Census 2021 to Personal Demographics Service linkage report

Methods used to link Census 2021 to the Personal Demographics Service (PDS) using deterministic and probabilistic methods.

Contact:
Rhiannon Brook, Daniel Cheung and Sarah Cummins
linkage.hub@ons.gov.uk

Release date:
23 August 2023

Next release:
To be announced

Notice

1 November 2023

We have corrected an error in 5. Quality assurance, under the heading Precision and recall. The previous version read “For large sample sizes with no errors, the lower bound was adjusted to 0 and the upper bound was calculated by dividing the sample size by 3”. It should have read “For large sample sizes with no errors, the lower bound was adjusted to 0 and the upper bound was calculated by dividing 3 by the sample size”. This happened because of human error.

Table of contents

1 . Main points

- Census 2021 was linked to the Personal Demographics Service (PDS) to update the Public Health Data Asset.
- Initially the datasets were linked deterministically, followed by probabilistic linkage on the residual records.
- The Census 2021 data used are taken from an early stage of processing, prior to editing and imputation, meaning that totals in this report may not correspond exactly to published census totals.
- Deduplication of Census 2021 resulted in the removal of approximately 1 million records, to remove individuals recorded at multiple addresses.
- The linkage rate by census ID level was 95.75%, as 55,105,336 census IDs linked to NHS numbers on the PDS.
- The estimated precision for the linkage was 99.95%; estimated recall was 99.99%.
- The estimated precision for deduplication of the census was 96.93%.
- Bias analysis has also been performed on this linkage, showing that the most underrepresented characteristics are: 20 to 29 years old, male, from an ethnic minority and living in London or Cardiff.

2 . Overview

This report summarises the linkage between the Census 2021 and the Personal Demographics Service (PDS). The linkage was commissioned by the Data and Analysis for Social Care and Health (DASCH) team in the Office for National Statistics (ONS) to assign an NHS number to Census 2021 records, allowing further linkage to other health datasets with the NHS number unique identifier. This lookup has been used to update the Public Health Data Asset (PHDA) 2011 cohort with characteristics from Census 2021.

The Public Health Data Asset 2011 cohort was established by linking data from the 2011 Census to hospitalisation records from the Hospital Episodes Statistics (HES) and death registration records. It was established to investigate inequalities in mortality and morbidity and further our understanding of the social determinants of health. The cohort uniquely combines detailed information on socio-demographic characteristics and health outcomes such as hospitalisation and death for nearly the whole population of England.

The linkage used both deterministic and probabilistic methods, with various mechanisms throughout the linkage to identify and remove duplicate census records. A quality assessment including bias analysis was carried out on the linkage and deduplication, detailed later in the report.

Please note that this linkage was a bespoke product created for DASCH and as such has been designed to their specifications. An agile delivery approach involved several versions of the linked data. Therefore, some of the final methodology was conducted on a previous iteration of the linkage (for example, the initial deduplication). This report discusses the final version of the linked data.

3 . Methods

Census 2021

The census, administered by the Office for National Statistics (ONS), happens every 10 years and gives us a picture of all the people and households in England and Wales. The most recent census day for England and Wales was on Sunday 21 March 2021.

The dataset used contained personal identifiable information, created for the purpose of linkage. The cut of census used for linkage was prior to any editing, imputation, estimation and statistical disclosure control.

The variables used for linkage included:

- forename
- middle name
- surname
- sex
- date of birth
- address
- postcode
- alternative postcode
- postcode from one year ago

PDS

The Personal Demographics Service (PDS) is an NHS dataset that covers England, Wales and the Isle of Man. Records are created for newborns or when a patient contacts an NHS service, primarily by registering with a General Practitioner (GP) practice, but also through accessing A&E or attending hospital. Records are updated when an individual updates their details at a GP or through other interactions with health services.

For the linkage, the PDS data were supplemented with the PDS movers and updates data, which provide information on people who have moved residences.

The variables used for linkage included:

- forename
- middle name
- surname
- sex
- date of birth
- address
- postcode
- postcode updates

Pre-processing

Both the census and PDS information underwent standardisation. Cleaning steps included:

- converting the sex variable to binary integers
- creating derived variables for day, month and year of birth, as well as postcode breakdowns by sector, district and area
- removing non-alphabetical characters, and spaces where appropriate
- splitting forename components into separate variables
- splitting surname components into separate variables
- inclusion of variables with alphabetically reordered forenames and surnames (alphaname)
- inclusion of a nickname lookup

Deterministic linkage

The deduplicated Census 2021 data were deterministically linked with the PDS using 37 matchkeys. Each matchkey consists of a set of rules or criteria that must be met to make a link.

To account for expected errors in the data, the criteria are loosened on different linkage variables. Matchkeys are applied hierarchically, starting at the strictest matching criteria, and becoming looser. Table 1 shows the full list of matchkeys used and the number of links from each one in the final linkage.

Table 1: Matchkey descriptions and number of links per matchkey for the Census 2021 and PDS linkage (England and Wales, 2021)

Matchkey	Description	Number of links
MK1	Concordant first name, middle name, surname, date of birth, postcode and sex.	24,043,512
MK2	Concordant first components of first name, middle name, surname, date of birth, postcode and sex.	121,489
MK3	Concordant first components of first name, middle name, first components of surname, date of birth, and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	1,376,396
MK4	Concordant first components of first name, middle name, first components of surname; concordant date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	50,209
MK5	Concordant alphaname for entire first name and surname, concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	24,080,344
MK6	Levenshtein distance less than 2 on first and second components of first name individually, concordant first components of surname, date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	18,363
MK7	Levenshtein distance less than 2 on first components of surname, concordant first components of first name and middle name. Concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	90,696
MK8	Levenshtein distance less than 2 on first components of both first name and surname, concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	1,869,547
MK9	Soundex on first components of first name, concordant first components of surname, concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	122,799
MK10	Soundex on first components of surname, concordant first component of first name and concordant date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	116,782
MK11	Soundex on first components of both first name and surname, concordant date of birth and sex. Levenshtein distance less than 2 or Jaro-Winkler greater than 0.7 for address. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	6,240
MK12	Jaro-Winkler greater than 0.7 on first components of first name, concordant first component of surname, date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	437,567
MK13	Jaro-Winkler greater than 0.7 on first components of surname, concordant first components of first name and concordant date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	111,903
MK14	Jaro-Winkler greater than 0.9 on first name and Jaro-Winkler greater than 0.7 on surname, concordant date of birth and sex. Allow concordant postcodes between census' s usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	54,782

MK15	Concordant first three characters of first name and surname, concordant middle name and date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	708
MK16	Nicknames on first component of first name, concordant surname and date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	41,316
MK17	Transposed first names between the first three components, concordant surname, date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	63,045
MK18	Transposed first names and surnames among the first three components of each, concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	256,419
MK19	Transposed first names and surnames (first three components of each), between different corresponding variables, concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	1,804
MK20	Concordant first components of both first name and surname, concordant sex, month and year of birth. Discordant day of birth allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	300,883
MK21	Transposed day and month of birth. Concordant first components of both first name and surname. Concordant year of birth and sex. Allow concordant postcodes between census' s usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	40,573
MK22	Discordant year of birth plus or minus 10 years and transposed day and month of birth. Concordant first name, first components of surname and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	4,437
MK23	Discordant year of birth and transposed day and month of birth. Concordant first components of first name and surname and concordant sex. Soundex or missing middlename. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	2,660
MK24	Levenshtein distance date of birth less than 2. Concordant first components of both first name and surname, concordant sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	237,560
MK25	Jaro-Winkler date of birth greater than 0.7 (missing year of birth) and Jaro-Winkler address greater than 0.7. Concordant first components of both first name and surname, concordant sex. Soundex, Jaro-Winkler greater than 0.88, or missing middle name. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	45,444
MK26	Postcode sector transpositions between former, current, alternative on census and current or updated on PDS. Concordant first components of both first name and surname, concordant date of birth and sex.	491,034
MK27	Postcode sector transpositions between former, current, alternative on census and current or updated on PDS. Concordant first components of both first name and surname, concordant date of birth. Discordant sex allowed.	353,690
MK28	Postcode area transpositions between former/current/alternative on census and current /updated on PDS. Concordant first components of both first name and surname, concordant date of birth and sex. Levenshtein distance less than 10 or Jaro-Winkler greater than 0.85 on address.	3,656
MK29	Levenshtein distance any postcodes less than 2. Concordant first components of both first name and surname, concordant date of birth. Discordant sex allowed.	14,454

MK30	Levenshtein distance less than 10 or Jaro-Winkler greater than 0.85 on address. Concordant first components of both first name and surname, concordant date of birth. Discordant sex allowed.	1,674
MK31	Transposition between first name, middle name and/or surname. Concordant date of birth and sex. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	233,694
MK32	Transposition between first name, middle name and/or surname. Concordant date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	1,692
MK33	Discordant sex and day and month of birth allowed. Concordant first components of first name and surname, concordant middle name and year of birth. Levenshtein distance less than 5 or Jaro-Winkler greater than 0.7 on address. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	1,307
MK34	Female only, Jaro-Winkler greater than 0.6 for surname. Concordant first name, middle name and date of birth. Jaro-Winkler greater than 0.7 on address. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	12,355
MK35	Missing year of birth. Concordant first name, middle name and surname. Discordant sex allowed. Jaro-Winkler greater than 0.7 on address. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	4,203
MK36	Soundex first components of first name, concordant surname and date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	2,857
MK37	Jaro-Winkler greater than 0.7 on first component of first name, concordant first component of surname and date of birth. Discordant sex allowed. Allow concordant postcodes between census's usual postcode, alternative postcode, 1 year ago postcode and PDS's standard postcode and updated postcodes.	4,284

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

Probabilistic linkage on residuals

Residual records of the deterministic linkage were linked probabilistically to pick up any additional links that could have been missed.

Splink 2 was used, which is a probabilistic linkage library implementing the Fellegi Sunter method and was developed by the Ministry of Justice. More information appears on their Splink GitHub pages.

Blocking passes utilised year of birth and postcode, alternative postcode and one year ago postcode variables. Values for *m* and *u* probabilities were derived using the online m and u probabilities tool. *m* values (agreement weights) are the probability that a variable agrees on two data sources given that they are a true match, so are a measure of data quality - how accurately the variable is recorded or freedom from error. *u* values (disagreement weights) are the probability that the variable agrees on both data sources given the pair are not a true match, so are a measure of distinguishing power or likelihood of matching by chance. Comparison columns included first name, middle name, surname, alphaname, day of birth, month of birth, address and sex.

Quality assurance and spot-checks were performed to establish a threshold for acceptance. To ensure that links were of high quality, those possessing a match probability of greater than or equal to 0.90 and a match weight of greater than or equal to 14.50 were accepted, while all other links were rejected.

Deduplication

While multiple responses (where there is more than one census return for a person at an address) are removed from the census microdata file, the data still contain duplicate entries where multiple returns have been made for a person at different addresses. Examples include but are not limited to higher education students, children of divorced parents, and those staying at temporary addresses for work reasons and interpersonal relationships. Such individuals may appear more than once at different addresses. For census processing, such cases are dealt with in the estimation process.

Various mechanisms throughout the linkage were used to identify and remove duplicate census records. First, Census 2021 underwent an initial deduplication process. This incorporated a combination of approaches:

- using an initial iteration of the deterministic Census 2021 to PDS lookup, cases of multiple census records that linked to one NHS number were identified as duplicates and the census record with the closest matching geographic information to PDS was retained alongside records that were not identified as duplicates

- isolating the residuals of the initial iteration of the deterministic Census 2021 to PDS lookup, the census records were deduplicated using deterministic matching methods and where duplicates were identified, the most complete records were retained

Additional deduplication was carried out after both the deterministic and probabilistic linkage. For the deterministic linkage, when many census IDs to one NHS number clusters formed, only the record pair with the highest degree of similarity between the Census 2021 and PDS was retained. For the probabilistic linkage, when many census IDs to one NHS number clusters formed, only the record pair with the highest match weight was retained.

For both stages of the linkage, cases of multiple NHS numbers to one census ID were generally found to reference the same person. In the absence of any intelligence that could be used to determine which NHS number was "active", they were all retained.

4 . Results

Deterministic linkage

There were 54,596,983 census IDs yielded following deterministic linkage retaining clustered census IDs by highest similarity to the Personal Demographics Service (PDS). The number of links made by matchkey can be found in Table 1.

Probabilistic linkage on residuals

There were 508,353 census IDs yielded following probabilistic linkage on residuals retaining clustered census IDs by highest match weight. The threshold for acceptance was a match weight of greater than or equal to 14.50, all of which had a match probability of greater than or equal to 0.90.

Final output

Results of the deterministic and probabilistic records were then appended, yielding 55,105,336 census IDs linked to the PDS. This represents a 95.75% linkage rate by census ID. Table 2 shows a summary of census ID counts following each stage.

Table 2: Summary of census ID counts following each stage of the process for the linkage of Census 2021 to PDS (England and Wales, 2021)

Description	ID count	Duplicates removed
Census 2021	58,623,712	–
Initial deduplication	58,420,462	203,250
Deterministic links	55,465,688	–
Resolving clusters by highest similarity with PDS	54,596,983	868,705
Census 2021 residuals	2,954,774	–
Probabilistic links	510,358	–
Resolving clustered IDs by highest match weight	508,353	2,005
Grand deduplicated total	57,553,762	1,073,960
Total census IDs linked	55,105,336	–

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

Notes

1. Probabilistic links counted here are those possessing a match probability greater than or equal to 0.90 and match weight greater than or equal to 14.50, followed by cluster resolution by match weight.

2. All values shown at ID level and not at record level.

5 . Quality assurance

Clusters analysis by type

Table 3 shows that most of the links (99.96%) consist of one census ID linked to one NHS number.

However, a small percentage (0.04%) consists of conflicting links in the form of one census ID to many NHS numbers. Following spot checking of samples, we did not observe any false positive links present in this kind of cluster. Despite this, it cannot be ruled out that false positives will inevitably exist in the dataset - it is an expected feature of data linkage. In all cases quality assured, one census ID was correctly linked to two NHS numbers - both Personal Demographics Service (PDS) records possessed corresponding and matching information.

Table 3: Summary of census ID counts by cluster type for the linkage of Census 2021 to PDS (England and Wales, 2021)

Cluster type	Count	Percentage composition
One census ID linked to one NHS number	55,081,298	99.96%
Many census IDs linked to one NHS number	0	0.00%
One census ID linked to many NHS numbers	24,038	0.04%
Many census IDs linked to many NHS numbers	0	0.00%
Total	55,105,336	100.00%

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

Notes

1. Percentages shown to 2 decimal places.

2. ID counts may not sum precisely to corresponding percentages because of rounding.

Precision and recall

The standard approach to estimate error in the linked data is to perform clerical review (manual checking) on a sample of links and rejected record pairs, to estimate the number of true positives (correct links), false positives (incorrect links) and false negatives (missed links). 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:

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

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

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

Clerical review was carried out in multiple stages: false positive analysis of deterministic results, false positive analysis of probabilistic results, and false negative analysis of all results (rejected probabilistic matches).

Sample sizes were derived with Statulator using a confidence level of 95.00%, an expected proportion (of false positives) of 0.05, and a margin of error (relative to expected proportion) of 0.577.

For the false positive analysis of deterministic links, 7,447 record pairs were sampled for clerical review, stratified by matchkey. Links resulting from matchkeys 1 and 2 were not included in the clerical review as they were assumed to be true matches. Forty-seven false positives were identified.

For the false positive analysis of probabilistic links, a sample of 4,232 links (with conflicts removed) were taken, stratified by match weight and clerically reviewed. There were 164 false positives identified in the sample.

The false negative analysis was carried out on the rejected matches from the probabilistic linkage (with conflicts removed). Buckets were created based on a combination of match weight and match probability, giving a total of 31 buckets. A sample of 5,909 pairs were clerically reviewed to detect false negatives and 365 false negatives were identified in the sample.

Overall precision and recall for the entire population were derived using total estimated errors. This is the sum of multiplying the error rate with the number of record pairs for each bucket and then aggregating up to the entire population. Estimated numbers of true positives, false positives and false negatives calculated in this process are shown in Table 4.

Confidence intervals (CIs) estimated for the population were derived using the Agresti-Coull method with a confidence level of 95.00% for each bucket, and then aggregated up to the population level. Upon aggregation of the overall CI of precision or recall, a value of 0 was applied for both the lower and upper bounds for each bucket where no false positives or false negatives were found and where the sample size was small. For large sample sizes with no errors, the lower bound was adjusted to 0 and the upper bound was calculated by dividing 3 by the sample size. Precision for the whole linkage was found to be 99.95% CI [98.54%, 99.98%]. Recall for the whole linkage was found to be 99.99% CI [99.92%, 99.99%].

Table 4: Summarised table of precision and recall for the overall linkage, showing record-level counts for the linkage of Census 2021 to PDS (England and Wales, 2021)

Overall linkage outcomes		
Matches or non-matches	**Linked records**	**Residual records**
Matches	Estimated true positives (TP) approximately equal to 55,972,394	Estimated false negatives (FN) approximately equal to 8,257
Non-matches	Estimated false positives (FP) approximately equal to 29,187	

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

False positives analysis of deduplication and cluster resolution

False positive analysis of the deduplication of Census 2021 and the conflict cluster resolution following linkage was carried out to assess quality of these processes. A sample of 2,982 was clerically reviewed and 76 false positives were found. The overall precision for the deduplication and cluster resolution, calculated using total estimated errors and the Agresti-Coull method, was estimated to be 96.93% CI [93.62%, 98.57%].

Bias analysis

Bias analysis is important for telling us about the representativeness of our linked data. It provides a measure of whether linkage failure is random or related to characteristics of those in the data. If there is bias in a linkage process which causes a certain demographic to be more or less likely to appear in the linked data, then any conclusions drawn from the linked data could be incorrect.

Bias was analysed using proportional discrepancy. Positive proportional discrepancies convey overrepresentation of the characteristic evaluated, while negative proportional discrepancy scores convey underrepresentation. Main points from this analysis include:

- females are overrepresented and males are underrepresented in both England and Wales
- 20- to 29-year-olds are the most underrepresented age group, and 70- to 79-year-olds are the most overrepresented age group
- when looking at ethnic groups, the White group is overrepresented, while other groups are underrepresented
- the most underrepresented areas in England are Inner and Outer London, and the most underrepresented area in Wales is Cardiff

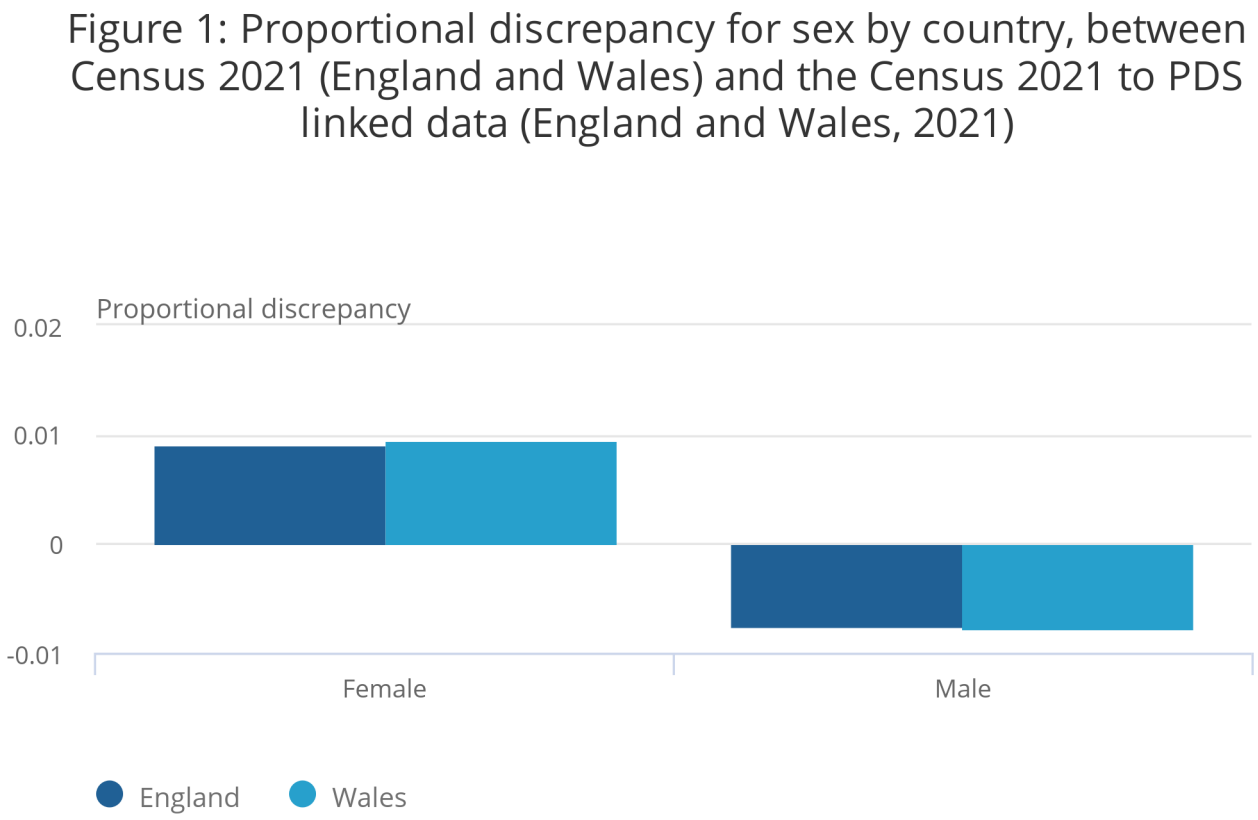
There are several possible reasons for these trends. Young males are typically not well represented in admin data because of lack of interaction with systems. Migrants (also typically young males) may not immediately register with a General Practitioner when immigrating into the country. Another reason for underrepresentation of ethnic minorities is that the linkage process is likely to be better at matching Western names than non-Western names. It is not surprising that all groups investigated are underrepresented in London as the population there are more likely to be young people and/or migrants.

Biases in sex

Figure 1 shows the examination of proportional discrepancy for sex by country, showing that females are overrepresented, while males are underrepresented in both England and Wales (to a similar degree for each country).

The mean proportional discrepancy score of 0.01 for females over England and Wales suggests that the linked data have matched 100.94% of the expected number of matches given the overall match rate. The mean proportional discrepancy score of negative 0.01 for males over England and Wales suggests that the linked data have only matched 99.23% of the expected number of matches given the overall match rate.

Figure 1: Proportional discrepancy for sex by country, between Census 2021 (England and Wales) and the Census 2021 to PDS linked data (England and Wales, 2021)



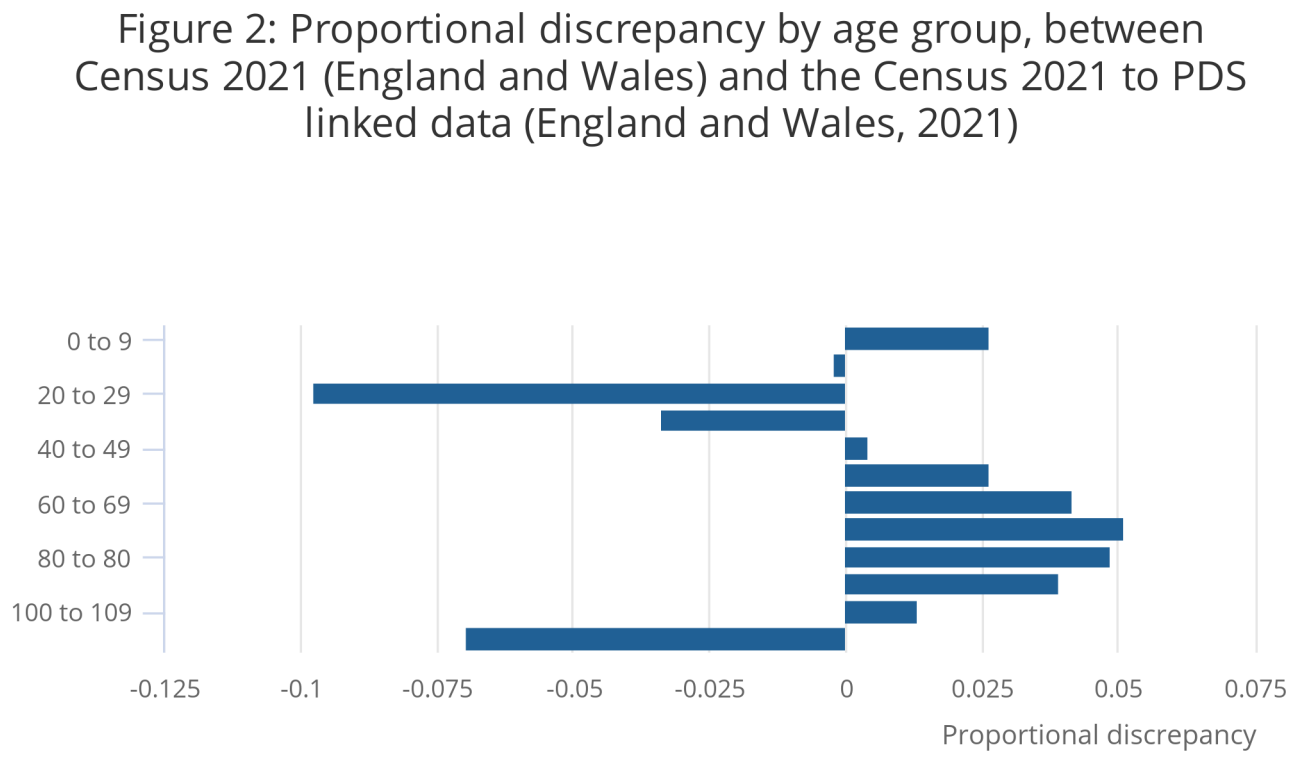
Source: Census 2021 to PDS linked data from the Office for National Statistics

Biases in age

Figure 2 shows proportional discrepancy data for age group, calculated using age on census day. The single most underrepresented group in these results are 20- to 29-year-olds, the proportional discrepancy score of negative 0.10 suggests that the linked data have only matched 90.24% of the expected number of matches given the overall match rate.

Our analysis also revealed that the most overrepresented group are 70- to 79-year-olds, the proportional discrepancy score of 0.05 suggests that the linked data have matched 105.13% of the expected number of matches given the overall match rate.

Figure 2: Proportional discrepancy by age group, between Census 2021 (England and Wales) and the Census 2021 to PDS linked data (England and Wales, 2021)



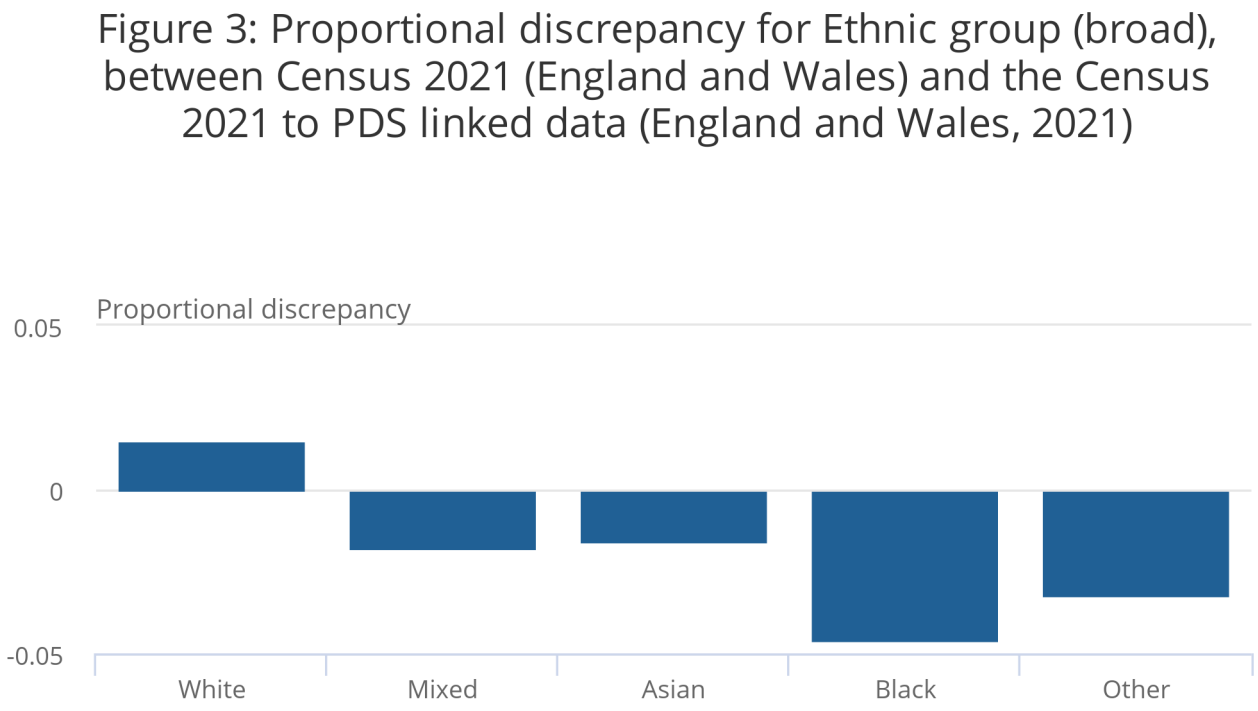
Source: Census 2021 to PDS linked data from the Office for National Statistics

Biases in ethnicity

Examination of proportional discrepancy for ethnic group, shown in Figure 3, shows that White is the only overrepresented group. The proportional discrepancy score of 0.01 for White suggests that linked data have linked 101.48% of the expected number of matches given the overall match rate.

All other groups (Asian, Mixed, Other and Black) are underrepresented. Black is the most underrepresented group, the proportional discrepancy score of negative 0.05 suggests that the linked data have matched 95.49% of the expected number of matches given the overall match rate.

Figure 3: Proportional discrepancy for Ethnic group (broad), between Census 2021 (England and Wales) and the Census 2021 to PDS linked data (England and Wales, 2021)

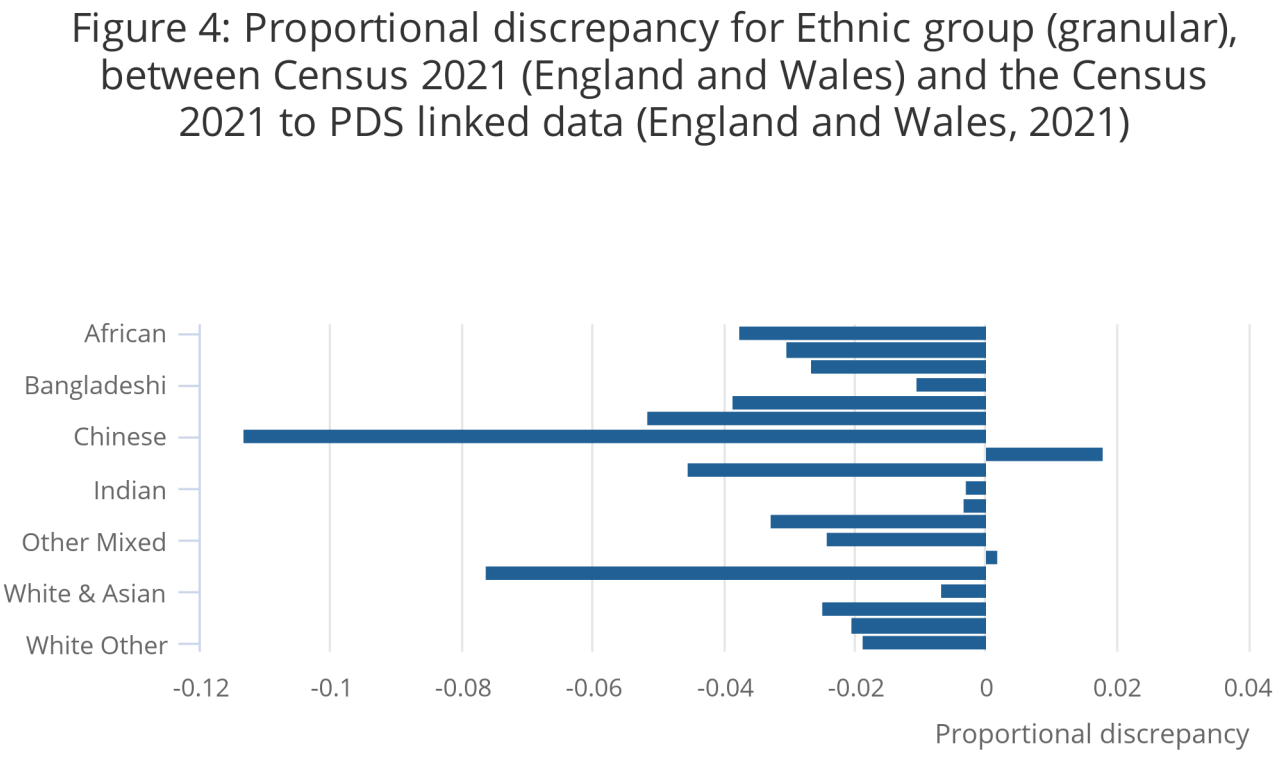


Source: Census 2021 to PDS linked data from the Office for National Statistics

Figure 4 shows a more granular breakdown of ethnic group. All groups are underrepresented except for English and Pakistani. The most underrepresented groups overall are Chinese, Roma and Caribbean.

English is the most overrepresented group, the proportional discrepancy score of 0.02 suggests that the linked data have matched 101.79% of the expected number of matches given the overall match rate. Chinese is the most underrepresented group, the proportional discrepancy score of negative 0.11 suggests that the linked data have only matched 88.70% of the expected number of matches given the overall match rate.

Figure 4: Proportional discrepancy for Ethnic group (granular), between Census 2021 (England and Wales) and the Census 2021 to PDS linked data (England and Wales, 2021)



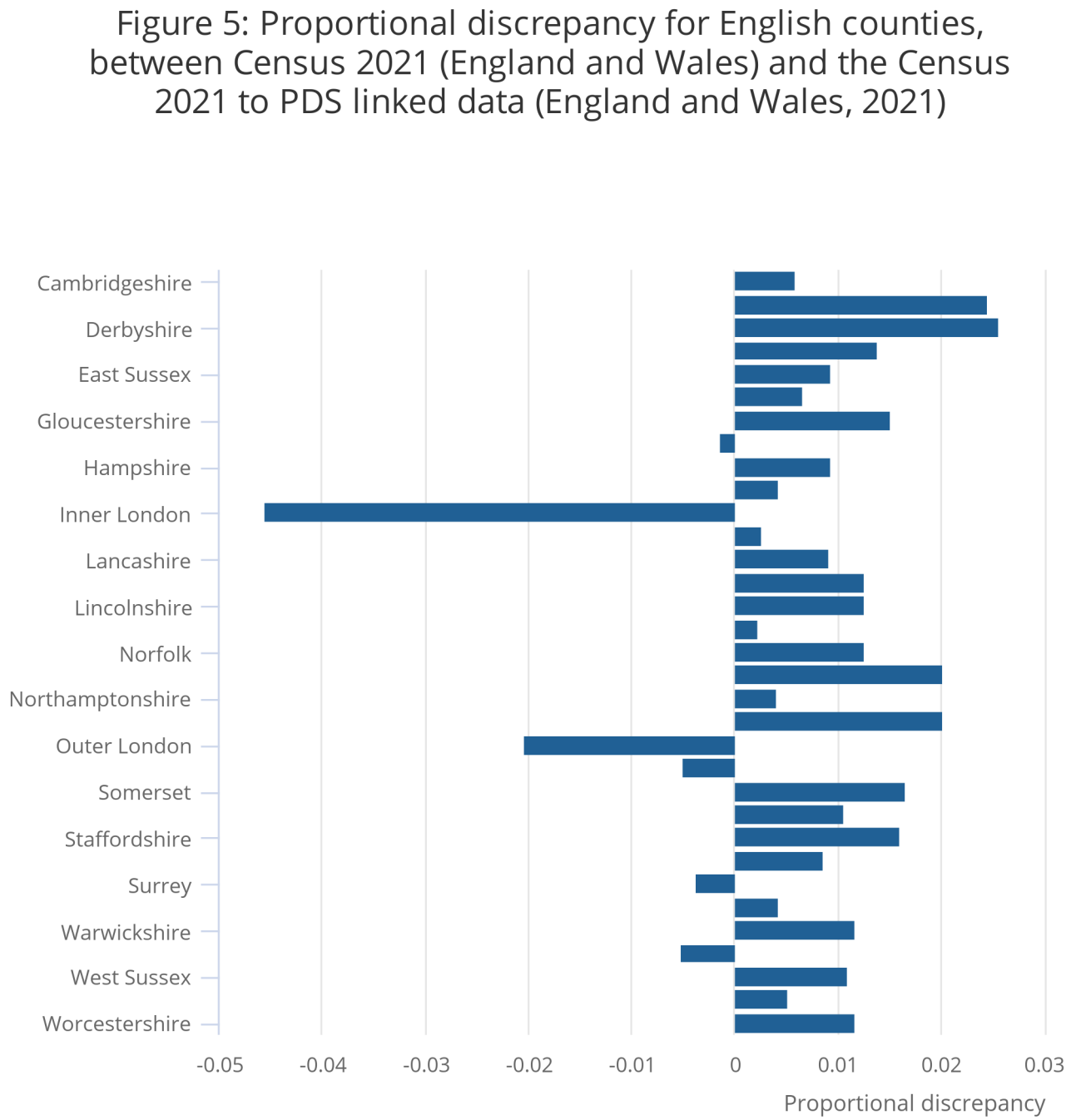
Source: Census 2021 to PDS linked data from the Office for National Statistics

Biases in geography

Proportional discrepancies for English counties are shown in Figure 5. From this it can be seen that Outer London and Inner London are the most underrepresented counties with proportional discrepancy scores of negative 0.02 and negative 0.05, respectively. This tells us that the linked data have matched 97.97% for Outer London and 95.45% for Inner London of the expected number of matches given the overall match rate.

West Midlands, Surrey, Oxfordshire and Greater Manchester are also underrepresented, but to a lesser extent than Inner and Outer London. The most overrepresented county is Derbyshire, which has a proportion discrepancy score of 0.03. This suggests that the linked data have matched 102.57% of the expected number of matches given the overall match rate.

Figure 5: Proportional discrepancy for English counties, between Census 2021 (England and Wales) and the Census 2021 to PDS linked data (England and Wales, 2021)

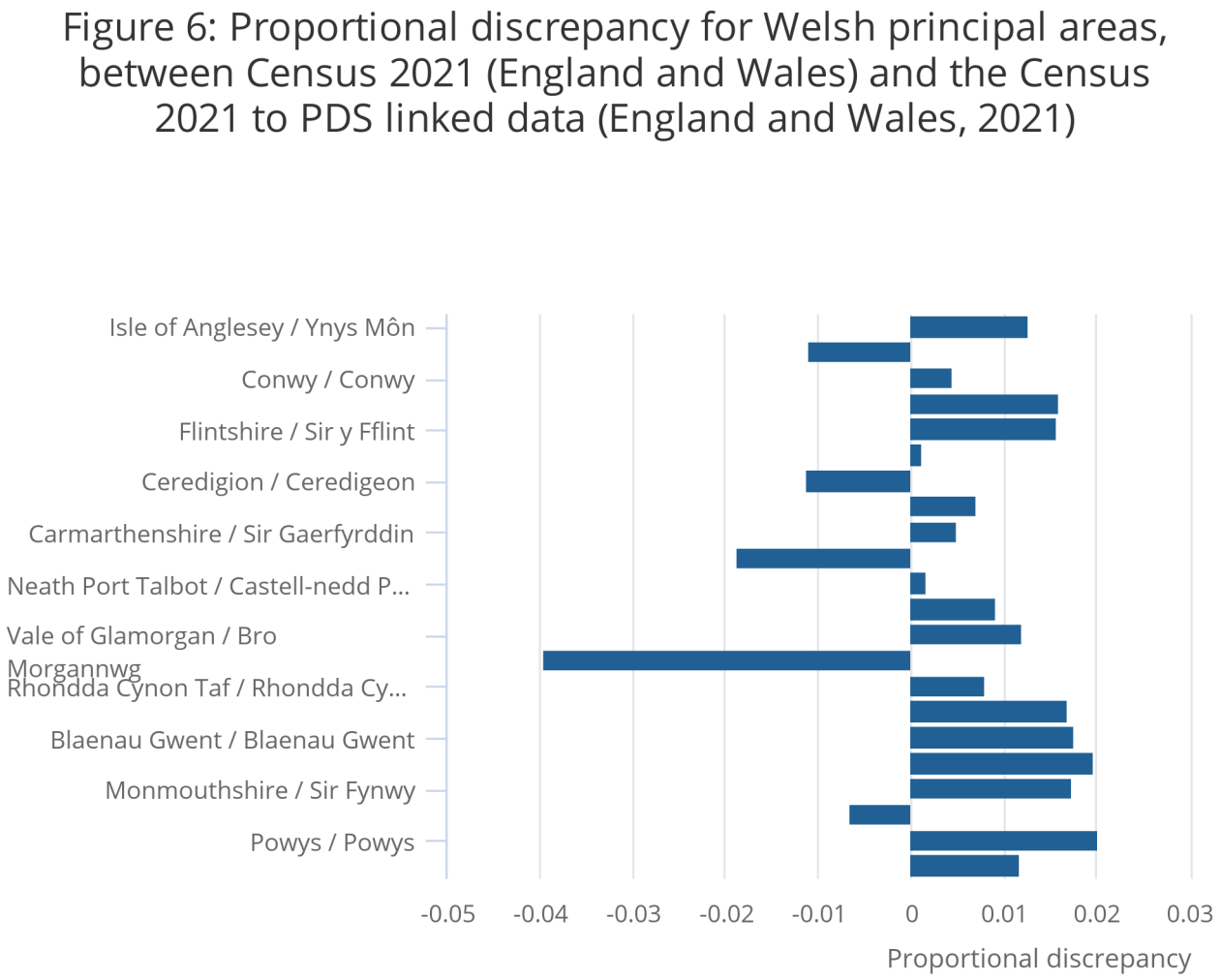


Source: Census 2021 to PDS linked data from the Office for National Statistics

Figure 6 shows overall proportional discrepancies for Welsh principal areas. Cardiff is by far the most underrepresented area, with a proportional discrepancy score of negative 0.04. This suggests that the linked data have matched 96.04% of the expected number of matches given the overall match rate.

Newport, Swansea, Ceredigion and Gwynedd are also underrepresented, but to a lesser extent than Cardiff. Powys is the most overrepresented principal area, with a proportional discrepancy score of 0.02, which suggests that the linked data have matched 102.01% of the expected number of matches given the overall match rate.

Figure 6: Proportional discrepancy for Welsh principal areas, between Census 2021 (England and Wales) and the Census 2021 to PDS linked data (England and Wales, 2021)



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

6 . Recommendations and limitations

The linkage between Census 2021 and Personal Demographics Service (PDS) has been shown to be very comprehensive, with a 95.75% linkage rate and estimated precision and recall of 99.95% and 99.99%, respectively.

Duplicated census records were identified and removed from the linked data, with a precision of 96.93%. A lookup of clustered census IDs and information about which were dropped or retained was also produced alongside the linked data. Duplicate PDS records were retained in the linked data to avoid arbitrarily removing random NHS numbers.

Efforts to account for biases in non-Anglo naming conventions included incorporating a nickname lookup for common non-Anglo names, as well as allowing looseness in transposed first names, middle names and surnames. Despite this, bias analysis on ethnicity suggests that individuals with Anglo-Western naming conventions are still overrepresented in this linkage.

7 . Related links

Previous studies using the Public Health Data Asset (PHDA) include:

Ethnic-minority groups in England and Wales-factors associated with the size and timing of elevated COVID-19 mortality: a retrospective cohort study linking census and death records
Article | Released 8 December 2020
Estimated population-level associations between ethnicity and coronavirus disease 2019 (COVID-19) mortality using a newly linked census-based dataset and investigated how ethnicity-specific mortality risk evolved during the pandemic.

Ethnicity, household composition and COVID-19 mortality: a national linked data study
Article | Released 24 March 2021
Estimated the proportion of ethnic inequalities explained by living in a multi-generational household.

Risk of suicide after diagnosis of severe physical health conditions: A retrospective cohort study of 47 million people
Article | Released 14 December 2022
Estimated whether a diagnosis of severe physical health conditions is associated with an increase in the risk of death by suicide using a dataset based on the 2011 Census linked to hospital records and death registration records.

Religious affiliation and COVID-19-related mortality: a retrospective cohort study of prelockdown and postlockdown risks in England and Wales
Article | Released 6 January 2021
Sought to understand the variation in risk of COVID-19-related death across religious groups in England and Wales both before and after the first national lockdown.

Deaths involving COVID-19 by self-reported disability status during the first two waves of the COVID-19 pandemic in England a retrospective, population-based cohort study
Article | Released 6 October 2021
Retrospective, population-based cohort study of adults aged 30 to 100 years living in private households or communal establishments in England, using population-level data to estimate the association between self-reported disability and death involving COVID-19 during the first two waves of the COVID-19 pandemic in England.

Occupation and COVID-19 mortality in England: a national linked data study of 14.3 million adults
Article | Released 27 December 2021
Estimated occupational differences in COVID-19 mortality and tested whether these are confounded by factors such as regional differences, ethnicity and education or due to non-workplace factors, such as deprivation or prepandemic health.

Further information on the methodology of the PHDA can be found in:

Deaths involving COVID-19 by religious group and ethnic group, England: methodology
Methodology | Released 14 May 2021
Detailed quality and methodology information for "Deaths involving COVID-19 by religious group, England: 24 January 2020 to 28 February 2021" and "Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 24 January 2020 to 31 March 2021".

8 . Cite this methodology

Office for National Statistics (ONS), released 23 August 2023, ONS website, methodology, Census 2021 to Personal Demographics Service (PDS) linkage report