Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021

Linkage methodology and quality information for Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021, to facilitate research into estimating the size and structure of the adult social care workforce.

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

Release date:
5 March 2025

Next release:
To be announced

Table of contents

1 . Main points

- To facilitate research into estimating the size and composition of the adult social care (ASC) workforce, Care Quality Commission (CQC) data were linked to the Inter-Departmental Business Register (IDBR) and Census 2021.
- Deterministic linkage was conducted using match-keys in two stages: first, linking CQC to IDBR data, then linking Census 2021 to the CQC-IDBR linked dataset.
- The linkages resulted in 9.38% of census residents (21.03% of working census residents) linking to a corresponding CQC and/or IDBR record.
- Further "uncertain" links obtained through looser match-keys increased the census link rate by 5.34 percentage points; these were used to provide more confidence in non-links.
- Overall precision of the census links to CQC-IDBR was calculated as 96.09% (excluding uncertain links); the precision of the "uncertain" links alone was calculated as 37.59%.
- Overall recall of the census links to CQC-IDBR was estimated as 99.97%.
- Analysis of CQC residuals (records not linked to census) indicated that providers for domiciliary care and supported living services were more prevalent, suggesting under-representation in the linked data.

2 . Background to the linkage

The Department of Health and Social Care (DHSC) commissioned the Office for National Statistics (ONS) Data and Analysis for Social Care and Health (DASCH) team to lead a project for exploring the feasibility of using census data to estimate the size and composition of the adult social care (ASC) workforce. To do this, DASCH requested the linkage of Care Quality Commission (CQC), Inter-Departmental Business Register (IDBR) and Census 2021 data from the ONS Bespoke Linkage team. This article describes the methodology used to link the datasets as well as quality information and limitations of the linkage.

The purpose of the linkage was to create a labelled dataset to train and test a series of models to identify and classify the ASC workforce. The models were trained on free text (write-in responses) from the census. Information from the CQC (service type and service users) and IDBR (industry codes) was used to label the dataset for training, as an independent source for the census respondent's organisation within the ASC sector (for example, residential or domiciliary care). The modelling used the CQC-IDBR-census links, CQC-census only links and IDBR-census only links detailed in this article.

3 . Data sources and pre-processing

Census 2021

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. The latest census took place on 21 March 2021.

For this project, the variables of interest were around individuals' employment. These included employer, workplace address and workplace postcode.

Care Quality Commission directory

The Care Quality Commission (CQC) directory contains a list of every care home, hospital, GP, dentist and home care agency in England. It also contains other types of service like ambulances, prison care services and hospices supplied by the CQC.

For linkage, the dataset closest to "Census Day" was selected (January to March 2021).

Inter-Departmental Business Register

The Inter-Departmental Business Register (IDBR) is a comprehensive list of UK businesses used by government for statistical purposes. It contains Value Added Tax (VAT) and Pay As You Earn (PAYE) data from HM Revenue and Customs (HMRC), as well as additional information from Companies House, Dun and Bradstreet, and Office for National Statistics (ONS) business surveys.

The datasets closest to "Census Day" were selected, then engineered into local level format, so that individual care homes could be identified in the data. Rows contained local unit, address data, and where applicable, Company Reference Number (CRN), VAT number and/or PAYE number.

Reference Data Management Framework

The Reference Data Management Framework (RDMF) is a tool produced by the ONS, which allows the ONS to link data consistently and securely. The RDMF consists of five "indexes", including information on locations, people and businesses.

The Business Index (BI) is one of the indexes and comprises a list of all UK businesses, containing information such as VAT and PAYE reference numbers from HMRC and CRN from Companies House. The table also includes a unique identifier (Business Index ID).

Another of the indexes is the Demographic Index (DI), which comprises people 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 unique identifier (Demographic Index ID).

The BI and DI were both used as part of the methodology to facilitate the linkage of census data to IDBR data, through the use of business variables CRN, VAT and PAYE (see Pre-processing). The following tables were used:

- Census-DI lookup table; a table which was created by linking Census 2021 to the DI using deterministic and probabilistic linkage of personal identifiers; the output contains census resident ID and Demographic Index ID
- Cross-Index Association (XIA) lookup table; a table which provides a lookup between the DI and the BI through HMRC PAYE Real Time Information (RTI) data, which is part of both DI and BI build; the output contains Demographic Index ID and Business Index ID

Pre-processing

The CQC, IDBR and census data were standardised and cleaned so that the variables used for linkage were in a consistent format. The following steps were taken:

- checking for invalid ID numbers and incorrect postcode length
- converting linkage variables to upper case, for consistency
- removing special characters and extra white space from linkage variables
- deriving postcode sector, district and area variables, from postcode
- deriving house name and number, and road names, where possible
- replacing entries of "rd" and "ave" with "road" and "avenue", where possible
- creating additional address variables, by concatenating address lines together
- creating additional variables for company names, which omitted "limited" and "ltd"
- adding relevant suffixes to column labels, to represent their corresponding sources

To facilitate linkage of census to the IDBR, the CRN, VAT and PAYE references for individuals' main workplace were joined onto the census data. This was done by first joining the census data to the DI using the census-DI lookup and second joining to the BI using the XIA lookup table. A summary of joined census records is shown in Table 1.

Table 1: Summary of Census 2021 IDs joined through pre-processing steps

	Number of Census IDs	Proportion of Census IDs
Census IDs joined to a Demographic Index ID	58,163,733	96.24%
Census IDs joined to a Business Index ID	35,105,808	58.09%
Census IDs joined to a CRN	30,349,061	50.22%
Census IDs joined to a VAT number	30,216,719	50.00%
Census IDs joined to a PAYE number	35,105,796	58.09%

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

4 . Deterministic linkage

Deterministic linkage is a rule-based linkage method that takes the form of a series of match-keys. Match-keys are a list of criteria on which records must agree to be declared a match. To account for expected errors in the data, the criteria is loosened on different linkage variables. For example, using partial agreement, which allows us to link record pairs that contain inconsistencies and/or error. Match-keys are applied hierarchically, starting at the strictest matching criteria and gradually become looser.

This project took a two-part approach: first, match-keys were used to link the Care Quality Commission (CQC) and Inter-Departmental Business Register (IDBR) data; then, match-keys were used to link Census 2021 data to IDBR and/or CQC records.

Linkage of the CQC to IDBR data

Thirty deterministic match-keys were developed, to link the CQC to the IDBR data (detailed table of match-key criteria available on request). The linkage variables were:

- Company Reference Number (CRN), location, provider, brand name, postcode, and address from the CQC directory

- CRN, Companies House name, name lines 1 to 3, postcode and address from the IDBR

The deterministic linkage resulted in 31,931 linked records pairs, with 17,441 unique CQC location IDs linked - a link rate of 34.81% of CQC location IDs. See Considerations for further detail on why this link rate is low.

Some CQC records linked to multiple IDBR entries. This is because some care providers have multiple entries in IDBR, because of duplication (for example, IDBR entries have the same CRN, name and geography, but slight variations in address naming conventions). It is also possible that these one-to-many links occurred because of linkage error.

Linkage of Census 2021 to the CQC-IDBR data

The linkage of Census 2021 to the CQC and IDBR data was conducted in two stages, both consisting of sets of deterministic match-keys. The linkage variables were:

- CRN, location, provider, brand name, postcode, Unique Property Reference Number (UPRN), and address from the CQC directory
- CRN, VAT, PAYE, Companies House name, name lines 1 to 3, postcode, and address from the IDBR
- CRN, VAT, PAYE, employer name, employer postcode, UPRN, and address from Census 2021 (joined to the Business Index)

The first stage consisted of a series of 89 strict deterministic match-keys (detailed table of match-key criteria available on request), some of which were designed to link CQC residuals directly to the census.

This first stage resulted in 11,385,108 linked record pairs, of which 5,671,644 were unique census resident IDs. The linkage yielded a match rate of 9.38% of census residents (N = 60,433,979), or 21.03% of working census residents (N = 5,669,579 out of 26,960,213). Note, working residents are defined as those who said they were working in the last week in the census. See Considerations for further detail on why this link rate is low.

For the second stage, a further seven deterministic match-keys were created with a looser criterion, to gather "uncertain" links (detailed table of criteria available on request). This second stage resulted in 33,019,333 additional linked record pairs, of which 3,230,171 were unique census resident IDs. Although these "uncertain" links introduced more scope for linkage error, it meant that there was less scope for missed links (false negatives) and more certainty that the unlinked records were correctly unlinked.

Table 2: Summary of links between Census 2021 and CQC-IDBR data, by match-key

Stage	Match-key number	Number of links achieved	Number of Census resident IDs linked	Census link rate
1	1 to 89	11,385,108	5,671,644	9.38%
2	90 to 96	33,019,333	3,230,171	5.34%
Total	1 to 96	44,404,441	8,901,815	14.73%

Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Notes

1. Discrepancies between sum of census link rates occur, as a result of rounding.

As expected, there were cases where multiple census resident IDs linked to one IDBR local unit (that is, where the unit had multiple employees). Because of the presence of multiple IDBR local units, there were also cases where one census resident ID linked to multiple IDBR local units. These cases were prevalent where match-key criterion loosened. Types of clusters are summarised in Table 3.

Table 3: Summary of links between Census 2021 and CQC-IDBR data

Type of link	Number of rows
One to one	80,194
One Census to many CQC/IDBR IDs	1,368,603
Many Census to one CQC/IDBR IDs	5,055,337
Many to many	37,900,307

Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Linkage output

The output consisted of a lookup table, which contained unique IDs for CQC, IDBR and Census 2021, alongside match-key number, which analysts could use to keep or remove different match-key groups.

A summary of the types of links provided to analysts can be found in Table 4. Note, the adult social care (ASC) modelling did not use the CQC-IDBR only links.

Table 4: Summary of different types of links between Census 2021, CQC and IDBR data

Row contains CQC ID	Row contains IDBR ID	Row contains 2021 Census ID	Count (N)
Yes	**Yes**		2,838
Yes		**Yes**	124,039
	Yes	**Yes**	42,989,795
Yes	**Yes**	**Yes**	1,290,607

Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Disclosure risk

The linkage did not use personal identifiers; however, it is important to note that individuals could input disclosive information in the free text boxes when completing their employer information. To minimise risk of disclosure, the output contained only the variables required by analysts (see Linkage output).

5 . Exploration of residual (unlinked) records

Over- and under-representation of population groups in the linked data could lead to bias in future data analysis, especially if the analysis focuses on particular groups. Therefore, it is important to understand the characteristics of the residual (unlinked) records, so that analysts can account for potential bias. The following quality assurance (QA) was carried out once linkage was completed:

- exploration of residual (unlinked) census records
- exploration of residual (unlinked) CQC records

Exploration of residual (unlinked) census records

There are two main reasons why a census record may not link to a Care Quality Commission (CQC) and/or Inter-Departmental Business Register (IDBR) record.

The first reason is that the census record does not have a corresponding workplace in the CQC or IDBR data; therefore no link exists. This may occur if the workplace recorded in census for a person is outside of the coverage of the CQC and/or IDBR data. For example, if there are no employees paid more than £123 a week, the business may not be Pay As You Earn (PAYE) registered, and not be in the IDBR. Further, if an organisation does not provide regulated health or social care, they would not appear in the CQC data.

The second reason is linkage error; where we were unable to identify the links from the linkage methodology used. Likely causes of linkage error include:

- missing identifiers for Company Reference Number (CRN), Value Added Tax (VAT) and PAYE references on census, where linkage to the Office for National Statistics (ONS) Business Index (BI) was not possible

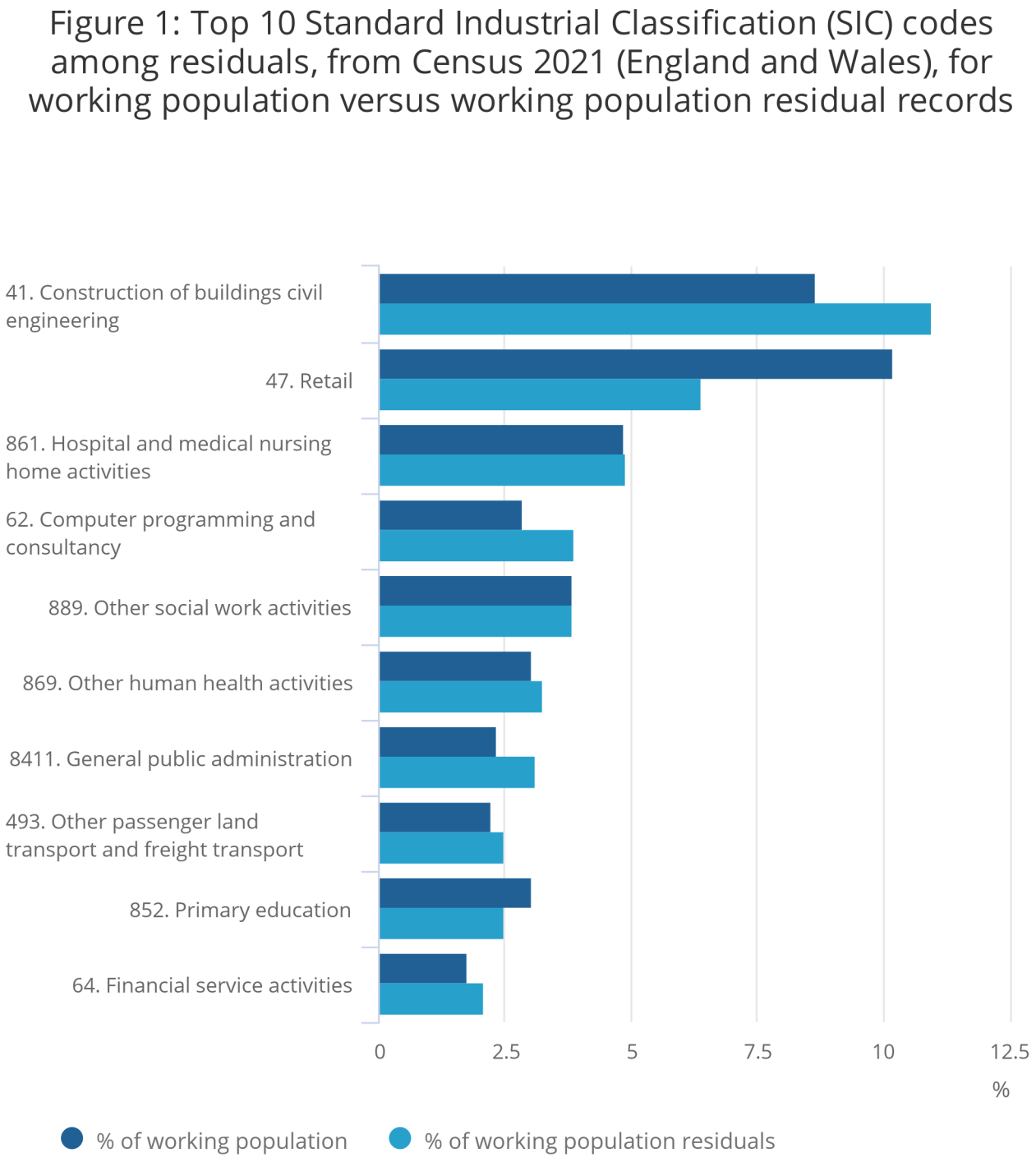
- challenges linking to the free text census variable on employer, for example, if the information provided is unclear, inaccurate or differs significantly with the corresponding CQC-IDBR workplace

- challenges linking to the correct local unit, or to the correct employer within an address (such as a hospital)

While it is difficult to disentangle coverage differences with linkage error, it is useful to examine the unlinked census residuals to understand whether linkage failure is random or related to characteristics of those in the data. To do this, the characteristics of residual census records were compared with characteristics of the census working population.

Note: Both groups were filtered to the working population (residual census records N = 18,076,239, census working population N = 26,960,213).

Figure 1: Top 10 Standard Industrial Classification (SIC) codes among residuals, from Census 2021 (England and Wales), for working population versus working population residual records



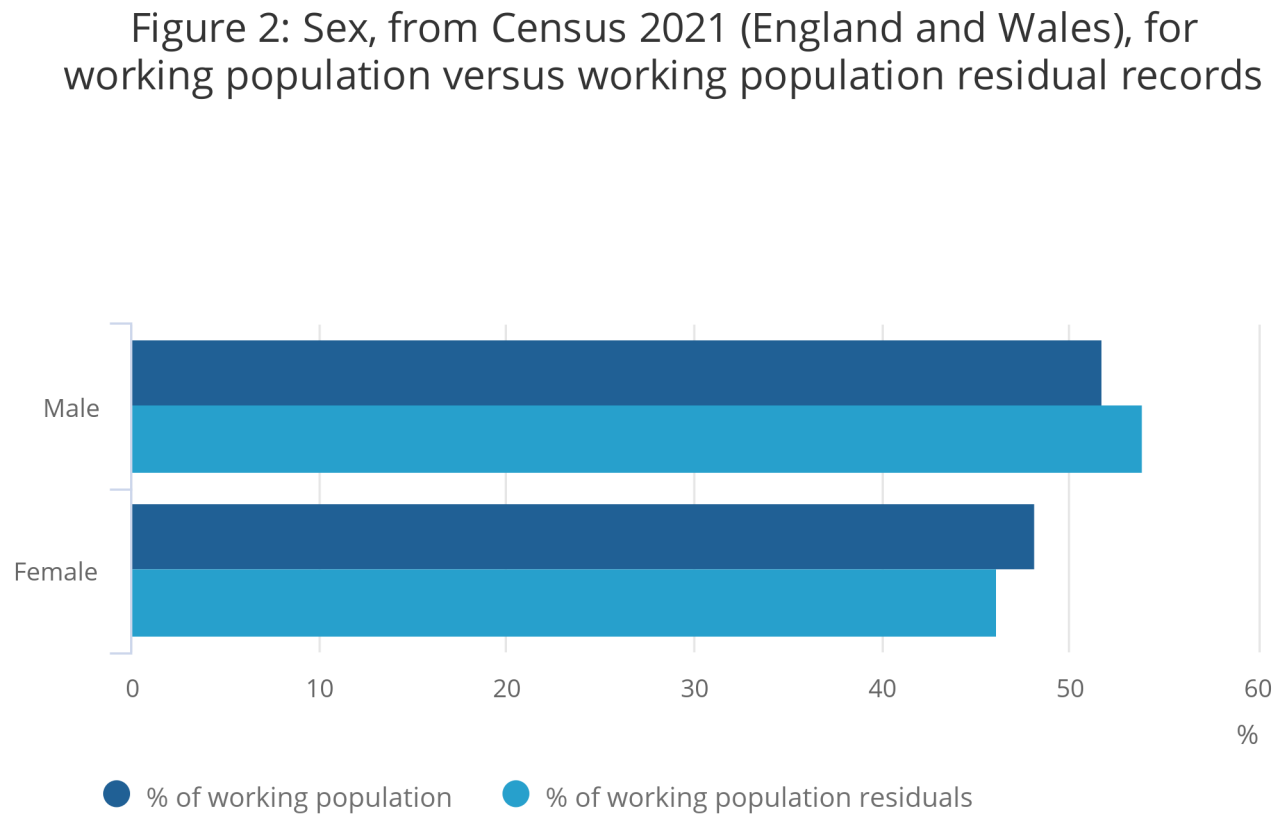
Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Residual census records were over-represented for those in the construction and civil engineering industry. Conversely, individuals working in retail were under-represented in the residuals versus the working population, indicating that they were over-represented in the linked data.

When looking specifically at health and social care-related occupations, there were minimal differences between comparison groups.

Further to this, there were minimal differences seen between comparison groups for Standard Occupational Classification (SOC) codes. The most notable discrepancies, however, were among retail workers, warehouse operatives and care workers, who were slightly under-represented in the residuals, indicating that they were slightly over-represented in the linked data.

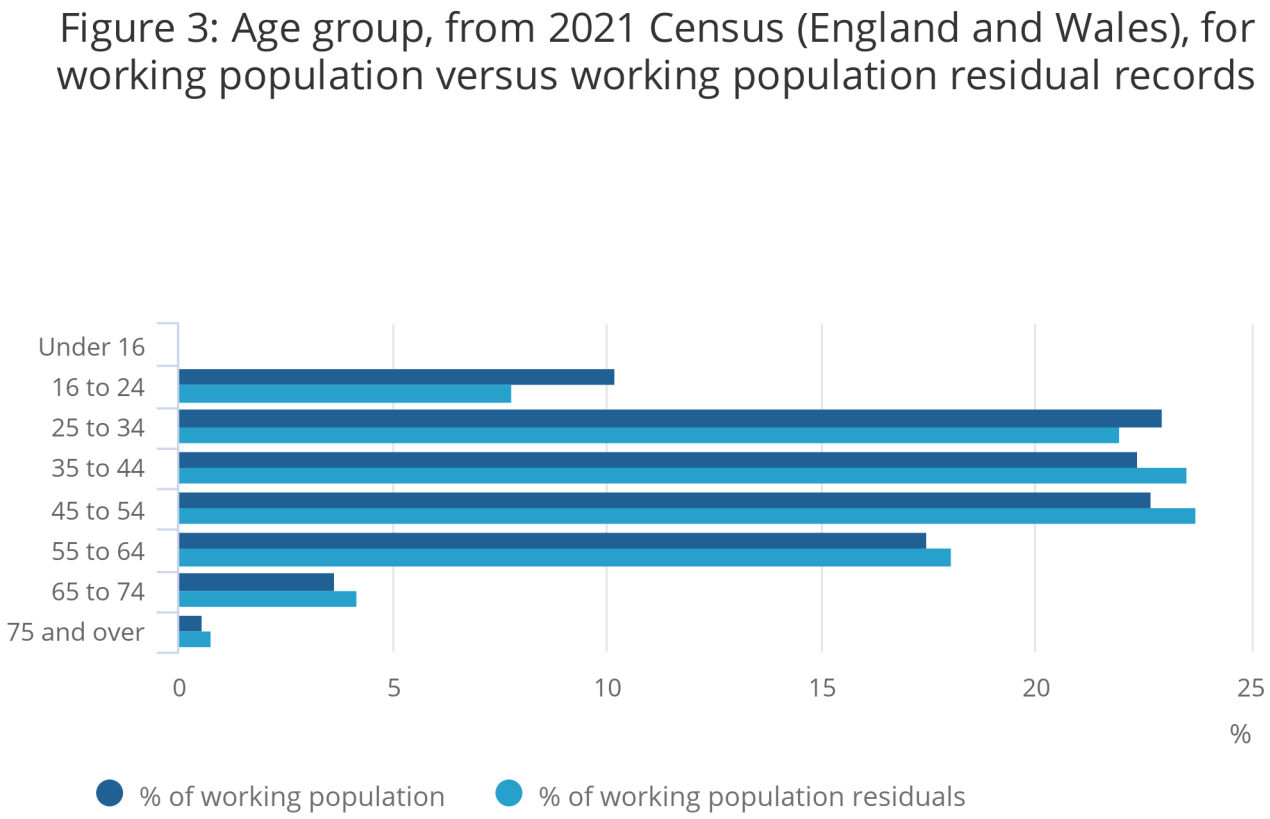
Figure 2: Sex, from Census 2021 (England and Wales), for working population versus working population residual records



Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Males were marginally over-represented in the residual census data versus the working population, indicating that they were slightly under-represented in the linked data.

Figure 3: Age group, from 2021 Census (England and Wales), for working population versus working population residual records



Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Individuals aged over 35 years were over-represented in the census residual data versus the working population, indicating that they were under-represented in the linked data.

Aside from the small differences in age group and some industries, the demographic characteristics (sex, ethnicity, country of birth, main language and disability status) of the census residuals versus working population were relatively similar, which suggests that the personal characteristics of individuals did not impact the likelihood to link to the CQC and IDBR data.

Exploration of residual (unlinked) CQC records

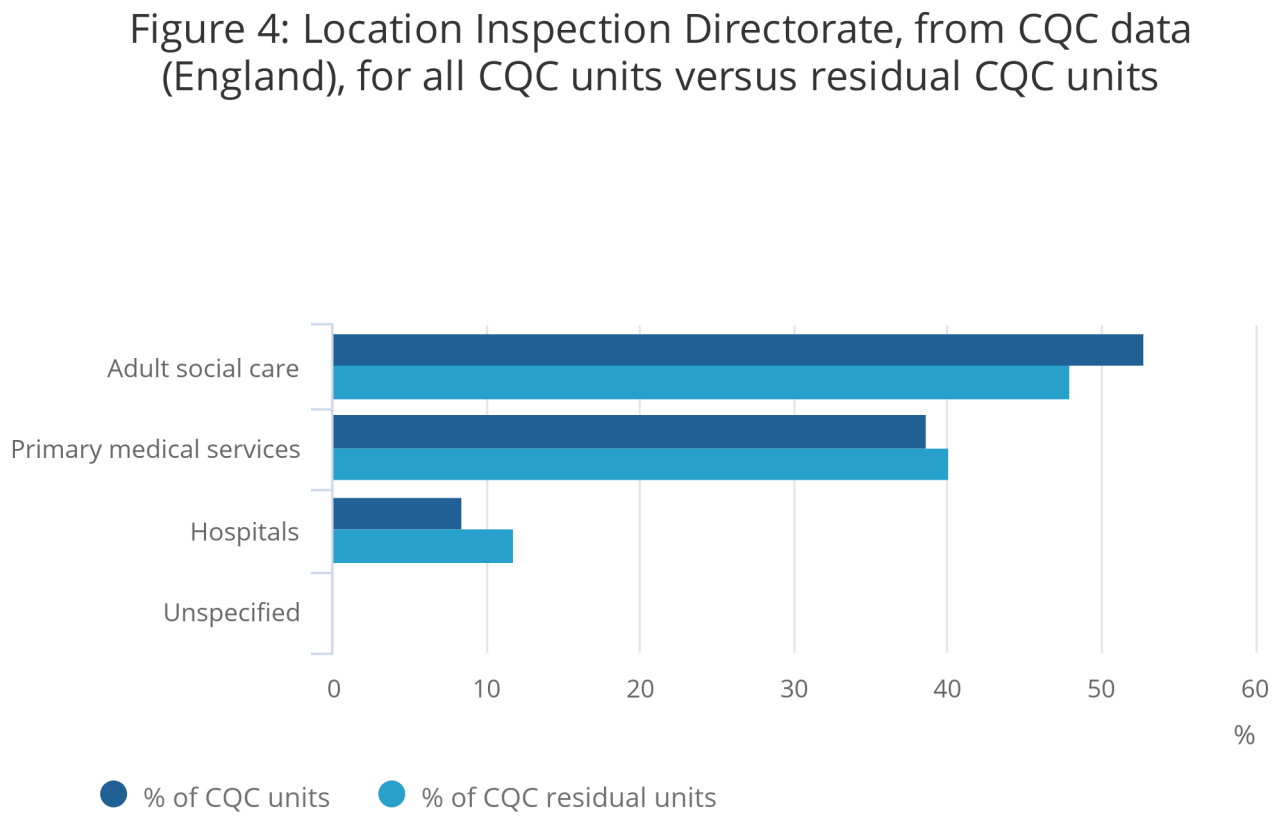
This section examines the CQC records that have not linked to a census record (CQC residuals). Like census residuals, there are two main reasons why a CQC record may not link to the census data: coverage differences and linkage error. An example of coverage differences includes where a provider closes but remains registered on CQC. Likely causes of linkage error include:

- failure to link to IDBR, making CQC linkage to census reliant on strict workplace name and geography
- challenges linking to the free text census variable on employer, for example, if the information provided is unclear, inaccurate or differs significantly with corresponding CQC variables

While it is difficult to disentangle coverage differences with linkage error, it is still useful to examine the CQC residuals to understand whether linkage failure is random or related to characteristics of the data.

To understand the potential impact of residual CQC records, the characteristics of this group were compared with the characteristics of the total CQC dataset (residual CQC units which did not link to the census N = 18,791, total CQC units N = 50,108).

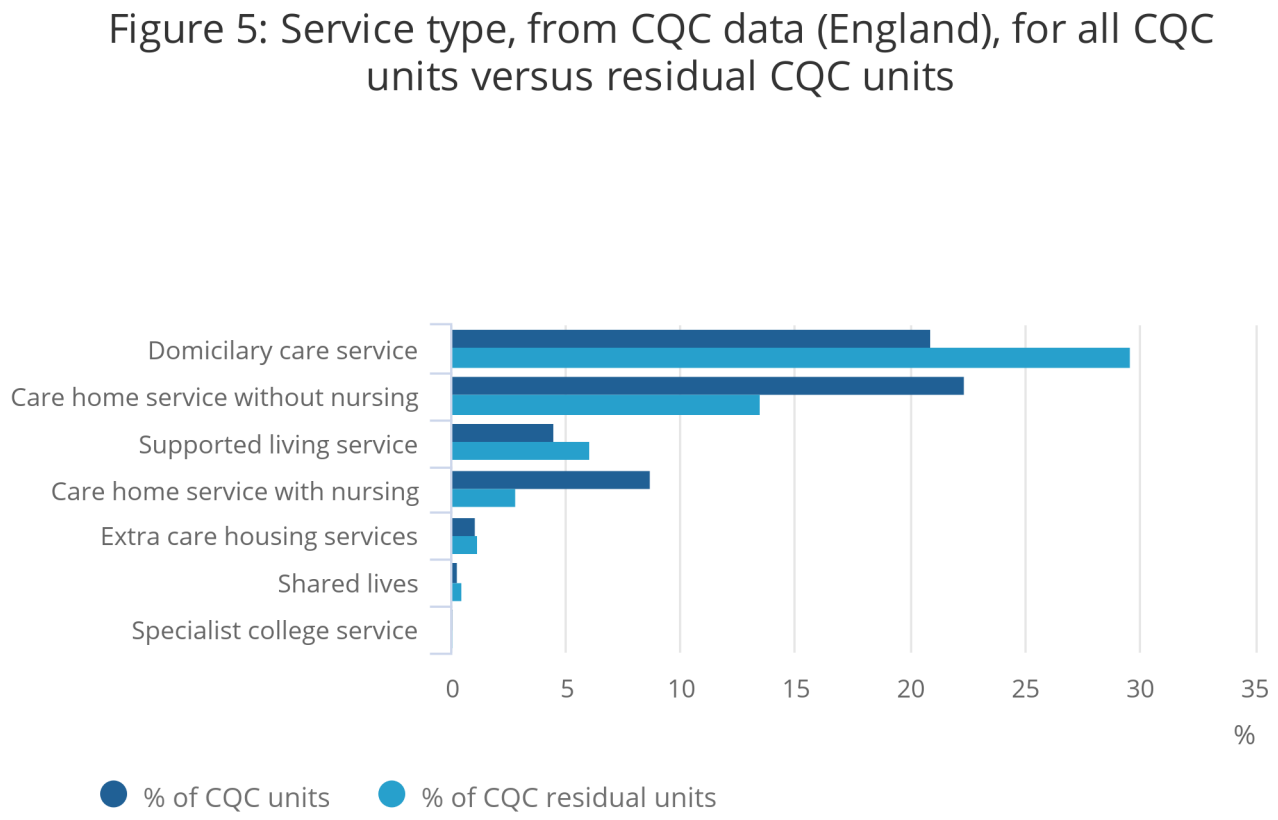
Figure 4: Location Inspection Directorate, from CQC data (England), for all CQC units versus residual CQC units



Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Residual CQC records saw over-representation among hospitals (12%, versus 8% of total CQC units). Manual exploration of the data indicated that it was more challenging to link hospitals at the correct local unit level, since hospitals are made up of different areas and appoint sub-contractors.

Figure 5: Service type, from CQC data (England), for all CQC units versus residual CQC units

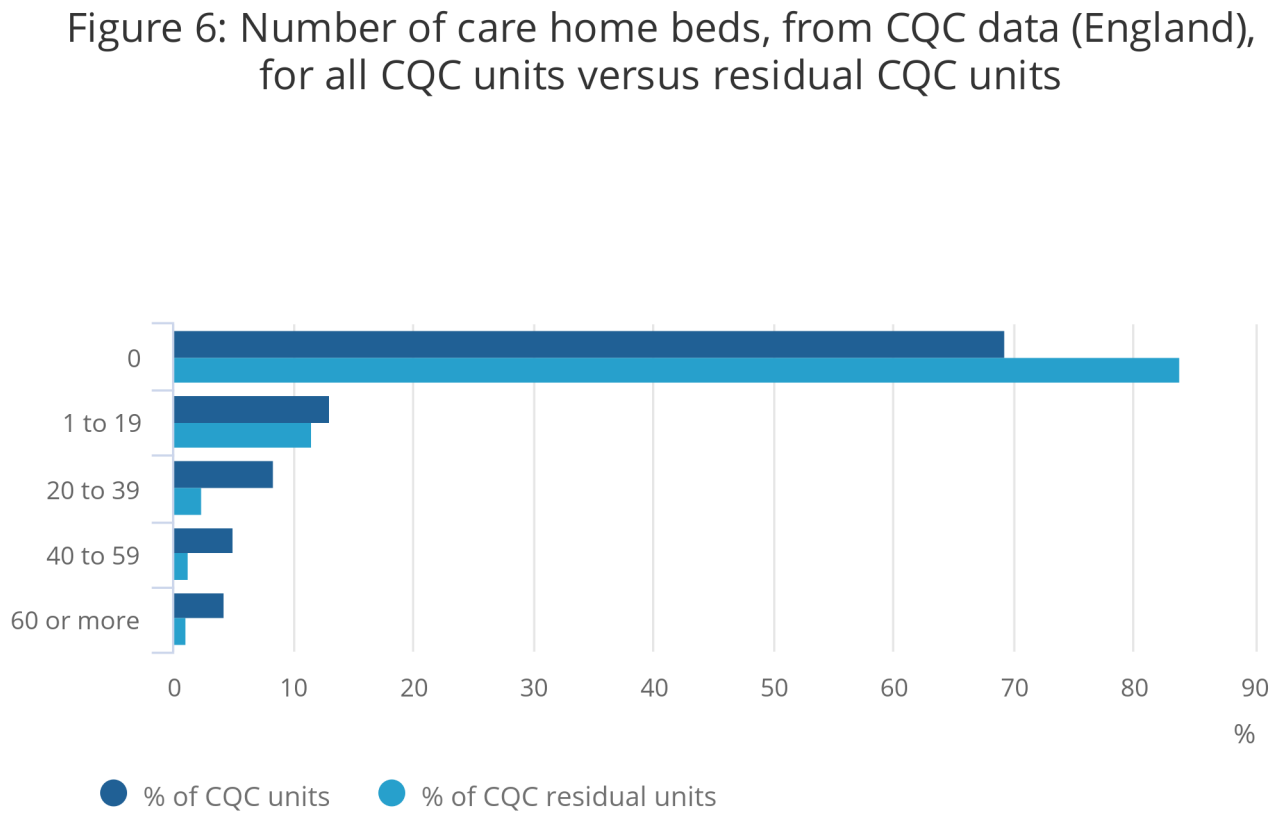


Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Residual CQC units were over-represented among domiciliary care and supported living services. This could be because of churn in the sector, for example, providers closing but remaining registered with CQC, and providers changing names; all of which would make it more difficult to link CQC to census.

Conversely, care home services (with and without nursing) were seen to be easier to link to the census.

Figure 6: Number of care home beds, from CQC data (England), for all CQC units versus residual CQC units



Source: Census 2021 to CQC-IDBR linked data from the Office for National Statistics

Residual CQC records saw over-representation among units with 0 care home beds (84%, versus 69% of total CQC units); 0 care home beds are likely found within supported living and domiciliary care services.

6 . Clerical review

The standard approach to estimate error in the linked data is to clerically review (manually check) a sample of links and a sample of rejected record pairs, to estimate the number of true positives (correct links), false positives (incorrect links) and false negatives (missed matches). In linkage, there is a trade-off between two types of error: precision and recall.

Precision is a measure of the accuracy of the matches that have been made:

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

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

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

Clerical review was carried out in three stages for this linkage.

- first, false positive analysis was conducted on a sample of the deterministically linked records between Care Quality Commission (CQC) and Inter-Departmental Business Register (IDBR) to estimate the precision of this linkage

- a second false positive analysis was conducted on a sample of linked census records to CQC and/or IDBR records to estimate the precision of the linkage to census

- finally, false negative analysis was conducted on a sample of unlinked census records with candidate IDBR and/or CQC links to estimate recall

Clerical review was performed on linked and candidate (unlinked) clusters of records from census, CQC and/or IDBR. Clusters refer to where an ID from one source has been linked to one or more IDs from another source. While some clusters contained one record from the CQC, census and the IDBR, there were some cases of larger clusters. For example, a census record being linked to multiple CQC-IDBR units. Consequently, the results of clerical review were split into total, partial and non-agreement within clusters.

Clerical review for false positives: CQC linked to IDBR

There were 1,639 clusters reviewed from match-key 1, which was calculated using the Statulator tool using a confidence level of 95%, an expected proportion (of false positives) of 0.05, and a relative proportion of 0.2. All linked records from match-keys 2 to 30 were reviewed, meaning that a total of 3,039 clusters were taken for the clerical review for false positives.

The results of the clerical review were categorised as follows:

- total agreement: all records in the cluster group were the same local unit

- partial agreement: the cluster contained some agreement, but not all records were the same local unit

- non-agreement: no records in the cluster belonged to the same local unit

Table 5: Estimated precision in the CQC-IDBR dataset

	Estimated proportion of total agreement (precision)	Estimated proportion of partial agreement	Estimated proportion of non-agreement
Total, unweighted	89.14%	3.52%	7.34%
Total, weighted	97.70%	1.02%	1.28%

Source: CQC to IDBR linked data from the Office for National Statistics

There was a high level of agreement across clusters reviewed. After weighting the data, it was estimated that 97.70% of clusters in the linked CQC-IDBR dataset contained total agreement (precision estimate); 1.02% of clusters were estimated to contain some agreement and 1.28% of clusters were estimated to be incorrectly linked.

Precision was estimated for every match-key. Match-keys 25 and 26 yielded lower levels of agreement (12.58%, 17.02%). Despite matching on name and geography, these match-keys contained cases where a CQC unit had been linked to a different local unit within the same large organisation, or a CQC unit had been linked to the large organisation itself (rather than the local unit).

Lower performing match-keys may be removed from the dataset for analysis, as appropriate, to improve confidence in links. Tables detailing match-key criteria, as well as precision estimates by match-key are available on request.

Clerical review for false positives: Census 2021 linked to CQC and IDBR

A sample of 17,502 clusters was taken for clerical review of the census linked to CQC-IDBR records. Approximately 200 clusters from each match-key were reviewed, unless the match-key produced fewer than 200 links (in which case, all links were reviewed). Although small samples would yield a lower level of confidence in our estimates, getting a precision estimate for each match-key was prioritised. Clerical reviewers were instructed to only review the links to census rather than review the links between CQC and IDBR.

The results of the clerical review are categorised as follows:

- total agreement: census records were the same workplace (local unit) as the CQC and/or IDBR records
- partial agreement: census records contained some agreement to the CQC and/or IDBR records, but not all records were the same workplace (local unit)
- non-agreement: no census records belonged to the same workplace (local unit) as the CQC and/or IDBR records

Table 6: Results from clerical review for false positives in Census 2021 to CQC-IDBR linked dataset

Match-key number	Match-key description	Number of linked clusters in Census 2021 -CQC-IDBR dataset	Estimated proportion of total agreement (precision), weighted	Estimated proportion of partial agreement, weighted	Estimated proportion of non-agreement, weighted
1 to 51	**Used only IDBR variables for linkage**	2,612,272	98.95%	1.05%	0.00%
52 to 77	**Introduced use of CQC variables for linkage**	196,281	99.04%	0.47%	0.49%
78 to 89	**Used only IDBR variables for linkage, with looser criteria (e.g. no employer name match needed, postcode sector)**	2,863,091	93.28%	1.55%	5.17%
90 to 96	**Loosest criteria, to produce “uncertain links”**	3,230,171	37.59%	1.80%	60.61%
1 to 89	**Total excluding “uncertain” links**	5,671,644	96.09%	1.28%	2.63%
1 to 96	**Total**	8,901,815	74.86%	1.47%	23.67%

Source: 2021 Census to CQC-IDBR linked data from the Office for National Statistics

For the "certain" links (match-keys 1 to 89), it was estimated that 96.09% of census records that linked to CQC and/or IDBR records contained total agreement, and 74.86% of all clusters (including "uncertain" links) contained total agreement.

As expected, the "uncertain" links introduced more linkage error. However, they were used to give more certainty that the unlinked records were correctly unlinked. Since precision was estimated for each match-key, the analysts can assess which links to keep or remove for their project. Tables detailing precision by match-key are available on request.

Clerical review for false negatives: Census 2021 linked to CQC and IDBR

To obtain the unlinked clusters, a series of loose match-keys were run on the census residuals (a table detailing deterministic match-key criteria is available on request). A sample of 6,671 clusters was taken for the clerical review for false negatives.

Sample sizes per match-key were calculated using the Statulator tool using a confidence level of 95%, an expected proportion (of false negatives) of 0.1, and a relative proportion of 0.2. Reviewers were instructed to assess whether the residual census record belonged to the same workplace (local unit) as the CQC and/or IDBR records.

Table 7: Results from clerical review for false negatives in Census 2021 to CQC-IDBR linked dataset

Match-key number	Sample of clusters taken for clerical review	Number of false negatives found in clerical review	Estimated number of false negatives in full dataset
1	836	4	61
2	845	1	23
3	71	0	0
4	388	0	0
5	754	1	4
6	743	4	16
7	267	0	0
8	1	0	0
9	129	0	0
10	855	4	187
11	823	2	20
12	1	0	0
13	1	0	0
14	106	0	0
15	851	37	1,359
Total	6,671	53	1,670

Source: 2021 Census to CQC-IDBR linked data from the Office for National Statistics

Based on the clerical review results, the estimated number of false negatives in the full dataset was 1,670. Therefore, recall is estimated as 99.97%.

7 . Considerations

Census 2021

A limitation of the census data is that the employer information is a free text question, which means that we have limited control over incorrect or insufficient information being inputted.

Further, respondents are asked to input details of their main job, or last main job (one's main job is which they usually work the most hours). Therefore, if a person's job in adult social care is a secondary job, it may not be captured in the data.

Linkage methodology

The link rates across the linkage stages were relatively low (34.81% of Care Quality Commission (CQC) location IDs, 9.38% to 14.73% of census). There are two main reasons for this.

Firstly, the methodology was designed to link businesses at a local unit level and to maintain this granularity; the linkage criteria were relatively strict. Therefore, it is possible that some links between sources were missed.

Secondly, there are reasons why matches may not exist between the sources. Examples include where the care provider or census workplace is outside of the coverage of the Inter-Departmental Business Register (IDBR) or where a care provider closes but remains registered on CQC. However, the assessment of recall provided confidence that a high proportion of the matches between census and CQC-IDBR were found using the linkage methodology.

The use of the lookup tables (as part of preprocessing) increases scope for error. For example, if incorrect links (false positives) exist in a lookup table, then they could create compounding error in subsequent linkages.

Quality review

When creating samples for the clerical review for false negatives, the loose match-key criteria combined with large size of the data files impacted the speed at which samples could be drawn. Project timing constraints meant that smaller samples of the census residual records were run through the loose match-keys. As a result, the actual proportion of links per match-key had to be estimated. For example, where 20% of residents were sampled from a match-key, the actual number of links for that match-key was estimated by multiplying the count by 5.

Further to this, there could be missed links (false negatives) in the data, which were not brought together through the looser match-keys and therefore did not have an opportunity to be sampled from. Therefore, the recall estimate may be an overestimate.

Another limitation of the quality review is that clerical review is used as a proxy for the true match status. Therefore, the estimates for precision and recall do not factor in the uncertainty resulting from subjective human decision making, which is constrained by the information available on each of the data sources.

8 . Cite this methodology

Office for National Statistics (ONS), released 5 March 2025, ONS website, methodology, Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021.