

Strategies to record and use ethnicity information in routine health data

Author Names

Ash Routen¹, Ashley Akbari², Amitava Banerjee³, Srinivasa Vittal Katikireddi⁴, Rohini Mathur⁵, Martin McKee⁶, Vahe Nafilyan⁷, Kamlesh Khunti¹.

Author Affiliations:

1. Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK
2. Population Data Science, Swansea University Medical School, Swansea University, Swansea, UK
3. Faculty of Population Health Sciences, Institute of Health Informatics, University College London, London, UK
4. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, UK
5. London School of Hygiene and Tropical Medicine, Faculty of Epidemiology and Population Health, London, UK
6. London School of Hygiene and Tropical Medicine, Faculty of Public Health & Policy, London, UK
7. Office for National Statistics, Government Buildings, Newport, UK

Corresponding author: Professor Kamlesh Khunti, Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK. Email:

kk22@leicester.ac.uk

Word Count: 2044

References: 26

Ethnicity information is often missing from health data, impeding action on inequalities. Recording and using ethnicity data will require training, efforts at standardisation, and policy changes, while engaging with patients and the public.

With a few exceptions, health systems worldwide may collect data on people's age and sex, but few ask patients about ethnicity or race. There are many reasons, from historical abuses of data on race or religion in the early 20th century, to simple inertia. Yet countries that collect these data comprehensively have shown that COVID-19 has been characterised by wide inequalities, with those from some ethnic groups disproportionately affected by the direct and indirect effects of the pandemic.

Data collection

To researchers who study ethnicity and health, these inequalities came as no surprise ¹. It has long been recognised that people from certain groups experience substantial barriers to accessing health care, as well as disproportionate rates of disease and its consequences, compounding their many other disadvantages in sectors such as employment, education, and housing, themselves risk factors for poor health.^{2,3} Therefore, from a health equity perspective, the case for measuring and understanding variation in health and its determinants in different ethnic groups is obvious. Effective planning of health, social care and public health services requires data on ethnicity to ensure they are culturally appropriate, allocate resources equitably, and evaluate the impact of policy, with the ultimate aim to reduce variations in outcomes⁴.

A lack of such data is a particular problem in research, where participants are often unrepresentative of those who will ultimately receive interventions being evaluated. This has been apparent throughout the COVID-19 pandemic, with many clinical trials failing to record or report results by ethnicity ⁵, in part because some countries prohibit the collection of data on ethnicity, citing concerns about privacy ⁶.

Ethnicity is a complex concept, incorporating notions of race (the classification of people based on physical appearance), religion, and culture, especially for individuals born into families with multiple heritages. When asked their ethnicity, a

patient's first instinct may be to question why they are being asked. Their answer may differ according to their previous experience, especially if they have several possible identities, as is increasingly the case, and where they perceive that answering the question may place them at risk of discrimination or even violence, as described, for example, among Roma in Central Europe⁷. There are also particular challenges in international comparisons, as context influences the response. For example, it may be quite different to be a Gujarati in India, Kenya, or the United Kingdom. Even where such data are collected, the quality may be variable. It has been suggested recently that inconsistencies in data on race (which is related to, but distinct from, ethnicity) in the USA delayed identification of groups at greatest need during the COVID-19 pandemic⁸.

The experience in countries that collect ethnicity data shows its importance, with many examples from the United Kingdom (UK), which is unique in that data on ethnicity are collected routinely in most interactions with public authorities. There is a legal requirement on the National Health Service (NHS) to do so, rather than, as in some other countries, limiting this to ad hoc research studies or using proxy measures, such as country of birth (which will miss second and subsequent generation immigrants), as in France.

Defining ethnic groups

The recording of ethnicity primarily involves collecting data on an individual's membership of an ethnic group, but related information on primary language, religion and country of birth may also be collected, depending on the country and healthcare sector. In the UK, information on ethnicity is collected across a wide range of routine electronic health record data sources, primarily captured from patients within NHS systems at the point of care. These data are collected primarily via self-reporting by patients and with/by health care practitioners.

Ethnic groups are socially constructed and distinct from genetic ancestry. Ethnicity is defined on the basis of a society's norms, attitudes and expectations, rather than being a readily measurable biological variable (like blood pressure) where there is widespread agreement on measurement^{9,10}. For this reason, the categorisation of ethnic groups can differ over time and place. In the UK, the broad ethnic group

'Asian' largely comprises people from the Indian subcontinent whereas in the USA the term often implies people from East Asia. Similarly, classifications often evolve over time, with the 'one-drop' rule defining a person as Black if they had any Black ancestry, information used for the purposes of segregation in 20th century USA¹¹. Terminology reflects social factors such as experiences of migration and broader historical processes.

Establishing meaningful ethnic groups to analyse health disparities is not a straightforward task. On the one hand, it is often preferable to study narrow ethnic groups lest important heterogeneity be masked^{10,12}. On the other hand, some minority ethnic communities may be relatively small, which can prevent robust statistical analysis and raise concerns around maintaining confidentiality. It is increasingly appreciated that ethnicity intersects with other characteristics, such as gender, sex, or socioeconomic position¹³. There can be considerable value in adopting an intersectionality perspective, but this may again require a trade-off against studying more disaggregated ethnic groups. In the UK, a pragmatic classification of 18 ethnic categories has been chosen¹⁴, which where available provides standardised categories across government and healthcare settings allowing for the monitoring of inequalities across health, policy and social care spheres.

Patient involvement

Although information on ethnicity is typically collected through patient self-report, minority ethnic communities have rarely been involved in the design, implementation, collection and use of routine healthcare data collection on ethnicity^{15,16}. The active involvement of minority groups in the various processes of collecting and utilising ethnicity data should be prioritised as a way to build trust in communities that are often hesitant to provide data due to wrongful use or past abuse of gathered information. The quality and accuracy of ethnicity data and the evolving relevance of ethnic categories will likely be improved by community members validating or sense-checking ethnic group coding standards¹⁵.

Community members have been mobilised in some inclusive ethnicity data collection or design processes. For example, in the 2018 census in Colombia, the national

statistical office undertook a wide ranging consultative exercise with various indigenous and minority ethnic groups. This consultation resulted in revised question wording and ethnic group response options, to align with the consulted communities' needs. Such public involvement should be undertaken by all organisations collecting routine ethnicity information in health and social care systems, who should also work with ethnic minorities on governance of data repositories and ownership of data ¹⁷.

Boosting collection

One barrier to ethnicity data collection (See Box 1 for more detail) is healthcare professionals' lack of knowledge about the importance and use of the data, including a reluctance to ask for ethnicity data and a lack of demonstrated need for the collection of data.¹⁸ Experiences of providing ethnicity information within healthcare have generally been reported as acceptable, although in some studies, participants have expressed dissatisfaction about being asked to provide their ethnicity on repeat visits ¹⁹, and there needs to be a clear explanation from the healthcare provider as to why the data is being collected and how and what it would be used for ¹⁹.

At the organisational level, a comparison of self-reported ethnicity and Hospital Episode Statistics (HES) coded ethnicity in England found that misclassification varied only by a small amount between ethnic groups, but varied by a greater degree between hospitals (the accuracy of coding of ethnicity across hospitals ranged from 67% to 100%) ²⁰. This suggests that processes within hospitals may influence coding accuracy, although whether this is driven by staff or organisational issues is unclear.

Incentives to record ethnicity can be effective. Ethnicity recording was introduced into UK primary care in 1991 and into Hospital Episode Statistics in 1995, but this was only financially incentivised under the Quality and Outcomes Framework (QOF) between 2006-2011. The completeness of ethnicity recording rose from 27% for individuals registered from 1990-2012 to 78% for individuals registered from 2006-2012, after it was incentivised, and the ethnic breakdown of CPRD participants was comparable to the UK population²¹. Challenges remain, however, as while there is high accuracy for people who self-identify as White British (97% accurate), there is poorer accuracy for minority ethnic groups (59% accurate)²⁰.

Training and standardization

Two key factors that affect collection of ethnicity data are staff training and knowledge, and variation in data collection procedures ^{18,22}. Standardised data collection protocols and the use of standardised ethnicity categories that can be harmonised across sectors would reduce this variation. Harmonisation across countries or continents may not be appropriate, as ethnicity is a social construct and varies significantly between countries.

Self-identification of ethnicity by the patient will avoid errors and emphasises the value of an individual's lived experiences, as will self-completion of developed data collection forms. This will require comprehensive staff training to address barriers such as lack of time, or patient capacity for self-report, which can present in pressurised clinical situations⁴. Training should be developed with patients and public members to ensure that developed protocols are acceptable to patients. The reasons for collecting ethnicity information are clear and justifiable for both patients and staff members, and conveyed in appropriate languages and spoken or written formats ¹⁹.

Public involvement will avoid pitfalls. During the COVID-19 pandemic, the initial practice in the UK was to use the ethnic grouping "Black, Asian and Minority Ethnic (BAME)" for early data analyses on outcomes, until this was discontinued following public input and feedback that this was not an appropriate description, as BAME groups together disparate ethnicities ²³. Community members should also be involved when disseminating information to the public on the importance of data linkage and collection more generally, as well as communications on how health and ethnicity data in particular are used and interpreted.

As well as collecting ethnicity information, data on the wider determinants of health will help to understand inequalities. Most of the differences between ethnic groups in COVID-19 outcome analyses are due to wider structural factors, such as housing and intergenerational living, poor quality employment and occupational exposure, and environmental support for health behaviours, that are imperfectly collected in electronic health records, if collected at all ²⁴.

Policy and professionalism

Newly developed protocols, guidance and ethnic grouping standards will require support across sectors. Policy changes will be needed to enshrine regular reviews of guidance, as well as routine monitoring and publication on the quality of ethnicity coding data⁴. There is an opportunity for ethnicity to be an exemplar for improving data quality overall; the power of ethnicity data will be increased if all data recording is also improved. EHR minimum standards should be introduced and health informatics, including the work of the Professional Records Standards Body and Faculty of Clinical Informatics in these areas should be professionalised, and research using electronic health records increased.

Data quality should be regularly reported at regional and national levels, with cycles of improvement for both completeness and quality. In addition, ethnicity collection reporting could include cross-disease and cross-country comparison; there was little or no mention of ethnicity in the Global Burden of Disease studies and other similar inter-country or global data collection exercises²⁵. Efforts to collate and improve comparability of ethnicity data globally are needed, as comparison of health inequalities data between countries is limited due to inconsistency in ethnicity data collection methods and variance in ethnic group categorisations. Addressing these issues will be complex as self-defined ethnicity will differ depending on the social and cultural context in which the individual is responding²⁶. In addition, any international comparison analyses must be clear on the different experiences and cultural context of comparable ethnic groups between countries, such as how they are or are not minoritised.

Ultimately, funding is required for health care systems and researchers, with supporting policy to ensure continued implementation and monitoring. These are not new arguments, but they are receiving renewed attention following the impact of the COVID-19 pandemic. Action is required to ensure that existing health inequalities based on ethnicity are not maintained or exacerbated. Improving the collection and reporting of ethnicity information in routine health data should be one part of a wider process to tackle health inequalities.

Box 1 – Common challenges of ethnicity data collection and recording in routine health data

- Healthcare professionals' lack of knowledge about the importance and use of ethnicity data, and understanding or confusion about ethnicity categories.
- Healthcare staff reluctance to ask for ethnicity data and lack of confidence to be able to do this in a sensitive and culturally appropriate manner.
- Pressures on healthcare staff and lack of time or opportunity to ask the patient about their ethnicity.
- Lack of central supporting resources to provide training and appropriate data collection materials, such as the costs of translation into multiple languages, and data collection templates.
- Processes for collecting patient ethnicity information, such as recording templates, vary within and between general practices and hospitals.
- Lack of appropriate, or use of outdated, ethnicity codes in electronic health records.

Box 2 – Recommendations to improve ethnicity recording in routine health data

- Embed ethnicity information collection within routine processes, such as during regular health checks within primary care.
- Creation and adoption of national standardised data collection protocols and standardised ethnicity categories.
- Develop comprehensive training to allow staff who collect ethnicity data in health and social care to be able to routinely ask about ethnicity in a sensitive manner, focusing on self-reported data collection.
- Ensure inclusion of minority ethnic communities throughout the development of staff training and the development of standardised data collection and reporting protocols.
- Facilitate linkage between primary and secondary care sectors to avoid repeated questioning of patients for ethnicity information.
- Ensure implementation of strict confidentiality controls, maintaining clear separation between data for health purposes from potential other uses, such as immigration controls.
- Regular reporting of data quality at regional, national and international levels, with cycles of improvement for both completeness and quality.
- Convey the value of data collection to the public in order to maintain public trust. This could involve a cycle of feedback and improvement between healthcare providers, data providers, and the public.
- Ensure the collection of wider determinants of health, which will include mechanisms for health inequalities between ethnic groups.
- Develop electronic health record minimum standards, professionalise informatics, and invest in electronic health record research.
- Raise the policy profile of ethnicity recording and associated research at national and international levels, as is being done by the UK Scientific Advisory Group for Emergencies (SAGE) ethnicity subgroup and Marmot reports.

Acknowledgements

The views expressed in this article are those of the author(s) and not necessarily those of UKRI, the NIHR, or the Department of Health and Social Care. SVK acknowledges funding from a NRS Senior Clinical Fellowship (SCAF/15/02), the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17). RM is funded by a postdoctoral fellowship from the Wellcome Trust (201375/Z/16/Z)

Author Contributions

All authors led on drafting, editing and revising the content. All authors approved the final version and are accountable for all aspects of this work.

Competing Interests

KK is a director of the University of Leicester Centre for Ethnic Health Research, trustee of the South Asian Health Foundation, chair of the Ethnicity Subgroup of the Scientific Advisory Group for Emergencies (SAGE). SVK was co-chair of the Scottish Government's Expert Reference Group on ethnicity and COVID-19 and a member of the Ethnicity Subgroup of SAGE. RM is a member of the Ethnicity Subgroup of SAGE.

References

1. Bhopal, R., *et al.* *Public Health* **126**, 241-244 (2012).
2. Lympelopoulou, K., *et al.* *Urban Studies* **54**, 2540-2560 (2017).
3. Marmot, M., Allen, J., Goldblatt, P., Herd, E., Morrison, J. Build Back Fairer: The COVID-19 Marmot Review. The Pandemic, Socioeconomic and Health Inequalities in England. . (Institute of Health Equity, London, 2020).
4. Scobie, S., *et al.* Ethnicity coding in English health service datasets. (Nuffield Trust, London, 2021).
5. Pan, D., *et al.* *EClinicalMedicine* **23**, 100404-100404 (2020).
6. Oltermann, P., *et al.* France and Germany urged to rethink reluctance to gather ethnicity data. in *The Guardian* (2020).
7. Csepeli, G., *et al.* *Journal of Ethnic and Migration Studies* **30**, 129-150 (2004).
8. Webb Hooper, M., *et al.* *JAMA* **323**, 2466-2467 (2020).
9. Katikireddi, S.V., *et al.* *Journal of Epidemiology and Community Health* **75**, 970 (2021).
10. Bhopal, R. *Migration, Ethnicity, Race, and Health in Multicultural Societies*, (Oxford University Press, Oxford, 2014).
11. Davis, F.J. *Who is Black? One Nation's Definition*, (Pennsylvania State University Press, Pennsylvania, 2010).
12. Bhopal, R., *et al.* *BMJ* **319**, 215 (1999).
13. Bauer, G.R. *Social Science & Medicine* **110**, 10-17 (2014).
14. ONS. Ethnic group, national identity and religion <https://www.ons.gov.uk/methodology/classificationsandstandards/measuringequality/ethnicgroupnationalidentityandreligion> (2022).
15. Balestra, C., *et al.* Diversity statistics in the OECD: How do OECD countries collect data on ethnic, racial and indigenous identity? (Organisation for Economic Co-operation and Development, 2018).
16. Farkas, L. Analysis and comparative review of equality data collection practices in the European Union: Data collection in the field of ethnicity. (Publications Office of the European Union, Luxembourg, 2017).
17. Network, M.D.S. Our Data, Our Sovereignty, Our Future <https://www.temanararaunga.maori.nz/> (2022).
18. Iqbal, G., *et al.* *Diversity and Equality in Health and Care* **6**, 267-285 (2009).
19. Iqbal, G., *et al.* *BMC Public Health* **12**, 243 (2012).
20. Saunders, C.L., *et al.* *BMJ Open* **3**, e002882 (2013).
21. Mathur, R., *et al.* *Journal of Public Health* **36**, 684-692 (2013).
22. Morrison, Z., *et al.* *2014* **21**, 14 (2014).
23. Khunti, K., *et al.* *BMJ* **371**, m4493 (2020).
24. Nafilyan, V., *et al.* *European Journal of Epidemiology* **36**, 605-617 (2021).
25. Vos, T., *et al.* *The Lancet* **396**, 1204-1222 (2020).
26. McGregor-Smith, R. Race in the workplace: The McGregor-Smith review <https://www.gov.uk/government/publications/race-in-the-workplace-the-mcgregor-smith-review> (2017).