

Article

Producing admin-based ethnicity statistics for England: methods, data and quality

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.

Contact:
Alison Reynolds
admin.based.
characteristics@ons.gov.uk
+44 (0)1329 447187

Release date:
6 August 2021

Next release:
To be announced

Notice

7 February 2023

This is not the latest research update. For the latest update on our admin-based ethnicity statistics feasibility research, please see our [Developing admin-based ethnicity statistics for England and Wales: 2020 article](#).

Table of contents

1. [Main points](#)
2. [Data used to produce the admin-based ethnicity statistics](#)
3. [Method used to produce the admin-based ethnicity statistics](#)
4. [Record-level comparisons with the 2011 Census](#)
5. [Glossary](#)
6. [Data sources and quality](#)
7. [Future developments](#)
8. [Related links](#)

1 . Main points

- To produce the [admin-based ethnicity statistics](#), the admin-based population estimates (ABPE) V3.0 dataset was used as the population base and ethnicity data linked on from the Hospital Episode Statistics (HES), English School Census (ESC) and Improving Access to Psychological Therapies (IAPT) administrative data sources.
- As people may appear multiple times within the administrative data sources, a set of rules was implemented to select one ethnicity per person.
- If an individual refused to state their ethnicity on the most recent data collection, we recorded their final ethnicity as refused rather than taking their last stated ethnicity; this affected 2.7 million individuals (5.0% of the ABPE).
- Of those with a stated ethnicity in the admin-based ethnicity statistics, 9.9% had more than one ethnicity recorded for them in the administrative data.
- To assess data quality, we conducted record-level comparisons between the administrative data and 2011 Census; the ethnicities matched for 90.7% of linked HES records, 92.4% of linked ESC records and 93.3% of linked IAPT records.
- There is substantial variation in the agreement rates by ethnic group between the administrative data and 2011 Census, with the White ethnic group having the highest agreement rate across all three administrative data sources and the Other ethnic group the lowest.
- Based on the linked 2011 Census-administrative data, refusal rates in the administrative data appear to be similar across the ethnic groups, suggesting that refusals should not have a substantial impact on the representativeness of the admin-based ethnicity statistics; however, this will be reviewed once Census 2021 data are available for analysis.

2 . Data used to produce the admin-based ethnicity statistics

Population base

The [2016 admin-based population estimates V3.0 dataset \(ABPE\)](#) was used as the population base for our analysis. The ABPE was created by combining multiple administrative data sources, including:

- Benefits and Income Datasets
- NHS Patient Register and Personal Demographic Service
- Higher Education Statistics Agency
- English School Census
- Welsh School Census
- Births Registrations

The ABPE is a record-level dataset, with individuals included if they met one or more "activity-based" rules, meaning they were deemed to be part of the usually resident population. The ABPE has undercoverage overall but some overcoverage, including for children under one year and those of school age. The [quality of the population base](#) will affect the quality of the admin-based ethnicity statistics, particularly if the level of coverage differs by ethnic group. More information about the coverage of the ABPE can be found in [this article](#).

As the [feasibility research](#) was conducted for England only, the ABPE was subset to individuals resident in England.

Administrative data sources

Three administrative data sources were used in our admin-based ethnicity statistics [feasibility research](#).

[Hospital Episode Statistics \(HES\)](#) is 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). For the [feasibility research](#), we used HES data covering the period 1 April 2009 to 31 March 2016. The patient's ethnicity is collected while they are in hospital and the [guidance for hospitals](#) states that the patient's ethnicity should be self-classified other than for specific exceptions.

[English School Census \(ESC\)](#) is a statutory data collection about pupils in Local Authority maintained primary, secondary, nursery and special schools in England. For the [feasibility research](#), we used data collected in January each year for the period 2011 to 2016. [Guidance for schools](#) states that the pupil's ethnicity must come from the parent/guardian or pupil and the school must not ascribe an ethnicity to a pupil.

[Improving Access to Psychological Therapies \(IAPT\)](#) is a dataset containing individuals accessing NHS psychological therapies in England. For the [feasibility research](#), we used data from 1 April 2012 to 31 March 2016. The [IAPT guidance](#) states that an individual's ethnicity must be obtained by asking the patient.

HES has the highest coverage of the three datasets overall, with 82.0% of ABPE records linked to a HES record. Coverage is generally higher for females than males, which is likely due in part to hospital attendances for pregnancy appointments and childbirth. Coverage is lowest for males in their 20s and 30s at 69.8%. This is likely because of the relatively good health of individuals in this age range.

Of the three data sources, ESC has the highest coverage for children, with 92.9% of children aged 5 to 15 years in the ABPE linked to an ESC record.

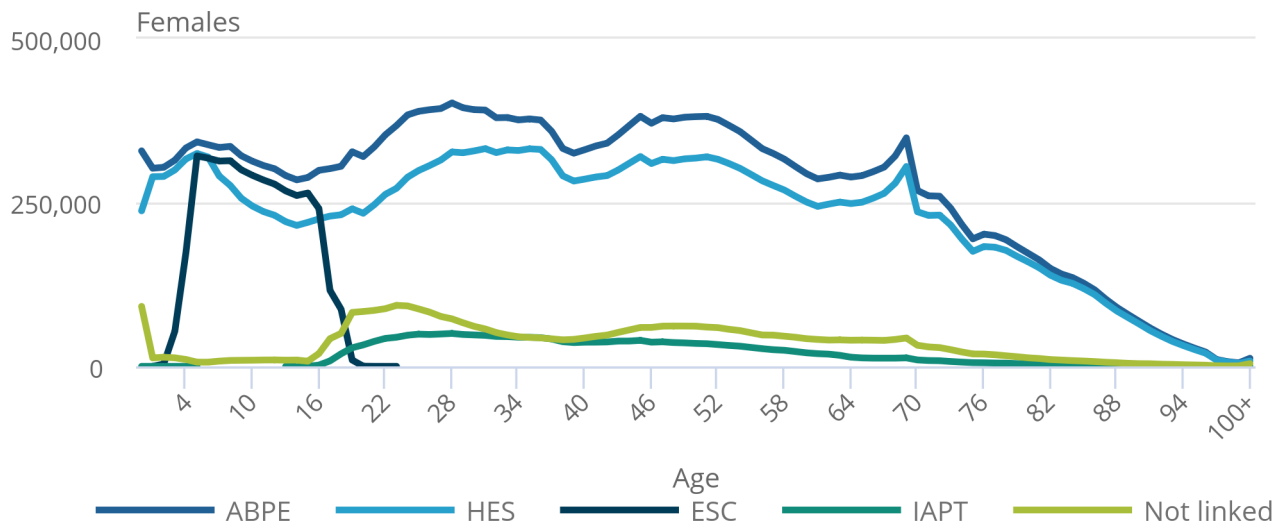
IAPT is the smallest dataset and therefore contributes the least to the overall coverage of the administrative data. However, around 176,000 individuals in the ABPE linked to IAPT but not HES or ESC so we were able to get ethnicity data for additional individuals through including it in the analysis. IAPT has higher coverage for females than males, with its highest coverage being for females in their 20s and 30s at 11.9% of ABPE records in this age range.

Figure 1: HES has the highest overall population coverage for females, but ESC has the highest coverage for school-aged children

Number of females by age linked to each administrative data source based on the ABPE as the population base, England, 2016

Figure 1: HES has the highest overall population coverage for females, but ESC has the highest coverage for school-aged children

Number of females by age linked to each administrative data source based on the ABPE as the population base, England, 2016



Source: Office for National Statistics

Notes:

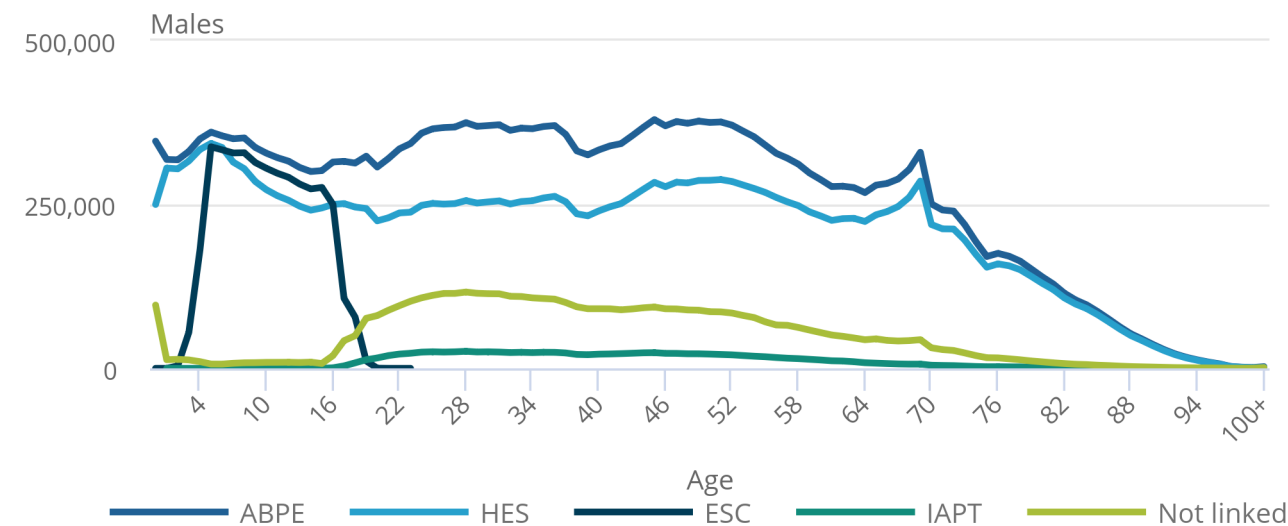
1. Not linked refers to individuals in the ABPE who have not been linked to HES, ESC or IAPT.
2. HES stands for Hospital Episode Statistics, ESC stands for English School Census, IAPT stands for Improving Access to Psychological Therapies and ABPE stands for admin-based population estimates.
3. In the data download, c means that the data have been suppressed for confidentiality reasons.

Figure 2: HES has the highest overall population coverage for males but this is lower for males in their 20s and 30s than males of other ages

Number of males by age linked to each administrative data source based on the ABPE as the population base, England, 2016

Figure 2: HES has the highest overall population coverage for males but this is lower for males in their 20s and 30s than males of other ages

Number of males by age linked to each administrative data source based on the ABPE as the population base, England, 2016



Source: Office for National Statistics

Notes:

1. Not linked refers to individuals in the ABPE who have not been linked to HES, ESC or IAPT.
2. HES stands for Hospital Episode Statistics, ESC stands for English School Census, IAPT stands for Improving Access to Psychological Therapies and ABPE stands for admin-based population estimates.
3. In the data download, c means that the data have been suppressed for confidentiality reasons.

Ethnic groups

The ethnic group response options differ between sources (Table 1), with the administrative data sources not aligned with the [GSS harmonised standard](#) for collecting ethnicity data. The main differences are for the Arab, Gypsy, Roma and Irish Traveller groups. These differences will affect the number of ethnic groups we are able to produce admin-based ethnicity statistics for.

Table 1: Ethnic group response options in the 2011 Census and administrative data sources

2011 Census higher level ethnic group (5- category)	Lower-level ethnic group (18-category)			
	2011 Census	HES	ESC	IAPT
Asian /Asian British	Bangladeshi	Bangladeshi	Bangladeshi	Bangladeshi
Asian /Asian British	Chinese	Chinese	Chinese	Chinese
Asian /Asian British	Indian	Indian	Indian	Indian
Asian /Asian British	Pakistani	Pakistani	Pakistani	Pakistani
Asian /Asian British	Any other Asian background	Any other Asian background	Any other Asian background	Any other Asian background
Black /African/ Caribbean/ Black British	African	African	Black – African	African
Black /African/ Caribbean/ Black British	Caribbean	Caribbean	Black Caribbean	Caribbean
Black /African/ Caribbean/ Black British	Any other Black/ African /Caribbean background	Any other Black background	Any other black background	Any other Black background
Mixed /multiple ethnic group	White and Asian	White and Asian	White and Asian	White and Asian
Mixed /multiple ethnic group	White and Black African	White and Black African	White and Black African	White and Black African
Mixed /multiple ethnic group	White and Black Caribbean	White and Black Caribbean	White and Black Caribbean	White and Black Caribbean
Mixed /multiple ethnic group	Any other Mixed/multiple ethnic background	Any other mixed background	Any other mixed background	Any other mixed background

White	English/ Welsh/ Scottish/ Northern Irish/ British	British	White – British	British
	Gypsy or Irish Traveller	X	Gypsy/Roma	X
White			Traveller of Irish heritage	
White	Irish	Irish	White – Irish	Irish
White	Any other White background	Any other White background	Any other white background	Any other White background
Other Ethnic Group	Arab	X	X	X
	Any other ethnic group	Any other ethnic group	Any other ethnic group	Any other ethnic group
Unknown	Unknown	Not known	Unknown	Not known
X	X	Not stated	Refused	Not stated

Source: Office for National Statistics

Notes

1. Ethnicities have been presented alphabetically at the 5-category level (except for Other) and then alphabetically within each 5-category ethnic group.
2. In HES and IAPT, Chinese is grouped into the higher category of Other ethnic group. In ESC, Chinese is classified as its own higher-level ethnic group. For our analysis, we are aligning with the 2011 Census so Chinese is included as part of the Asian ethnic group.
3. X – response option not available.
4. HES stands for Hospital Episode Statistics, ESC stands for English School Census, IAPT stands for Improving Access to Psychological Therapies and ABPE stands for admin-based population estimates.

3 . Method used to produce the admin-based ethnicity statistics

To produce the [admin-based ethnicity statistics](#), we combined the annual extracts for each administrative data source across time. We used the NHS number as the unique identifier within Hospital Episode Statistics (HES) and Improving Access to Psychological Therapies (IAPT) and the Pupil Matching Reference Number within English School Census (ESC). Individuals could appear multiple times within a single dataset and some had different ethnicities recorded on different records. We implemented a set of rules to select a final ethnicity per person per data source (Figure 3).

Figure 3: Flow chart for selecting a final ethnicity per person



Source: Office for National Statistics

Notes:

1. We used the NHS number as the unique identifier within HES and IAPT and the Pupil Matching Reference Number within ESC.
2. To determine the most recent record, instead of dates, a sequential and unique record number was used in IAPT, with the highest record number for an individual indicating the most recent record. This left no conflicting responses to be resolved.

The general approach was to take the ethnicity from the most recent record. A date variable was used to do this in HES and ESC, and a sequential record number used in IAPT. If the ethnicity on the most recent date was unknown, the last stated ethnicity or refusal was selected where available, otherwise their ethnicity was coded as unknown. If an individual refused to provide an ethnicity on the most recent date, their ethnicity was coded as refused, regardless of whether they had previously provided their ethnicity. We decided to take this more cautious approach to dealing with refusals for the initial feasibility research following discussion at an Ethnic Group Assurance Panel meeting, where a consensus view was not reached on the ethics of looking back past a refusal. We will explore other options in future.

In HES and ESC, some individuals had more than one ethnicity recorded on the latest date. In ESC, unless one of these was a refusal (as above), the individual's ethnicity was coded as unresolved. In HES, a dataset hierarchy of Admitted Patient Care (APC), Accident and Emergency (AE), and Outpatients (OP) was implemented to choose between the records. However, if this did not resolve the ethnicity conflict (for example, if two different ethnicities were recorded in APC on the latest date), the ethnicity was coded as unresolved.

By combining the datasets over time, it was likely that we had included individuals who had subsequently died or emigrated, or who were only short-term visitors. To subset the data to usual residents, the HES, IAPT and ESC data were linked to the [admin-based population estimates \(ABPE\) V3.0 2016](#) dataset. ESC was linked to the ABPE using the Pupil Matching Reference Number and HES and IAPT were linked using NHS Number. Any individuals not linked to the ABPE were dropped. More information about the unlinked records can be found in [Section 6](#).

As some individuals were present in more than one data source, we used a similar process as above to select a final ethnicity for each person. A year variable was used instead of the date or record number variables and a hierarchy of IAPT, ESC, HES was used when individuals had different ethnicities recorded in different data sources in the same year. We based this hierarchy on the findings from record-level comparisons with the 2011 Census ([Section 4](#)).

After implementing these rules, 70.2% of ABPE records had a stated ethnicity, 9.3% had refused, 4.9% had unknown or unresolved and 15.6% were not linked to ESC, HES or IAPT.

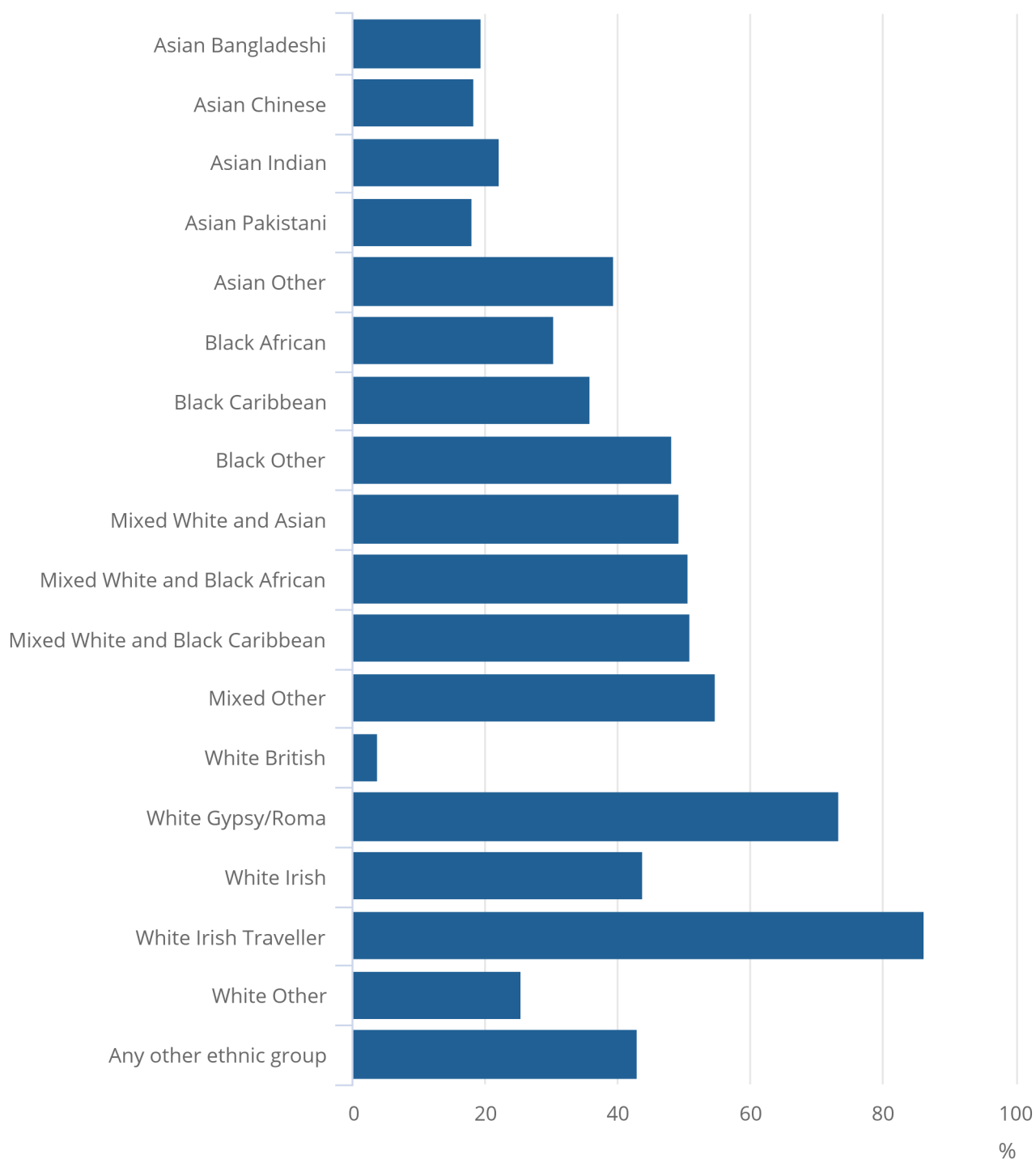
Of those with a stated ethnicity, 9.9% of records had more than one ethnicity recorded within the administrative data. However, this varied greatly by ethnic group (Figure 4), from 3.9% for individuals with a final ethnicity of White British to 86.3% for individuals with a final ethnicity of White Irish Traveller. The high proportion for the White Irish Traveller ethnic group, and also for the White Gypsy/Roma ethnic group, is to be expected, given that these categories are only found in the ESC data. Proportions were also over 40% for individuals in the Black Other, White Irish and Other ethnic groups, and all of the Mixed ethnicity sub-groups.

Figure 4: Individuals in the Mixed, White Gypsy/Roma and White Irish Traveller ethnic groups were most likely to have more than one recorded ethnicity in the administrative data

Proportion of records with that ethnicity in the admin-based ethnicity statistics that had more than one ethnicity recorded for them in the administrative data sources included in this research, England, 2016

Figure 4: Individuals in the Mixed, White Gypsy/Roma and White Irish Traveller ethnic groups were most likely to have more than one recorded ethnicity in the administrative data

Proportion of records with that ethnicity in the admin-based ethnicity statistics that had more than one ethnicity recorded for them in the administrative data sources included in this research, England, 2016



Notes:

1. The ethnicity is based on the final selected ethnicity after implementing the rules outlined in [section 3](#).

Looking at those with refused as their final ethnicity, 53.1% of individuals had previously stated their ethnicity in one of the administrative data sources. If we were to include the ethnicities for these individuals, we could get a stated ethnicity for an additional 2.7 million individuals (5.0% of the ABPE) which could improve the quality of our admin-based ethnicity statistics. However, we need to consider this from an ethical perspective, as there are varying viewpoints on recording a previously stated ethnicity where they have since refused.

4 . Record-level comparisons with the 2011 Census

To further understand the quality of each of the data sources, we compared the ethnic groups recorded in the administrative data against the 2011 Census at the record level. We did this for Hospital Episode Statistics (HES) data from April 2009 to March 2011, English School Census (ESC) data for 2011 and Improving Access to Psychological Therapies (IAPT) data for April 2012 to March 2013 (the closest available data to the 2011 Census). We selected one ethnic group per person within each administrative dataset using the rules outlined in [Section 3](#). More information about the linked dataset used to conduct this analysis can be found in [Section 6](#).

For individuals with a stated ethnicity in both the administrative data and 2011 Census, we calculated an agreement rate (the proportion with the same ethnicity on both sources). Table 2 shows that every dataset has a higher agreement rate at the 5-category level than at the 18-category level, as may be expected. IAPT has the highest agreement rate across both levels and HES has the lowest.

Table 2: Agreement rates by administrative data source in comparison to the 2011 Census, for 18 and 5 ethnic groups

Data source	18 ethnic groups	5 ethnic groups
HES	90.7%	96.1%
ESC	92.4%	96.6%
IAPT	93.3%	96.9%

Source: Office for National Statistics

Looking at the agreement rate by age for 18 ethnic groups, within HES, there is a general trend of increasing agreement rate with age. It ranges from 85.4% for those under 1 year of age to 94.9% for those aged 90 years and over. This may be because of the ethnic composition of the population varying by age, as analysis by ethnic group showed different agreement rates for different ethnic groups. IAPT has agreement rates of over 92.0% for all age groups and agreement rates for ESC range from 88.8% to 93.3%.

Breaking down the data by local authority, agreement rates are generally lowest in London local authorities. There is a strong positive correlation between the proportion of the population in the local authority that are White British and the agreement rate between the administrative data and 2011 Census.

Table 3 shows that for HES, the White ethnic group has the highest level of agreement with the 2011 Census, with 98.7% of linked individuals recorded as White in the 2011 Census also recorded as White in HES. The Asian and Black ethnic groups also have high levels of agreement.

The Mixed and Other ethnic groups have much lower agreement rates. Only 36.9% of linked individuals who recorded their ethnic group as Mixed on the 2011 Census are also recorded as Mixed on HES. Just over a quarter of those who recorded their ethnic group as Other on the 2011 Census are also recorded as Other on HES. Both are more likely to be recorded as White on HES than the ethnic group they answered on the 2011 Census. Those from the Other ethnic group are also more commonly recorded as Asian than Other on HES.

Table 3: Level of agreement for higher ethnic groups between the 2011 Census and 2009 to 2011 HES

2011 Census ethnic group					
	Asian	Black	Mixed	White	Other
Asian	87.9%	1.5%	6.1%	0.1%	30.7%
Black	0.7%	84.7%	9.8%	0.2%	5.6%
HES ethnic group Mixed	1.6%	5.2%	36.9%	0.2%	5.7%
White	3.7%	4.7%	39.3%	98.7%	30.6%
Other	5.9%	4.0%	7.8%	0.8%	27.5%

Source: Office for National Statistics

Notes

1. The 2011 Census ethnic group totals have been used as the denominators when calculating the percentages.
2. The percentages are based on linked individuals with a stated ethnic group on the 2011 Census and HES.
3. The ethnic group from HES is the ethnic group after using the rules to select one ethnic group per person.
4. HES stands for Hospital Episode Statistics.

ESC has a higher level of agreement with the 2011 Census than HES across all five high-level ethnic groups. The White ethnic group has the highest agreement rate (98.9%) and the agreement rates for the Asian and Black ethnic groups are also high. Nearly three-quarters of linked individuals from the Mixed ethnic group on the 2011 Census are also recorded as Mixed on ESC. However, of those with an ethnic group of Other on the 2011 Census who were linked to ESC, less than a third are also recorded with an ethnic group of Other in ESC.

Table 4: Level of agreement for higher ethnic groups between the 2011 Census and 2011 ESC

2011 Census ethnic group					
	Asian	Black	Mixed	White	Other
Asian	93.0%	0.6%	3.2%	0.1%	26.4%
Black	0.2%	93.6%	5.2%	0.1%	8.7%
ESC ethnic group					
Mixed	2.0%	3.9%	73.5%	0.6%	10.3%
White	0.6%	0.8%	15.1%	98.9%	22.3%
Other	4.1%	1.0%	2.9%	0.3%	32.4%

Source: Office for National Statistics

Notes

1. The 2011 Census ethnic group totals have been used as the denominators when calculating the percentages.
2. The percentages are based on linked individuals with a stated ethnic group on the 2011 Census and ESC.
3. The ethnic group from ESC is the ethnic group after using the rules to select one ethnic group per person.
4. ESC stands for English School Census.

IAPT shows a similar pattern to HES and ESC with high agreement rates for the Asian, Black, and White ethnic groups and lower agreement rates for the Mixed and Other ethnic groups. Those recorded as Mixed on the 2011 Census are commonly recorded as White in IAPT and those in the Other ethnic group on the 2011 Census commonly recorded as Asian or White in IAPT.

Table 5: Level of agreement for higher ethnic groups between the 2011 Census and 2012 to 2013 IAPT

2011 Census ethnic group					
	Asian	Black	Mixed	White	Other
Asian	88.8%	0.9%	4.2%	0.1%	24.0%
Black	0.8%	85.7%	7.7%	0.1%	3.6%
IAPT ethnic group					
Mixed	2.4%	5.8%	50.3%	0.5%	9.1%
White	3.4%	5.4%	33.7%	98.9%	32.6%
Other	4.6%	2.2%	4.1%	0.4%	30.7%

Source: Office for National Statistics

Notes

1. The 2011 Census ethnic group totals have been used as the denominators when calculating the percentages.
2. The percentages are based on linked individuals with a stated ethnic group on the 2011 Census and IAPT.
3. The ethnic group from IAPT is the ethnic group after using the rules to select one ethnic group per person.
4. IAPT stands for Improving Access to Psychological Therapies.

Looking at the 18-category ethnic groups, agreement rates are generally lower for the other sub-groups than the specific ethnic groups. For example, of those recorded as Black Other in the 2011 Census, 50.0% were recorded as Black African, 23.1% as Black Caribbean and only 16.0% as Black Other in ESC. This compares with agreement rates of 86.7% and 81.1% respectively for the Black African and Black Caribbean ethnic groups.

The exception to this is within the White ethnic group. For example, in HES, agreement rates are 96.7% for White British and 64.1% for White Other but only 39.5% for White Irish. Of the linked individuals recorded as White Irish in the 2011 Census, 55.9% are recorded as White British in HES.

Other ethnic group

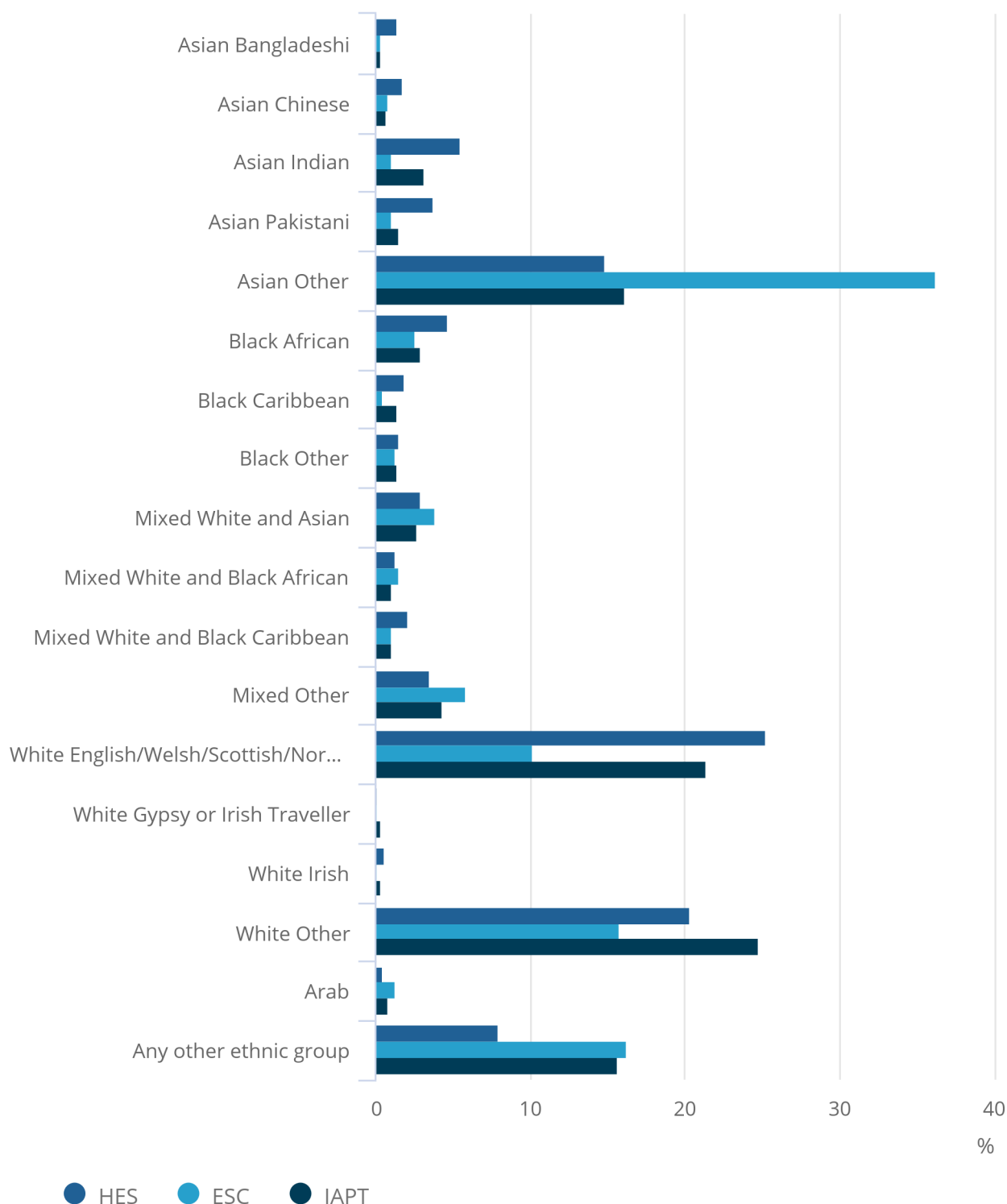
Our [feasibility research](#) found that a higher proportion of individuals are recorded in the Other ethnic group in the admin-based ethnicity statistics than in the 2011 Census. Looking at the linked Census-admin dataset, only 8.6% of those recorded as Other in HES were also recorded as Other in the 2011 Census (including Arab). These proportions are 17.5% and 16.5% for ESC and IAPT respectively. These individuals were commonly recorded as Asian Other, White British or White Other in the 2011 Census.

Figure 5: Those recorded as belonging to the Other ethnic group in the administrative data are commonly recorded as Asian Other, White British and White Other in the 2011 Census

Proportion recorded in each ethnic group in the 2011 Census, out of those recorded as Other in the administrative data, England

Figure 5: Those recorded as belonging to the Other ethnic group in the administrative data are commonly recorded as Asian Other, White British and White Other in the 2011 Census

Proportion recorded in each ethnic group in the 2011 Census, out of those recorded as Other in the administrative data, England



Notes:

1. The percentages have been calculated using the number of people in the Other ethnic group in the administrative data as the denominator.
2. The percentages are based on linked individuals with a stated ethnic group in the administrative data and 2011 Census.
3. The HES data are from 1 April 2009 to 31 March 2011, the ESC data from January 2011 and the IAPT data from 1 April 2012 to 31 March 2013.
4. HES stands for Hospital Episode Statistics, ESC stands for English School Census, IAPT stands for Improving Access to Psychological Therapies and ABPE stands for admin-based population estimates.

Refusals

Within each administrative source, some individuals refused to provide their ethnicity. The refusal rates differ by source but are similar across the ethnic groups within each source (Table 6). This suggests that refusals should not have a substantial impact on the representativeness of the admin-based ethnicity statistics. However, this may change over time so we will review this again when Census 2021 data are available for analysis.

Table 6: Refusal rate by administrative data source and ethnic group as per the 2011 Census

	2011 Census ethnic group				
	Asian	Black	Mixed	White	Other
HES	6.2%	6.5%	6.6%	5.3%	7.3%
ESC	0.4%	0.8%	1.7%	0.4%	1.4%
IAPT	12.6%	11.8%	14.4%	12.9%	13.4%

Source: Office for National Statistics

Notes

1. The 2011 Census ethnic group totals have been used as the denominators when calculating the percentages.
2. The percentages are based on all those included in the linked Census-admin dataset (including unknown responses and refusals)
3. Refused refers to the individual having a refusal on their most recent record in the administrative data.
4. HES stands for Hospital Episode Statistics, ESC stands for English School Census, IAPT stands for Improving Access to Psychological Therapies and ABPE stands for admin-based population estimates.

5 . Glossary

Agreement rate

Of those with a stated ethnicity on the administrative data source and 2011 Census, percentage of linked records where the ethnicity in the administrative data and 2011 Census are the same.

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), if a parent/guardian or pupil has declined to provide ethnicity data, this is recorded as "refused". In Hospital Episode Statistics (HES) and Improving Access to Psychological Therapies (IAPT), where a patient chooses not to identify their ethnic group, the code "Z - Not Stated" 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 ESC, where the ethnicity has not yet been collected, this is recorded as "NOBT" (information not yet obtained). In HES and IAPT, the default code "99 Not Known" is used where the person's ethnicity is unknown.

Ethnicity unresolved

Where multiple ethnicities were recorded on the latest date (and for HES, a dataset hierarchy of Admitted Patient Care, Accident and Emergency, and Outpatients didn't resolve the conflict), these have been coded as "unresolved".

Not linked

This refers to individuals who are in the admin-based population estimates V3.0 for 2016 but have not been linked to ESC, HES or IAPT.

Refusal rate

Percentage of linked records where the individual did not provide their ethnicity data.

6 . Data sources and quality

Unlinked records

As explained in [Section 3](#), the [admin-based population estimates \(ABPE\) V3.0 2016](#) dataset was subset to people living in England and used as the population base for the admin-based ethnicity statistics. After being combined across time, Hospital Episode Statistics (HES), English School Census (ESC) and Improving Access to Psychological Therapies (IAPT) data were linked to the ABPE and any records not linked were dropped. Around 8.5 million HES records, 3.3 million ESC records, and 260,000 IAPT records could not be linked to the ABPE.

For ESC, because of 2011 to 2015 ESC data not being used in the creation of the ABPE, records could only be linked to the ABPE if the individual was in school in January 2016. This means that those in the 2011 to 2015 ESC data who left school before 2016 are in the ESC dropped records, even though they may be in the ABPE (due to being in another data source). We will be looking to link 2011 to 2015 ESC records to the ABPE in future so that we can utilise the ethnicity information on these records. We will also be working to understand the non-links across all three data sources as part of our future work programme.

Linked Census-admin dataset

We conducted record-level comparisons with the 2011 Census using a dataset where the administrative data sources had been linked to the 2011 Census. To create this dataset, the 2011 English School Census, 2011 Welsh School Census, 2011 Patient Register and 2010 to 2011 Higher Education Statistics Agency data were first linked based on personal identifiable information using exact, deterministic, and probabilistic matching. An anonymous identifier called the ONS ID was then created to represent an individual, which grouped together the dataset IDs that related to said individual. The 2011 Census dataset was then linked on using the same matching techniques. The final product was a linkage key containing the dataset IDs (e.g. NHS number), ONS ID and Census ID.

We joined the ethnicity variables from HES and IAPT to the 2011 Census using the NHS number. We joined the ethnicity variables from ESC using the Pupil Matching Reference Number.

Some HES, ESC and IAPT records could not be linked to a 2011 Census record. These could be false negatives where the individual was in the administrative data and 2011 Census, but a link could not be made because of incorrect, missing or out of date personal identifiable information. Other non-links may be individuals not in the 2011 Census data because of the different time periods covered by the datasets, Census under-enumeration or because we only looked at individuals on the 2011 Census who were usually resident and some individuals in the administrative data may not meet the usual residence definition. Additionally, some records may have been linked when they were actually for different people (false positives). In total, 84.8% of HES records, 90.7% of ESC records and 83.8% of IAPT records were linked to the 2011 Census and included in the analysis.

7 . Future developments

This is an initial exploration of the potential to produce admin-based ethnicity statistics and the methods outlined above have produced promising results. However, this is not a finalised method and we will continue to explore alternate methods and data sources to improve the admin-based ethnicity statistics. These include:

- trialing alternative methods for handling multiple recorded ethnicities and refusals
- incorporating additional data sources to improve the population coverage
- combining the administrative data with survey data using the Generalised Structure Preserving Estimator (GSPREE), building on [previous work](#) using this method
- continuing [research](#) into producing survey-based ethnicity statistics, to provide a more robust comparator and an improved survey source to feed into GSPREE
- producing admin-based ethnicity statistics for Wales
- producing admin-based ethnicity statistics for other years
- exploring the potential to produce multivariate statistics on ethnicity by other characteristics
- engaging with existing efforts within the health sector to improve data collection practices
- collaborating with external experts and peer organisations conducting research in this area

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.

8 . Related links

[Admin-based ethnicity statistics for England, feasibility research: 2016](#)

Article | Released 06 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.

[Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016](#)

Article | Released 21 June 2019

Research into developing a new methodology to create population estimates from administrative data. These estimates are not official statistics on the population.