UK Health Data Research Alliance

Diversity in Data - Ethnicity coding standards working group

Workshop Report – 24th May 2023

On 24th May 2023, we held our fourth Diversity in Data Ethnicity coding standards working group session at College Court Conference Centre, in Leicester. The session was chaired by Kamlesh Khunti (Leicester University) and Ashley Akbari (Swansea University).

The purpose of this meeting was to share the overall outputs from this series of workshops and agree on content for a final set of recommendations to call for improvements in recording of ethnicity data. Once finalised, these recommendations will be presented at one of the next Alliance Council meetings.

Around 60 participants attended the in-person meeting and had the opportunity to share their thoughts and feedback. Following an initial introduction by the Chairs, we had 6 key presentations, each followed by group discussions covering the proposed emerging recommendations.

Below we summarise the main points highlighted during the meeting.

Recommendation 1 – The purpose and value of capturing ethnicity data

Presentation by Marta Pineda Moncusi (Oxford University)

It is recommended that collection of ethnicity data as well as wider determinants (such as socio-economic status, religion, language) and individual characteristics should be sought consistently across the health and social care sectors to enrich completeness of information to tackle health inequalities.

This group discussion highlighted the need and value of collecting and recording information about ethnicity to ensure reliability in research, to inform healthcare practice and deliver equitable care. While the positive aspects were prominent, some considerations were outlined, including the need of being open about why this data is being collected and how it can be used for public benefit.

Summary of most common responses

Why collect?

- Identify where differences exist and their scale.
- Explore and understand what causes these differences.
- Address inequity of access.

- Address inequity of resource allocation.
- Guide development of new technologies (including digital and AI) to be more inclusive and fair.

Collection of other wider determinants

- Wider determinants in addition to ethnicity would be useful to collect as proxies: ie: race, country of origin, nationality, language spoken at home, lineage. Immigration and cultural identity are also important.
- Asking for ancestry as an additional means to ascertain meaningful information.
- Considerations whether collection of other wider determinants could cause a data collection burden.

Recommendation 2 – Ethnicity data standardisation

Presentation by Vahé Nafilyan and Rose Drummond (Office for National Statistics)

Standardised ethnicity categories such as the Office for National Statistics (ONS) 19 ethnic group categories should be adopted across the four nations; and all research data and routine collected data should collect the highest possible level and report at a minimum of five levels. More on the ONS 19 ethnic group categories here

Summary of most common responses

Terminology and definitions

- There is confusion around concept of 'race' and 'ethnicity' and the various terminologies (e.g. British vs American). We need clear definitions for race and ethnicity (see 'A glossary for the first World Congress on Migration, Ethnicity, Race and Health' and ONS guidance).
- There are many interpretations of key terms (geographic, genetic, country of origin, ethnicity, race) which can lead to misclassification. Definitions needed for each category.
- Disaggregating within 'Mixed' where possible in existing data collections and better designed collection options in the future.
- We should consider making 'Other' categories more meaningful. For example, other, in comparison to prefer not to say, not available or unknown needing a clear distinction.
- Important to balance self-expression vs the need to aggregate/harmonise within and across data sources. Should free text be used more widely in future? For example ONS/census includes free text option: 'Any other ethnic group, please describe?

Categorisation and coding

- It is challenging to get the number of categories right. There is a need to find a balance between granularity and accessibility vs. reproducibility and usability of the data.
- Wide use of categorisations is useful to have an appropriate sample size for research. Self-reporting is key.
- Challenges might be introduced by differences between local and national levels. Is ONS 18 harmonised standards representative of regional devolved nations of Scotland, Wales and N. Ireland?

Mapping/grouping

- Transparency of why we are asking for this data and how data will be re-mapped and for what reason is key.
- There is a need for reproducible and consistent mapping – A standards to facilitate this would be useful. A suggestion is to record it at most granular level, then map it to a higher level afterwards (for research purposes).
- Groupings must be self-determined.

Other considerations

- Challenges over time - ethnicity data may change over time as with people's identities and other sociodemographic data. We might need to consider an evolving standard and balance the need to update codes vs challenge of data shift and longitudinal capturing.
- There is a conflict between what standardisation needs vs. what individual deems appropriate.
- Standardisation is not necessarily the solution.
- Importance of recording *when* ethnicity was coded as categories may have changed since registered at GP or coded elsewhere.
- Consider that data is collected for multiple purposes – it is difficult to balance clinical and research utility.
- Sample size might be limited for certain ethnicities when stratified in 19 groups, specially in less diverse areas/Nations. However, starting form a much granular level always allows aggregation of the data (e.g., using 9 ethnic groups rather than 19).

Recommendation 3 – Communication and Transparency

Presentation by Jo Palmer, Joseph Alderman, Elinor Laws STANDING Together (no slides available)

High-quality and reliable research based on use of routinely collected data is strictly dependent on patients understanding the purpose of data collection and providing accurate information. A concerted effort should be made across the healthcare system to ensure the positive value of using data for research and innovation for societal benefit is widely articulated. Clear communication and explanations of the reasons for asking about personal information, including ethnicity, is also crucial.

Summary of most common responses

- Rephrasing of the sentence in recommendation, “strictly dependant on patients”, was suggested It’s *our* responsibility to *earn* public trust.
- The recommendation needs to include the privacy element and not just communications of benefits. Public narrative around privacy needs to be addressed. Dangers of data availability and misuse – we need to unpick that conflict.
- Need to be open and transparent to patients how we protect their identity.
- Communication of positive value of using data and articulate how data has been used for good is crucial.
- If we don’t address the knowledge gap, we can’t move on and engage people more.
- We need to be clear about why data is important for research and innovation and communicate the positive value of using data.

- We should explore opportunities at the point of collection to say why we are asking for personal information, what the data could be used for and what the possible public benefits would be.
- Using public service announcements / information campaign across UK nations will be important.
- We shouldn't put the onus on the public to learn about the data – it is our collective responsibility to be transparent and offer opportunity for engagement.

Recommendation 4 – Training and guidance for ethnicity data collection

Presentation by Jonathan Valabhji (NHS England)

Healthcare professionals play a key role in data collection. But there is some reluctance or lack of knowledge about the importance of data collection. It is recommended that standard guidance around data collection, including information on ethnicity, is distributed across NHS settings for healthcare professionals and other NHS staff to consider in their interactions with patients. Staff training for standardisation of recording of ethnicity data will have a role to play. Any training material and guidance should be developed with input from ethnic minority public contributors

Summary of most common responses

Is there a need for training?

- Yes – Training should focus on *why* it's important to capture high quality data and how this data can be used for research.
- Training needs to include an explanation of why we are collecting this data in a culturally appropriate manner. Cultural competency training will therefore also be important
- The need for training extends beyond ethnicity data – we need to increase general understanding on importance of data in healthcare.
- Training focus on the pathway from capturing ethnicity to how it would impact clinical care.
- Framing as to *improve* health rather than to stop inequality – positive message to public.

Who needs the training?

- Healthcare professionals/NHS Staff – those asking the questions and those directly collecting and entering the data – they need to understand why the data is being collected so they can explain to patients.
- Data users / researchers.
- General public.
- However, healthcare professionals are extremely busy, how can we balance the need for training and what is practical in NHS settings?
- An option could be to give the public the opportunity to self-report and update their information (e.g. via NHS App)

Other considerations

- Quantity vs quality - You can affect change around quantity of data collection – what about quality?
- How to measure impact of training?
- Does mandating data entry increase 'other' or 'unknown.
- Providing printed resources costs money. Incorporate with EDI training?

- Consider sensitivity training for staff.
- Ambassador in public spaces like leaders of community groups and religious groups.
- Need to respect that there are many reasons for non-completion: NHS staff not asking, public fatigue [of being asked], patient apathy, past experiences of discrimination/racism, broken trust etc.

Recommendation 5 – Data linkage to improve data completeness

Presentation by Angela Wood (BHF Data Science Centre, University of Cambridge)

Linkages of datasets from different data sources can help enrich the information needed and data completeness. For ethnicity and other determinants of health, as well as other protected characteristics, linkage can be used as strategy to increase data quality. Efforts to ensure system interoperability between settings across the UK should be made, leveraging work around NHS England Secure Data Environments and in line with the Goldacre recommendations.

Summary of most common responses

- Data sources - Census data is considered the "gold standard" in the past. However, it was previously hard to link with other data sources like routinely-collected electronic health record data sources. We can observe improved coverage but not necessarily quality.
- Some systems do not communicate with each other. Interoperability is key, however might be collected for clinical care and not research.
- Need for caution in denominators - Numerator/denominator bias.
- Biases – missing at random or not?
    - Differences in data completeness between sources.
    - Different sources have different data collection purposes which may introduce bias.
    - Misclassification bias.
    - To deal with bias, always be transparent, carry out sensitivity analysis and report limitations of analysis. Getting better metadata is important and knowing why data was collected.
- Minority distribution and definition varies by region.
- Requires lots of tailoring to be meaningful on new country context.
- Often the rate of missingness is too high to allow for valid imputation.
- Ethnicity recorded over time – why restrict it to EHR, why not get self-report from patient via NHS App?
- Should there be a flag to update in GP records? (i.e. 5yrs)
- Informed patient consent - opt out in direct patient care, is it then appropriate to link all your data together so that it bypasses consent?

UK Health Data Research Alliance

Considerations on international harmonisation

Presentation by Alastair Denniston (University of Birmingham)

Summary of most common responses

- Harmonisation and standardisation in data collection practices might be useful, but highly complex.
- In their 'Human rights-based approach to data', the United Nations (UN) identifies six principles for data collection including self-identification, both at individual and group level.
- The concept of ethnicity has different relevance in different settings, with the same individual potentially defining their ethnicity differently in different contexts. Ethnicity is a social concept, so we need to consider it might not be harmonised between cultures. Think about what ethnicity means in your dataset/study?
- There would be value in greater international coordination to identify differences in concepts of ethnicity and map categories where this is possible and appropriate. There is a need to map current initiatives (e.g. data_collection_in_the_field_of_ethnicity.pdf (europa.eu)).
- Consideration should be given to who benefits from the harmonisation. This should be a partnership between countries and not assumed or imposed from the outside.
- In some countries, ethnicity may have a different relationship (or even no relationship) to inequity;
- in these contexts countries may prioritise other drivers of inequity, and this may also reduce the relevance of harmonising across countries.
- In some countries, ethnicity coding is restricted by law, usually for historical reasons linked to systematic persecution along racial or ethnic lines. Whilst well-intentioned there is a risk that this hides a problem rather than addresses it.

Conclusion and next steps

The interactive discussions at this last in person event, combined with insights from previous online sessions, have demonstrated a broad interest in improving collection of data on ethnicity, as well as other personal demographic information. Most participants recognised that high quality data collection can benefit people through data-driven research, innovation, and policy decisions. But improvements and action from key decision makers across the four nations of the UK are needed. The UK Health Data Research Alliance can offer a forum to bring forward the discussion and drive adoption of best practice.

Following the Ethnicity coding standards working group series, attendees and other experts in this field have been invited to contribute to an Academic Position paper to outline proposed recommendations covering some of the aspects discussed during these meetings, with the view to raise awareness around the importance of consistency in the collection of ethnicity data and the use of established coding standards, and to call for action. Over the coming weeks we will be in contact with those who have expressed an interest in contributing to this final piece of work.

The recommendations paper will also be presented at one of the next Alliance Council meetings for endorsement and adoption of good practice.

UK Health Data Research Alliance

Appendix

Time	Item	Lead and Slides
10:00 - 10:20	**Arrival and welcome refreshments**	
10:20 - 10:30	**Introduction from the Chairs: The Alliance Diversity in Data working group and our work so far**	Kamlesh Khunti (Leicester University) & Ashley Akbari (Swansea University)
10:30 - 10:40	**Presentation 1 – The purpose and value of capturing ethnicity data and other determinants of health**	Marta Pineda Moncusí (Oxford University)
10:40 - 11:05	Group Interactive session	
11:05 - 11:15	**Presentation 2 – Ethnicity data standardisation**	Vahe Nafilyan & Rose Drummond (ONS)
11:15 - 11:40	Group Interactive session	
11:40 - 11:50	**Presentation 3 – Communication and transparency around collection of personal data**	Jo Palmer, Joseph Alderman, Elinor Laws (STANDING Together)
11:50 - 12:15	Group Interactive session	
12:15 - 13:00	**Lunch break**	
13:00 - 13:10	**Presentation 4 – The need for training and guidance for data collection**	Jonathan Valabhji (NHS England)
13:10 - 13:35	Group Interactive session	
13:35 - 13:45	**Presentation 5 – Data linkage to improve data completeness**	Angela Wood (University of Cambridge)
13:45 - 14:10	Group Interactive session	
14:10 - 14:20	**Presentation 6 – International harmonisation**	Alastair Denniston (INSIGHT)
14:20 - 14:45	Group Interactive session	
14:45 - 15:00	**Summary and close**	Paola Quattroni, Kamlesh Khunti, Ashley Akbari
15:00 - 15:30	**Refreshments and networking**	