Office for National Statistics

Article

Quality of ethnicity data in health-related administrative data sources by sociodemographic characteristics, England: May 2024

Comparing the quality of ethnicity data recorded in health-related administrative data sources with Census 2021.

Contact:
Health and Society team
health.data@ons.gov.uk
+44 1329 444110

Release date:
3 May 2024

Next release:
To be announced

Table of contents

1 . Main points

- We use non-identifiable data (all personal details are removed) to make person-level comparisons between ethnicity information recorded in administrative data sources, and we compare these with the ethnicity recorded in Census 2021 across different sociodemographic characteristics; this is widely regarded as the most robust population-level source of ethnicity information.
- General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) and Talking Therapies (TT) had the highest agreement overall, compared with Hospital Episode Statistics (HES) and Ethnic Category Information Asset (ECIA).
- There was no clear pattern in agreement across the different health data sources when examining each sociodemographic characteristic.
- Across most sociodemographic characteristics, the ethnic categories with the highest level of agreement were White British, Bangladeshi, Pakistani, Chinese, Indian, and Black African.
- Across most sociodemographic characteristics, the ethnic groups with the lowest level of agreement are the Mixed groups, Other groups and Gypsy and Irish Traveller.
- Index of Multiple Deprivation (IMD) quintile in GDPPR had the largest range in agreement; the White British category had an average agreement of 96.7% with Census 2021 data across its quintiles, and the Gypsy or Irish Traveller category had an average agreement of 4.6%

2 . The quality of ethnicity data by sociodemographic characteristics

The aim of this release is to explore whether the quality of ethnicity data is poorer within certain population groups. In our previous Quality of ethnicity data in health-related administrative data sources, England: November 2023 article we showed that across all NHS health data sources, the White British category consistently had the highest level of agreement with Census 2021 ethnicity recording (greater than 95%). This was followed by the Bangladeshi (greater than 92%), Pakistani (greater than 86%), Indian (greater than 82%), and Chinese (greater than 79%) categories.

The ethnic category with the lowest agreement across the Ethnic Category Information Asset (ECIA) and General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) datasets was the Gypsy or Irish Traveller category (less than 7%). This category was not available in the Hospital Episode Statistics (HES) or Talking Therapies (TT) datasets. For these datasets, the lowest level of agreement was for the Other Mixed (less than 35%), Any Other Ethnic Group (less than 26%) and Other Black (less than 20%) categories. However, whether the agreement differs across certain population groups remains unknown.

This article extends our previous release and adds to the broader collaborative research programme by grouping our analysis by sociodemographic characteristics. By grouping our analysis and investigating the quality of ethnicity data across NHS health administrative data sources, we can identify differences and patterns in ethnicity recording in certain population groups. The sociodemographic variables we include in our analysis are:

- age group
- sex
- region
- country of birth
- religion
- self-reported general health
- English language proficiency
- highest qualification
- household deprivation
- self-reported disability status
- urban or rural classification
- index of multiple deprivation (quintile)
- index of multiple deprivation (decile)

The data sources, methodologies and analysis applied in this article are the same as our previous release therefore, we do not discuss our methodologies in detail in this publication. For information on our data sources and methodologies, please refer to our previous release. We provided one example for each data source used (based on the methodology which reported the highest agreement in our previous release). The details of the data sources in this release are:

Ethnic Category Information Asset

- recency methodology
- no reallocation applied

General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR)

- modal methodology
- unknown only reallocation

Hospital Episode Statistics (HES)

- modal methodology
- unknown only reallocation

NHS England's Talking Therapies, for anxiety and depression (TT)

- recency methodology
- unknown only reallocation

Figure 1 allows users to select a data source and sociodemographic characteristic. This displays heat maps showing the agreement of ethnic recording in the health data source with Census 2021.

All underlying data are available in the accompanying dataset. This includes additional information on the counts in our cohort, results for five-category ethnic groupings and the 18-category ethnic grouping data presented in this article. Data for a sensitivity analysis are also available in this accompanying dataset. This is a complete case analysis which restricts the cohort to those with available data (that is, no missing data), for any of the sociodemographic characteristics included.

Figure 1: There are differences in agreement across ethnic categories and sociodemographic characteristics

Agreement between health datasets and Census 2021 using 18-category ethnicities, by sociodemographic characteristics, England

Notes:

1. Agreement is based on linked individuals with a stated ethnicity in the relevant health dataset and Census 2021. The population included is therefore different for each data source. The percentages in this interactive are calculated on rounded and suppressed values.
2. For each source, the health data ethnic group totals have been used as denominators when calculating percentages.
3. The missingness for each sociodemographic variable is different therefore the population for each sociodemographic characteristic may be different. Further details regarding this can be found in the accompanying dataset.
4. The Arab and Traveller ethnic group categories are not available in HES or NHS TT, so agreement for these categories are only presented for ECIA and GDPPR. The Roma ethnic group is not available for any health administrative dataset.
5. For GDPPR, HES and TT data sources, these data refer to when the Unknown only reallocation methodology has been applied.
6. Categories with a total population of less than 40 have been denoted with an asterisk (*).

Download the data

Downloac

While this research programme has the specific purpose of developing guidance for analysts to improve coherence of statistics of ethnic health disparities using different sources, it was carried out within the context of our broader strategic aim; this aim is to explore the use of administrative data to produce population statistics, including characteristics such as ethnicity.

This work includes our Developing admin-based ethnicity statistics for England and Wales: 2020 article, which uses ethnicity information from a range of sources, including HES, TT, education and birth registration data. However, access to the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) was acquired under limited use conditions as part of coronavirus (COVID-19) pandemic response planning. Where appropriate, methods have been aligned. We hope this work to understand the quality of a range of additional health sources will help inform improvements to the admin-based ethnicity statistics (ABES).

3 . Data

Quality of ethnicity data in health-related administrative data sources, by sociodemographic characteristics
Dataset | Released 3 May 2024
Comparing the quality of ethnicity data recorded in health-related administrative data sources with Census 2021, by sociodemographic characteristics.

Quality of ethnicity data in health-related administrative data sources where population was restricted to those with data for all sociodemographic characteristics
Dataset | Released 3 May 2024
Agreement rates between ethnicity data recorded in health-related administrative data sources with Census 2021 by sociodemographic characteristics, where population was restricted to those with data for all sociodemographic characteristics.

4 . Glossary

Agreement

Agreement is calculated as the percentage of linked records where the ethnicity in the health administrative data source and Census 2021 are the same. This is based on records with a stated ethnicity in each health administrative data source linked to Census 2021.

Ethnicity stated

Ethnicity stated refers to the ethnicity being recorded as a specific ethnic group and not recorded as being "Not Stated" or "Not Known".

Ethnicity unresolved

Where multiple ethnic categories were recorded on the latest date, or there were other conflicts as previously described in our previous release, these have been coded as unresolved. Additionally, for Hospital Episode Statistics (HES) if a dataset hierarchy of Admitted Patient Care (APC), Accident and Emergency (A&E) and Emergency Care Data Set (ECDS), and Outpatients (OP) did not resolve the conflict then this was coded as “Unresolved”.

For General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) data, we derived the most recent ethnicity recording by taking it from either the GP-Journal (SNOMED codes) or GP-Patient (ETHNIC column) tables. Priority was given to the GP-Journal table recording in instances of conflict in recordings on the same most recent date between sources.

5 . Data sources and quality

Administrative data sources

We used multiple administrative data sources to make person-level comparisons between ethnicity information. These sources were:

- Hospital Episode Statistics (HES)
- General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR)
- Ethnic Category Information Asset (ECIA)
- Talking Therapies for anxiety and depression (TT)

Sociodemographic characteristics

The sociodemographic characteristics used in this analysis used data from:

- Census 2021, from the Office for National Statistics (ONS)

Quality

Limitations regarding our methodology, which are applicable to this current release, are detailed in our previous release in Section 7: Data sources and quality.

6 . Related links

Quality of ethnicity data in health-related administrative data sources, England: November 2023
Article | Released 6 November 2023
Comparing the quality of ethnicity data recorded in health-related administrative data sources with Census 2021.

Understanding consistency of ethnicity data recorded in health-related administrative datasets in England: 2011 to 2021
Article | Released 16 January 2023
Comparisons showing differences in the recording of ethnicity data between health administrative data sources and the 2011 Census.

Methods and systems used to collect ethnicity information in health administrative data sources, England 2022
Article | Released 16 January 2023
Findings from semi-structured qualitative interviews that assess the quality of ethnicity data and identify sources of bias across three health data sources in England.

Developing admin-based ethnicity statistics (ABES) for England and Wales
Article | Released 7 February 2023
Research update on producing population statistics by ethnic group for England and Wales from administrative data, with comparisons with Census 2021 estimates.

7 . Cite this article

Office for National Statistics (ONS), released 3 May 2024, ONS website, article, Quality of ethnicity data in health-related administrative data sources by sociodemographic characteristics, England: May 2024