

2011 Census linkage to DWP master key and encrypted NINo

Linkage methodology and quality information for 2011 Census linkage to DWP (Department for Work and Pensions) master key and encrypted NINo (National Insurance number).

Contact:
Data Linkage and Integration
Hub
linkage.hub@ons.gov.uk

Release date:
6 December 2024

Next release:
To be announced

Table of contents

1. [Main points](#)
2. [Background to the linkage](#)
3. [Linkage methodology: Stage 1](#)
4. [Linkage methodology: Stage 2](#)
5. [Quality information](#)
6. [Limitations](#)
7. [Related links](#)
8. [Cite this methodology](#)

1 . Main points

- To allow health and labour market analysis projects, the 2011 Census was linked to the Department for Work and Pensions (DWP) and HM Revenue and Customs (HMRC) data.
- The linkage was conducted in two stages: first, the 2011 Census was linked to the Office for National Statistics (ONS) Demographic Index (DI), via the Patient Register (PR); then, residual 2011 Census records were linked to the DI, via the Demographic Index Matching Service (DIMS).
- The first stage of linkage resulted in 89.7% of census IDs linking to a DWP master key or encrypted National Insurance number (NINo); through the second stage, the match rate was subsequently increased to 96.7% of census IDs linking to a DWP master key or encrypted NINo.
- The linkage outputs contained lookups between the 2011 Census and DWP master key, and the 2011 Census and encrypted NINo identifiers.
- The overall precision estimate was calculated to be 97.76% and recall was estimated to be 99.95%.

2 . Background to the linkage

The purpose of this linkage was to bring census, Department for Work and Pensions (DWP) and HM Revenue and Customs (HMRC) data together as part of health and labour market analysis projects. The linkage of the 2011 Census to DWP and HMRC data will allow for census data to be integrated with health and economic datasets:

- the link between 2011 Census and DWP master key will enable Data and Analysis for Social Care and Health (DASCH) to integrate the DWP Benefits and Income Dataset with the [Public Health Data Asset](#)
- the link between 2011 Census and encrypted National Insurance number (NINo) will enable the integration to economic data, in particular HMRC Pay As You Earn (PAYE) data

Analysis of these linked datasets will allow for research into the relationship between health conditions and intervention programmes, with labour market outcomes. Please see [blog post](#) for more detail, and [examples of research outputs](#).

Note: A linkage between 2021 Census, DWP and HMRC data was also conducted, using a separate linkage methodology. See the [methodology report](#).

3 . Linkage methodology: Stage 1

2011 Census

Every 10 years, the census provides a detailed snapshot of all the people and households in England and Wales. The census provides information that government needs to develop policies, plan and run public services, and allocate funding.

2011 Census-PR lookup

This project involved linking the 2011 Census-Patient Register (PR) data to the Demographic Index (DI), via NHS number, to obtain Department for Work and Pensions (DWP) master key and encrypted National Insurance number (NINo). The 2011 Census-PR lookup was used as it contained a link from the 2011 Census to NHS number, which is a variable that can be directly indexed to the DI. Linkage using NHS number also allowed us to avoid personal information (PI) which may have contained differences because of the time lag between the 2011 Census and the DI (data from 2016).

Demographic Index

The Demographic Index (DI) is part of the Reference Data Management Framework (RDMF), which is a set of tables and services that allow the Office for National Statistics (ONS) to link data to produce more useful analyses in a secure way. The RDMF is a tool produced by the ONS that is made up of five "indexes" (datasets or tables), including information on locations, businesses and people.

The DI attempts to provide an entry for each person in England and Wales. It contains longitudinally linked administrative data to provide information on the population who interact with admin data sources. A person's records are de-identified to ensure people cannot be directly identified and referenced with a "Demographic Entry ID", which becomes that person's unique identifier.

The Demographic Entry ID then references a cluster of all the admin data records that belong to that specific person throughout the years (from 2016 to present). Included in the clustered admin data are Department for Work and Pensions (DWP) master key and encrypted National Insurance number (NINo) identifiers.

The DI was used because of its coverage of the population; it covers a spread of different admin sources that most of the population should have interacted with. However, it is important to note the data limitations of the DI, as explained in [Section 6: Limitations](#).

Linkage between 2011 Census and the Demographic Index, via Patient Register

The 2011 Census was joined to the DI using NHS number (via 2011 Census-PR lookup) to obtain DWP master key and encrypted NINo.

The 2011 Census was used as a spine, meaning that any residual census records that did not link to a DWP master key or encrypted NINo were retained in the data. Cleaning steps were also conducted, for example, removing duplicate rows, splitting DWP master key and encrypted NINo into separate tables, and removing NHS number from the final tables.

Summary of Stage 1 linkage

Table 1: 2011 Census linked to DWP master key, summary of Stage 1 outputs

| | | Number of census IDs | As percentage of total number of census IDs (%) |
|----------------|--|-----------------------------|--|
| Group 1 | Census IDs without master key or Demographic Entry ID (census residuals) | 4,868,461 | 9.1 |
| Group 2 | Census IDs with Demographic Entry ID only | 659,487 | 1.2 |
| Group 3 | Census IDs with master key and Demographic Entry ID | 47,955,508 | 89.7 |
| Total | Total census IDs in linked data frame (including census residuals) | 53,483,456 | - |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Table 2: 2011 Census linked to encrypted NINo, summary of Stage 1 outputs

| | | Number of census IDs | As percentage of total number of census IDs (%) |
|----------------|--|-----------------------------|--|
| Group 1 | Census IDs without NINo or Demographic Entry ID (census residuals) | 4,868,461 | 9.1 |
| Group 2 | Census IDs with Demographic Entry ID only | 654,143 | 1.2 |
| Group 3 | Census IDs with NINo and Demographic Entry ID | 47,960,852 | 89.7 |
| Total | Total census IDs in linked data frame (including census residuals) | 53,483,456 | - |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

4 . Linkage methodology: Stage 2

Linkage of residual census records via DIMS

A second phase of linkage was undertaken, involving the Demographic Index Matching Service (DIMS).

DIMS is a linkage pipeline developed to index datasets against the Demographic Index (DI) to return a Demographic Entry ID for records in the dataset. DIMS uses personal data across the sources to link datasets to the DI via deterministic and probabilistic methods.

The residual census records that remained (N = 4,868,461) were indexed to the DI using DIMS, with the aim to obtain more links and therefore to improve the quality of the linkage. A deidentified census link table was created by removing personal data and retaining a Demographic Entry ID where a link was made to the DI.

Summary of Stage 1 and 2 linkages

Table 3: 2011 Census linked to DWP master key, summary of Stages 1 and 2 outputs

| | | Number of census IDs | As percentage of total number of census IDs (%) |
|----------------|--|-----------------------------|--|
| Group 1 | Census IDs without master key or Demographic Entry ID (census residuals) | 1,008,164 | 1.9 |
| Group 2 | Census IDs with Demographic Entry ID only | 741,051 | 1.4 |
| Group 3 | Census IDs with master key and Demographic Entry ID | 51,734,241 | 96.7 |
| Total | Total census IDs in linked data frame (including census residuals) | 53,483,456 | - |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Table 4: 2011 Census linked to encrypted NINo, summary of Stages 1 and 2 outputs

| | | Number of census IDs | As percentage of total number of census IDs (%) |
|----------------|--|-----------------------------|--|
| Group 1 | Census IDs without NINo or Demographic Entry ID (census residuals) | 1,008,164 | 1.9 |
| Group 2 | Census IDs with Demographic Entry ID only | 733,637 | 1.4 |
| Group 3 | Census IDs with NINo and Demographic Entry ID | 51,741,655 | 96.7 |
| Total | Total census IDs in linked data frame (including census residuals) | 53,483,456 | - |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

5 . Quality information

Quality Assurance (QA) was carried out once linkage was completed. This included:

- exploration of residual (unlinked) records
- exploration of clustered records in the linked data
- clerical review of samples from the linked data, to estimate precision of linkage
- clerical review of samples from the residual (unlinked) records, to estimate recall of linkage

It is important to note, there are two types of error that can occur when indexing to the Demographic Index (DI):

- error within the DI itself (as part of the methods used to create the DI)
- error in the linkage to the DI

This quality assurance is not an assessment of error within the DI, but an assessment of the linkage to the DI.

Exploration of residual (unlinked) records

There are two main reasons why a census record may not have linked to a cluster in the DI.

The first reason is that the person in the 2011 Census does not have a corresponding cluster in the DI. While we expect a relatively high level of coverage overlap between the two sources, because of the temporal differences between the 2011 Census and DI sources (2016 onwards), we expect that there may be coverage differences because of deaths and emigration prior to 2016.

The second reason is linkage error, where because of data quality issues we were unable to identify the links from the linkage methodology used. While we were unable to separate these coverage differences from linkage error, to understand the potential for linkage errors to cause bias in the linked data, the demographic characteristics of the following three groups were compared.

Table 5: Groups for comparison

| Group Description | Proportion of census IDs (%) |
|--|-------------------------------------|
| 1 Census records that did not link to the Demographic Index (residuals) | 1.9 |
| 2 Census records that linked to the Demographic Index, but did not obtain a DWP master key / encrypted NINo | 1.4 |
| 3 Census records that linked to the Demographic Index and obtained a DWP master key / encrypted NINo | 96.7 |

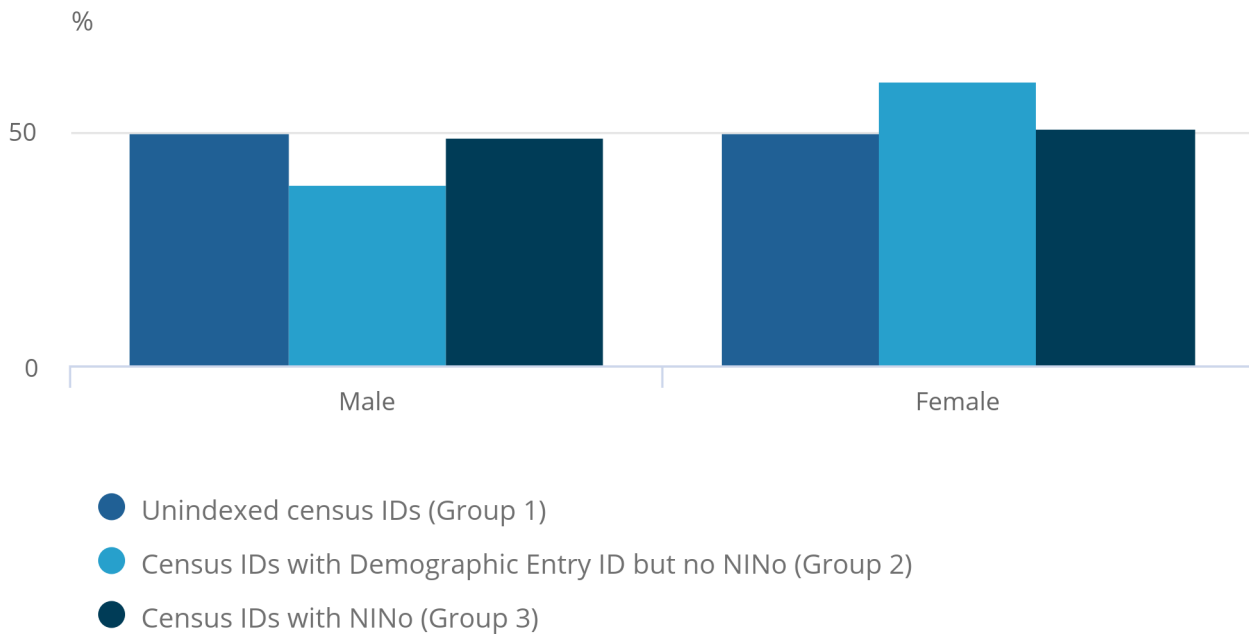
Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Over-representation of population groups in Groups 1 or 2 could lead to bias in the analysis of the linked data, particularly if the analysis focuses on particular population groups. However, the residual records make up only a small proportion of the overall population, so any bias is expected to be small.

Note: The following comparisons were seen in the 2011 Census-National Insurance number (NINo) dataset. However, consistent findings were seen for the 2011 Census-Department for Work and Pensions (DWP) dataset.

Figure 1: Sex, from 2011 Census (England & Wales), by groups for comparison

Figure 1: Sex, from 2011 Census (England & Wales), by groups for comparison



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

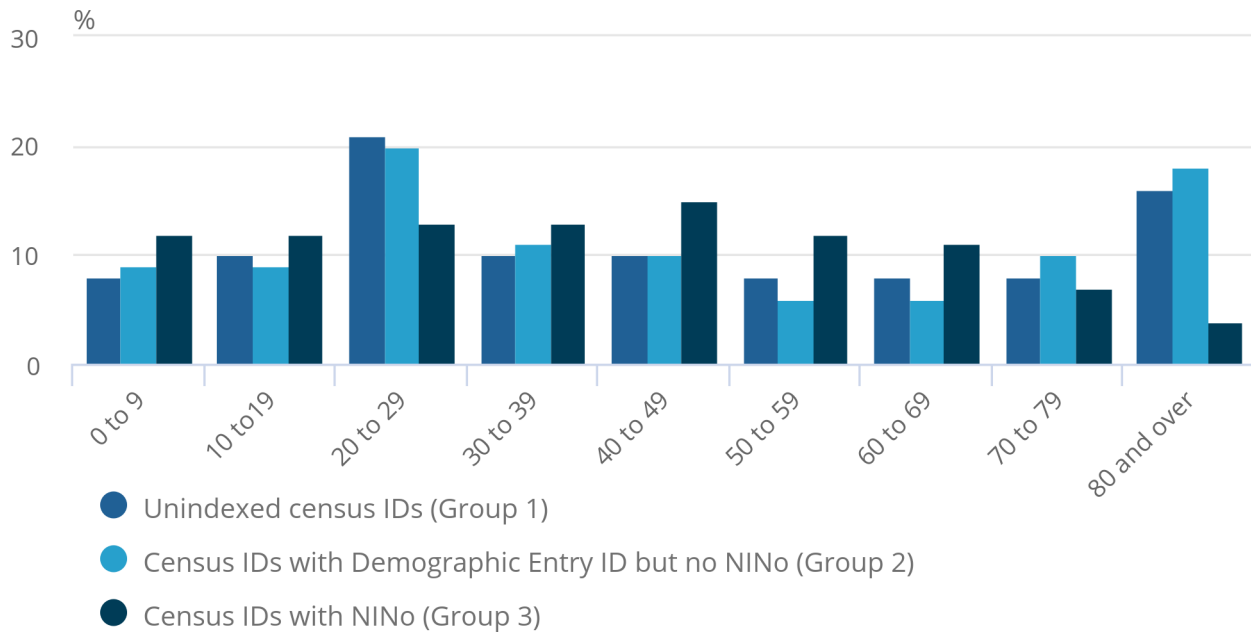
Notes:

1. N = 1,008,164 Group 1, 733,637 Group 2, 51,741,655 Group 3.

When comparing the characteristics of Group 1 with Group 3, males and females were well represented in the linked data, as there were minimal differences in sex composition between groups. However, males were under-represented in Group 2 (39%, versus 49% of Group 3).

Figure 2: Age group, from 2011 Census (England & Wales), by groups for comparison

Figure 2: Age group, from 2011 Census (England & Wales), by groups for comparison



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Notes:

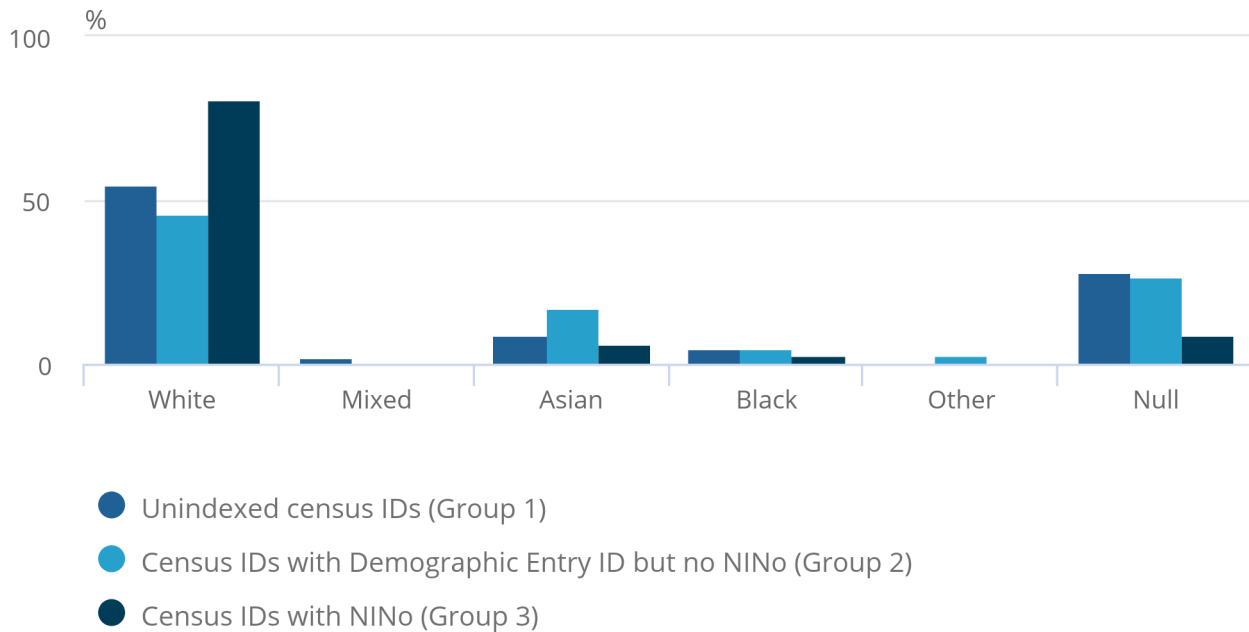
1. N = 1,008,164 Group 1, 733,637 Group 2, 51,741,655 Group 3.

Through comparison of the demographic profiles of Groups 1, 2 and 3, it was seen that 20- to 29-year-olds were over-represented in Groups 1 and 2 (21% of Group 1, 20% of Group 2, versus 13% of Group 3). These records may reflect characteristics of individuals who are less likely to interact with the admin data that make up the DI.

Further to this, individuals aged 80 years and over were also over-represented in Groups 1 and 2 (16% of Group 1, 18% of Group 2, versus 4% of Group 3), which could be because of deaths before 2016 (meaning that these individuals were not included in the DI).

Figure 3: Ethnicity, from 2011 Census (England & Wales), by groups for comparison

Figure 3: Ethnicity, from 2011 Census (England & Wales), by groups for comparison



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Notes:

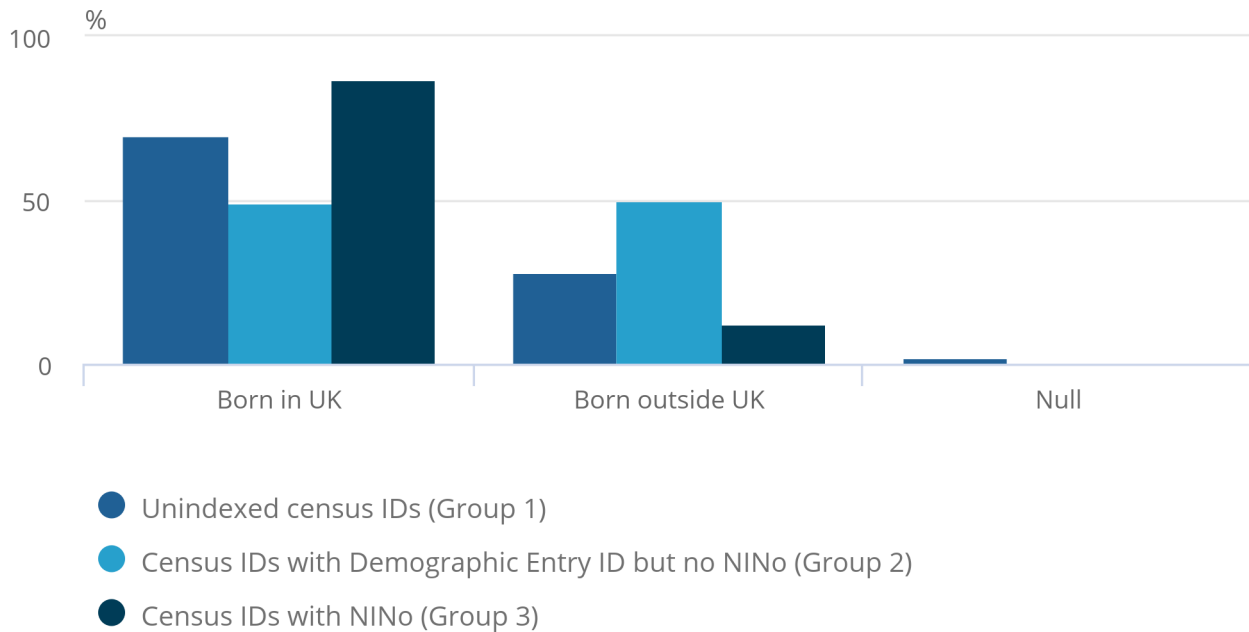
1. N = 1,008,164 Group 1, 733,637 Group 2, 51,741,655 Group 3.

When comparing Group 2 with Group 3, Asian ethnicities were over-represented (17%, versus 6% of Group 3), as were those with a null ethnicity (27%, versus 9% of Group 3). This indicates that some individuals of Asian ethnicity or without an ethnicity recorded may not have been allocated a NINo. This could be because of individuals immigrating to the UK to study, for example.

It was also seen that null ethnicities were over-represented in Group 1 (28%, versus 9% of Group 3). This suggests that these census records could contain less information, so may have made them more difficult to link.

Figure 4: Country of birth, from 2011 Census (England & Wales), by groups for comparison

Figure 4: Country of birth, from 2011 Census (England & Wales), by groups for comparison



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

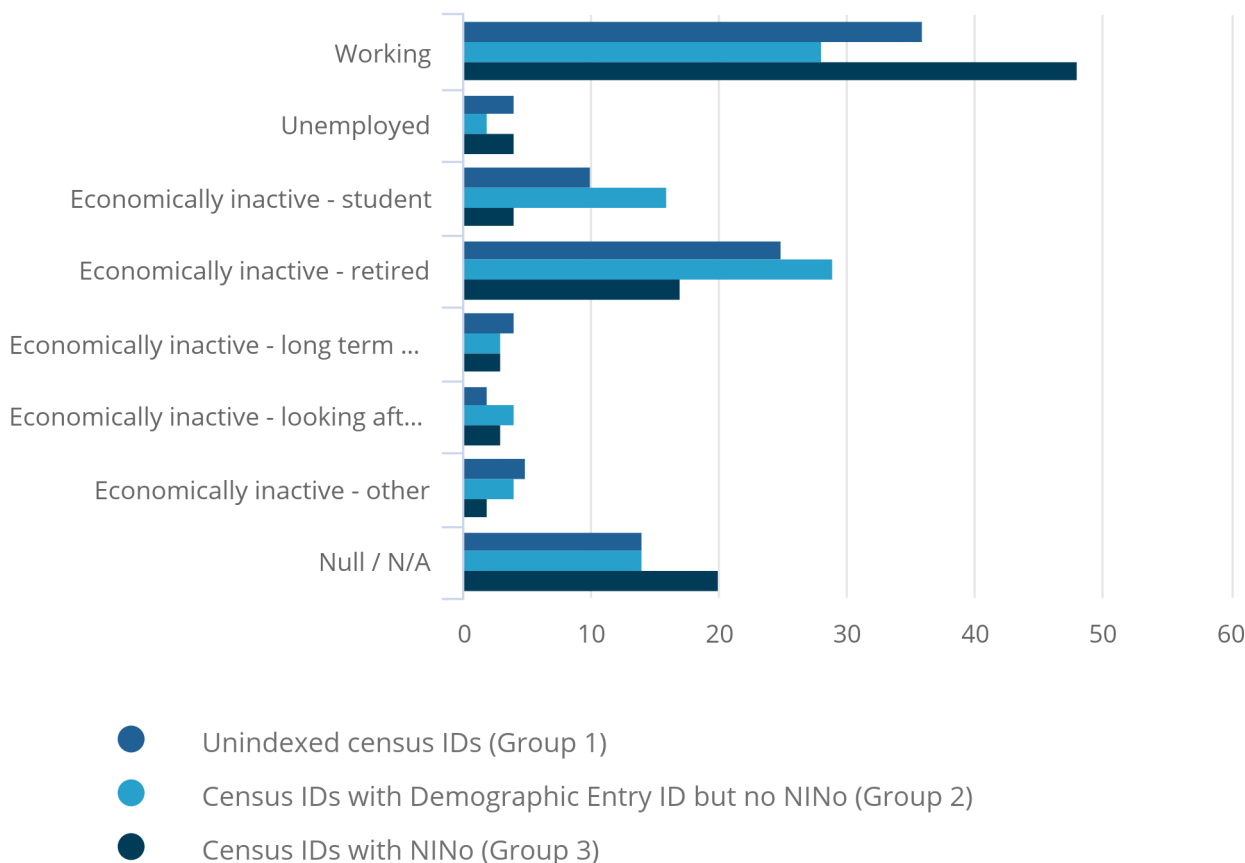
Notes:

1. N = 1,008,164 Group 1, 733,637 Group 2, 51,741,655 Group 3.

Groups 1 and 2 were also likely to be born outside the UK (28% of Group 1, 50% of Group 2, versus 12% of Group 3), which could be because of reasons such as our linkage methods being better suited to matching Western names, individuals moving into the country but not interacting with admin sources, or individuals emigrating prior to 2016.

Figure 5: Activity last week, from 2011 Census (England & Wales), by groups for comparison

Figure 5: Activity last week, from 2011 Census (England & Wales), by groups for comparison



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Notes:

1. N = 1,008,164 Group 1, 733,637 Group 2, 51,741,655 Group 3.

It was seen that records in Groups 1 and 2 reflected characteristics of individuals who are less likely to interact with HM Revenue and Customs (HMRC) or update the admin data that make up the DI. For example, these groups saw over-representation among economically inactive individuals (46% of Group 1, 56% of Group 2, versus 29% of Group 3), particularly:

- full-time students (10% of Group 1, 16% of Group 2, versus 4% of Group 3)
- retired individuals (25% of Group 1, 29% of Group 2, versus 17% of Group 3)

Exploration of conflicting clusters

Where an ID from one source has been linked to multiple IDs from another source, this is described as a conflicting cluster. Two types of conflicting clusters can be seen in the linked data.

Census IDs with multiple DWP master keys or NINos

These conflicting clusters are likely to reflect where more than one master key or NINo is clustered in a Demographic Entry ID as part of the DI build. This does not necessarily indicate that an error in clustering has occurred, although this could happen in a small proportion of cases. The exact processes that lead to multiple DWP master keys or NINos being given to one person is unknown.

DWP master keys or NINos with multiple census IDs

These conflicting clusters are likely to occur when multiple NHS numbers have been clustered in a Demographic Entry ID, bringing together census IDs as duplicates, and further census IDs being identified as duplicates in the residuals.

Table 6: Counts of two types of conflicting clusters observed in the DWP and HMRC linked data frames

| | [DWP master keys] | [NINos] |
|---|--------------------------|----------------|
| Number of census IDs with multiple [IDs] | 121,924 | 172,674 |
| Number of [IDs] with multiple census IDs | 908,268 | 910,233 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

As shown in Table 6, the number of conflicting clusters in the linked data frames is relatively small when comparing with the entirety of census IDs.

Table 7: Size of conflicting clusters in linked data, where census IDs have linked to multiple DWP master keys or NINos

| | [DWP master keys] | [NINos] |
|---|--------------------------|----------------|
| Number of census ID which linked to 2 [IDs] | 116,839 | 165,738 |
| Number of census ID which linked to 3 [IDs] | 4,411 | 6,120 |
| Number of census ID which linked to 4 [IDs] | 500 | 614 |
| Number of census ID which linked to 5+ [IDs] | 174 | 202 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Table 8: Size of conflicting clusters in linked data, where DWP master keys or NINos have linked to multiple census IDs

| | [DWP master keys] | [NINos] |
|--|--------------------------|----------------|
| Number of [IDs] which linked to 2 census IDs | 895,166 | 897,063 |
| Number of [IDs] which linked to 3 census IDs | 11,959 | 12,018 |
| Number of [IDs] which linked to 4 census IDs | 962 | 970 |
| Number of [IDs] which linked to 5+ census IDs | 181 | 182 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Tables 7 and 8 indicate that the majority of conflicting clusters are small in size. There are minimal cases of conflicting clusters of 5 and over in size.

It is also important to note that there is a relatively high proportion of overlap between the two types of conflicting clusters. Approximately 10% of records with multiple census IDs also had multiple master keys or NINos. It is possible for some of these records that two census IDs from two different people have been brought together as a result of a clustering error in the DI.

Characteristics of conflicting clusters

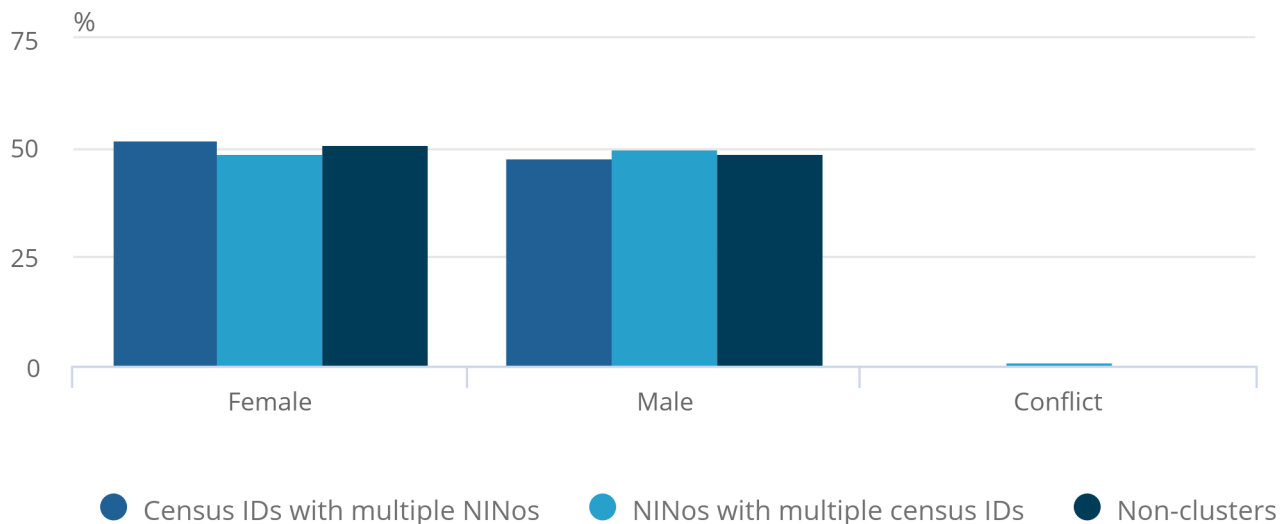
To understand the characteristics of conflicting clusters, we compared the demographic information of the two different types of conflicting clusters versus non-clusters.

Analysis of characteristics of conflicting clusters versus non-clusters strongly indicated a pattern of the types of people who are in each type of cluster. While the groups are relatively small in size, exclusion of conflicting clusters could lead to biases in analysis.

Note: The following comparisons were seen in the 2011 Census-NINo dataset. However, consistent findings were observed for the 2011 Census-DWP dataset.

Figure 6: Sex breakdown comparing conflicting clusters versus non-clusters

Figure 6: Sex breakdown comparing conflicting clusters versus non-clusters



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

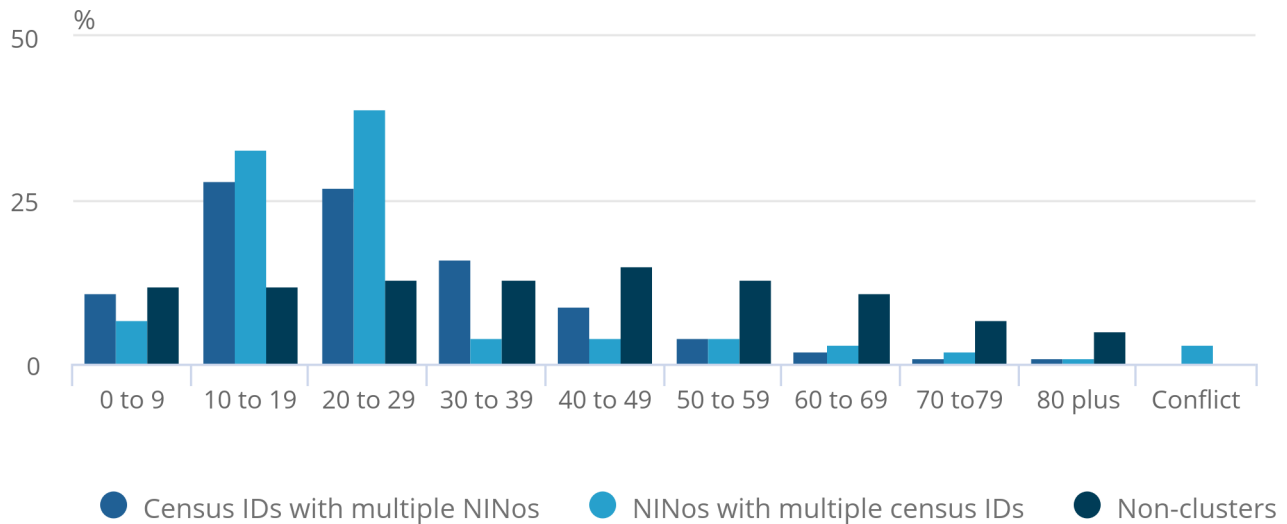
Notes:

1. N = 172,674 census IDs with multiple NINOs, 910,233 NINOs with multiple census IDs, 51,675,363 non-clusters.
2. NINOs categorised as "Conflict" contain multiple census records, of which, have different sexes (not counting "null").

As shown in Figure 6, there were minimal differences in sex breakdown between conflicting clusters and non-clusters.

Figure 7: Age group breakdown comparing conflicting clusters versus non-clusters

Figure 7: Age group breakdown comparing conflicting clusters versus non-clusters



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Notes:

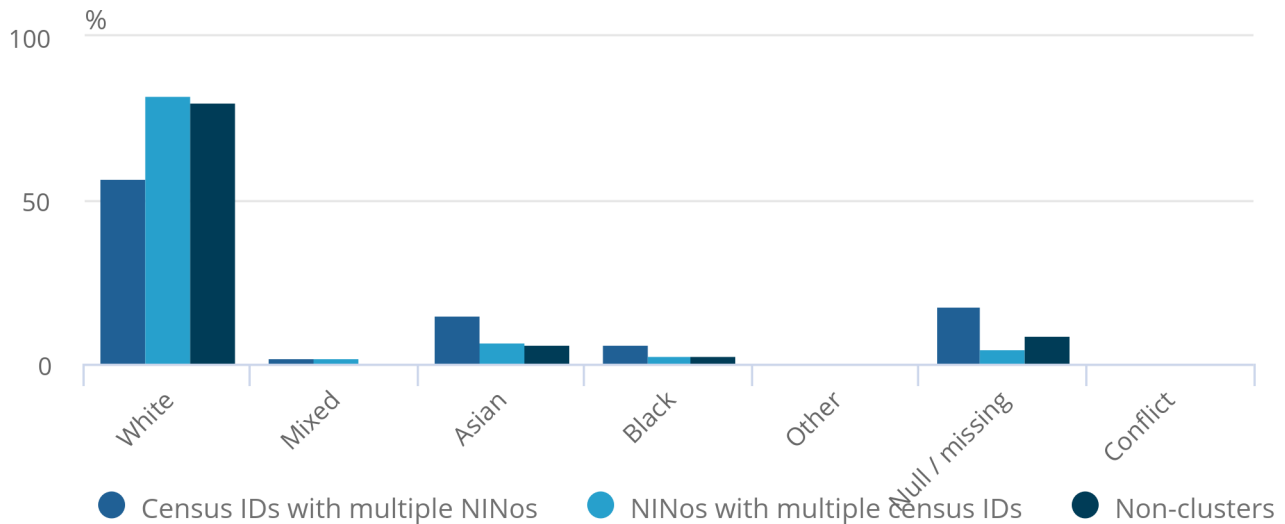
1. N = 172,674 census IDs with multiple NINOs, 910,233 NINOs with multiple census IDs, 51,675,363 non-clusters.
2. NINOs categorised as "Conflict" contain multiple census records, of which, have different age groups (not counting "null").

As shown in Figure 7, census IDs with multiple NINOs were most likely to be aged between 10 and 39 years.

NINOs with multiple census IDs were most likely to be aged between 10 and 29 years. These conflicting clusters may reflect cases where individuals are recorded in the census multiple times, for example, where students are enumerated at both their home address and term-time address, in two different IDs.

Figure 8: Ethnicity breakdown comparing conflicting clusters versus non-clusters

Figure 8: Ethnicity breakdown comparing conflicting clusters versus non-clusters



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

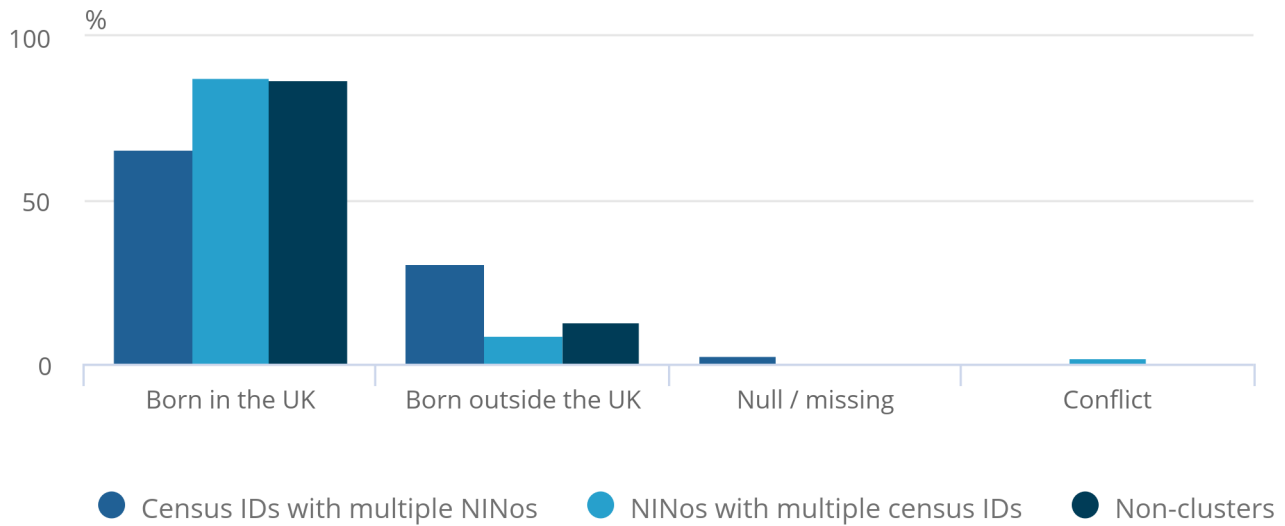
Notes:

1. N = 172,674 census IDs with multiple NINOs, 910,233 NINOs with multiple census IDs, 51,675,363 non-clusters.
2. NINOs categorised as "Conflict" contain multiple census records, of which, have different ethnicities (not counting "null").

As shown in Figure 8, census IDs which contain multiple NINOs are over-represented for non-White ethnicities compared with census IDs with a single NINo.

Figure 9: Country of birth (UK versus outside UK) breakdown comparing conflicting clusters versus non-clusters

Figure 9: Country of birth (UK versus outside UK) breakdown comparing conflicting clusters versus non-clusters



Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Notes:

1. N = 172,674 census IDs with multiple NINOs, 910,233 NINOs with multiple census IDs, 51,675,363 non-clusters.
2. NINOs categorised as "Conflict" contain multiple census records, of which, have different country of birth classifications (not counting "null").

As shown in Figure 9, census IDs which contain multiple NINOs are over-represented for those born outside the UK compared with census IDs with a single NINo. This indicates that individuals born outside the UK could be given more than one NINo, though the process is unknown.

Clerical review for false positives: 2011 Census linked to encrypted NINo (Stages 1 and 2)

False positive (FP) analysis estimates how many of the links made are incorrect. In other words, it calculates a type of linkage error, which occurs when records belonging to different individuals are erroneously linked together. To determine if such errors occurred in the linkage, samples of record pairs were clerically reviewed, and the number of incorrect pairs seen were counted.

Because of the personal data for DWP being hashed, it was decided to only quality assure the 2011 Census-NINo links.

The review of conflicting cluster characteristics highlighted differences between conflicting clusters and "non-clusters". Further to this, it was anticipated that there would be differences in the quality of links, dependent on the different linkage methods used. Therefore, samples from the following groups were taken for quality assessment.

Table 9: Breakdown of six groups for quality assessment

| | No conflicts ("non-clusters") | Census IDs with multiple NINos | NINos with multiple census IDs |
|--|--------------------------------------|---------------------------------------|---------------------------------------|
| Stage 1 linkage (where census ID has Demographic Entry ID and NINo) | 46,995,820 | 160,250 | 91,258 |
| Stage 2 linkage (where census ID has Demographic Entry ID and NINo) | 2,937,742 | 12,424 | 818,975 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Sampling approach

A sample of 4,278 was taken for the clerical review for false positives.

Table 10: Samples taken for the clerical review for false positives

| | No conflicts ("non-clusters") | Census IDs with multiple NINos | NINos with multiple census IDs | Total |
|------------------------|--------------------------------------|---------------------------------------|---------------------------------------|--------------|
| Stage 1 linkage | 713 | 713 | 713 | 2,139 |
| Stage 2 linkage | 713 | 713 | 713 | 2,139 |
| Total | 1,426 | 1,426 | 1,426 | 4,278 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

It is worth noting that the overlap between "census IDs with multiple NINos", and "NINos with multiple census IDs" was taken into consideration when drawing samples. We did not want to clerically review the same record twice, but we did not want to exclude them completely. Therefore, records that fell into both groups were still included in the clerical samples, however, any records that were drawn twice were removed from one of the clerical samples (so they would only be clerically reviewed once). This only affected a small proportion of records.

Clerical Resolution Online Widget (CROW)

Linked records were reviewed using the CROW tool. Records were presented in pairs; showing one row from the 2011 Census and one row from the DI. However, so that reviewers were able to view multiple entries for the same ID (and make an informed decision on whether IDs had been correctly linked), records were presented in arrays.

Findings: estimation of false positives

Table 11: Results of the clerical review for false positives, Stage 1 only

| | No conflicts ("non-clusters") | Census IDs with multiple NINOs | NINOs with multiple census IDs |
|---------------------------------------|--|---|---|
| Sample reviewed | 713 | 713 | 713 |
| True positives (TP) in sample | 703 | 686 | 706 |
| False positives (FP) in sample | 10 | 27 | 7 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Table 12: Results of the clerical review for false positives, Stage 2 only

| | No conflicts ("non-clusters") | Census IDs with multiple NINOs | NINOs with multiple census IDs |
|---------------------------------------|--|---|---|
| Sample reviewed | 713 | 713 | 713 |
| True positives (TP) in sample | 682 | 583 | 711 |
| False positives (FP) in sample | 31 | 130 | 2 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Table 13: Results of the clerical review for false positives, Stages 1 and 2

| | No conflicts ("non-clusters") | Census IDs with multiple NINOs | NINOs with multiple census IDs |
|---------------------------------------|--|---|---|
| Sample reviewed | 1426 | 1426 | 1426 |
| True positives (TP) in sample | 1385 | 1269 | 1417 |
| False positives (FP) in sample | 41 | 157 | 9 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Precision is a measure of the accuracy of the matches that have been made. To calculate the precision of our outputs, our error estimates were inputted into the linkage error grid (Table 14). Using the error grid, precision was calculated using:

$$precision = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}}$$

Table 14: Estimated error grid of the linkage between the 2011 Census and the Demographic Index, Stages 1 and 2

| | No conflicts ("non-clusters") | Census IDs with multiple NINOs | NINOs with multiple census IDs |
|---------------------------------------|--|---|---|
| Estimated true positives (TP) | ~49,146,706 | ~336,362 | ~1,822,950 |
| Estimated false positives (FP) | ~786,856 | ~17,053 | ~12,081 |
| Precision estimate | 98.42% | 95.17% | 99.34% |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

As shown, the estimated error rates for different conflicting cluster types were relatively low. Census IDs with multiple NINos yield the lowest estimated precision (95.17%).

After weighting the data to account for the actual proportions of these groups, the overall precision estimate was calculated to be 97.76%. The estimated precision (97.76%) indicates the percentage of links classified as true matches.

Table 15: Estimated error grid of the linkage between the 2011 Census and the Demographic Index, by stage

| | Stage 1 | Stage 2 |
|---------------------------------------|-------------|------------|
| Estimated true positives (TP) | ~47,452,092 | ~6,170,484 |
| Estimated false positives (FP) | ~676,518 | ~551,341 |
| Precision estimate | 98.59% | 91.80% |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

As shown, the highest proportion of true positives was seen in the first stage (2011 Census-PR lookup). Running the residual records through DIMS yielded a lower precision estimate, however, this only accounted for 12% of links.

Clerical review for false negatives: 2011 Census linked to encrypted NINo

False negative (FN) analysis estimates how many true matches exist between the datasets but were missed by the linkage methods. To determine if such errors occurred in the linkage, samples of unlinked record pairs were clerically reviewed, and the number of pairs incorrectly labelled as non-matches were counted.

Sampling approach

A sample of 7,309 was taken for the clerical review for false negatives. The clerical review involved comparing personal data from two records and deciding if they were a match or whether they should remain unlinked. A probabilistic data linkage algorithm (Splink) was applied to the non-links and record pairs were grouped by score. Sampled record pairs were reviewed using the CROW tool, as per the false positive review.

Table 16: Samples taken for the clerical review for false negatives

| | Group 1 (Highest scoring) | Group 2 | Group 3 | Group 4 (Lowest scoring) | Total |
|---|------------------------------|---------|---------|-----------------------------|-------|
| Total record pairs | 1,924 | 1,968 | 1,732 | 1,685 | 7,309 |
| Record pairs of which contain NINo | 1,537 | 1,537 | 1,537 | 1,537 | 6,148 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Findings: estimation of false negatives

Table 17: Results of the clerical review for false negatives

| | Total record pairs | Record pairs of which contain NINo |
|--------------------------------|--------------------|------------------------------------|
| Sample reviewed | 7,309 | 6,148 |
| True negatives (TN) in sample | 6,745 | 5,649 |
| False negatives (FN) in sample | 564 | 499 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Table 18: Estimated error grid of the linkage between the 2011 Census and the Demographic Index, overall level

| | Records matched | Records not matched (residuals) |
|------------------|-----------------------------------|---------------------------------|
| Links | True positive (TP) ~53,622,576 | False negative (FN) ~25,929 |
| Non-links | False positive (FP) ~1,227,859 | True negative (TN) ~982,235 |

Source: 2011 Census to Demographic Index linked data from the Office for National Statistics

Notes

1. Base: Record pairs of which contain NINo

Recall is a measure of the proportion of matches that have been made from all the possible matches. Recall was calculated using:

$$recall = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

The recall estimate (99.95%) is the percentage of true matches that were classified as links.

Note: the recall estimate is only for the 2011 Census-NINo links.

6 . Limitations

It is important to take into consideration the limitations of this linkage.

Demographic Index (DI)

Although the DI has good coverage, across a spread of admin sources, records only date back to 2016, so may not contain records for everyone in the 2011 Census.

The DI is created using admin data, so there may be issues surrounding the quality of source data (for example, missingness, input errors). This can lead to errors in the clustering process.

The DI contains some cases of conflicting clusters, where one Demographic Entry ID is linked to multiple external identifiers (for example, a Demographic Entry ID with multiple National Insurance numbers (NINOs)). In most cases, this is because of one person having multiple external identifiers. However, sometimes it can be indicative of an error in clustering. Cases where a census ID has linked to multiple master keys (N=113,127) or NINOs (N=160,250) have been flagged in the linked data.

2011 Census-PR linkage

One limitation of the linkage method is potential for error in the 2011 Census-Patient Register (PR) lookup. The lookup was created using deterministic and probabilistic linkage methods, meaning that there is scope for incorrect links (false positives) and missed links (false negatives). The estimated false positive rate for the 2011 Census-PR linkage was 0.21%, and estimated false negative rate was 1.1%.

Further to this, the methods used for the 2011 Census-PR linkage only allowed for one-to-one links. This means that duplicate census records (people who filled out a census form at different locations, such as students at their home and term-time address) were not accounted for and were likely to be highly concentrated in the residuals of the 2011 Census-PR linkage. Many of these were picked up in the Stage 2 linkage and therefore we recommend use of Stage 1 and Stage 2 links together.

7 . Related links

[2021 Census linkage to DWP master key and encrypted NINo](#) Methodology | Released 6 December 2024
Linkage methodology and quality information for 2021 Census linkage to DWP (Department for Work and Pensions) master key and encrypted NINo (National Insurance number).

8 . Cite this methodology

Office for National Statistics (ONS), released 6 December 2024, ONS website, methodology, [2011 Census linkage to DWP master key and encrypted NINo](#).