

Diversity in Data - Ethnicity coding working group

Workshop Report - 13th January 2022

Background

The importance of Diversity in Data was highlighted by the [UK Health Data Research Alliance](#) (The Alliance). It was acknowledged that improving data quality and **increasing representativeness of data used for research and innovation is crucial** for both custodians/collectors of data and researchers, to ensure that research is of high quality and based on complete data with the ultimate aim of improving patient outcomes. A working group was set up to discuss, understand and potentially address concerns relevant to the quality of available data, missingness, accessibility and harmonisation of data, which can hinder the reliability of scientific insights.

While there are many different aspects to consider within the diversity data field, this working group focuses on challenges and opportunities in the use of data reporting ethnicity. To start the conversations, we are **convening the community to share insights from ongoing work and understand the gaps that should still be addressed.**

The ethnicity coding working group

On 13th January 2022, we held our first Diversity in Data - Ethnicity coding working group session, chaired by Ashley Akbari and Kamlesh Khunti. The session provided an opportunity for stakeholders working in this area to share their research and intelligence experiences, with relevant talks from Alastair Denniston ([Standing Together initiative](#), Insight Hub), Matt Bosworth (Office for National Statistics) and Malorie Perry (Public Health Wales). 68 participants attended the meeting, with people from research and custodian communities, along with members of the public. The agenda, plus links to the presentations and video recording of the session can be found in the Appendix.

Existing work, ongoing challenges and opportunities were discussed around ethnicity data, with suggestions for possible future priorities to be taken forward by this group.

During the interactive breakout sessions, participants covered:

- Current projects using ethnicity data.
- Challenges and limitations in accessing and working with ethnicity data.
- Possible focus for future development of this group.

The next ethnicity coding working group sessions are scheduled for:

- Wednesday 6th April 2022 @1pm

Below we summarise the main points highlighted during the meeting.

Challenges around data capture and completeness

Many health-related data sources either do not include any data on ethnicity coding, or they are varied in their collection and recording, leading to challenges of data quality. The discussion highlighted:

- Incompleteness or missingness in coding and recordings.
- Uncertainty of what coding means or changes over time and/or between data sources. It would be good to have a consistent ethnicity standard.
- Lack of consistency in recording/reporting data and ethnicity over time can change. Is there consistency across nations about how ethnicity is captured?
- The new Census data (2021) will capture new information and categories (to be published soon).
- Some groups are investigating the completeness and representativeness of ethnicity data in primary and secondary care linked data sources and looking at the best practices to categorise and standardise coding information.
- It was acknowledged that it is difficult to harmonise data due to changes to coding standards.
- Categorisation: there are regional variations, and categories can be too broad.
- The reason for collecting this data can vary, which impacts why they were collected and how an individual may declare their ethnicity – administrative, insurance, health, survey etc.

Public trust and social aspect

- There is a need to agree on definitions and what do we mean by ethnicity; i.e. ethnicity or race or ancestry.
- What data to collect – race or ethnicity and how interchangeably researchers use these terms.
- There is a need to capture cultural data alongside ethnicity data – culture and ethnicity are not the same. We could be missing things by not understanding the cultural impact of health inequalities.
- There could be a bias depending on whose data is captured – how do we ensure we access data from all groups, so everyone is represented.
- Better understanding between genetic determinants and cultural background. There needs to be an understanding on why the distinction is important.
- In data collection, we are reliant on the public to provide their data and details – some members of the public may be reluctant to offer this information (this could be for various reasons; cultural sensitivities regarding collecting this information; mistrust as people don't understand why we collect data, they may fear how data is used or who has access etc.)
- Whose responsibility is it to capture ethnicity data? When should it be captured? How frequently are the general public asked? In some situations, it is difficult for healthcare practitioners to ask for this information. What is the role of practitioners in this?
- When working with Census data which is self-reported, and health data which a health care professional records, we need to consider there might be differences between the ways data has been recorded, and potential reporting biases introduced.

Possible priorities for this group

The group highlighted a number of possible areas of work:

1. **Improving public awareness and trust:**
 - Transparency and communication are crucial for the public to be aware of the type of research that can be done using ethnicity information and improve public trust.
 - Public awareness activities need to be accessible and delivered in culturally appropriate formats.
2. **Pulling resources together:**
 - It would be helpful to have a single portal to share resources relevant to work already ongoing; standards and codes currently used; definitions and categorisations; sharing knowledge across UK nations.
 - What data sources/systems exist that contain ethnicity data, and can we improve these existing sources/systems.
3. **Harmonising data standards and codes across nations:**
 - There is a need to agree on definitions. Define what we mean/what data we are using – i.e. ethnicity/ancestry/country of origin/language/race/religion – it can be hard to understand what people mean without accurate descriptions. It can be incredibly subjective without guidelines.
 - Granularity vs consistent harmonisation – appreciate the value and need for both, i.e. highly granular data, which would have high variability between data sources/systems, compared to harmonised data which would be low in granularity but low in variability.
 - Link to developing AI methodologies.
 - Could we look at capturing info from birth as part of existing or a new standardised national data collection, which captures the range of possible details (ethnicity/ancestry/country of origin/language/race) so some elements would not change throughout your lifetime –but others could be asked over an individuals lifetime.
4. **The role of health care professionals**
 - Engaging health care professionals as they must understand the importance of capturing the data and how to tackle their bias and adopt and use standardised ethnicity coding, which would improve data collection and reporting, and lead to improvements in the reductions of potential health inequalities for their patients.



Next steps

We will hold at least two more working sessions to focus on specific aspects of working with ethnicity data. It was suggested the Alliance would be well placed to keep momentum and the conversation live. As a convener of data holding organisations, the Alliance could leverage expertise to explore aspects such as data quality, the accuracy of data capture at source and agreement on definitions used. Suggestions from the group included:

- The Alliance could be a vehicle to increase awareness among the public about inequalities highlighted from using this data.
- Health care professional involvement – with a need for primary care, secondary care, and community (*e.g.* pharmacy) lens to be represented, there is a need to facilitate conversations with this group.
- Working in partnership with key organisation stakeholders, such as the National Health Service of England, Northern Ireland, Scotland and Wales, and the Office for National Statistics, with an outcome from this group being a recommendation and a framework that policymakers should address.
- Raise awareness that this problem is not confined to the UK – We could look at what is ongoing globally and learn from it.
- The Alliance Diversity in Data working group is important in maintaining feedback loop, which could be facilitated by sharing knowledge (increase the visibility of resources, provide regular updates, offer a forum for discussion).

Appendix

Agenda

Time	Item	Lead and Slides
2 pm	Welcome and introductions	Chairs
2.05 pm	Importance of standardising ethnicity coding – why we are here	Chairs
2.15 pm	Breakout groups <ul style="list-style-type: none"> • What are the key reasons for using ethnicity data? • What is ongoing around the UK in different groups? • What are the main challenges that you are experiencing? • What are the challenges and limitations in accessing and working with ethnicity data? • What current research is there around understanding ethnicity data as part of COVID and non-COVID studies? 	Facilitators
2.35 pm	Discussion	All
2.50 pm	Inequalities in coverage of COVID-19 vaccination in Wales	Malorie Perry (Public Health Wales)
2.55 pm	Investigating inequalities by ethnic group during the coronavirus pandemic in England using linked administrative data	Matt Bosworth (Office for National Statistics)
3 pm	STANDING Together initiative	Alastair Denniston (Insight Hub)
3.05 pm	Q&A	
3.20 pm	Breakout session <ul style="list-style-type: none"> • What priorities are not currently being addressed that could be in the future? • What should the ongoing focus of this group be? • What other organisations and groups should be involved in the conversation? 	Facilitators
3.40 pm	Discussion	Chairs
3.55 pm	AOB and Close	



Outputs

[Video Recording](#) of the session

[Importance of standardising ethnicity coding](#) - Presented by Ashley Akbari (Swansea University) and Kamlesh Khunti (University of Leicester)

[Inequalities in coverage of COVID-19 vaccination in Wales](#) - Presented by Malorie Perry (Public Health Wales)

[Investigating inequalities by ethnic group during the coronavirus pandemic in England using linked administrative data](#) - Presented by Matt Bosworth (Office for National Statistics)

[STANDING Together initiative](#) - Presented by Alastair Denniston (Insight Hub)