Office for National Statistics

Article

Quality of ethnicity data in health-related administrative data sources, England: November 2023

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

Contact:
Cameron Razieh, Bethan Cairns, Alicja Januszkiewicz and Rose Drummond
health.data@ons.gov.uk
+44 1329 444110

Release date:
6 November 2023

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 in Hospital Episode Statistics (HES), General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), the Ethnic Category Information Asset (ECIA), and Talking Therapies for anxiety and depression (TT) administrative data sources, and we compare these with ethnicity as recorded in Census 2021, which is widely regarded as the most robust population-level source of ethnicity information.

- Across all health administrative data sources, the White British category consistently had the highest level of agreement with Census 2021 (greater than 95%), 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 ECIA and GDPPR datasets was the Gypsy or Irish Traveller category (less than 7%); this category was not available within HES or TT, where the lowest level of agreement within these sources was for the Other Mixed (less than 35%), Any Other Ethnic Group (less than 26%) and Other Black (less than 20%) categories.

- A set of reallocation methodologies was applied to assess the impact of reallocating the Not Known, Any Other Ethnic Group and Not Stated ethnic categories in GDPPR, HES and TT on agreement with Census 2021, with these methodologies having little impact on agreement.

- We assessed the agreement of each individual HES sub-dataset with Census 2021.

- We have released a tool to help inform expert decisions when mapping detailed ethnicity codes recorded in General Practitioner (GP) data to harmonised ethnicity categories for analysis; this tool is designed as an aid and not intended to replace expert judgement.

2 . About our research on ethnicity data quality

Collecting high-quality ethnicity data within administrative data sources has become of great interest to governments, data providers and the public over recent years. Electronic health records (EHRs) have increasingly been used to produce statistics and analysis on health inequalities across ethnic groups.

The urgent need for robust statistics on health outcomes for different ethnic groups was emphasised during the coronavirus (COVID-19) pandemic, where people from minority ethnic groups were found to be at higher COVID-19 mortality risk. The limited research on the quality of the recording of ethnicity across different EHRs indicates that missingness (absence of data) is relatively high and varies across sources. However, the accuracy of the recorded ethnicities remains unknown.

The Office for National Statistics (ONS) is collaborating with Wellcome on a programme of research to explore the quality of ethnicity information recorded in different health data sources, and to examine the potential bias caused by inconsistencies. The goal is to improve analysts' understanding of the limitations of the data, test potential solutions and develop guidance to improve the comparability of analyses based on different sources.

Analysis published so far in the programme of research includes:

- our Methods and systems used to collect ethnicity information in health administrative data sources article, which explores the process of collecting ethnicity data in healthcare settings and why differences in ethnicity data might occur

- our Understanding consistency of ethnicity data recorded in health-related administrative datasets in England article, a quantitative analysis assessing the differences in ethnicity recording between sources

- our How ethnicity recording differs across health data sources and the impact on analysis blog, which summarises the results of both analyses

- a complementary analysis published by Wellcome in collaboration with the Race Equality Foundation on Improving the recording of ethnicity in health datasets based on focus groups exploring how minority ethnic communities are asked about their ethnicity and how this in turn is recorded

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 has been carried out within the context of the ONS's broader strategic aim of exploring the use of administrative data to produce population statistics including characteristics such ethnicity. This work includes developing admin-based ethnicity statistics for England and Wales. Where appropriate, methods have been aligned.

3 . Method for comparing ethnicity information across sources

This article builds upon our previous release and adds to the broader collaborative research programme by:

- updating census information to use Census 2021 as a comparator

- updating and extending the NHS England (NHSE) health data sources included

- testing additional reallocation methodologies to derive a single ethnicity per person from episodic health data

- comparing agreement of each individual Hospital Episode Statistics (HES) sub-dataset with Census 2021

- creating a tool for experts to use alongside other evidence when deciding how to map the detailed ethnicity codes recorded in General Practitioner (GP) data to the harmonised ethnic categories used for national-level analyses

Data sources

The current analysis utilises the same data sources as in the previous publication, but additionally updates and includes:

- a full extract of the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) dataset, where previously only a subset was used

- Hospital Episode Statistics (HES), which is made up of three sub-datasets, including the Accident and Emergency (A&E) sub-dataset which was superseded by the Emergency Care Dataset (ECDS) in April 2020; the current analysis now additionally includes the ECDS

- NHSE's Talking Therapies, for anxiety and depression (TT), formerly Improving Access to Psychological Therapies (IAPT), is a new addition to this analysis; this dataset was developed to monitor and evaluate an NHSE programme aimed at improving the delivery of, and access to, evidence-based, psychological therapies for adults with depression and anxiety disorders

Ethnicity definitions within each data source

Ethnic categories vary across data sources, and the wording of categories also varies, even when they align across data sources. Census 2021 includes 19 ethnic categories, including a newly implemented Roma category, whereas the health administrative data sources have either 18 categories (GDPPR and the Ethnic Category Information Asset (ECIA)) or 16 categories (HES and TT). Table 1 of our accompanying dataset compares the ethnic categories for each source included in the analysis, and the mapping used for comparisons. For more information, see GOV.UK's List of ethnic groups web page and the Government Analysis Function's Ethnicity harmonised standard web page.

Handling multiple ethnicity records per person

GDPPR, HES and TT contain information about all interactions a patient has with the relevant health service, so generally contain multiple records per patient. Within these data sources, some individuals have multiple recorded ethnicities within the same data source at different episodes.

A set of rules was therefore implemented to select a single ethnicity per person for comparison with Census 2021. The ECIA contains a single ethnicity per person, based on the most recent ethnicity recorded in either GDPPR or HES. Full details of the methodology used to determine this have been published by NHSE.

We applied two methods to derive an individual's ethnicity within GDPPR, HES and TT sources: the most common (modal) and most recent (recency) ethnicity recorded for each person. These methods are explained in our previous publication. Because of the way the extract of TT data available to us was structured and processed, no modal definition was possible.

Our extract of TT data was pre-processed and assigns the most recent ethnicity recording per year and per service provider. To derive the most recent ethnicity recording within TT, we selected the most recent ethnicity recording from the most recent service provider in the most recent year available. If there were two or more different ethnicity recordings on the same most recent date, they were classified as "Unresolved". The categories Data Not Recorded and Value Outside of National Code were treated as the Not Known category.

Reallocating ethnicity records

We applied additional methodologies to test whether reallocating Not Known and Not Stated ethnic categories improved agreement with Census 2021 by improving the coverage of people with a stated ethnic category. Any Other Ethnic Group was also reallocated because of evidence suggesting there is likely over-coding of this ethnic group. For more information, see Nuffield Trust's Ethnicity coding in English health service datasets report.

Once a single ethnicity recording was derived for each person in GDPPR, HES and TT using recency and modal methodologies, a set of reallocation rules were applied, where certain ethnic categories were reallocated if alternative ethnic categories were available within their records. This was done even if these records were older or less frequent. Where an individual's record contained only one or more recording of the same ethnicity category, their record was kept as is and not reallocated.

The ethnic categories that were sequentially reallocated were:

- Not Known
- Not Known; Any Other Ethnic Group (where a person only had Not Known and Any Other Ethnic Group categories recorded, Any Other Ethnic Group was chosen as the reallocation destination)
- Not Known; Any Other Ethnic Group; Not Stated (where a person only had either Not Known or Not Stated, or both, and Any Other Ethnic Group categories recorded, Any Other Ethnic Group was chosen as the reallocation destination)

Based on our specific research question, we have reallocated the Not Stated category to assess whether increasing coverage of people with a stated ethnic category within health data sources improves agreement with Census 2021. However, other Office for National Statistics (ONS) work takes a different approach as it uses other data sources to improve coverage. For further details, see our Developing admin-based ethnicity statistics for England and Wales article.

Data linkage

To enable comparisons of ethnicity recorded in each health administrative data source with Census 2021, people enumerated in Census 2021 were linked securely to the NHS Personal Demographics Service (PDS) to obtain their NHS number (with 95.75% of persons in the census probabilistically and deterministically matched to persons in the PDS). For more information on the linkage methodology, see our Census 2021 to Personal Demographics Service linkage report.

Our Census 2021 study population included 55.1 million people enumerated in England and Wales for whom we could obtain an NHS number. We then excluded individuals who were resident in Wales at the time of the census (2.8 million), those who had not answered the ethnicity question (0.5 million) and those who were not usual residents in England (0.4 million). Therefore, a total of 51.3 million individuals from England were included in our analysis, covering 90.8% of the population of England on Census Day (21 March 2021), which was estimated to be 56.5 million. For further details on this estimate, see our Population and household estimates, England and Wales bulletin.

Individuals with available ethnicity data from each health administrative data source were then linked to the census using NHS number.

For GDPPR, HES and TT, we included all available ethnicity records recorded up to and including 29 January 2022 (the most recent date within ECIA).

Table 1: Count of people in the linked datasets created to compare the quality of ethnicity recording in health data sources with that in Census 2021, England

	Count of people in each linked dataset		Count of people in linked dataset with a stated ethnicity in both health and census sources	
Linked dataset, ethnicity allocation method	**Millions**	**Percentage of the population of England on Census Day 2021**	**Millions**	**Percentage of the population of England on Census Day 2021**
Linked census-ECIA	47.4	83.9	47.4	83.9
Linked census-GDPPR, modal	43.5	77.0	40.1	71.0
Linked census-GDPPR, recency	43.5	77.0	42.2	74.7
Linked census-HES, modal	47.8	84.6	40.1	71.0
Linked census-HES, recency	47.8	84.6	39.7	70.3
Linked census-TT, recency	6.3	11.2	5.4	9.6

Source: Office for National Statistics

Notes

1. A stated ethnicity in both sources excludes individuals who could not be linked ("Not linked") or whose ethnicity from the health data was "Not Known", "Not Stated" or "Unresolved" after applying the recency or modal rules.

2. For GDPPR, HES and TT data sources, these data refer to when the Unknown only reallocation methodology has been applied.

4 . Person-level cross tabulations

Person-level cross tabulations of each health administrative data source with Census 2021

To explore the consistency of ethnicity information across data sources, we produced 18-category and 5-category ethnic group cross tabulations of Census 2021 with each health administrative data source. This enabled us to examine the distribution between each ethnic category assigned in the Ethnic Category Information Asset (ECIA), General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics (HES) and Talking Therapies (TT) sources, and the ethnic category an individual was assigned in Census 2021, as counts and proportions. See Tables 7 to 18 for 18-category comparisons and Tables 19 to 30 for 5-category comparisons in our accompanying dataset.

5 . Person-level agreement

Person-level agreement in ethnicity coding in each health administrative data source compared with Census 2021

To summarise the information within the cross tabulations, we presented the agreement for each health administrative data source compared with Census 2021. For each person, the ethnic category recorded in the census and each respective data source were compared and classified as:

- 1, if the recorded ethnicities were the same
- 0, if they were different

Where the ethnic categories used in the health administrative sources data did not exactly match with the Census 2021 categories, ethnic categories were matched with the most aligned Census 2021 ethnicity category. Only those with a stated ethnicity category in both data sources were included in the agreement calculations; the Not Stated, Not Known, and Unresolved categories were not included in agreement calculations. Arab and Gypsy or Irish Traveller ethnic categories are not available within Hospital Episode Statistics (HES) and Talking Therapies (TT), and therefore no agreement was calculated for these categories in HES and TT. For further details on how agreement was calculated, see the Methods tab of our accompanying dataset.

Table 2: Overall agreement by health data source in comparison with Census 2021, using 18-category and 5-category ethnic categories, England

Dataset, ethnicity allocation method	Overall agreement for 18-category ethnic groups	Overall agreement for 5-category ethnic groups
ECIA	86.7	94.0
GDPPR, modal	89.9	95.7
GDPPR, recency	87.0	94.6
HES, modal	87.6	94.1
HES, recency	86.4	93.3
TT, recency	92.4	96.4

Source: Office for National Statistics

Notes

1. Data are presented as percentages.
2. The percentages are based on individuals with a stated ethnicity in both the health data source and Census 2021; those whose ethnic information was "Not linked", "Not Stated", "Not Known" or "Unresolved" in either source were excluded.
3. For GDPPR, HES and TT data sources, these data refer to when the Unknown only reallocation methodology has been applied.

Overall agreement ranged from 86.4% for HES-recency to 92.4% for TT for the 18-category ethnic groups. Similar patterns were seen with the 5-category ethnic groups, but agreement was higher for all sources (ranging from 93.3% for HES-recency to 96.4% for TT).

Agreement by ethnic group

Figure 1 shows how, across all data sources, the White British category consistently showed the highest level of agreement with Census 2021 (greater than 95%). The Bangladeshi category showed the second highest levels of agreement across all sources (greater than 92%), with the same level of agreement as the White British category for General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) and Talking Therapies (TT) sources. Pakistani, Indian and Chinese categories showed the next highest levels of agreement across all data sources (greater than 86%, 82% and 79%, respectively). Black African and Black Caribbean showed agreement with Census 2021 ranging from 70% to 86% across all data sources.

The ethnic category with the lowest agreement across the Ethnic Category Information Asset (ECIA) and GDPPR datasets was the Gypsy or Irish Traveller category (less than 7%). The Gypsy or Irish Traveller ethnic group was not available within HES or TT, which use 16 ethnic categories for reporting. The ethnic categories with the lowest level of agreement within HES and TT data sources were the Any Other Ethnic Group and Other Black categories (less than 26% and 20%, respectively). Agreement was generally lower for all Mixed and Other ethnic categories across all data sources (less than 76% and 72%, respectively).

Figure 1: Agreement was lowest for Gypsy or Irish Traveller, Other Black and Any Other Ethnic Group categories (25% or less), and highest for White British and Bangladeshi categories (93% or more), for all sources and methods

Percentage of agreement between health datasets and Census 2021 using 18-category ethnicities, 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.

2. For each source, the health data ethnic group totals have been used as denominators when calculating percentages.

3. The Arab and Traveller ethnic group categories are not available in HES or NHS TT, so agreement rates for these categories are only presented for ECIA and GDPPR. The Roma ethnic group is not available for any dataset.

4. For GDPPR, HES and TT data sources, these data refer to when the Unknown only reallocation methodology has been applied.

Downloac

To understand the extent to which differences may occur between different high-level ethnic groups, we conducted analysis using 5-category ethnic groups (see Tables 19 to 30 in our accompanying dataset).

Patterns of agreement were similar when aggregating ethnicity to 5-category ethnic groupings. For White, Asian and Black categories, the 5-category ethnic groupings showed agreement of 88% or higher across all sources, meaning differences in ethnicity recording are predominantly within the same 5-category grouping (for example, Black African, Black Caribbean and Other Black). Mixed and Other category agreement was mostly higher compared with the 18-category results of the same disaggregated categories, but still showed low agreement overall, meaning the differences in ethnicity recording are less likely to be within the same 5-category ethnic groupings (for example, Other Asian and Any Other Ethnic Group).

Figure 2: Agreement was lowest for the Other ethnic category (39% or less) and highest for White, Asian and Black ethnic categories (88% or more), for all sources and methods

Percentage of agreement between health datasets and Census 2021 using 5-category ethnicities, 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.

2. For each source, the health data ethnic group totals have been used as denominators when calculating percentages.

3. For GDPPR, HES and TT data sources, these data refer to when the Unknown only reallocation methodology has been applied.

Downloac

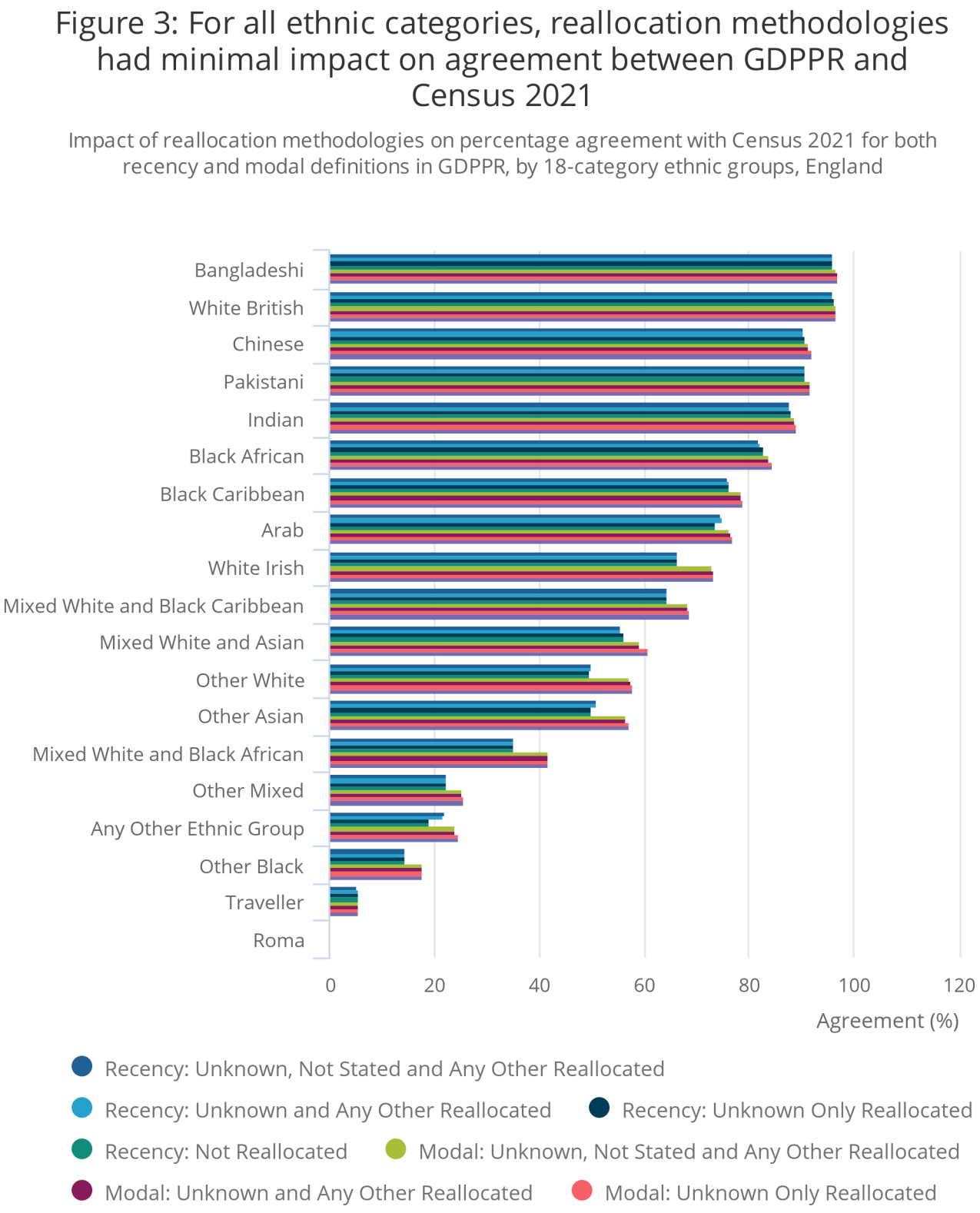
We conducted two sensitivity analyses where we restricted the back series of ethnicity data to 1 April 2015 (aligning the first date within our extract of ECIA) and restricted the population to only those who had a stated ethnic category in each of the GDPPR, HES and Census 2021 datasets. We did this to assess the extent to which agreement was affected by differences in coverage and populations between GDPPR and HES. Results were similar to our main analysis and can be found in Tables 31 to 34 of our accompanying dataset.

Reallocation of ethnicity in episodic health administrative datasets (GDPPR, HES and TT)

For GDPPR, HES and TT, no notable changes in agreement with Census 2021 were seen between any of the reallocation methodologies applied for either modal or recency definitions, with agreement per ethnic group being similar for all levels of reallocation. The modal approach could not be applied for TT.

Figure 3: For all ethnic categories, reallocation methodologies had minimal impact on agreement between GDPPR and Census 2021

Impact of reallocation methodologies on percentage agreement with Census 2021 for both recency and modal definitions in GDPPR, by 18-category ethnic groups, England



Source: Office for National Statistics

Notes:

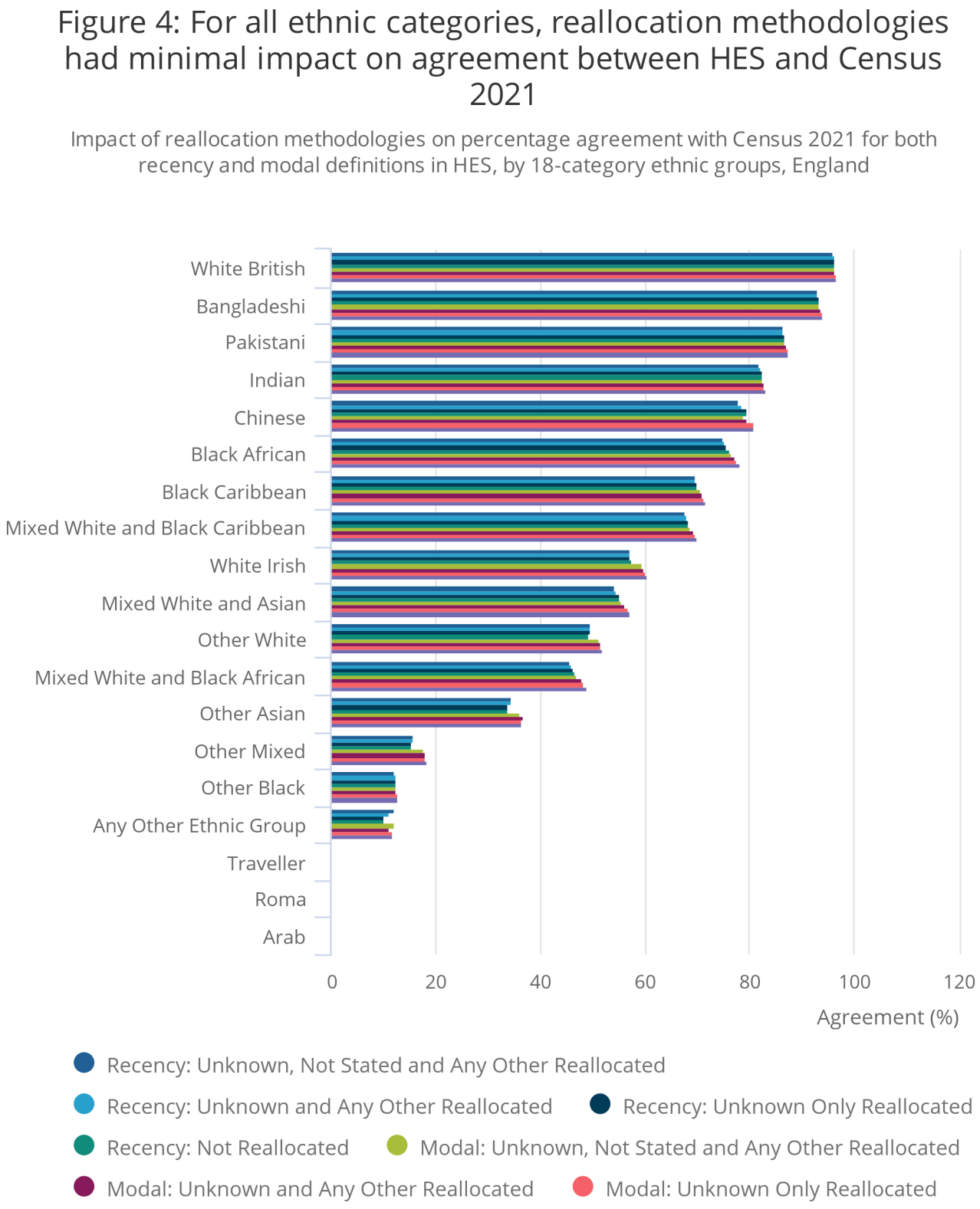
1. Agreement is based on linked individuals with a stated ethnicity in the relevant reallocation methodology GDPPR dataset and Census 2021. The population included is therefore different for each data source.

2. For each reallocation methodology, the health data ethnic group totals have been used as denominators when calculating percentages.

3. The Roma ethnic group category is not available in GDPPR.

Figure 4: For all ethnic categories, reallocation methodologies had minimal impact on agreement between HES and Census 2021

Impact of reallocation methodologies on percentage agreement with Census 2021 for both recency and modal definitions in HES, by 18-category ethnic groups, England



Source: Office for National Statistics

Notes:

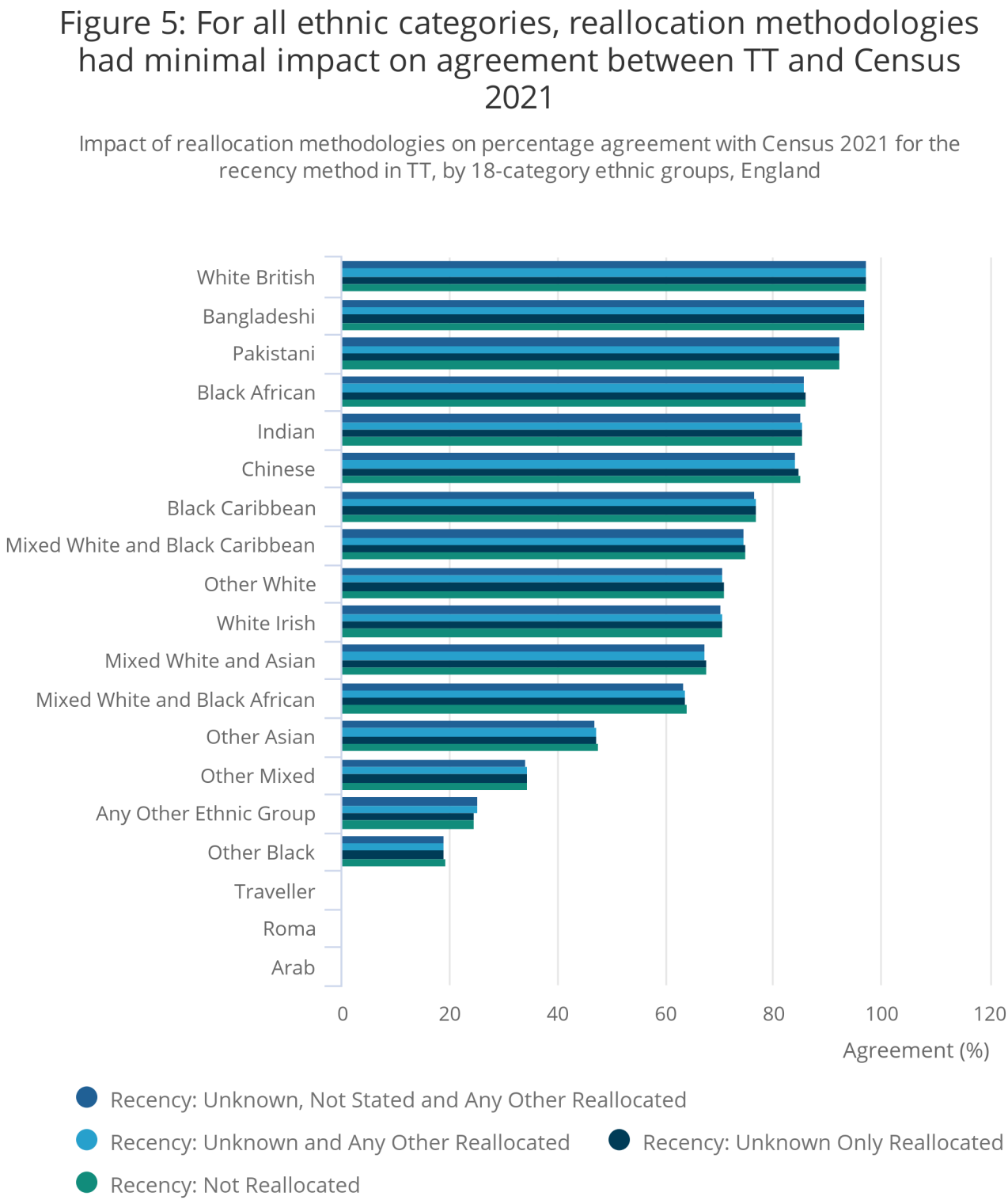
1. Agreement is based on linked individuals with a stated ethnicity in the relevant reallocation methodology HES dataset and Census 2021. The population included is therefore different for each data source.

2. For each reallocation methodology, the health data ethnic group totals have been used as denominators when calculating percentages.

3. The Arab, Traveller and Roma ethnic group categories are not available in HES.

Figure 5: For all ethnic categories, reallocation methodologies had minimal impact on agreement between TT and Census 2021

Impact of reallocation methodologies on percentage agreement with Census 2021 for the recency method in TT, by 18-category ethnic groups, England



Source: Office for National Statistics

Notes:

1. Agreement is based on linked individuals with a stated ethnicity in the relevant reallocation methodology TT dataset and Census 2021. The population included is therefore different for each data source.

2. For each reallocation methodology, the health data ethnic group totals have been used as denominators when calculating percentages.

3. The Arab, Traveller and Roma ethnic group categories are not available in TT.

Agreement with Census 2021 in each of the individual HES sub-datasets

We further assessed the agreement between each sub-dataset within HES and Census 2021. These sub-datasets are:

- Admitted Patient Care (APC)
- Accident and Emergency (A&E) and Emergency Care Dataset (ECDS)
- Outpatients (OP)

The results can be found in Tables 1 to 8 of our accompanying dataset.

6 . Mapping detailed ethnicity SNOMED codes to harmonised categories

The ethnicity information within the journal tables of the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) is recorded using SNOMED CT health terminology. The SNOMED CT structured clinical vocabulary contains several hundred codes for ethnicity (for more information, see Section 8: Glossary).

NHS England (NHSE) have published a mapping between SNOMED ethnicity codes and the 18 harmonised ethnic categories. However, previous Office for National Statistics (ONS) research identified a risk whereby this mapping could match individuals to an ethnic category they would not self-identify with. For further details, see our Methods and systems used to collect ethnicity information in health administrative data sources article. Identifying improvements to this mapping could potentially improve person-level agreement rates with Census 2021 for the GDPPR data.

The ONS has created a tool to facilitate comparisons between the NHSE SNOMED ethnicity code mapping and how those individuals self-identified their ethnicity in Census 2021. This information can be used as part of the decision-making process when mapping SNOMED ethnicity codes to harmonised ethnicity categories for analysis, but is not intended to replace expert judgement.

SNOMED CT tool findings

Overall, our tool shows that NHSE mapping is broadly consistent with how individuals self-identified their ethnicity in Census 2021. However, consistency varies between ethnic groups. When using the recency method to identify a single SNOMED code per person, 38% of SNOMED codes do not match how those individuals self-identified their ethnicity in Census 2021. However, these SNOMED codes accounted for only 6% of the population included in the analysis.

For most ethnic categories (11 out of 18), the SNOMED codes that did not match to the self-identified Census 2021 ethnic category accounted for no more than 1% of the population matched to that ethnic group (see Table 5 in our accompanying dataset). However, for four of the ethnic groups (Other Black, Gypsy or Irish Traveller, White and Black African, and Any Other Ethnic Group), the SNOMED codes which did not match to the self-identified Census 2021 ethnic category accounted for more than 80% of the population currently matched to that ethnic group.

7 . Data

Quality of ethnicity data in health-related administrative data sources, England
Dataset | Released 06 November 2023
Comparing the quality of ethnicity data recorded in health-related administrative data sources with Census 2021.

Quality of ethnicity data in Hospital Episode Statistics sub-datasets, England
Dataset | Released 06 November 2023
Comparing the quality of ethnicity data recorded in individual Hospital Episode Statistics (HES) sub-datasets, including Admitted Patient Care (APC), Outpatients (OP), and Accident and Emergency (A&E) and Emergency Care Dataset (ECDS), with Census 2021.

Mapping detailed SNOMED ethnicity codes to harmonised Census 2021 ethnic categories, England
Dataset | Released 06 November 2023
Comparing NHS England SNOMED code mapping with how individuals self-identified their ethnicity in Census 2021.

8 . Glossary

Agreement

Of those records with a stated ethnicity in each health administrative data source linked to Census 2021, agreement is calculated as the percentage of linked records where the ethnicity in the health administrative data source and Census 2021 are the same.

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 not stated

In the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics (HES) and Talking Therapies (TT) data sources, an individual can choose to not identify their ethnic group. In these instances, the code "Z – Not Stated" is recorded.

Ethnicity not known

In the HES and GDPPR data sources, if an individual's ethnicity is unknown, the code "X (prior to 2013) or 99 (post-2013) – Not Known" is recorded.

Ethnicity unresolved

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

For 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.

Ethnicity data not recorded

In the TT data source, if an individual's ethnicity was not recorded, the code "-1 – Data Not Recorded" is recorded.

Ethnicity value outside of national code

In the TT data source, if an individual's ethnicity was recorded as a category outside of the national code, the code "-3 – Value Outside of National Code" is recorded.

Not linked

“Not linked” refers to individuals who have a stated ethnicity in Census 2021 but could not be linked to the administrative data source, regardless of whether they had a stated ethnicity in the GDPPR, HES, the Ethnic Category Information Asset (ECIA) or TT data sources.

SNOMED code

SNOMED codes are the clinical coding standards used with General Practitioner (GP) records. Further information about SNOMED codes and how ethnicity is recorded within different fields and tables within GDPPR can be found on NHS Digital's GPES data for pandemic planning and research (COVID-19) web page.

9 . Data sources and quality

Census 2021 as a comparator

Although self-reported ethnicity may be prone to certain biases, it is generally considered one of the most robust methods to collect ethnicity information. Census data are the most complete source of self-reported ethnicity information for the whole population, and therefore widely regarded as the most reliable source of ethnicity data for England.

Self-reported ethnicity may change with time and age. However, the impact of this on our analysis is limited because of Census 2021 data being the most up to date ethnicity data available for the entire population of England at the time of analysis. Further, we were able to identify individuals with imputed census ethnicity and remove them from the analysis.

Some ethnicity responses in the census data may be provided by a proxy, for example, a parent on behalf of a child who cannot respond for themselves. It is not only census data that is affected by proxy reporting; the health data sources are likely to also contain some proxy responses affecting the comparisons. It has been reported that ethnic category is sometimes recorded by NHS staff without asking the patient or there is a reluctance from staff to ask about ethnicity within healthcare environments.

All individuals within our analysis had to have a stated ethnicity recorded in the Census 2021. Therefore, an important limitation of our analysis is that it excludes people who did not take part in the census (estimated to be 3% of the population), recent migrants, and people who could not be linked to the NHS Personal Demographic Service, which may affect representativeness of the population used. However, our dataset included 90.8% of the population living in England on Census Day.

Data linkage

A limitation of the linkage approach used is that linkage rates vary between ethnic group. However, this methodology does result in a linked population with a high coverage of England that is implemented in many other Office for National Statistics (ONS) publications. Linkage between sources may sometimes be imperfect and result in false positive linkage. For more information on linkage rates varying between ethnic groups, see our Ethnic differences in life expectancy and mortality from selected causes in England and Wales: 2011 to 2014 article.

Comparisons between data sources

The Not Stated category can be interpreted as a refusal to provide an ethnic category, in line with the methodology used in our previous Producing admin-based ethnicity statistics for England: methods, data and quality article. However, we have treated the Not Stated category in General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics (HES) and Talking Therapies (TT) as an invalid ethnicity in our final reallocation method. We have reallocated this category if there was another available stated ethnic category within a person's back series of data to assess the impact of reallocating it.

A further limitation is that the Gypsy or Irish Traveller and Arab ethnic categories do not exist within HES or TT. Therefore, a comparison of these categories within these sources is not possible. In addition, as NHS England have published code on methods to derive ethnicity from health data, we were able to replicate methods used by other analysts.

10 . Future developments

This work is part of a wider programme of research investigating the quality of ethnicity recording between health administrative data sources. Future research will assess the potential bias in mortality estimates based on different ethnicity recordings in different health data sources and develop solutions to mitigate any observed biases. We will use the findings from this programme of research to inform guidance for analysts using ethnicity data from health administrative data sources.

11 . Related links

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.

Producing admin-based ethnicity statistics for England: methods, data and quality
Article | Released 6 August 2021
An overview of methods, data sources and data quality for the feasibility research on producing statistics on the population by ethnic group from Hospital Episode Statistics, English School Census and Improving Access to Psychological Therapies data.

Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 10 January 2022 to 16 February 2022
Article | Released 7 April 2022
Estimates of COVID-19 mortality rates by ethnic group using linked data from the Office for National Statistics Public Health Data Asset.

12 . Cite this article

Office for National Statistics (ONS), released 6 November 2023, ONS website, article, Quality of ethnicity data in health-related administrative data sources, England: November 2023