Office for National Statistics

Article

Methods and systems used to collect ethnicity information in health administrative data sources, England: 2022

Findings from semi-structured qualitative interviews which assess the quality of ethnicity data and identify sources of bias across three health data sources in England.

Contact:
Gemma Quayle and Rose Drummond
health.data@ons.gov.uk
+44 1329 444110

Release date:
16 January 2023

Next release:
To be announced

Table of contents

1 . Main points

- This article uses findings from semi-structured qualitative interviews to explore potential sources of error and bias in the process of collecting ethnicity information across three NHS data sources: General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics (HES) and Improving Access to Psychological Therapies (IAPT).

- We identified sources of potential error and bias across data collection, data processing and quality assurance processes; similar issues were apparent across all three data sources.

- Qualitative analysis revealed three main themes which can result in bias and inaccuracies in ethnicity data recorded: data infrastructure challenges, human challenges, and institutional challenges.

- This article is one of a series that aims to improve understanding of the quality of ethnicity data in important NHS sources; future releases will build on the findings of this report and develop solutions to mitigate biases.

2 . Overview

There is significant interest in understanding health inequalities, and robust statistics on health outcomes for different ethnic groups became increasingly important during the coronavirus (COVID-19) pandemic. People from minority ethnic groups were found to be at higher COVID-19 mortality risk, and important data gaps were exposed.

The coronavirus pandemic has highlighted ethnicity data gaps as an important area for the health statistics system to focus on. This desk review is part of a wider research collaboration between the Office for National Statistics (ONS), Wellcome Trust and the Race Equality Foundation. The wider research project aims to improve understanding of the quality of ethnicity data in important NHS sources and develop solutions to mitigate biases.

To complement the findings from this desk review, the wider research project also includes:

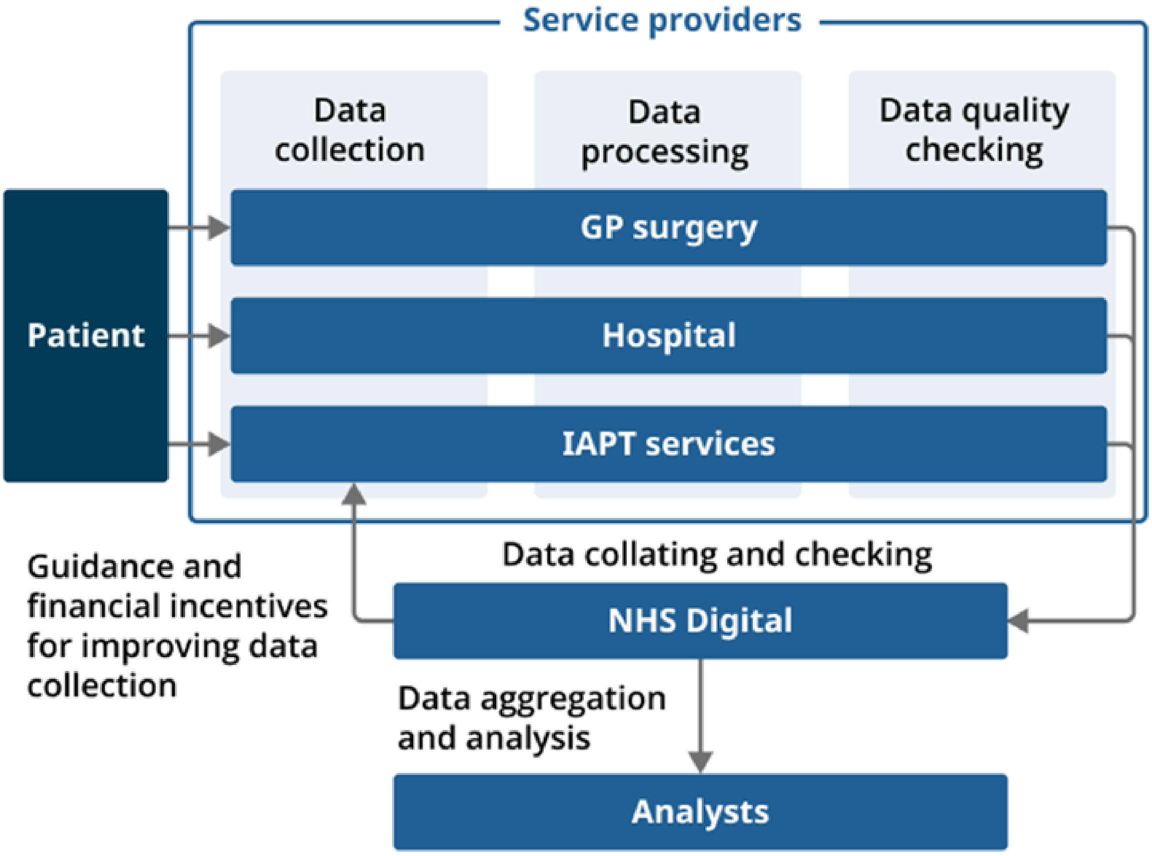
- quantitative person-level comparisons of ethnicity information recorded in hospital and General Practice data, compared with census ethnicity information (which is widely regarded as the most robust ethnicity data source covering the whole population)

- focus groups to provide further insights from the public and healthcare staff on their experiences of collecting ethnicity information data in practice, and potential for errors or bias to be introduced

3 . Summary of main findings

This summary highlights how ethnicity data are collected from a patient, and move through a process of data collection, processing, and quality assurance across all three data sources to prepare the data for analysis. Potential sources of error and bias identified throughout this process, and across all three data sources, are also summarised.

Diagram 1: Summary of the data collection process for three health data sources in England



Source: Office for National Statistics - Methods and systems used to collect ethnicity information in health administrative data sources, England: 2022

Summary of potential sources of error and bias throughout the data collection process

Our research identified ways in which data quality could be affected at different stages of the data collection process shown in Diagram 1. These are summarised as follows.

Patient

- For some sources, patients can opt out of data sharing, leading to missingness in data for analysis.

Service providers

Data are collected by each healthcare provider (GP surgeries, hospital departments and Improving Access to Psychological Therapies (IAPT) service providers). For all three data sources, participants described multiple methods for collecting ethnicity data, some of which are self-completion (for example, an online or paper form) and some are completed by a third party (for example, staff asking patients face-to-face or on the phone). Service providers also carry out some data processing and quality checks.

1. Data collection

- Variation in detailed ethnicity categories collected locally.
- Different data collection modes.
- Inconsistent use of residual categories: “Not stated”, “Not known”.
- Staff understanding of value of ethnicity data varied.
- Different staff may collect data slightly differently.
- Same data collected multiple times; little known about reliability of different entries.

2. Data processing

- Mapping detailed ethnicity categories to aggregated harmonised categories loses granularity.
- Complex and subjective coding processes for some data sources.

3. Data quality checking

- Trade-off between high-quality data and burden on service providers.
- Staff understanding and implementation of quality checks varied.

NHS Digital

- NHS Digital quality checks focus on completeness; checking accuracy is more challenging.
- Guidance on data collection and data processing limited for some sources.
- Any financial incentives focus on completeness.

4 . Main findings

Data infrastructure challenges

The organisation of ethnicity data differs throughout the data collection process and can therefore introduce barriers to data accuracy.

Variation in ethnicity categories collected and used locally

NHS Digital specifies data standards for ethnicity data, including a standardised list of ethnicity categories, which can be seen in the NHS data dictionary, and which are based on the 16 ethnicity categories used in the 2001 Census.

For all three sources, we found evidence of more detailed ethnicity categories used locally for data collection than categories required by NHS Digital. For example, in General Practice Extraction Service (GPES) IT systems, ethnicity information is stored using SNOMED CT health terminology, and 489 SNOMED codes were identified for ethnicity entry as of April 2022.

Service providers in each health setting may collect data at the granular level for the purposes of serving local population health needs. This applies between and within data sources whereby individual providers for the same data source can use additional, bespoke categories. However, this does not correspond with the wider purpose of data collection by NHS Digital, where aggregations are used for national level analysis. Consequently, the granularity of ethnicity information collected at the point of data entry is lost throughout the data collection process.

The breadth of ethnicity categories used between and across different health data sources has implications for accuracy because:

- the response options presented to patients vary between and across health settings; it can be inferred that the differences in the way ethnicity categories are presented could influence an individual’s response
- it can be inferred that the way that data are processed to aggregate categories to harmonised categories is not necessarily consistent between and across health settings, potentially affecting the quality of aggregated data used for analysis

Human challenges

Interviews provided evidence that human behaviour can influence the collection of ethnicity data in several ways.

Training and guidance

Variations in training and guidance can affect the capability and confidence of staff collecting and inputting data across different health data sources.

Differences were identified across the three data sources between:

- staff concerns over a lack of guidance and actively seeking more help (GPES)
- staff perceptions around the simplicity of data collection meaning that existing guidance is sufficient (Hospital Episode Statistics (HES))
- comprehensive training and guidance provided for staff meets their needs (Improving Access to Psychological Therapies (IAPT))

Specifically, a lack of clear guidance in GP data was reported to affect service providers ability to collect accurate data and meet financial incentives.

Understanding the value of collecting ethnicity data

Staff understanding of the value of collecting ethnicity information varied across data sources.

For example, it was reported that for GPES and IAPT, staff understood the value of collecting ethnicity data in terms of:

- its use for funding
- understanding and addressing population needs
- monitoring inequalities

In contrast, hospitals are a more challenging environment in which to collect ethnicity data, as providing urgent medical care takes priority over other tasks such as ethnicity data collection. The extent of this prioritisation varied between emergency and outpatient departments.

Evidence showed that differences in how staff value the collection of ethnicity data has implications for:

- the level of effort made in ensuring accurate, complete data collection, for example to what extent standardised data collection methods are used
- the amount of quality assurance that is conducted

Variation in how staff collect ethnicity data

The role of staff in each health setting can affect how data are collected and submitted to NHS Digital. Specifically, the interaction between patients and staff presented barriers to collecting accurate data.

For example, participants reported that staff inputting patients' responses at GP surgeries created opportunities for bias through subjective interpretation, if the response provided by the patient did not exactly match the ethnicity category options available. There was also evidence of ethnicity being conflated with country of birth. For HES, participants indicated that staff may feel reluctant to ask questions about ethnicity because of the perceived sensitivity of the topic.

In addition, varied staff understanding and use of residual categories such as "Not stated" and "Not known" were a concern for HES and IAPT data sources. Using such categories consistently to distinguish refusals from missing data enables missing data to be identified and followed up, and therefore has important implications for completeness.

It can be inferred that variations in staff practice could result in:

- lower quality of ethnicity information because of varying levels of subjectivity during data collection
- incomplete ethnicity data because of misuse of residual categories and a reluctance from staff to engage with patients about ethnicity and follow up missing data

Mapping bespoke ethnicity categories

The coding of ethnicity data by staff ready for submission to NHS Digital has implications for data accuracy.

For GP and IAPT data sources, mapping bespoke ethnicity categories was considered complex, particularly if the categories on the data collection form do not match the IT system. Participants raised the potential consequences of this complex mapping process, including:

- the potential over-coding of "Other" if staff are unsure of how to map bespoke categories to high level ethnic groups
- lack of transparency with patients on how their data entry could be aggregated

Institutional challenges

Centralised initiatives and incentive schemes around NHS data collection focus on data completeness, which could have implications on the accuracy of ethnicity data.

Quality checks and reports

Quality assurance processes conducted by both service providers and NHS Digital prioritise data completeness. A clear example of this was the finding that for HES, automated quality checks focused on checking the file structure is valid rather than the file content.

Data quality reports from NHS Digital also present metrics of data completeness and score service providers accordingly.

This focus on data completeness is grounded in the complex balance of ensuring high quality data versus reducing data collection burden on service providers. Concerns from participants over a focus on data completeness having an indirect impact on data accuracy were reported. However, checking accuracy was considered difficult because it requires verification of ethnicity by individual patients.

Financial incentives

There was variation in awareness of financial incentives available for collecting ethnicity data across health data sources.

Financial incentives described by participants were not specific to ethnicity, and similarly prioritised data completeness where only certain codes met funding criteria. For GPES, financial incentives could be confusing to implement, because of the long list of both legacy and up-to-date ethnicity codes included on IT systems.

Evidence showed service providers are motivated by financial incentives and viewed the schemes as a means of improving data collection. However, it can be inferred that focusing on completeness could have implications for accuracy. This is because limited guidance on how to meet financial incentives could result in errors during data collection.

5 . Glossary

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

A database comprised of patient-level information collected at General Practices (GP) in England only. It consists of patient demographic information and coded medical information.

Hospital Episode Statistics (HES)

A database comprised of patient-level information collected during a patient’s time in NHS hospitals in England. It is made up of several sources, however all datasets contain records which hold a range of information including administrative, clinical, geographical, and patient information.

Improving Access to Psychological Therapies (IAPT)

A programme developed to organise and improve evidence-based treatments for people with depression and anxiety disorders, delivered by a single clinician and typically managed by the GP. The IAPT dataset is a patient-level dataset which re-uses the clinical and operational data collected by service providers for the purpose of data analysis and reporting.

6 . Data sources and quality

Data sources

Analysis covered three important NHS administrative data sources in England:

- General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR)
- Hospital Episode Statistics (HES)
- Improving Access to Psychological Therapies (IAPT)

The sources of error and bias affecting these three sources provide insight into the types of issues that could be relevant to other electronic administrative health records. However, these three sources are not intended to be representative of issues affecting all health admin data sources.

Research design

A desk review was conducted for each dataset to inform the research questions for interviews.

Semi-structured interviews were the selected data collection method. Participants were recruited purposively and through snowball sampling measures. Participants were selected for their knowledge and experience of working with the dataset under investigation.

We interviewed a total of 21 participants and one additional participant provided comment through email. Five participants worked with the GDPPR data source, eight participants worked with the IAPT data source (one of whom provided email information only) and nine participants worked with the HES data source. Participants worked in a variety of roles and across a variety of employers including NHS England, NHS Digital service providers and IT systems suppliers.

Researchers designed and peer-reviewed topic guides for the semi-structured interviews. These consisted of a brief introduction informing the participant of the background and aims of the research, followed by questions asking about data collection, data submission, data processing, analysis, data collection standards, and coverage. The contents of the topic guides were informed by the literature review and were piloted.

Interviews were conducted by a lead interviewer on Microsoft Teams and lasted between 40 to 80 minutes. A second researcher attended the interview as an observer to take notes of responses. Interviews were recorded and digitally transcribed using the Microsoft Teams transcription function. All data were held in secure folders accessible only to the research team involved.

Approach to analysis

Recordings, transcriptions, and observer notes formed the data generated from the semi-structured interviews. All data collected was anonymised, and transcripts of the interviews were reviewed against their respective recordings as quality assurance.

A framework was created which followed the general structure of the topic guide, from which transcripts were sorted, and coded. Thematic analysis took place in a session with all interviewers and observers, where observer notes were reviewed, emerging themes were discussed and recorded, and interviewer reflection took place. Further analysis was then conducted using the organised data within the framework to explore emerging themes across interviews and produce high-level findings.

Strengths and limitations

The main strengths of this research are:

- the qualitative research design enabled a better understanding of the real-world circumstances surrounding ethnicity data collection, and answered knowledge gaps following the literature review

- purposive sampling allowed for relevant data collection, as well as the collection of a range of experiences from those in different roles

- semi-structured interviews afforded the opportunity to structure conversations around our research aims while also giving the flexibility for participants to describe individual experience

- topic guides were informed by the literature review and were peer-reviewed and piloted for quality

- interviews were attended by two researchers, one of whom took notes, as well as recording and transcribing the interviews to minimise any errors in data collection

- qualitative analysis was collaborative, reflective, and peer-reviewed at all stages

The main limitations of this research are:

- the sample size was small and, while saturation did appear to be achieved as a whole, one dataset had a smaller sample than the others

- there were some parts of the end-to-end process of data collection that participants did not have knowledge on, and resulted in some outstanding knowledge gaps

- the qualitative research represents the thoughts, views and experiences of our sample, and is not designed to be representative

- findings are limited to the experiences of staff from organisations involved with health data collection, as opposed to those using health services

- changes in wider circumstances for NHS data collection, for example updating ethnicity categories following Census 2021, may affect the relevance of the findings

7 . Future developments

This article is one of a series that aims to improve understanding of the quality of ethnicity information in health admin sources. Analysis released so far describes the extent to which ethnicity recording in different health admin data sources differs, and in this article we explore reasons for why that might be.

Future analysis will look at solutions and methods analysts can use to produce more reliable estimates despite the differences between sources.

8 . Related links

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

9 . Cite this article

Office for National Statistics (ONS), released 16 January 2023, ONS website, article, Methods and systems used to collect ethnicity information in health administrative data sources, England: 2022