Office for National Statistics

Article

Producing admin-based ethnicity statistics for England: changes to data and methods

An overview of the changes to data sources and methods in the feasibility research on producing statistics on the population by ethnic group from administrative data.

Contact:
Alison Morgan
admin.based.characteristics@ons.gov.uk
+44 1329 447 187

Release date:
23 May 2022

Next release:
To be announced

Table of contents

1 . Main points

- We have produced a new set of admin-based ethnicity statistics for 2016 to 2020 as part of our feasibility research.
- We have incorporated three additional administrative data sources, which are the Higher Education Statistics Agency (HESA), Birth Notifications, and the Emergency Care Data Set (ECDS), and linked on additional Hospital Episode Statistics (HES) and English School Census (ESC) data.
- We have changed our method for dealing with multiple recorded ethnicities for an individual to include an additional step in the process for those with a most recently recorded ethnicity of Any other ethnic group; this affected a small proportion of people with a stated ethnicity (1.0% in 2016) but improved the admin-based ethnicity statistics.

2 . Overview

In August 2021, we published findings from our initial feasibility research on producing statistics on the population by ethnic group for England from administrative data. The research was based on linking ethnicity data from Hospital Episode Statistics (HES), English School Census (ESC) and Improving Access to Psychological Therapies (IAPT) to a 2016 admin-based population base and implementing a set of rules to deal with multiple recorded ethnicities for an individual. A set of admin-based ethnicity statistics were produced for 2016 based on the proportion of people in each ethnic group; we refer to these as version 1. Full details of the previous method can be found in our methods article published alongside the research outputs.

Our initial admin-based ethnicity statistics showed early promise, but had the following issues:

- the proportion of people with a stated ethnicity varied greatly by age, sex and local authority
- the proportion of people in the Other ethnic group was much higher in the admin-based ethnicity statistics than was expected based on comparisons with the 2011 Census
- there was under-representation of the Asian, Black and Mixed ethnic groups in those aged 20 to 24 years, those aged 25 to 29 years, and those aged 30 to 34 years

This article provides information on the work we have been doing to improve our method and accompanies two other articles. One article presents a new set of admin-based ethnicity statistics for 2016 (version 2) and compares them with those produced in version 1. The other article presents admin-based ethnicity statistics for 2016 to 2020, based on version 2, and explores population coverage and change in ethnicity over time.

3 . Incorporating additional data sources

Higher Education Statistics Agency

The Higher Education Statistics Agency (HESA) student record is a census-based individualised data collection covering students at publicly funded higher education institutions and the University of Buckingham. Coverage extends to all students undertaking a course or programme leading to a qualification or institutional credit, unless studying overseas for the duration of their course. There are around three million students in each annual HESA dataset.

To produce the admin-based ethnicity statistics, we combined HESA data from the year ending July 2011 (the earliest year available) up to the year of interest. The combined dataset for 2016 included 7.9 million people, rising to 11.8 million people for the 2020 dataset. However, not all HESA records could be linked to the admin-based population estimates (ABPE) v3.0; 3.4 million linked to the 2016 ABPE and 5.3 million linked to the 2020 ABPE. There are four main reasons for this:

1) HESA data cover the UK, but we are only covering people usually resident in England.

2) Some people who attended university in the UK from 2010 onwards will have emigrated or died prior to the year of interest.

3) Anyone in the ABPE who attended university exclusively between 2011 to 2012 and 2014 to 2015 will not have been linked to their HESA record because HESA data have only been linked to other admin datasets for 2010 to 2011 and for 2015 to 2016 onwards; it is estimated that 2.1 million people are affected by this.

4) We have only been able to link individuals in the 2010 to 2011 HESA data to the ABPE if their HESA record had been successfully linked to the Patient Register and the ABPE had a record of their NHS number; it is estimated that 1.5 million people are affected by this.

We will look to address points 3 and 4 in future iterations of the research.

Ethnicity categories

The ethnicity categories used in the HESA data collection vary across the countries of the UK and have changed over time (see the accompanying dataset). Ethnicity is self-reported by the student, and providers are expected to reconfirm the ethnicity with the student at least annually. The HESA data do not have the full breakdown of the White ethnic group as per the GSS harmonised standard. When combining the HESA data with the other administrative datasets, code 10 (White) was re-coded as White not specified. The Arab response option was introduced for the academic year ending in July 2013.

From August 2022, the ethnicity categories used in the HESA data collection are changing to align with the latest census data collections in each country of the UK.

The completeness of the ethnicity field has slightly improved over time, with the proportion of records with an unknown ethnicity decreasing from a high of 8.4% in 2013 to 2014 to 6.4% in 2019 to 2020. The level of refusals has also decreased, from 3.5% in 2010 to 2011 to 1.8% in 2019 to 2020.

Comparisons with the 2011 Census

As was done for other administrative data sources in the previous publication, HESA data from 1 August 2010 to 27 March 2011 (Census Day) were linked to the 2011 Census to compare the ethnicities recorded for individuals. HESA had slightly higher agreement rates with the 2011 Census data than the other data sources across all ethnic groups. Like the other data sources, HESA had the highest agreement rate with the 2011 Census for the White ethnic group (99.3%) and lowest for the Other ethnic group (31.4%).

Table 1: Ethnicity recorded in the HESA data compared with the 2011 Census by five-category ethnic group, England

		2011 Census				
		Asian	**Black**	**Mixed**	**White**	**Other**
HESA	**Asian**	96.4	0.2	0.7	0.7	2.0
	Black	0.2	96.6	2.1	0.8	0.4
	Mixed	3.8	4.2	77.5	11.0	3.5
	White	0.1	[low]	0.4	99.3	0.2
	Other	21.6	6.3	9.3	31.5	31.4

Source: Office for National Statistics

Notes

1. Proportions may not sum to 100% because of rounding.

2. HESA ethnic group totals have been used as the denominators when calculating the percentages. ,The percentages are based on linked individuals with a stated ethnic group on the 2011 Census and in HESA data. ,[low] denotes that the proportion is less than 0.05.

Birth Notifications

The Birth Notifications data contain information about babies born in England, Wales, and the Isle of Man. The data are collected by a midwife or other medical professional within 36 hours of a birth. Ethnicity is reported by the mother.

To produce the admin-based ethnicity statistics, we combined data from 1 January 2006 (the earliest year available) up to 30 June of the year of interest. There are around 700,000 births recorded in the Birth Notifications data each year, which total to 10 million records from 1 January 2006 to 30 June 2020. Of these, 8.7 million were linked to the 2020 ABPE. As most births from 1 April 2009 onwards are captured in the Hospital Episode Statistics (HES) data and most children appear in the English School Census (ESC) data once they reach school age, only 36,000 individuals in the 2020 ABPE appeared in the Birth Notifications data only.

Ethnicity categories

The Birth Notifications data contain 105 different ethnicity codes (excluding not known and not stated), which we aggregated to 17 ethnic groups (see the accompanying dataset). There is not an Arab category in the Birth Notifications data.

The proportion of records in the Birth Notifications data with an unknown ethnicity is very low, at less than 1.0% across all years. The level of refusals has decreased over time, from 10.0% in 2006 to between 3.0% and 4.5% for all years from 2010 onwards.

Comparisons with the 2011 Census

The Birth Notifications data for 1 January 2006 to 27 March 2011 (Census Day) were linked to the 2011 Census and ethnicities compared at record level. Agreement rates by ethnic group were similar to the other administrative data sources.

Table 2: Ethnicity recorded in the Birth Notifications data compared with the 2011 Census by five-category ethnic group, England

		2011 Census				
		Asian	**Black**	**Mixed**	**White**	**Other**
Birth Notifications	**Asian**	93.0	0.3	3.5	1.6	1.5
	Black	0.7	87.1	8.9	2.2	1.1
	Mixed	3.6	5.4	69.5	20.3	1.2
	White	0.3	0.1	2.2	97.3	0.2
	Other	23.6	5.9	18.5	43.0	8.9

Source: Office for National Statistics

Notes

1. Proportions may not sum to 100% because of rounding.
2. The Birth Notifications ethnic group totals have been used as the denominators when calculating the percentages.
3. The percentages are based on linked individuals with a stated ethnic group on the 2011 Census and Birth Notifications.

Emergency Care Data Set

The Emergency Care Data Set (ECDS) contains information about people who have attended emergency departments in England. The ECDS was introduced in April 2019 to replace the A&E part of the Hospital Episode Statistics (HES) data. From April 2019 to March 2020, both datasets were available, and from April 2020, only the ECDS was.

For the admin-based ethnicity statistics, we used ECDS data for 1 April to 30 June 2020, which included 2.6 million people. Of these, 2.4 million were linked to the ABPE.

As with the Birth Notifications data, the ECDS ethnicity codes were aggregated (see the accompanying dataset), but this time to 16 ethnic groups. The ECDS does not have response options for the Gypsy, Roma, Irish Traveller or Arab ethnic groups. In the ECDS data for 1 April to 30 June 2020, 7.1% of records had an unknown ethnicity and 6.0% had the ethnicity refused.

Hospital Episode Statistics (HES)

The 2016 ABPE has a reference date of 30 June 2016. However, in the first iteration of the admin-based ethnicity statistics feasibility research, we only included HES data up to 31 March 2016, which is the end date for the 2015 to 2016 annual dataset. The main impact of this was on those aged under one year, as babies born between 1 April and 30 June 2016 were captured in the population base but not linked to a HES record.

Since 1 April 2020, we have been receiving a monthly supply of HES data. This means that we can incorporate HES data for the April to June period each year without having to wait for the next annual supply. For the 2020 admin-based ethnicity statistics in the time series publication, we incorporated HES monthly data for 1 April to 30 June 2020. For 2016 to 2019, we simulated the monthly data by linking on HES records from the next annual dataset with a reference date between 1 April and 30 June. This resulted in the proportion of those aged under one year in the ABPE linking to the HES data increasing to 96.8% for 2016, which is an increase of nearly 25 percentage points.

The level of unknown ethnicities in HES decreased dramatically between 2009 to 2010 (84.8%) and 2012 to 2013 (7.1%) and has been fairly similar since. The level of refusals has increased over time, from 1.0% in 2009 to 2010 to 10.3% in 2019 to 2020.

English School Census additional links

In the initial feasibility research, despite having ESC data for earlier years, we were only able to link ESC records to the ABPE if the individual was in school in January 2016. There are still limitations on the linkage that we can do, but we have linked an additional 2.6 million ESC records to the ABPE using a 2011 Patient Register to ESC linked dataset. This linkage has significantly increased the proportion of people aged 17 to 21 years in 2016 linked to one of the admin-based ethnicity data sources.

The level of completeness of the ethnicity information in the ESC data has been consistently high. For all years, less than 2.0% of records had an unknown ethnicity and less than 0.6% had a refusal.

Incorporating the 2011 Census

In addition to producing the admin-based ethnicity statistics using administrative data only, we have produced a set of figures based on incorporating the 2011 Census as an additional data source. This is because we want to make the best use of all available data sources and the 2011 Census is the most complete source of ethnicity data as at Census Day. It also demonstrates what may be possible in future using Census 2021 data to ensure we maximise the utility of this rich data source.

Incorporating 2011 Census data means that for anyone not in the administrative data post-2011 Census, whose 2011 Census record we successfully managed to link to the ABPE, we took their ethnicity from the 2011 Census. If someone's ethnicity was unknown in their post-census admin data, their 2011 Census ethnicity would also be used. By incorporating the 2011 Census, an ethnicity was recorded for an additional 5.4 million people in 2016 (10.0% of people in the ABPE).

4 . Exploring alternative ethnicity selection rules

As people may appear multiple times within and across the administrative data sources, sometimes with different ethnicities recorded for them, we implemented a rules-based approach for selecting one ethnicity per person in our initial research. The original approach was to take the most recently recorded ethnicity for an individual, with additional rules for refusals, unknowns and multiple recorded ethnicities on the latest date. However, this resulted in the proportion of people in the Other ethnic group being much higher than was expected based on comparisons with the 2011 Census data. This could reflect changes in the population over time, but research by the Nuffield Trust on ethnicity coding in English health services datasets identified overuse of the Any other ethnic group code. As 80.4% of our ethnicity records are from health datasets, this overuse is likely to be contributing to our higher-than-expected proportion of people in the Other ethnic group.

We have trialled alternative approaches for dealing with multiple recorded ethnicities for an individual and have implemented an additional step in the process for those with a most recently recorded ethnicity of Any other ethnic group (Figure 1). This affected 1.0% of people with a stated ethnicity in the 2016 admin-based population estimates (ABPE).

The new approach is still largely based on taking the most recently recorded ethnicity for an individual but includes some additional steps. The first step is to link all admin data ethnicity records for an individual to the ABPE.

The next step is to select the ethnicity on the most recent date, except:

- if the ethnicity on the most recent date is unknown, in which case you select the last stated ethnicity or refusal where available, or otherwise code their ethnicity as unknown
- if an individual refused to provide an ethnicity on the most recent date, in which case you code their ethnicity as refused, regardless of whether they have previously provided their ethnicity
- if there are multiple stated ethnicities recorded on the latest date, in which case you code the ethnicity as unresolved

If the most recently stated ethnicity is Any other ethnic group and an individual has at least one alternative ethnicity recorded, take the next most recent. This means that:

- if this next ethnicity is a stated ethnicity (not unknown or refused), take this ethnicity
- if this next ethnicity is refused, keep Any other ethnic group as the ethnicity
- if this next ethnicity is unknown, see whether there are ethnicities prior to that and follow the approach for either a stated ethnicity or a refused ethnicity
- if this next ethnicity is a conflict (multiple ethnicities on the same date), keep Any other ethnic group as the ethnicity

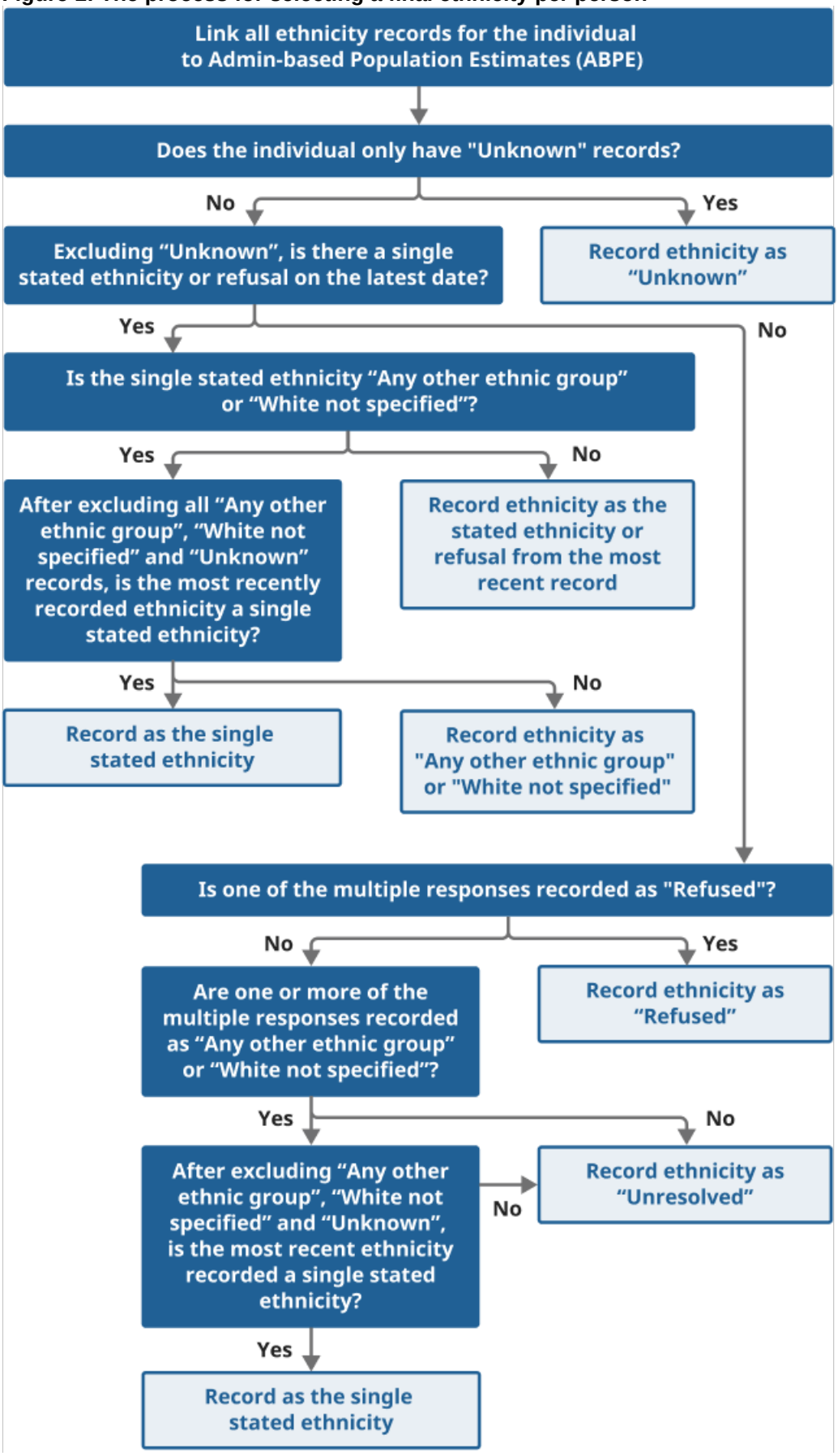
If the most recent ethnicity is a conflict of ethnicities, and one of the ethnicities is Any other ethnic group, there are two options. In this case:

- if removing the Any other ethnic group record resolves the conflict, assign the non-Any other ethnic group stated ethnicity as their new ethnicity

- if removing the Any other ethnic group record does not resolve the conflict, code their ethnicity as unresolved

As outlined in Section 3, the HESA data collection for England and Wales does not have the full breakdown of the White ethnic group as per the GSS harmonised standard. For those with White not specified as their most recently recorded ethnicity, the next most recent ethnicity has been considered following a similar process as for Any other ethnic group.

Figure 1: The process for selecting a final ethnicity per person



Source: Office for National Statistics

Any other ethnic group

After implementing the new ethnicity selection rules for the 2016 admin-based ethnicity statistics, the proportion in the five-category Other ethnic group decreased from 2.1% to 1.2%, which is much closer to the 1.0% in the 2011 Census. For most local authorities, the proportion in the Other ethnic group was closer to the 2011 Census proportion after implementing the new ethnicity selection rules than when using the original approach (Figure 2).

Figure 2: The admin-based ethnicity statistics for the Other ethnic group are closer to the 2011 Census estimates after implementing the new ethnicity selection rules

Proportion of people in the Other ethnic group in each local authority in the 2016 admin-based ethnicity statistics produced using the new and original ethnicity selection rules, versus the 2011 Census, England

Notes:

1. The admin-based ethnicity proportions have been calculated out of those with a stated ethnicity.
2. Local authority boundaries are as of 2021.
3. Both sets of figures are based on using Hospital Episode Statistics, English School Census, Improving Access to Psychological Therapies, Higher Education Statistics Agency and Birth Notifications data.

Download the data

.xlsx

We linked the 2016 admin-based ethnicity dataset to the 2011 Census to explore the census ethnicities for those with an assigned ethnicity of Any other ethnic group using the original approach, but a different ethnicity using the new ethnicity selection rules. Of those where we were able to establish a 2011 Census ethnicity:

- the new ethnicity matched the 2011 Census for 64.6% of people
- 6.8% were recorded as being in the Any other ethnic group on the 2011 Census
- the new ethnicity did not match the 2011 Census for 28.5% of people, but they were also not recorded as being in the Any other ethnic group in the 2011 Census

This analysis suggests that the new ethnicity selection rules improve the accuracy of the admin-based ethnicity statistics. We therefore decided to implement it for our latest set of admin-based ethnicity statistics research outputs and welcome feedback on this.

Further tables comparing the admin-based ethnicity statistics using the original and new approaches are available in the accompanying dataset.

White not specified

For 2016, the additional step in the ethnicity selection process meant that we were able to assign a more specific ethnicity to 47.9% of people with White not specified as their most recently recorded ethnicity. Of those, 99.2% were assigned to a White ethnicity, mostly to White British (92.3%) or White Other (6.3%). Of those that we were able to link to a 2011 Census record, the new ethnicity matched the 2011 Census ethnicity in 95.4% of cases.

Treatment of refusals

In addition to trialling additional rules for those with a most recently recorded ethnicity of Any other ethnic group, we also tested out two other options for changing the ethnicity selection rules.

Our original approach for refusals was to record the individual's ethnicity as refused if they refused to provide their ethnicity on the latest date. We outlined in the previous publication that there were 2.7 million people with a final ethnicity of refused who had a stated ethnicity previously recorded in the admin data. We explored the impact of assigning these people to their last stated ethnicity prior to the refusal. We found that it made minimal difference to the proportion of people in each ethnic group at national level, or by age, sex or local authority.

We plan to do further research and public acceptability testing on the treatment of refusals. In the meantime, we have decided to continue to record the final ethnicity as refused for these people.

Most frequent

The Office for Health Improvement and Disparities use ethnicity data from Hospital Episode Statistics for coronavirus (COVID-19) analysis. Their method for dealing with multiple recorded ethnicities for an individual uses the most frequently recorded ethnicity rather than the most recent, but with the second most frequent taken for those with a most frequent ethnicity of Any other ethnic group.

We explored using the most frequent approach and found that the proportions of people in each ethnic group were similar to those when using the most recent approach at national level and by age, sex and local authority. The most recent and most frequent approaches had similar levels of agreement with the 2011 Census at record level. As the most recent approach would pick up changes in identity more quickly, we have decided to continue with this as the basis for our approach. We will review this once Census 2021 data are available.

5 . Glossary

Ethnic group

The self-reported ethnic group of the individual, according to their own perceived ethnic group and cultural background.

Ethnicity refused

In the English School Census (ESC), it is recorded as "refused" if a parent or guardian, or pupil has declined to provide ethnicity data. In Hospital Episode Statistics (HES), the Emergency Care Data Set (ECDS), Birth Notifications and Improving Access to Psychological Therapies (IAPT) data, where a patient chooses not to state their ethnicity, the code “Z - Not Stated” is recorded. In the Higher Education Statistics Agency (HESA) data, the code "98 Information Refused" is recorded.

Ethnicity stated

Ethnicity stated refers to the ethnicity being recorded as a specific ethnic group and not refused or unknown.

Ethnicity unknown

In the ESC data, where the ethnicity has not yet been collected, this is recorded as NOBT (information not yet obtained). In HES, ECDS, IAPT and Birth Notifications data, the default code “99 Not known” is used where the person's ethnicity is unknown. All blank and null ethnicity values in Birth Notifications were also treated as unknown. In HESA data, “90 Not known” is used.

In this article, the unknown category also includes individuals with multiple recorded ethnicities where the rules did not lead to a final ethnicity being selected. These have been termed “ethnicity unresolved”.

Ethnicity unresolved

Where multiple ethnicities were recorded on the latest date, these have been coded as “unresolved” and grouped into the “unknown” category for the analysis in this article.

Not linked

This refers to individuals who are in the admin-based population estimates (ABPE) v3.0 but have not been linked to any sources of ethnicity data.

Usually resident

As defined in our latest ABPE publication, we are currently adopting the UN definition of "usually resident". This is the place at which a person has lived continuously for at least 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months (United Nations, 2008).

Version 1

Version 1 refers to the admin-based ethnicity statistics produced using HES, IAPT and ESC data and published in August 2021.

Version 2

Version 2 refers to the admin-based ethnicity statistics produced using HES, ECDS, IAPT, ESC, HESA and Birth Notifications data and with the new ethnicity selection rules.

6 . Future developments

The admin-based ethnicity statistics are still in the development phase, and we will continue to explore how we can further improve them. The planned next steps include:

- incorporating additional data sources to improve the population coverage for England and expand coverage to Wales
- exploring the potential to produce multivariate statistics on ethnicity by other characteristics
- exploring methods to adjust for missingness in the admin data
- exploring the potential to produce admin-based ethnicity statistics for smaller geographic areas
- engaging with data suppliers to better understand and improve data collection practices
- combining the administrative data with survey data using the Generalised Structure Preserving Estimator (GSPREE), building on previous work using this method
- conducting public acceptability testing on ethnicity selection methods
- using Census 2021 data to further assess the quality of the admin-based ethnicity statistics

Feedback

We welcome feedback on the method used to produce the admin-based ethnicity statistics and the planned future developments. Please email your feedback to Admin.Based.Characteristics@ons.gov.uk.

7 . Related links

Developing admin-based ethnicity statistics for England: 2016
Article | Released 23 May 2022
Update on feasibility research on producing statistics on the population by ethnic group for England from administrative data, following the inclusion of additional data sources and method changes. These research outputs are not official statistics.

Change over time in admin-based ethnicity statistics, England: 2016 to 2020
Article| Released 23 May 2022
Feasibility research on producing statistics on the population by ethnic group for England for 2016 to 2020 from administrative data, exploring change over time. This builds on work to improve the 2016 admin-based ethnicity statistics. These research outputs are not official statistics.

Admin-based ethnicity statistics for England, feasibility research: 2016
Article | Released 6 August 2021
The Office for National Statistics (ONS) does not currently produce annual statistics by local authority on the population by ethnic group and the last official statistics were from the 2011 Census. This feasibility research combines ethnicity data from English School Census, Hospital Episode Statistics and Improving Access to Psychological Therapies to explore whether administrative data can be used to produce statistics on the population by ethnic group for 2016 at national and local authority level for England, by five-year age group and sex. These research outputs are not official statistics.

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.