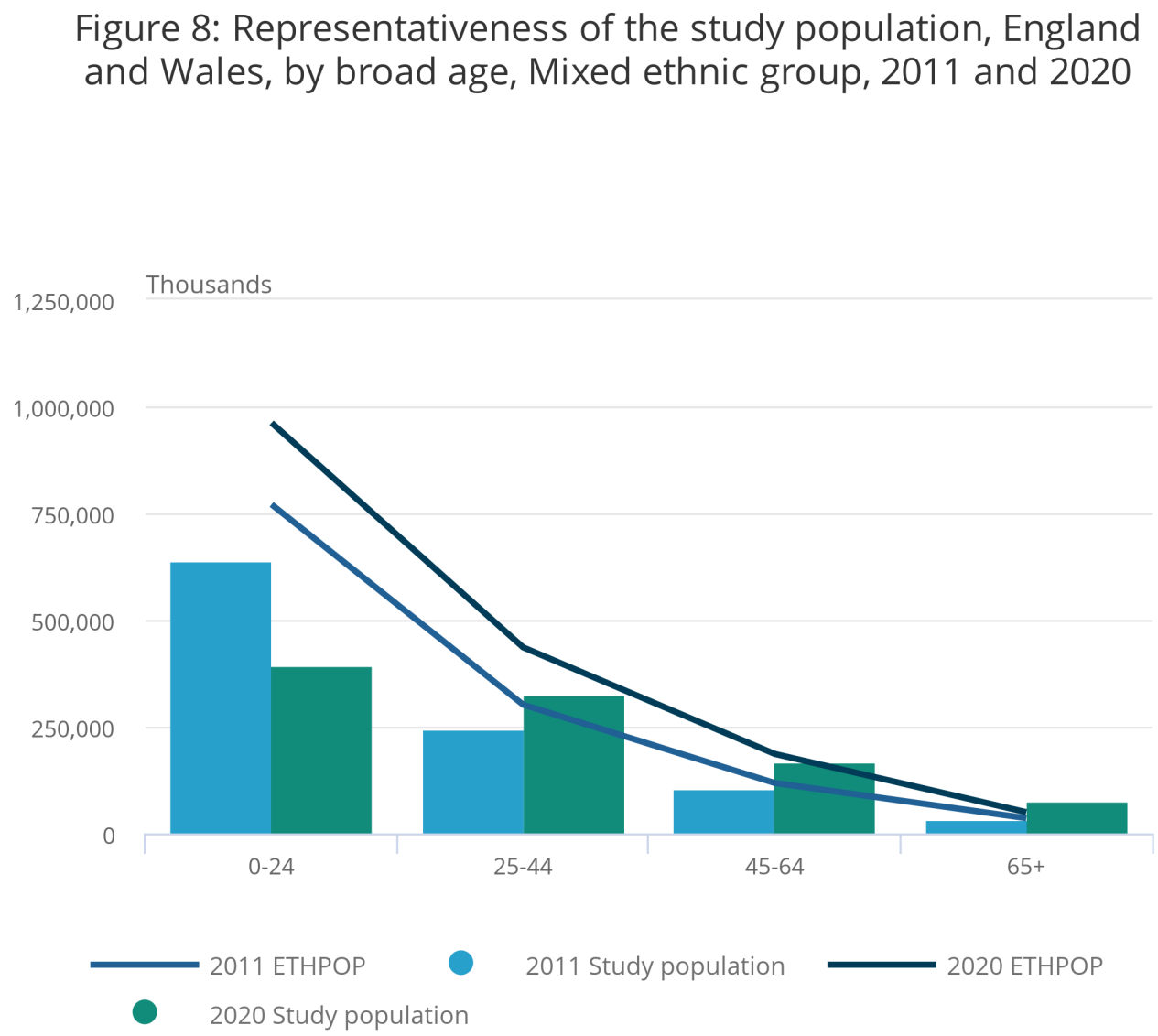
We developed a hybrid approach that built on the strength of the relationship between the LS (usual resident-based) rates and the IPS-based rates for 2012 to 2013 through to 2015 to 2016. This addressed the shortcomings of LS rates only being available up to 2015 to 2016 and IPS-based rates to 2018 to 2019, and our doubts about the accuracy using the IPS-based rates for the 2011 to 2012 study population.

The methodology takes the mean of LS-based and IPS-based rates for years 2012 to 2013 through to 2015 to 2016 and extends this for the decade:

- for 2011 to 2012 the ratio of LS to LS/IPS mean rates in 2012 to 2013 was applied to LS rates in 2011 to 2012 to derive a new mean for 2012 to 2013
- for 2016 to 2017 through to 2019 to 2020 the absolute difference between the IPS/LS mean and the IPS rate in 2015 to 2016 was applied to IPS rates for 2016 to 2017, 2017 to 2018 and 2018 to 2019 to extrapolate new means for each group in these years; the 2018 to 2019 rates were repeated for 2019 to 2020

# 9 . Appendix B: Representativeness of the study population between 2011 and 2020 by broad and broad ethnicity

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

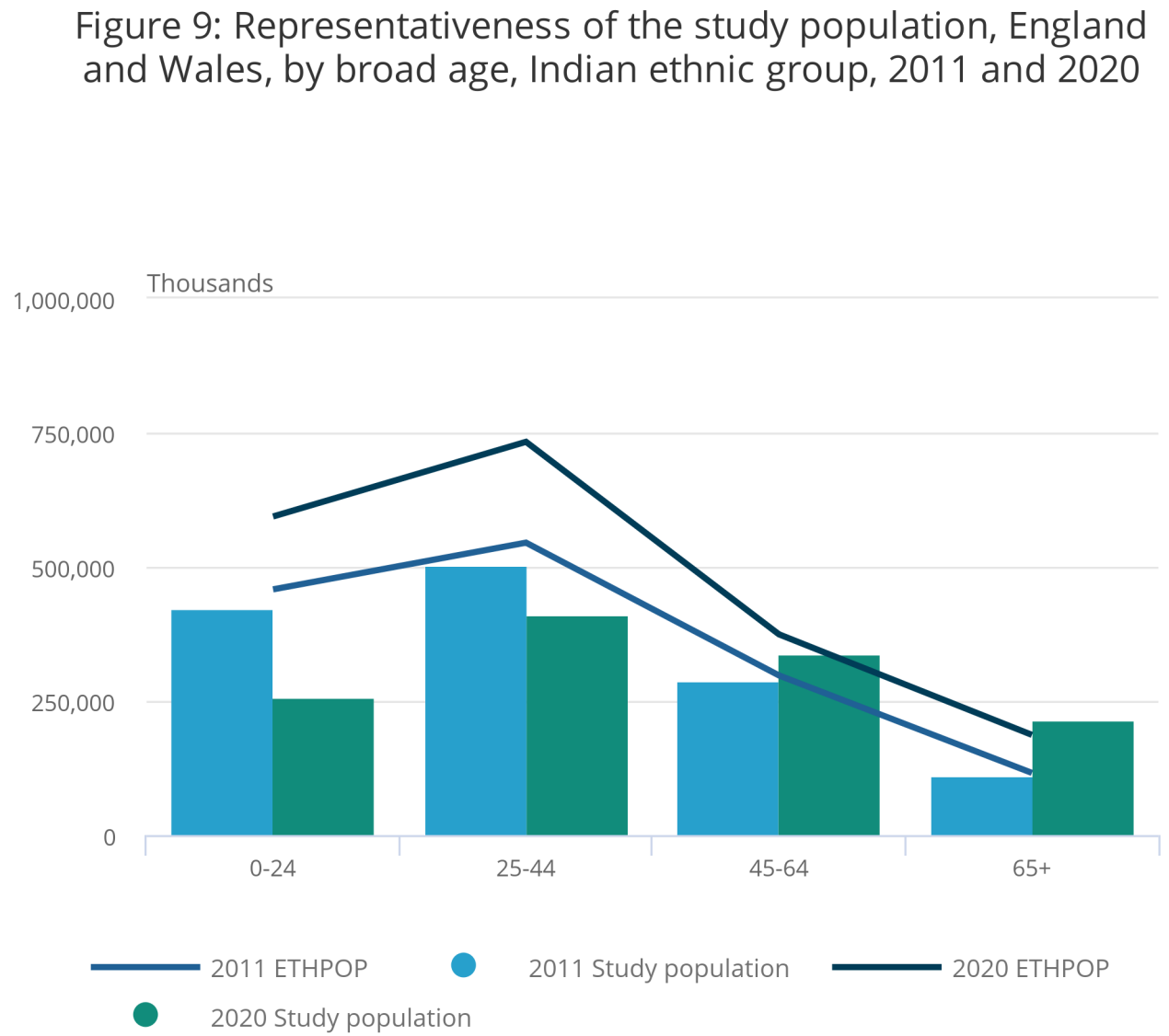



**Source: Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data.**

**Notes:**

1. ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates

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

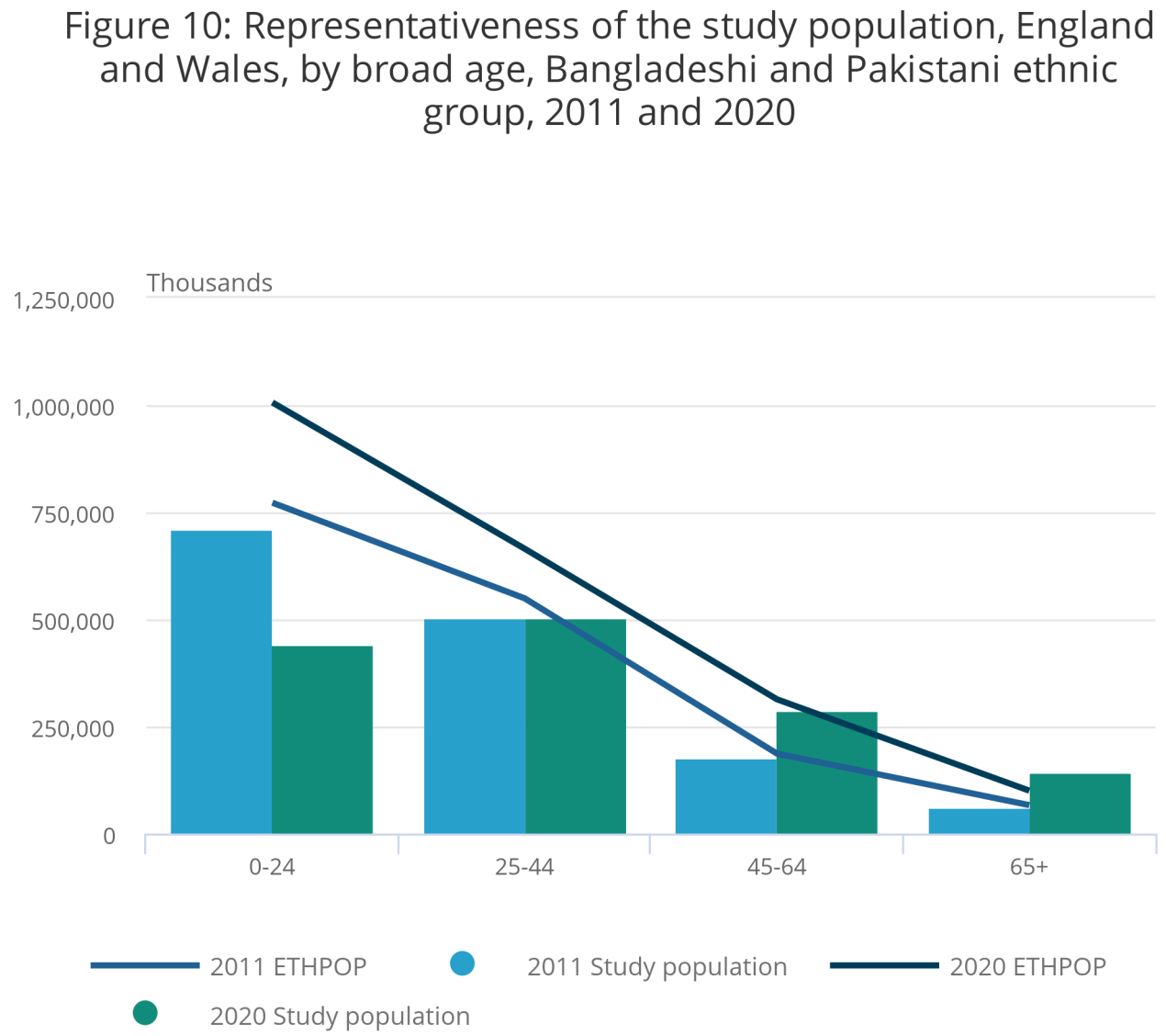



**Source: Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data.**

**Notes:**

1. ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates.

**Figure 10: Representativeness of the study population, England and Wales, by broad age, Bangladeshi and Pakistani ethnic group, 2011 and 2020**

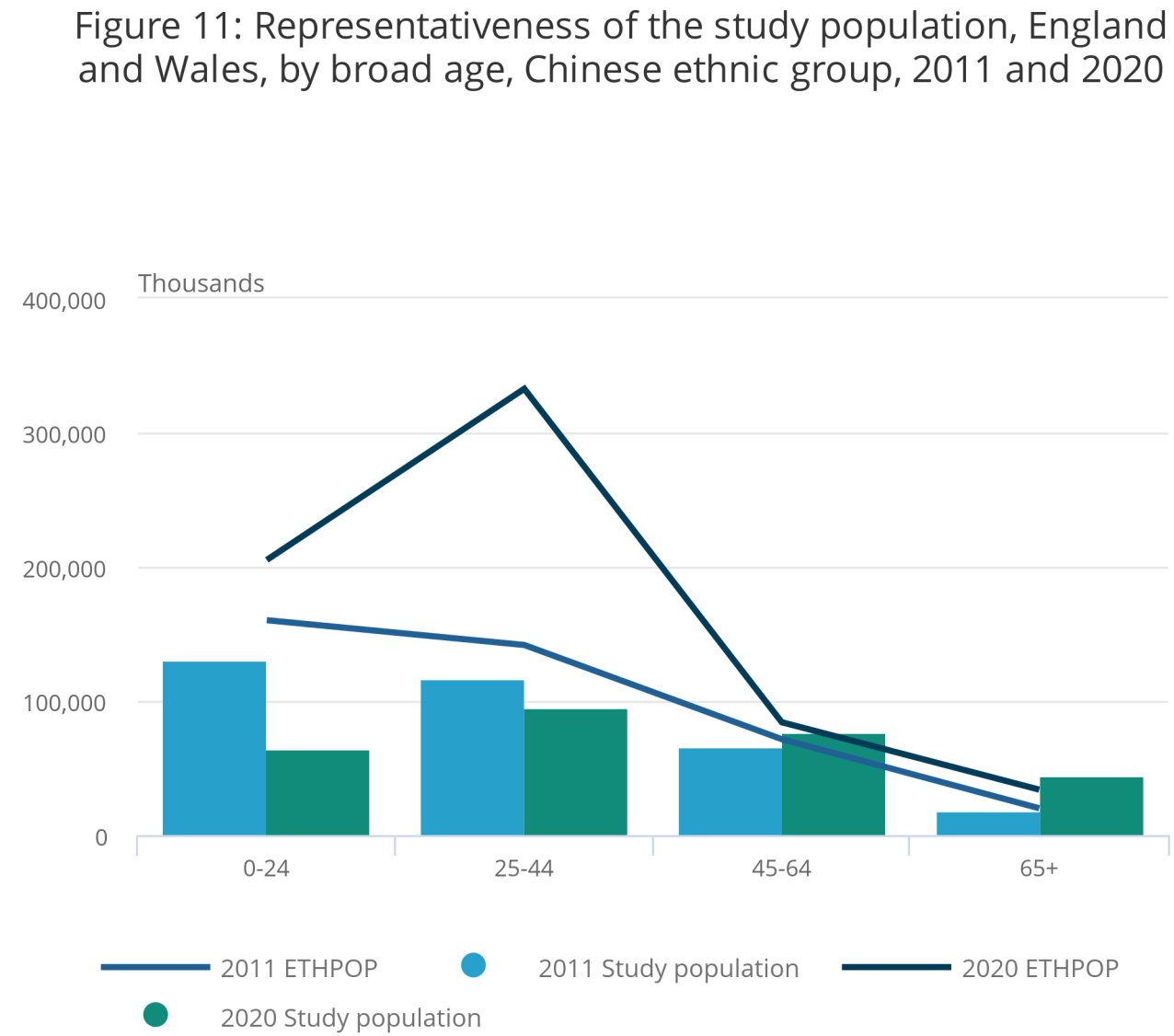



**Source: Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data.**

**Notes:**

1. ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates .

**Figure 11: Representativeness of the study population, England and Wales, by broad age, Chinese ethnic group, 2011 and 2020**

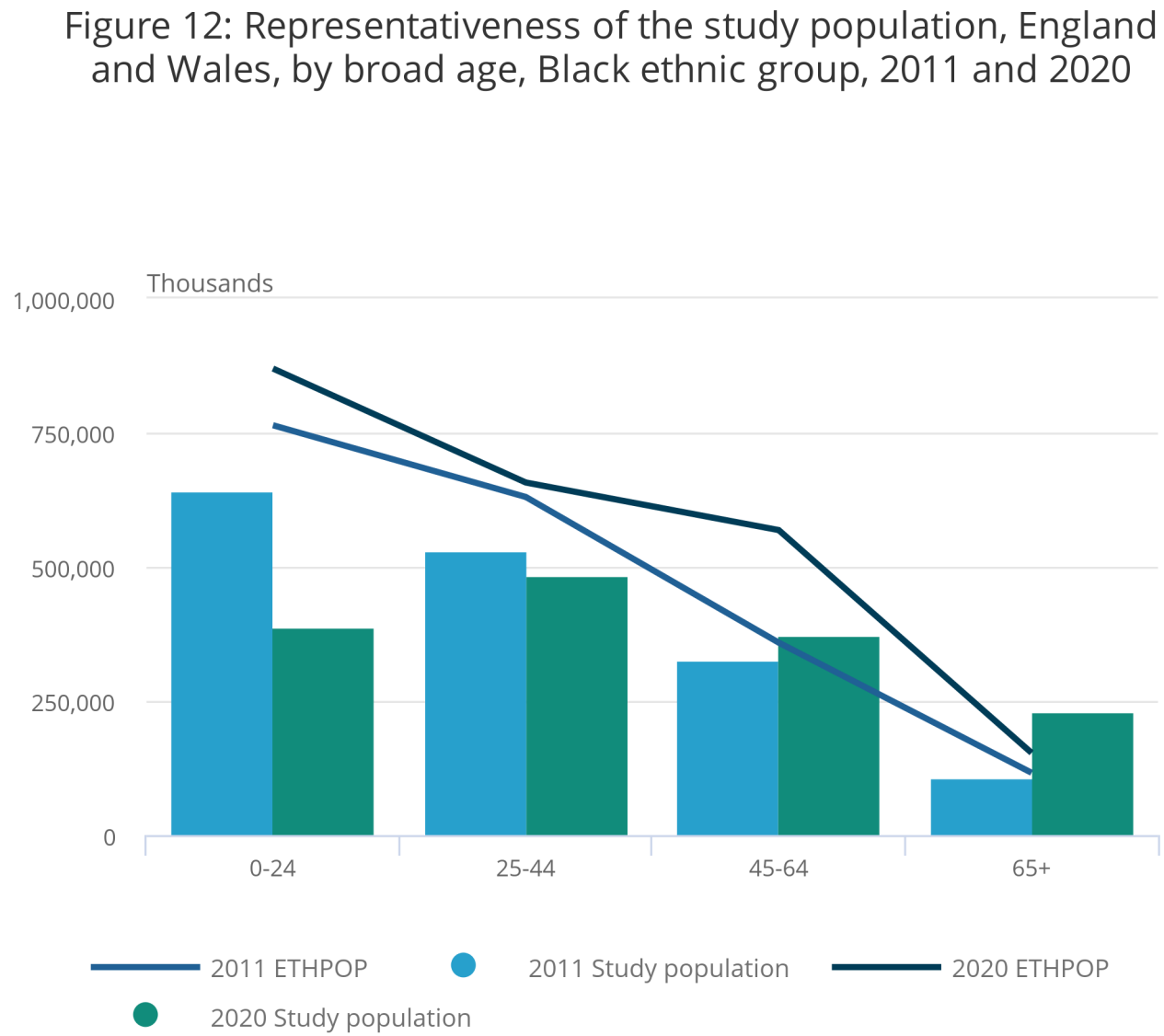



**Source: Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data**

**Notes:**

ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates.

**Figure 12: Representativeness of the study population, England and Wales, by broad age, Black ethnic group, 2011 and 2020**

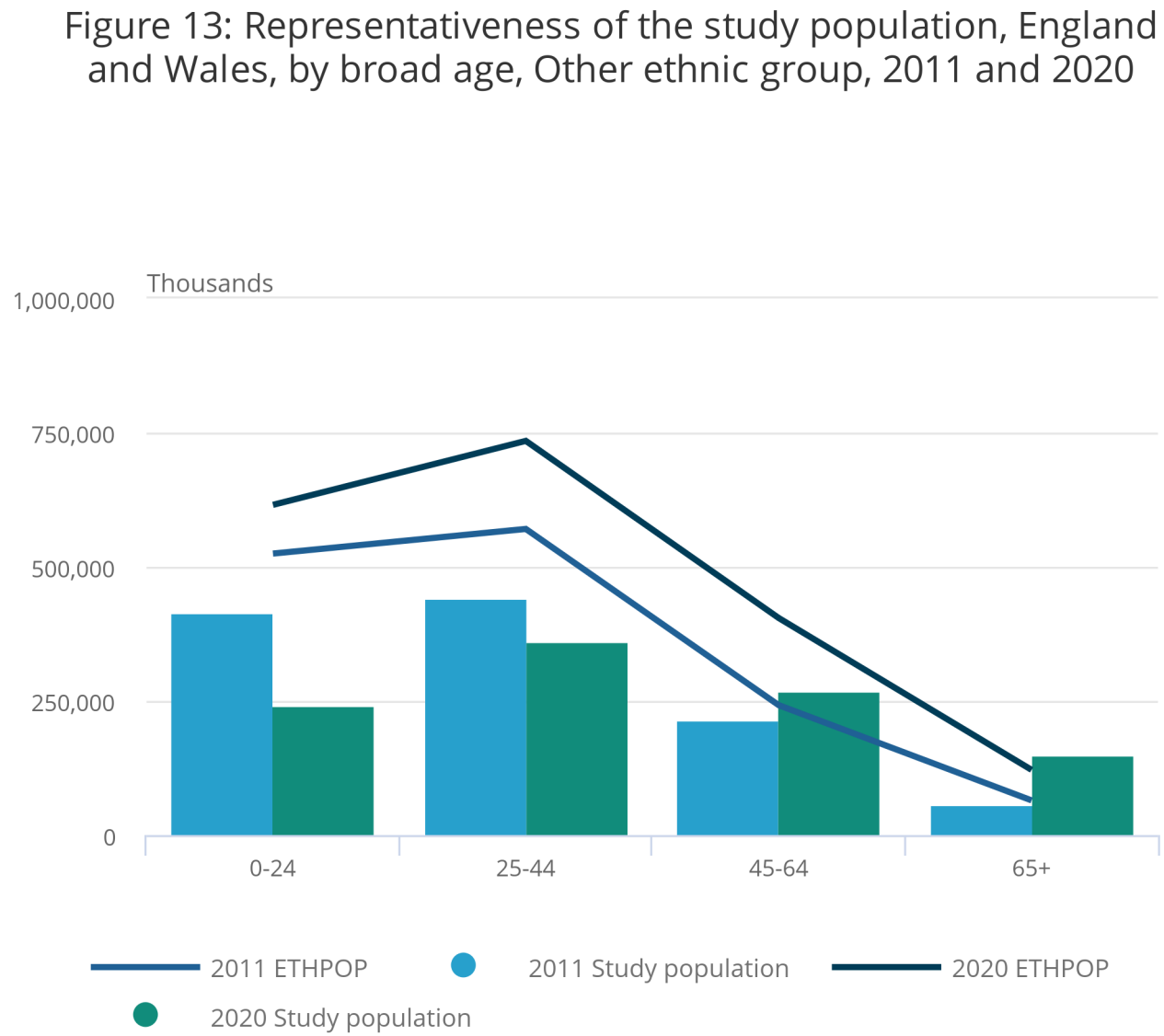



**Source: Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data**

**Notes:**

1. ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates .

**Figure 13: Representativeness of the study population, England and Wales, by broad age, Other ethnic group, 2011 and 2020**




**Source: Office for National Statistics – 2011 Census Microdata and Leeds University ETHPOP data**

**Notes:**

1. ETHPOP data are population projections based on Census, survey, and Mid-Year Estimates.