2021 Census linkage to DWP master key and encrypted NINo

Linkage methodology and quality information for 2021 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

4. Quality information

5. Limitations

6. Related links

7. Cite this methodology

1 . Main points

- To allow health and labour market analysis projects, the 2021 Census was linked to the Department for Work and Pensions (DWP) and HM Revenue and Customs (HMRC) data.
- The linkage was conducted by indexing the 2021 Census data to the Office for National Statistics (ONS) Demographic Index (DI).
- The linkage resulted in 96.7% of census IDs linking to a DWP master key or encrypted National Insurance number (NINo).
- The linkage outputs contained lookups between the 2021 Census and DWP master key, and the 2021 Census and encrypted NINo identifiers.
- The overall precision estimate was calculated to be 99.87% and recall was estimated to be 99.86%.

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 2021 Census to DWP and HMRC data will allow for census data to be integrated with health and economic datasets:

- the link between 2021 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 2021 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 2011 Census, DWP and HMRC data was also conducted, using a separate linkage methodology. See the methodology report.

3 . Linkage methodology

2021 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.

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, explained in Section 5: Limitations.

Demographic Index Matching Service (DIMS)

DIMS is a linkage pipeline developed to index datasets against the 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.

Linkage between 2021 Census and the Demographic Index

The 2021 Census was indexed to the DI using DIMS. 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.

The 2021 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 and splitting DWP master key and encrypted NINo into separate tables.

Table 1: 2021 Census linked to DWP master key, summary of 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)**	459,980	0.8%
Group 2	**Census IDs with Demographic Entry ID only**	1,490,074	2.5%
Group 3	**Census IDs with master key and Demographic Entry ID**	56,673,658	96.7%
Total	**Total census IDs in linked data frame (including census residuals)**	58,623,712	-

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Table 2: 2021 Census linked to encrypted NINo, summary of 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)**	459,980	0.8%
Group 2	**Census IDs with Demographic Entry ID only**	1,446,963	2.5%
Group 3	**Census IDs with NINo and Demographic Entry ID**	56,716,769	96.7%
Total	**Total census IDs in linked data frame (including census residuals)**	58,623,712	-

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

4 . 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 2021 Census does not have a corresponding cluster in the DI. While this is possible, we expect a high level of coverage overlap between the two sources.

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 3: Groups for comparison

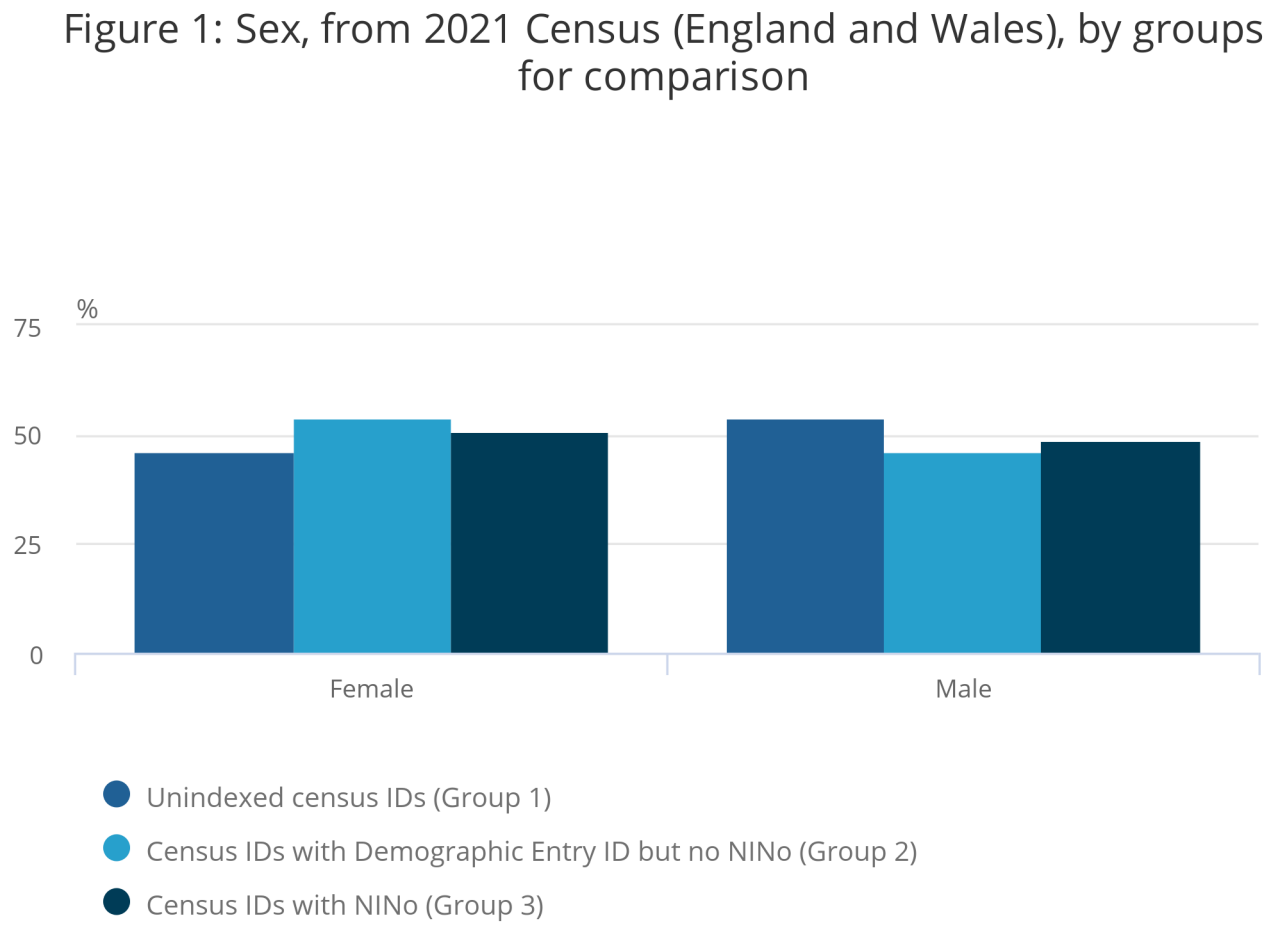
Group	Description	Proportion of census IDs
1	**Census records that did not link to the Demographic Index (Census residuals)**	0.8%
2	**Census records that linked to the Demographic Index, but did not obtain a DWP master key / encrypted NINo**	2.5%
3	**Census records that linked to the Demographic Index and obtained a DWP master key / encrypted NINo**	96.7%

Source: 2021 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 Census-National Insurance number (NINo) dataset. However, consistent findings were seen for the Census-Department for Work and Pensions (DWP) dataset.

Figure 1: Sex, from 2021 Census (England and Wales), by groups for comparison



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

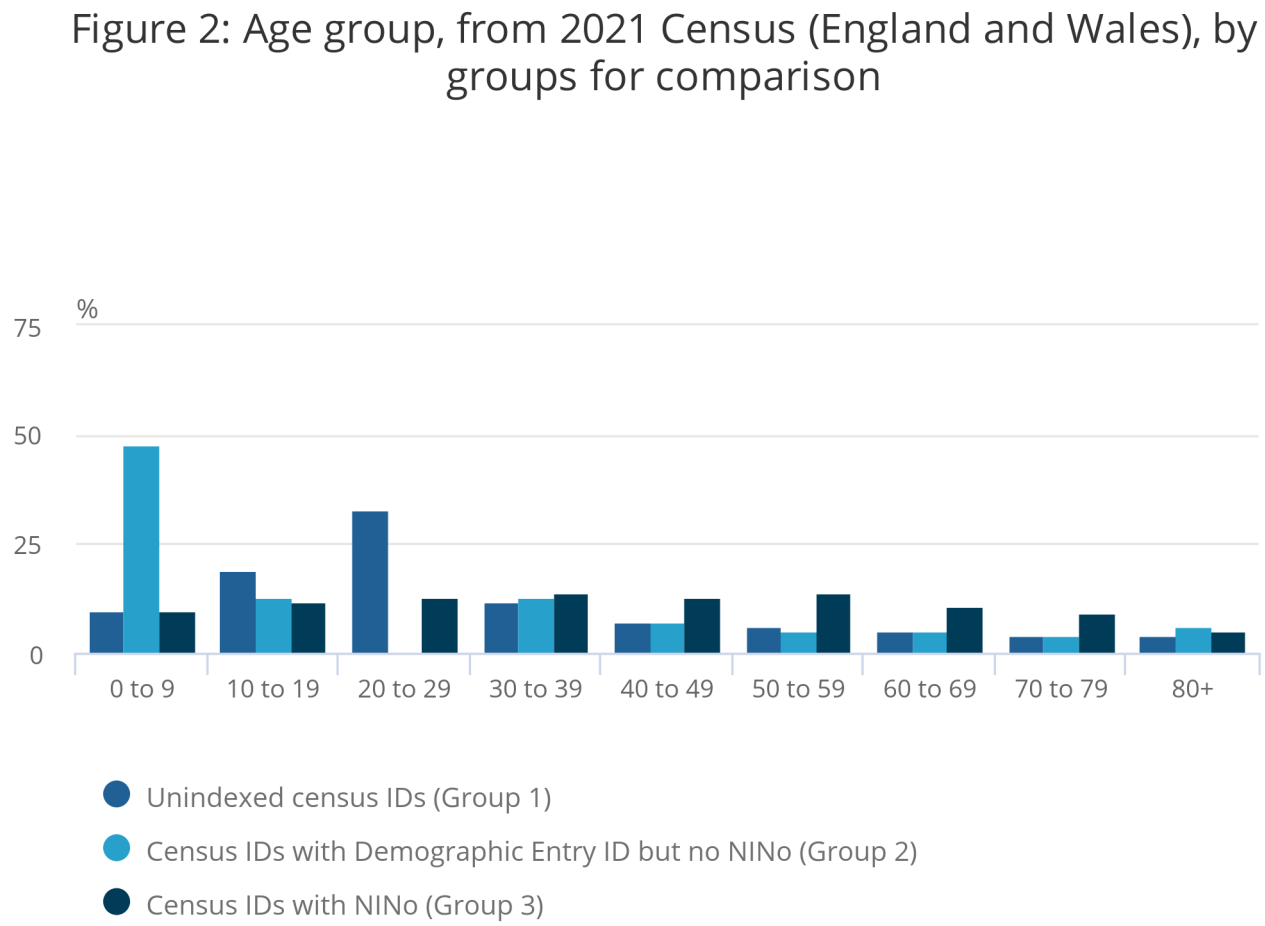
Notes:

1. N = 459,980 Group 1, 1,446,963 Group 2, 56,716,769 Group 3.

Through comparison of the demographic profiles of Groups 1 and 3, it was seen that males were marginally over-represented in Group 1 (54%, versus 49% of Group 3).

When comparing the demographic profiles of Groups 2 and 3, it was seen that females were marginally over-represented in Group 2 (54%, versus 51% of Group 3).

Figure 2: Age group, from 2021 Census (England and Wales), by groups for comparison



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

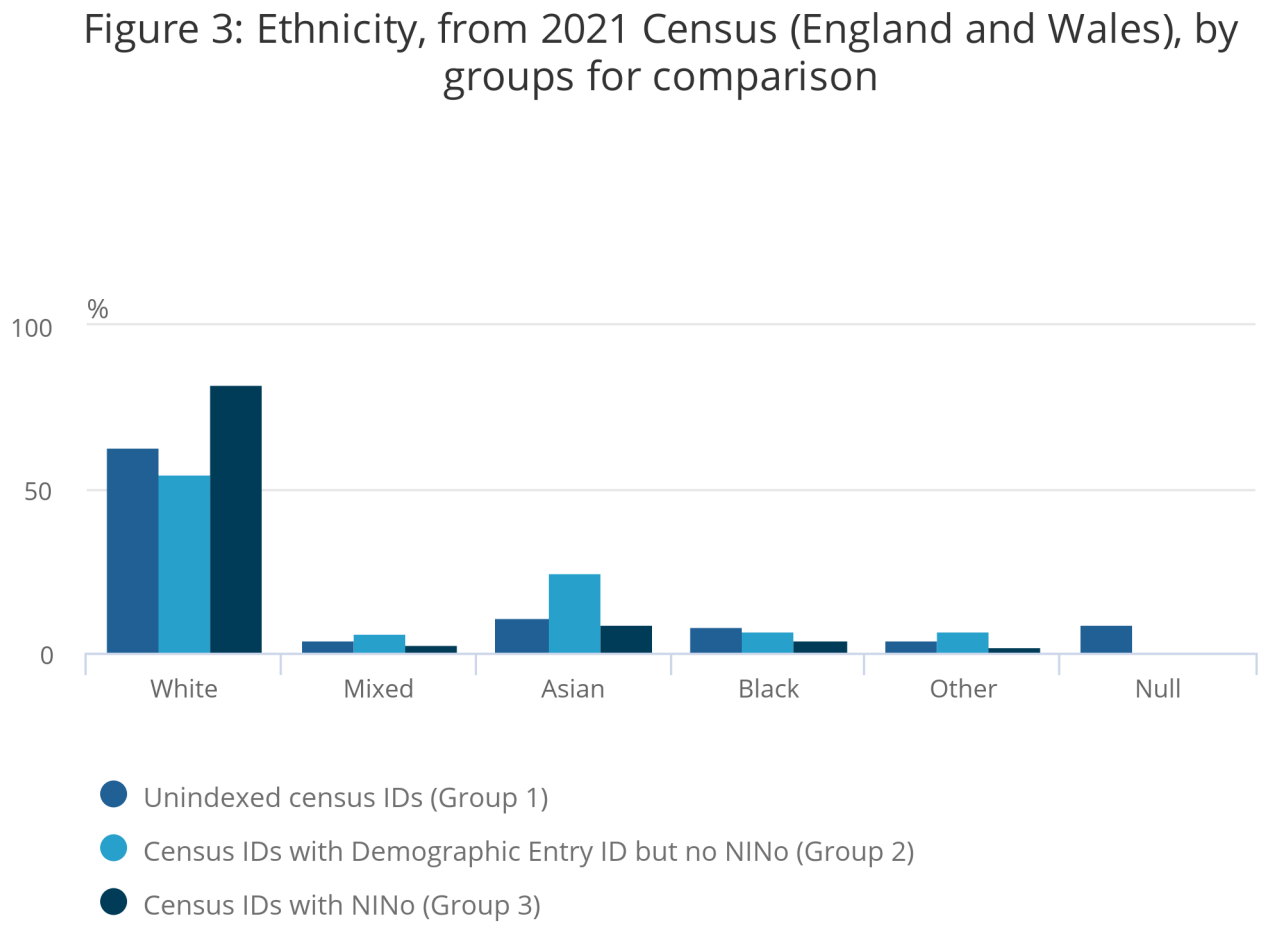
Notes:

1. N = 459,980 Group 1, 1,446,963 Group 2, 56,716,769 Group 3.

Through comparison of the demographic profiles of Groups 1 and 3, it was seen that 20- to 29-year-olds were over-represented in Group 1 (33%, versus 13% of Group 3). Of the characteristics assessed, age profile contained the biggest discrepancy between the residuals (Group 1) and linked data (Group 3). These residual records roughly reflect characteristics of individuals who are less likely to interact with the admin data that make up the DI.

When comparing Group 2 with Group 3, Group 2 were likely to fall into the under-10 years age band (48%, versus 10% of Group 3), which reflects the fact that the majority of this group will not have been allocated a NINo.

Figure 3: Ethnicity, from 2021 Census (England and Wales), by groups for comparison



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

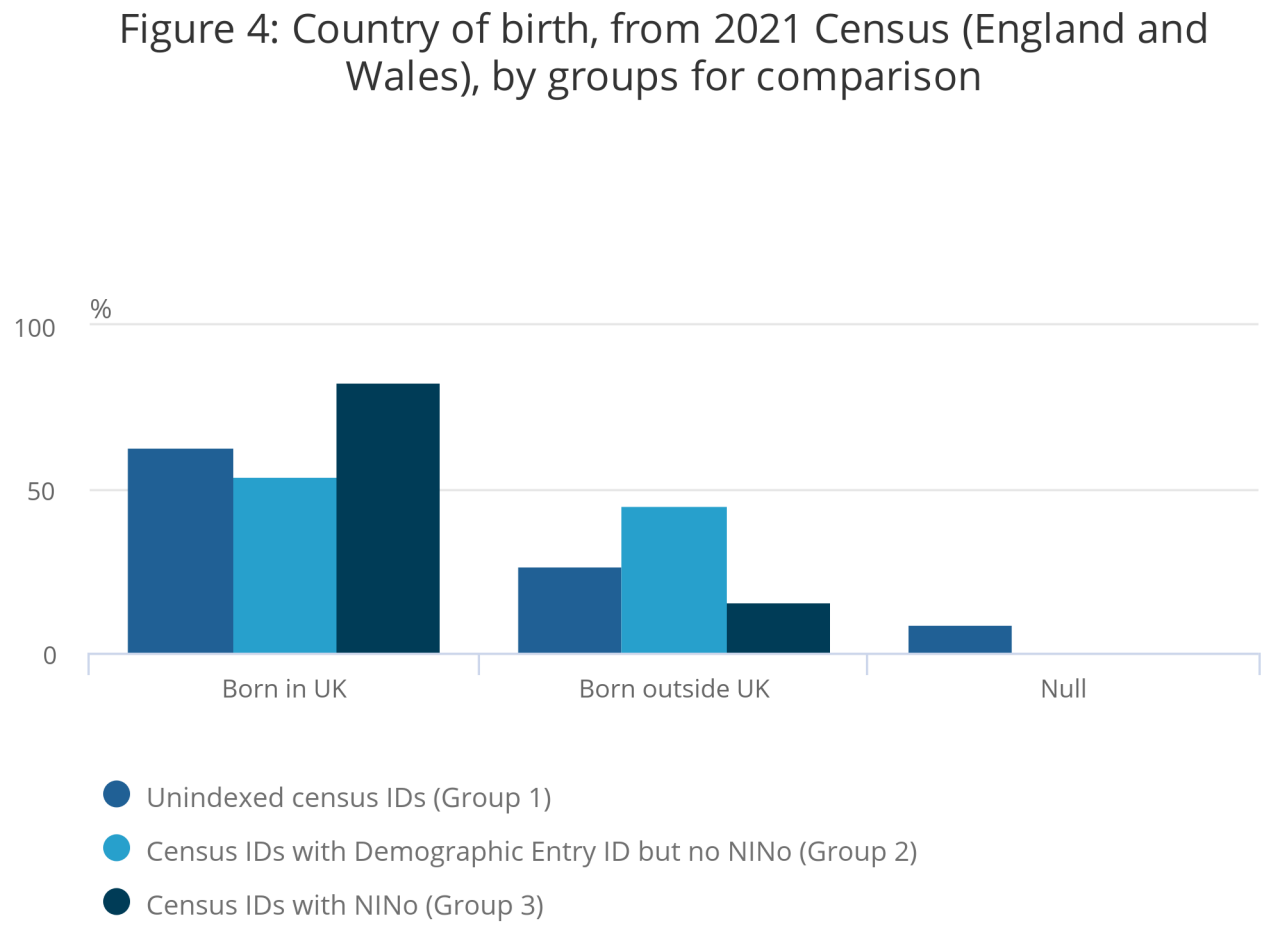
Notes:

1. N = 459,980 Group 1, 1,446,963 Group 2, 56,716,769 Group 3.

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

When comparing Group 2 with Group 3, Asian ethnicities were over-represented (25%, versus 9% of Group 3), indicating that some individuals of Asian ethnicity have not been allocated a NINo. This could be because of immigrating to the UK to study, for example.

Figure 4: Country of birth, from 2021 Census (England and Wales), by groups for comparison



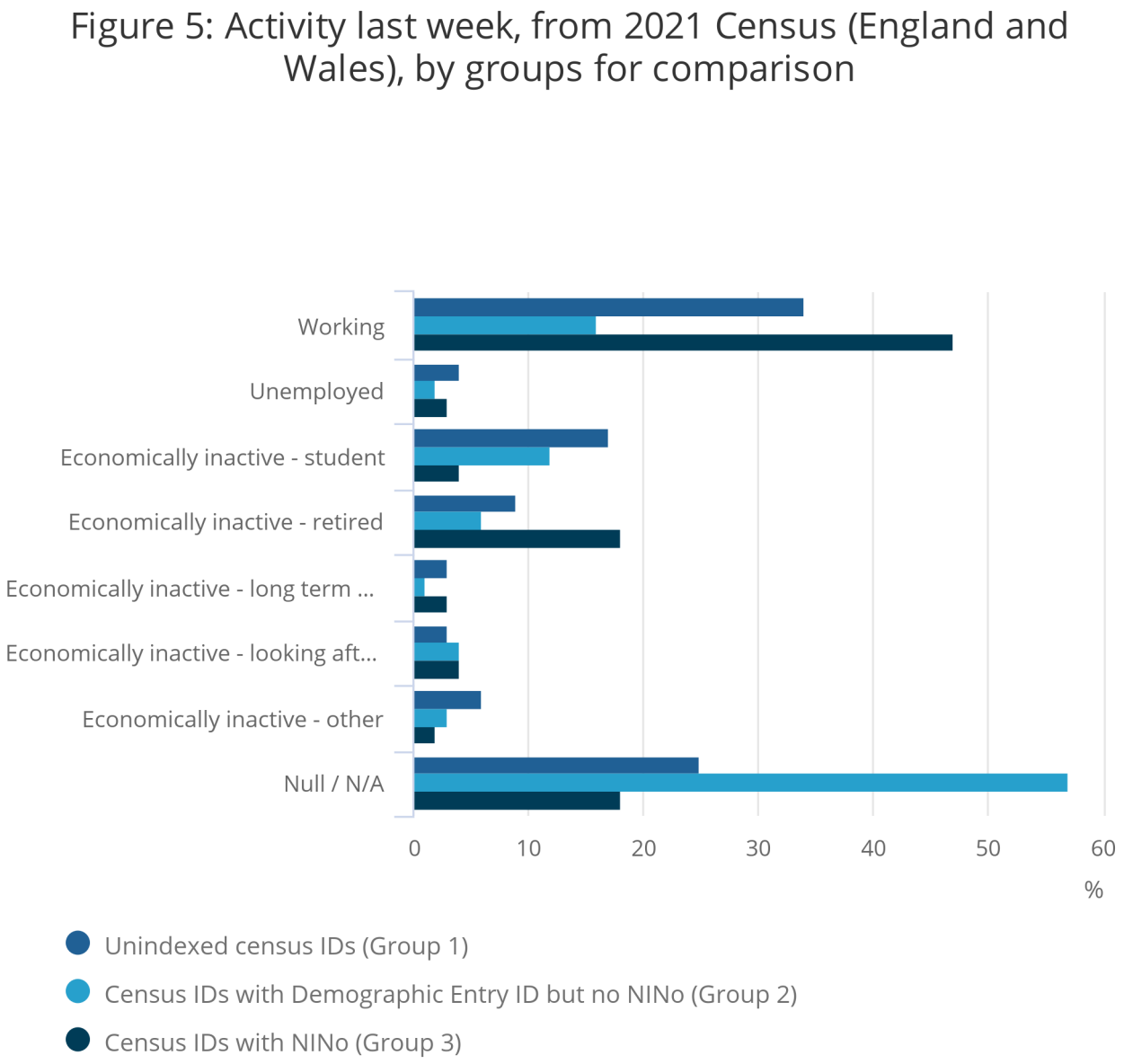
Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Notes:

1. N = 459,980 Group 1, 1,446,963 Group 2, 56,716,769 Group 3.

Groups 1 and 2 were also likely to be born outside the UK (27% of Group 1, 45% of Group 2, versus 16% of Group 3), which could be because of reasons such as our linkage methods being better suited to matching Western names, or individuals moving into the country but not interacting with admin sources.

Figure 5: Activity last week, from 2021 Census (England and Wales), by groups for comparison



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Notes:

1. N = 459,980 Group 1, 1,446,963 Group 2, 56,716,769 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. Group 1 saw over-representation of economically inactive individuals (37%, versus 32% of Group 3), and both groups saw over-representation among full-time students (17% of Group 1, 12% of Group 2, versus 4% of Group 3). Further to this, there was also a high proportion of "null or N/A" values seen among Group 2 (57%, versus 18% 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 reflect anyone in the census who was enumerated at multiple addresses (for example, students at their term and home address). This type of duplication was not reconciled at a record level and was dealt with at the estimation stage. However, there is also a possibility that multiple census records have been clustered incorrectly because of linkage error.

Table 4: 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]	147,636	217,439
Number of [IDs] with multiple census IDs	814,790	817,306

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

As shown in Table 4, the number of conflicting clusters in the linked data frames is relatively small when comparing with the entirety of census IDs.

Table 5: 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]	141,991	209,717
Number of census ID which linked to 3 [IDs]	4,823	6,720
Number of census ID which linked to 4 [IDs]	598	733
Number of census ID which linked to 5+ [IDs]	224	269

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Table 6: 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	798,239	800,642
Number of [IDs] which linked to 3 census IDs	15,118	15,205
Number of [IDs] which linked to 4 census IDs	1,281	1,304
Number of [IDs] which linked to 5+ census IDs	152	155

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Tables 5 and 6 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 11.6% 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.

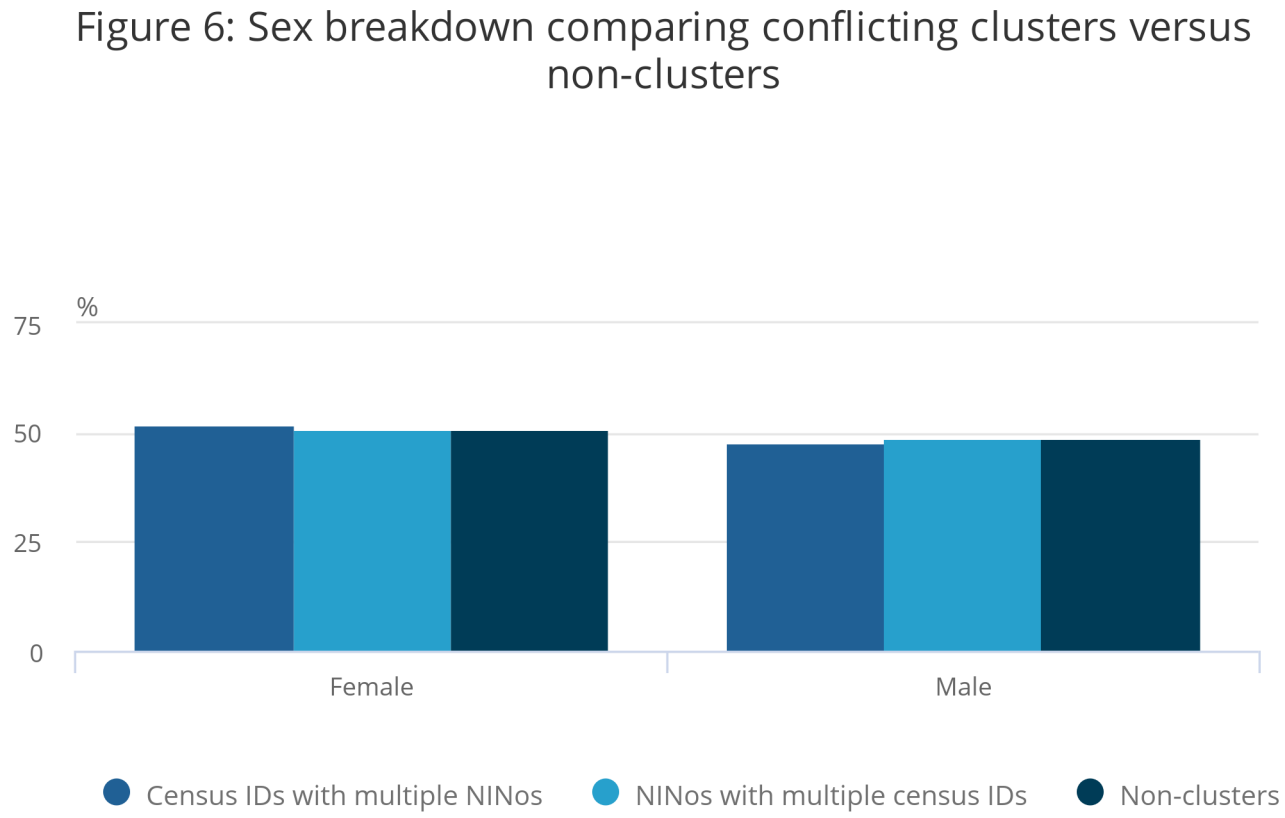
Characteristics of conflicting clusters

To understand the characteristics of conflicting clusters, we compared the demographics 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 2021 Census-NINo dataset. However, consistent findings were observed for the Census-DWP dataset.

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



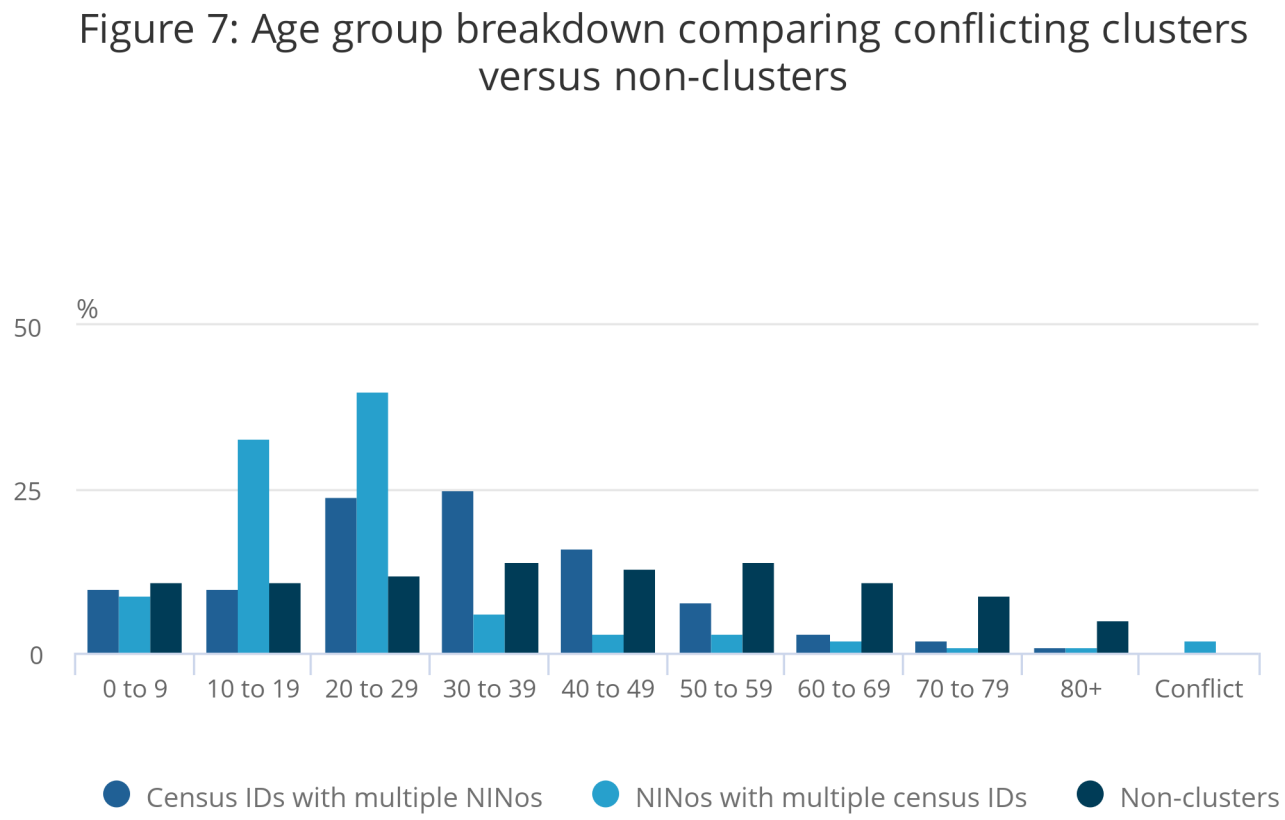
Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Notes:

1. N = 217,439 census IDs with multiple NINos, 817,306 NINos with multiple census IDs, 56,966,545 non-clusters.

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



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Notes:

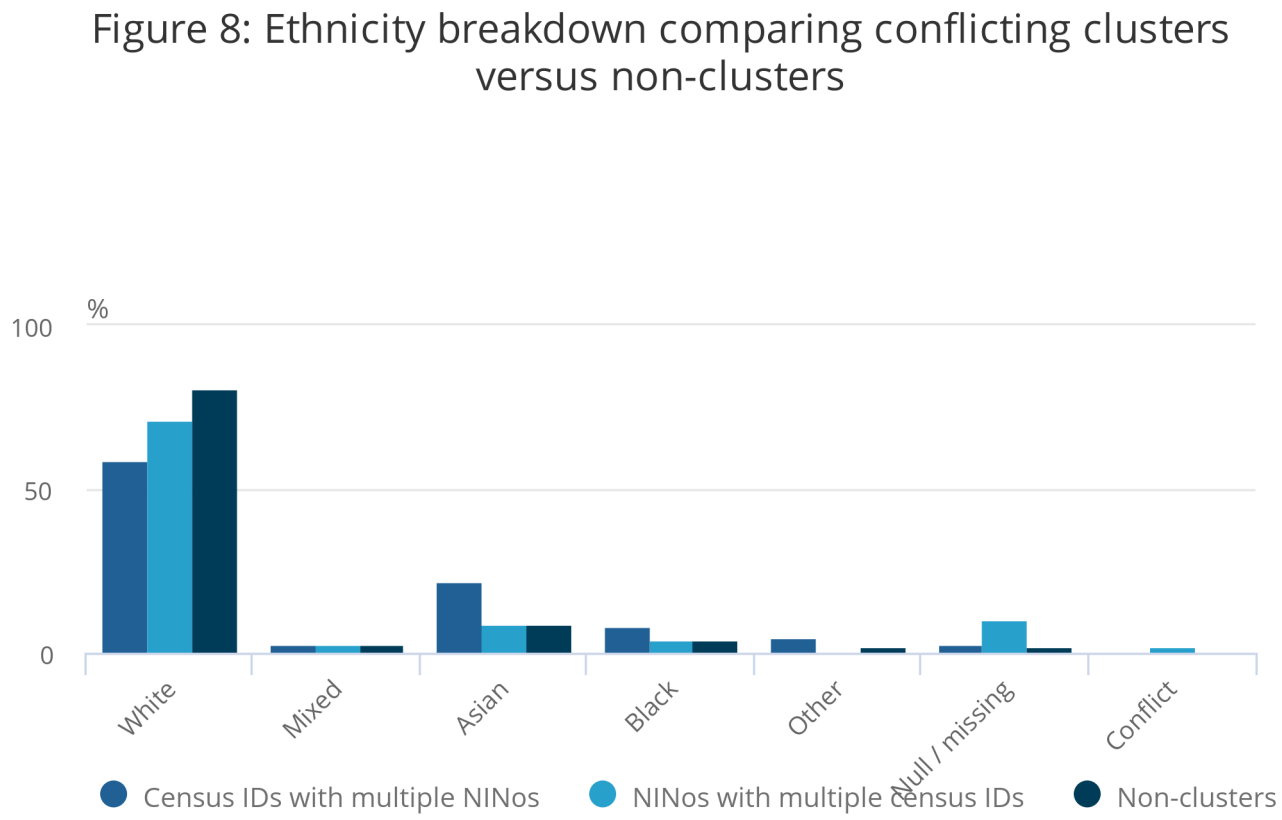
1. N = 217,439 census IDs with multiple NINos, 817,306 NINos with multiple census IDs, 56,966,545 non-clusters.

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 which have multiple NINos contain higher proportions of individuals aged between 20 and 49 years.

NINos which have multiple census IDs contain higher proportions of those aged between 10 and 29 years. These conflicting clusters can be explained by duplicate entries in the census, 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



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Notes:

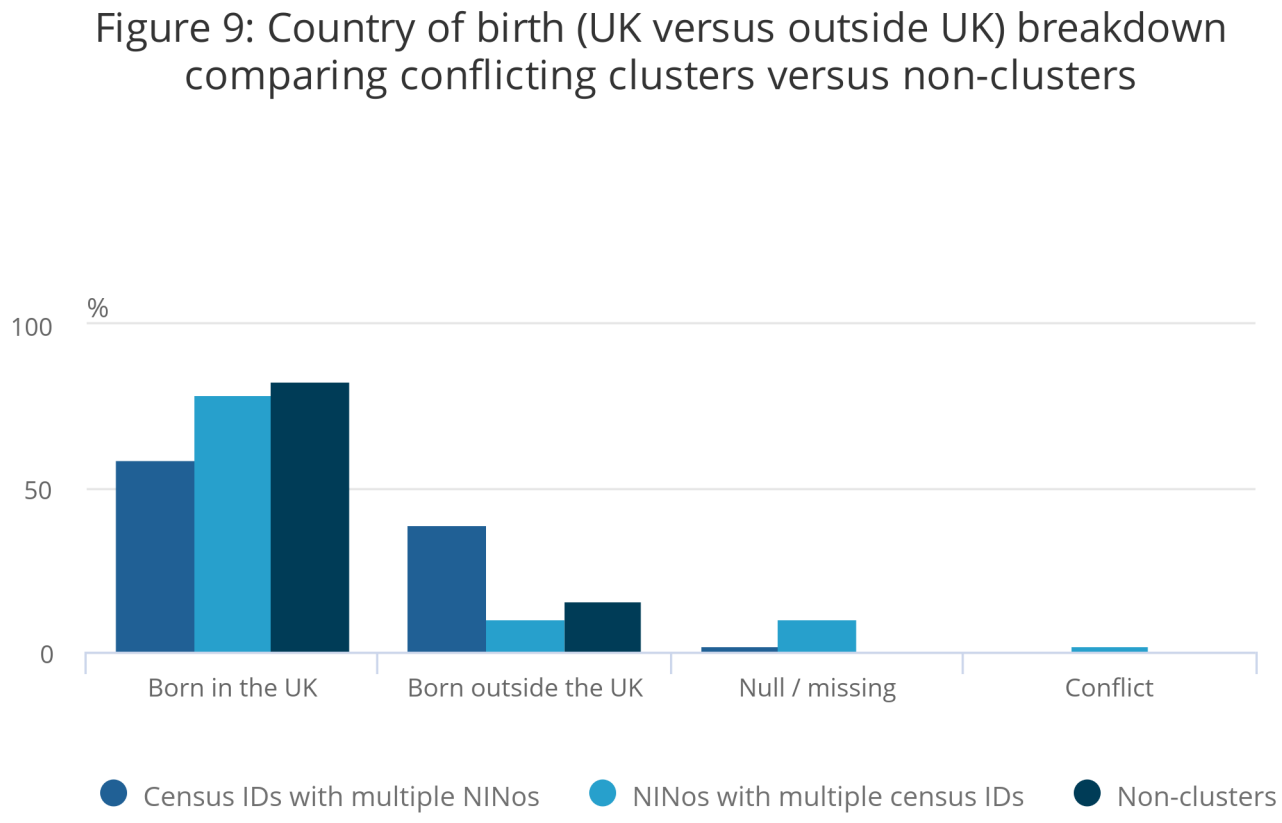
1. N = 217,439 census IDs with multiple NINos, 817,306 NINos with multiple census IDs, 56,966,545 non-clusters.

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 generally over-represented for non-White ethnicities compared with census IDs with a single NINo.

Of NINos which contain multiple census IDs, 10% have null ethnicity compared with 2% of records with a single census ID.

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



Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Notes:

1. N = 217,439 census IDs with multiple NINos, 817,306 NINos with multiple census IDs, 56,966,545 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.

Of NINos which contain multiple census IDs, 10% have null country of birth.

Clerical review for false positives: 2021 Census linked to encrypted NINo

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 2021 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 7: Breakdown of nine groups for quality assessment

	No conflicts ("non-clusters")	Census IDs with multiple NINos	NINos with multiple census IDs
Exact matches	45,712,103	183,120	538,407
Near exact/ deterministic matches	9,290,093	33,156	269,828
Probabilistic matches	57,406	1,163	9,071

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Sampling approach

A sample of 5,253 was taken for the clerical review of false positives. Note: for the exact matches, small samples were taken as no errors were expected.

Table 8: Samples taken for the clerical review for false positives

	No conflicts ("non-clusters")	Census IDs with multiple NINos	NINos with multiple census IDs	Total
Exact matches	137	419	419	975
Near exact/ deterministic matches	713	713	713	2,139
Probabilistic matches	713	713	713	2,139
Total	1,563	1,845	1,845	5,253

Source: 2021 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 2021 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 9: Results of the clerical review for false positives

	No conflicts ("non-clusters")	Census IDs with multiple NINos	NINos with multiple census IDs
Sample reviewed	1563	1845	1845
True positives (TP) in sample	1467	1770	1829
False positives (FP) in sample	96	75	16

Source: 2021 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 10). 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 10: Estimated error grid of the linkage between the 2021 Census and the Demographic Index

	No conflicts ("non-clusters")	Census IDs with multiple NINos	NINos with multiple census IDs
Estimated true positives (TP)	~ 54,987,127	~ 215,791	~ 815,640
Estimated false positives (FP)	~72,475	~1,648	~1,666
Precision estimate	99.87%	99.24%	99.80%

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

As shown, the estimated error rates for different conflicting cluster types were low. Census IDs with multiple NINos yielded an estimated precision of 99.24%, which is only marginally below the estimated precision for non-clusters.

After weighting the data to account for the actual proportions of these groups, the overall precision estimate was calculated to be 99.87%. The estimated precision (99.87%) indicates the percentage of links classified as true matches.

Table 11: Estimated error grid of the linkage between the 2021 Census and the Demographic Index, by indexing method

	Exact matches	Near exact/ deterministic matches	Probabilistic matches
Estimated true positives (TP)	~ 46,986,249	~ 9,812,274	~ 68,521
Estimated false positives (FP)	~ 0	~ 68,547	~ 7,743
Precision estimate	100.00%	99.31%	89.85%

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

As shown, the highest proportion of true positives was seen in the exact matching phase, followed by deterministic methods. Probabilistic methods yielded a lower precision estimate; however, they only accounted for less than 1% of links.

Clerical review for false negatives: 2021 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 7,115 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 12: 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	1,740	1,933	1,770	1,672	7,115
Record pairs of which contain NINo	1,537	1,537	1,537	1,537	6,148

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Findings: estimation of false negatives

Table 13: Results of the clerical review for false negatives

	Total record pairs	Record pairs of which contain NINo
Sample reviewed	7,115	6,148
True negatives (TN) in sample	4,958	4,209
False negatives (FN) in sample	2,157	1,939

Source: 2021 Census to Demographic Index linked data from the Office for National Statistics

Table 14: Estimated error grid of the linkage between the 2021 Census and the Demographic Index, overall level

	Records matched	Records not matched (residuals)
Links	True positive (TP) ~ 56,867,044	False negative (FN) ~ 81,051
Non-links	False positive (FP) ~ 76,290	True negative (TN) ~ 378,929

Source: 2021 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 (99.86%) is the percentage of true matches that were classified as links. Note: the recall estimate is only for the 2021 Census-NINo links.

5 . 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, 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=147,636) or NINos (N=217,439) have been flagged in the linked data.

There are also cases where one master key or NINo has been linked to multiple census IDs. This was largely because of students in the census who were enumerated at both their home address and term-time address in two different census IDs, but in some cases could be indicative of an error in clustering. This may lead to two census IDs being assigned one master key (N=814,790) or NINo (N=817,306). These cases have also been flagged in the linked data.

Linkage method

It is important to be aware that although deterministic and probabilistic linkage methods allow for error in the matching variables (and can therefore yield more matches than exact matching alone), there is potential for false positives (incorrect matches) to be introduced.

6 . Related links

2011 Census linkage to DWP master key and encrypted NINo
Methodology | Released 6 December 2024
Linkage methodology and quality information for 2011 Census linkage to DWP (Department for Work and Pensions) master key and encrypted NINo (National Insurance number).

7 . Cite this methodology

Office for National Statistics (ONS), released 6 December 2024, ONS website, methodology, 2021 Census linkage to DWP master key and encrypted NINo.