Office for National Statistics

Article

Understanding consistency of ethnicity data recorded in health-related administrative datasets in England: 2011 to 2021

Comparisons showing differences in the recording of ethnicity data between health administrative data sources and the 2011 Census.

Table of contents

1 . Main points

- We use non-identifiable data (all personal details 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) and Ethnic Category Information Asset (ECIA) administrative data sources compared with ethnicity recorded in the 2011 Census, which is widely regarded as the most robust population level source of ethnicity information (Census 2021 data were not available for analysis at the time this investigation took place).

- Across all three health administrative data sources, the White British category consistently reported the highest level of agreement with the 2011 Census (greater than 96%), with South Asian, which includes Bangladeshi (greater than 91%), Pakistani (greater than 86%) and Indian (greater than 81%), and Chinese (greater than 81%) categories also reporting high agreement rates across these sources.

- Across all three health administrative data sources, agreement was lower for all Mixed ethnic groups (less than 67%) and other ethnic groups, including Other Asian (less than 60%), Other White (less than 55%), Other Mixed (less than 21%), Other Black (less than 16%) and Any Other ethnic group (less than 15%).

- For instances where health administrative data sources contained more than one ethnicity per person, separate analyses compared the most recent and most common ethnicity with the 2011 Census; overall agreement rates were similar for both methods.

- Results based on the subset of GDPPR ethnicity data available for analysis suggests that overall agreement rates are similar across primary care and hospital data, though there is some variation between ethnic groups.

2 . Work to improve estimates of ethnic health disparities

There is significant interest in understanding health inequalities and robust statistics on health outcomes for different ethnic groups because of increased importance during the coronavirus (COVID-19) pandemic. This is because people from minority ethnic groups were found to be at higher COVID-19 mortality risk, as highlighted in our Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 10 January 2022 to 16 February 2022 article. At the same time, the coronavirus pandemic exposed gaps in data available on ethnicity; researchers do not all have access to the same data sources, and the quality and completeness of ethnicity data vary across data sources, which can lead to differences in estimates.

Ethnicity data gaps are an important area for the health statistics system to focus on, as highlighted in the Office for Statistics Regulation's Improving health and social care statistics: lessons learned from the COVID-19 pandemic report. To help improve the comparability of estimates based on different sources, we are working with the Wellcome Trust and Race Equality Foundation. This will be to produce a series of studies exploring the quality of ethnicity data in important NHS databases, examining whether these quality issues may bias epidemiological and public health studies, and developing solutions to mitigate biases in the underlying data.

This article compares ethnicity information recorded in several NHS data sources with ethnicity information from the 2011 Census (Census 2021 data were not available for analysis at the time this investigation took place), which is widely regarded as the most robust ethnicity data source covering the whole population of England.

As part of the wider research programme we are also undertaking a desk review, exploring potential sources of error and bias in the process of collecting NHS ethnicity information, and coordinating focus groups to provide further insights from the public and healthcare staff on their experiences of ethnicity data collection. For more information, see our Methods and systems used to collect ethnicity information in health administrative data sources, England: 2022 article.

3 . Method for comparing ethnicity information across sources

Data Sources

This analysis compares ethnicity recorded in three health administrative data sources for England with ethnicity recorded for residents in England in the 2011 Census (the most recent census data available at the time of analysis). The data sources used are:

- a subset of the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), which contains information on all active patients registered at a GP practice in England on 1 November 2019; note that the GDPPR extract available to the Office for National Statistics (ONS) for this analysis contained incomplete ethnicity information (see Section 8 for more information on the subset and potential impact on the analysis)

- the Hospital Episode Statistics (HES), a database containing details of all attendances at NHS hospitals in England; it is made up of three sub-datasets: Accident and Emergency (A&E), Admitted Patient Care (APC), and Outpatients (OP)

- the NHS Digital Ethnic Category Information Asset (ECIA) – this source combines ethnicity data from GDPPR and HES, making it the most complete NHS source of ethnicity information for England

Although it is a decade out of date compared with the administrative data, 2011 Census data were used as a "gold standard" comparator. Apart from some exceptions (such as parents responding on behalf of children), we can generally be confident the ethnicity data in census are self-reported.

Ethnicity definitions within each data source

Ethnic categories vary across data sources, and so does the wording of categories even where they do align across data sources (see Table 1 in our accompanying dataset for more details). The 2011 Census, GDPPR and ECIA all have 18 ethnic categories (GDPPR and ECIA categories are based on the 2011 Census). By contrast, HES data only contains 16 ethnic categories; it does not include "White: Gypsy or Irish Traveller" or "Other ethnic group: Arab" categories. The HES categories were updated in April 2001 to represent the ethnic categories as defined in the 2001 Census. In the 2011 Census, the Chinese ethnic group moved from the "Other" ethnic group to the "Asian" ethnic group, and new groups for "Gypsy or Irish Traveller" and "Arab" were added. In all health administrative data sources, the Chinese ethnic group is still in the "Other" ethnic group. For more information, see GOV.UK's List of ethnic groups.

Handling multiple ethnicity records per person

The ECIA contains a single ethnicity per person, based on the most recent ethnicity recorded in either GDPPR or HES. NHS Digital have published full details of the [methodology used to create the ECIA](#). In contrast, HES and GDPPR datasets contain information about all interactions a patient has with the relevant health service, so generally contains multiple records per patient. In GDPPR and HES, some individuals have multiple recorded ethnicities within the same data source, so a set of rules were implemented to select a single ethnicity per person for comparison with the 2011 Census.

We tested two methods to derive an individual's ethnicity within GDPPR and HES sources. These were the most common (modal) and most recent (recency) ethnicity recorded for each person. All data made available to analysts were non-identifiable.

Recency classification was derived by selecting the most recently recorded ethnic category (even if this was "Not Known") to determine an individual's ethnicity. If there were multiple ethnicities recorded on the same most recent date, the records were prioritised according to the sub-dataset for HES, in order of HES-APC, HES-AE, HES-OP. This prioritisation order in HES is based on the completeness of data from each source, for more information see the [Journal of Public Health's Completeness and usability of ethnicity data in UK-based primary care and hospital databases article](#). If conflicts still existed on the same date for the same sub-dataset, the ethnicity was classified as unresolved.

For GDPPR, we derived the most recent and most frequently reported ethnicity definitions from the ethnicity [SNOMED codes](#) (the clinical coding standards used with GP records) listed in our GDPPR data, therefore no hierarchy was applied. The total number of SNOMED codes per person were identified and then a most recent ethnicity was identified. For both GDPPR and HES, a "Not Stated" ethnic category may be interpreted as a refusal. For this analysis, "Not Stated" was treated as a valid ethnic category, meaning if the most recent record was "Not Stated", this record was selected. This methodology is consistent with the one previously used in our [Producing admin-based ethnicity statistics for England: methods, data and quality article](#). The "Not Known" category, which is used in HES to denote missing information, was treated as a valid ethnicity. For GDPPR, there were no ethnicity SNOMED codes available in our data extract that identified a "Not Known" ethnic category. Therefore, there was no "Not Known" ethnic category within this dataset.

Modal classification was derived by selecting the most frequently reported ethnicity category per data source. If there was more than one most common ethnicity, the ethnicity was classified as "Unresolved". "Not Stated" and "Not Known" were treated as valid ethnic categories, meaning if the most common record was "Not Stated" or "Not Known", this record was selected.

Data Linkage

To enable comparisons of ethnicity recorded in each health administrative source with the 2011 Census, people enumerated in the 2011 Census were linked to the General Practice Patient Register to obtain the NHS number for each person enumerated in the census (with 94.6% of persons in the census probabilistically and deterministically matched to persons in the Patient Register). Our 2011 Census study population included 53.5 million people enumerated in England and Wales that we could obtain an NHS number for. We then excluded individuals who were residing in Wales at the time of the 2011 Census (2.9 million), those remaining in our study population in the 2011 Census who we could not obtain an NHS number for (3.3 million), and those with a recorded ethnic category of "No code" in the 2011 Census (0.3 million). Therefore, a total of 47.1 million individuals from England in the 2011 Census dataset were included in our analysis. For context, on Census Day 2011 the population of England was estimated to be 53.0 million.

We then were able to link individuals' ethnicity from each health data source to the 2011 Census using an NHS number. Note that all data made available to analysts for this study were non-identifiable.

Table 1: Count of people in the linked datasets created to compare consistency of ethnicity recording in health sources with the 2011 Census, England, 2011 to 2021

Linked dataset	Count of people in each linked dataset		Count of people in linked dataset with a stated ethnicity in both health and Census sources	
	Millions	As a percentage of the population of England on Census Day 2011	Millions	As a percentage of the population of England on Census Day 2011
Linked Census-ECIA	39.1	73.7%	39.1	73.7%
Linked Census-GDPPR modal	21.4	40.4%	20.5	38.6%
Linked Census-GDPPR recency	21.4	40.4%	21.0	39.6%
Linked Census-HES modal	43.3	81.7%	35.0	66.0%
Linked Census-HES recency	43.3	81.7%	34.2	64.5%

Source: Office for National Statistics – Consistency of ethnicity data recorded in health-related administrative datasets in England

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. GDPPR data extract used in this analysis is a subset of the complete GDPPR dataset.

For GDPPR and HES, we included ethnicity records recorded between 27 March 2011 (Census Day) and 30 September 2021. Linked totals for ECIA, HES and particularly GDPPR are lower than the number of 2011 Census records that were linked to an NHS number. This is because there are people within census who cannot link to each data source, or have missing ethnicity data within the health administrative data sources. This is particularly relevant for the GDPPR dataset, as the dataset used in this analysis is a subset of the overall GDPPR dataset. Our GDPPR data extract is based upon the ethnicity SNOMED cluster criteria that were requested, and therefore our GDPPR data reflect the SNOMED clusters that were received.

4 . Person-level cross tabulations

Person-level cross tabulations of each health administrative source with 2011 Census

To explore the consistency of ethnicity information across data sources, we produced 18- and 5-category ethnic group cross tabulations of 2011 Census with each health administrative data source. This enabled the examination of 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) and Hospital Episode Statistics (HES) sources, and the ethnic category an individual was assigned in the 2011 Census, both as a count and as a percentage. See Tables 4 to 13 for 18-category comparisons and Tables 14 to 23 for 5-category comparisons in our accompanying dataset.

5 . Person-level agreement rates

Person-level agreement rate in ethnicity coding in each health administrative data source with 2011 Census

To summarise the information contained in the cross tabulations, we calculated agreement rates for each health data source compared with the 2011 Census. For each person, the ethnic category recorded in the 2011 Census and each respective data source were compared and classified as either: "Agree" if the ethnic classification were the same, or "Disagree" if they were recorded as different. Where the ethnic categories used in the health administrative sources data did not exactly match with the 2011 Census categories, ethnic categories were matched with the most applicable 2011 Census ethnicity category. Only those with a stated ethnicity category in both data sources were included in the agreement rate calculations, with Not Stated, Not Known and Unresolved ethnic categories not included in agreement rate calculations. Arab and Traveller ethnic categories are not available within Hospital Episode Statistics (HES), and therefore no agreement rates were calculated for these ethnic categories in HES. Details on how the agreement rates were calculated can be found in the Methods tab of our accompanying dataset.

Table 2: Agreement rates by health data source in comparison with the 2011 Census, using 18-and 5-category ethnicity categories, England, 2011 to 2021

	Overall agreement rate for 18-category ethnic groups	Overall agreement rate for 5-category ethnic groups
ECIA	89.1	95.6
GDPPR-Modal	92.7	97.0
GDPPR-Recency	91.2	96.4
HES-Modal	90.7	95.8
HES-Recency	89.7	95.2

Source: Office for National Statistics – Consistency of ethnicity data recorded in health-related administrative datasets in England

Notes

1. Data are presented as percentage (%).

2. The percentages are based on individuals with a stated ethnicity in both the health data source and the 2011 Census; those whose ethnic information was "Not linked", "Not Stated", "Not Known" or "Unresolved" in either source were excluded.

3. GDPPR data extract used in this analysis are a subset of the complete GDPPR dataset.

Table 2 shows how overall agreement rates are high and similar across all sources and methods, ranging from 89.1% for the Ethnic Category Information Asset (ECIA) to 92.7% for the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR)-modal, for the 18-category ethnic groups. Our finding that agreement rates for HES against the 2011 Census are similar using both recency and modal methods is consistent with the findings in our Producing admin-based ethnicity statistics for England: changes to data and methods article. These findings used a combined administrative dataset, including HES ethnicity records up to 2016, and were compared with the 2011 Census ethnicity using both methods. For both GDPPR and HES, the agreement rate was slightly higher when using the modal method compared with recency, however this could be because of the 2011 Census comparator being recorded a decade ago (compared with modal, the recency method is better able to capture changes in self-identified ethnicity over time). It would be valuable to repeat this analysis using Census 2021 data when available.

Agreement rates by ethnic group

Figure 1 shows how across all data sources, the White British category consistently reported the highest level of agreement. South Asian (Bangladeshi, Indian and Pakistani) and Chinese categories also reported high levels of agreement. The ethnic category with the lowest agreement across the linked ECIA and GDPPR datasets was the Traveller category, which consistently reported the lowest levels of agreement. The Traveller ethnic group was not available within HES as it only includes 16 ethnic categories. The ethnic group with the lowest level of agreement within the linked HES datasets was the "Other: Any Other Ethnic Group". Agreement was generally lower for all Mixed or multiple ethnic groups and Other ethnic group categories.

While patterns of agreement with census were similar across all health administrative data sources, the GDPPR modal generally had the highest levels of agreement for each ethnic category, except for "Mixed: White and Black African" and "Mixed: White and Black Caribbean", where HES modal reported the highest levels of agreement with the 2011 Census. For most ethnic groups, agreement rates for HES were lower than GDPPR.

Figure 1: For all sources and methods, agreement rates were lowest (21% or less) for Traveller, Any Other Ethnic Group, Any other Black background and Any other Mixed background groups, and highest (92% or more) for White British and Bangladeshi groups, England, 2011 to 2021

Agreement rates between health datasets and 2011 Census ethnicity

Notes:

1. Agreement rates are based on linked individuals with a stated ethnicity on the relevant health dataset and the 2011 Census. 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, so agreement rates for these categories are only presented for ECIA, GDPPR recency and GDPPR modal.

4. GDPPR data extract used in this analysis are a subset of the complete GDPPR dataset.

Download the data

.xlsx

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 14 to 23 of the accompanying dataset). Patterns of results for the 5-category agreement rates were similar to that of the 18-category results, with White (greater than 98%) and Asian (greater than 92%) ethnic groups reporting high agreement with the 2011 Census across all health administrative data sources. The Black ethnic group reported higher (greater than 86%) agreement using the 5-category definitions compared with the 18-category definitions. Mixed (less than 57%) and Other (less than 27%) categories reported the lowest agreement with the 2011 Census for the 5-category definitions, with the Other category reporting particularly low agreement across all data sources.

6 . Consistency of ethnicity data recorded in health-related administrative datasets in England data

Comparing ethnicity data recorded in three health admin data sources (ECIA, GDPPR, HES) with the 2011 Census, England: 2011 to 2021
Dataset | Released 16 January 2023
Comparisons showing differences in the recording of ethnicity data between three health administrative data sources (GDPPR, ECIA, HES) and the 2011 Census.

7 . Glossary

Agreement rate

Of those records with a stated ethnicity in the health administrative data source and the 2011 Census, the percentage of linked records where the ethnicity in the health administrative data source and the 2011 Census 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 Hospital Episode Statistics (HES) and General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) data sources, an individual can choose to not identify their ethnic group. The code "Z – Not Stated" is recorded.

Ethnicity not known

In 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 (and for HES, a dataset hierarchy of Admitted Patient Care, Accident and Emergency, and Outpatients did not resolve the conflict), these have been coded as "unresolved".

Not linked

This refers to individuals who have a stated ethnicity in the 2011 Census but could not be linked to the administrative data source, regardless of whether they had a stated ethnicity in the HES, GDPPR or Ethnic Category Information Asset (ECIA) data sources.

SNOMED code

SNOMED codes are the clinical coding standards used with GP records. Further information about SNOMED codes and how ethnicity is recorded within different fields and tables within the GDPPR and HES datasets can be found in NHS Digital's GPES data for pandemic planning and research and Hospital Episode Statistics page.

8 . Data sources and quality

Person-level comparisons

This study created novel linked data to enable person-level comparisons between ethnicity information recorded in three health administrative data sources, and ethnicity recorded in the 2011 Census. At the time of publication, the 2011 Census is widely regarded as the most robust source of ethnicity information covering the whole population in England. This was the first-time ethnicity data from the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) and the Ethnic Category Information Asset (ECIA) sources have been linked to census at person-level to assess the accuracy of ethnicity recording in these data sources.

2011 Census as a comparator

Approximately 94% of the 2011 population of England and Wales are included in the 2011 Census. This large study population enabled detailed analysis at a granular level for the differences in ethnicity recording between health administrative data sources and the census 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.

However, it is noted 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. Proxy reporting does not only affect census data; the health data sources are likely to also contain some proxy responses affecting the comparisons. In addition, it was not possible to identify ethnicity information that had been imputed during data processing for those who responded to the census but did not provide a valid response to the ethnicity question. The number of respondents in England who had their ethnicity imputed in the 2011 Census was 1.5 million (3.1%) out of 50.0 million. For more information, see our Response and imputation rates methodology.

At the time of conducting this analysis, Census 2021 data had been collected but were not yet available for analysis. It is noted that the ethnicity variable used for the 2011 Census was 10 years old at the time of this analysis. While ethnicity is somewhat less likely to change over time than other sociodemographic factors (for example, occupation), self-reported ethnicity may change with time and age, which may introduce some bias. However, previous analysis of ethnicity change from the 2001 to 2011 Census have shown that those from White, Black or South Asian backgrounds, which are the largest ethnic groups within the UK, may be less likely to change their ethnicity over time. Those from Mixed or Other ethnic categories are more likely to change over time and have more than one ethnicity listed in their records. For more information, see the Royal Statistical Society's The stability of ethnic identity in England Wales 2001-2011 article.

Health data sources and methods

The analysis includes three important health administrative data sources widely used for health analyses. However, only a subset of GDPPR data were available for analysis. This may bias our results because of our subset potentially not being representative of the entire GDPPR dataset in terms of ethnic category agreement to the 2011 Census.

We were able to replicate methods used by other analysts to derive a single ethnicity per person from health administrative data sources. This enabled us to draw conclusions about the comparability of ethnicity information used by other analysts and as an extension, were able to test the impact of modal and recency ethnicity definitions on accuracy. For more information on the methods replicated, see NHS Digital's git repository.

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. Further, 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

As the Not Stated category can be interpreted as a refusal, in line with the methodology used in our previous Producing admin-based ethnicity statistics for England: methods, data and quality article, we classified the Not Stated ethnicity in GDPPR and HES data sources as a valid ethnicity. Treating the Not Stated category in this manner may introduce bias in our agreement rates, which could differ if Not Stated was treated in the same way as missing data.

Separate linked datasets were created for each data source, as well as for recency and modal methods. This enables agreement rates to be calculated on the largest possible data source. However, the demographic characteristics of the linked data sources may vary, which could explain some of the observed differences in agreement rates.

GDPPR

General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) is a data extract from the General Practice Extraction Service created in response to the coronavirus (COVID-19) pandemic. The data contain information on all active patients registered at a GP practice in England on 1 November 2019. The GDPPR data extract available for this analysis contained incomplete ethnicity information, which may introduce bias to our results. However, we were able to derive an ethnic category for 32.0 million people in GDPPR prior to linkage to the 2011 Census. The version of GDPPR available for this analysis had 254 unique SNOMED codes for ethnicity out of a total of 489 listed on the NHS Digital git repository. The list of the 254 SNOMED codes available for this analysis can be found in Table 3 of our accompanying dataset.

HES

Hospital Episode Statistics (HES) is a database containing details of all attendances at NHS hospitals in England. Patients' ethnicity is collected while they are in hospital and the guidance for hospitals states that the patient's ethnicity should be self-classified.

Ethnic Category Information Asset

NHS Digital created an Ethnic Category Information Asset (ECIA), which is an amalgamation of GDPPR and HES; combining ethnic category data from these sources increases coverage of patient ethnicity data compared with either source individually. The ECIA has published details on the order it prioritises ethnicity data from GDPPR and HES sources, with prioritisation in the following order:

- GDPPR-Journal

- GDPPR-Patient

- HES-APC

- HES-AE

- HES-OP

While the census is classified as being the "gold standard" for ethnicity, the ECIA was created because the 2011 Census data was nearly 10 years old (potentially no longer reflecting the ethnic breakdown of the current population) and cannot be shared with NHS Digital. The ECIA provides a near population (England only) level view of ethnic category.

As the ECIA is an amalgamation of HES and GDPPR sources, Traveller and Arab ethnic categories are included.

9 . Future developments

This work is part of a wider programme investigating ethnicity consistency between data sources. These results will inform future research to assess the bias in estimates of mortality risk based on ethnicity recorded in health administrative sources, and to develop solutions to mitigate biases. Additional future work could include repeating this analysis with Census 2021 data, and a more complete extract of General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) data.

10 . Related links

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.

11 . Cite this article

Office for National Statistics (ONS), released 16 January 2023, ONS website, article, Understanding Consistency of ethnicity data recorded in health-related administrative datasets in England: 2011 to 2021