UK Health Data Research Alliance

Enhancing diversity and quality in health data:

Progress and recommendations from the UK Health Data Research Alliance's Special Interest Group on Ethnicity Coding Standards

UK Health Data Research Alliance

Status of the document

NOTE: Page to be removed in the published article.

White Paper v2.0 dated 14.05.24

This paper summarises key insights from the Alliance Special Interest Group on Ethnicity data coding standards and outlines a set of recommendations for improving ethnicity data recording.

We are very grateful for the contribution of over 100 workshop participants, covering a broad range of stakeholder views, including patient and public representatives, NHS and other data providers, academia, researchers, and system-level stakeholders.

The paper will be published as a White Paper on behalf of the Alliance Special Interest Group on Ethnicity data coding standards.

ACTION FOR MEMBERS

The paper is currently in the final draft stage. We invite Alliance Council members to note the progress to date and to provide final comments by 19 June 2024.

In particular, we ask members to:

1. Consider the Executive summary and recommendation section ;
2. Highlight any key questions or concerns around the framing of the recommendations;
3. Highlight any key missing points;
4. Acknowledge the publication of the paper as an output from the Alliance Special Interest Group on Ethnicity data coding standards.

UK Health Data Research Alliance

Executive summary

ACTION FOR COUNCIL MEMBERS: A Foreword is being considered for this section. We invite members to provide suggestions.

Data drives decisions across healthcare, ensuring the quality and representativeness of this data is crucial.

This paper presents the collective insights and recommendations from the Special Interest Group on Diversity in Data and Ethnicity Coding Standards, established by the UK Health Data Research Alliance in 2022. This group was formed in response to growing concerns about poor data quality and missing information in health and social care data research and specifically addresses the challenges and opportunities in the collection and use of ethnicity data. In this paper, we acknowledge the progress achieved in biomedical domain in the past few years, summarise the findings of the meetings that brought together experts from different disciplines, propose a set of key recommendations, and suggest future considerations to achieve data homogeneity.

The recommendations are:

1. **Improve consistency and completeness in data collection.**

The collection of ethnicity data, as well as wider determinants such as socio-economic status, language or marital status, and individual characteristics, should be sought and documented consistently across the health and social care sectors to enrich the completeness of information to tackle health and social inequalities.

2. **Align approaches to standardisation and categorisation across the UK.**

There should be UK-wide adoption of a standardised coding framework of ethnicity aligned to the one presented by the Office for National Statistics (ONS) guidance[1], which includes 19 core categories. All research and new routine data should carefully consider the specific levels that they should present as options. We recommend at a minimum the 19-level categorisation in static data collection forms, and the 6-level categorisation in verbal data collection. Data collection should be based on self-reporting wherever possible. The source and mechanism of the data capture should be contemporaneously recorded.

3. **Improve transparency and communication about the reason for collecting ethnicity data.**

Good practice should include starting data collection locally, working with trusted partners, co-developing materials, thinking about where and when the optimal efficiency of data collection can be captured, keeping the lines of communication open, embedding participatory governance and trying to make it diverse.

4. **Develop national guidance and training for data collectors in NHS and social care settings.**

Guidance on best practices for the collection of data relating to ethnicity and other personal characteristics should be distributed across health and social care settings for professionals and other staff to consider in their interactions with patients and individuals whose data is being captured. Staff training to support the standardisation of recording and to overcome some of the perceived barriers to data collection is required. Training material and guidance should be developed with input from ethnic minority public contributors.

5. **Facilitate data linkage to enrich the information available on ethnicity.**

Combining information from across multiple existing data resources through data linkage will achieve higher levels of coverage, completeness and accuracy on information about ethnicity.

UK Health Data Research Alliance

Introduction

The COVID-19 pandemic brought into sharp focus the effect of longstanding health inequalities and disproportionate disease burden among certain ethnic minority communities[2]. Accurate and complete ethnicity data is essential to facilitate health improvements for individuals from ethnic minorities. Without accurate and comprehensive ethnicity data, variations in operational and clinical practice between ethnic groups cannot be identified through research, meaning downstream harms may go unrecognised.

Despite the improvements of the past few years,[3] evidence suggests that the completeness and quality of ethnicity data still vary across data sources and are often limited. Ethnicity is not frequently reported in published work despite being a key demographic variable[4]. Ethnicity data quality in terms of completeness and consistency of capture within electronic health records (EHR) is poor, with frequent mismatches between the ethnicity recorded in EHRs and individuals' self-assigned information, especially for particular ethnic groups[5]. There are also issues relating to the perceived trustworthiness of institutions and the lack of transparency around data use, which exacerbate this problem. Multiple ethnicities are often recorded for the same patients, with a disproportionate number of records coded in "other" categories, making it difficult to accurately interpret the data in research[6 7].

The UK Health Data Research Alliance (the Alliance) recognised these issues and established a Special Interest Group, with a specific focus on ethnicity coding standards. During the working group sessions conducted from January 2022 to May 2023, it was recognised that while standards for ethnicity coding exist[8 9], it was agreed that further work is required to encourage organisations to understand which standards are appropriate and how these should be implemented consistently through the use of a national framework. In the summer of 2022, the Office for National Statistics (ONS) ran a consultation to seek input on their revised standards for ethnicity data and their proposals for monitoring and expanding the standards, recognising the need to improve the way public bodies record, understand and communicate ethnicity data. The Alliance working group highlighted the need for recording and using ethnicity data in a trustworthy manner. Comprehensive training, establishing standardised protocols, and effecting critical policy changes need to be established while actively and meaningfully involving the public.

On behalf of the Alliance Special Interest Group on ethnicity data coding standards, we propose a set of recommendations to include in a national framework to guide the use and collection of ethnicity data. We call on decision-makers in the UK to drive the implementation and adoption of recommended ethnicity data standards consistently.

The UK Health Data Research Alliance and its commitment to Diversity in Data

The Alliance is a growing network of leading healthcare and research organisations united to establish best practices to enable the ethical use of health data for research and innovation at scale[10]. It brings together

diverse organisations and stakeholders from around the UK, including public bodies, NHS trusts, biobanks, medical charities, cohorts and academic centres, to promote the development of standards, policies and tools and enable access to data in a secure, trustworthy and ethical way. The Alliance aims to create a trustworthy, federated and coordinated approach to health data research infrastructure to accelerate improvements in human health and care.

All partner organisations share the commitment to improving their processes to provide safe access to data, to make the data they hold Findable, Accessible, Interoperable and Reusable (FAIR) for research and to embed patient and public involvement in their works [11].

To date, the Alliance has developed standards and best practices in areas including technology services[12], trust and transparency[13], data usability[14] and information governance[15]. Outputs produced by community members in the area of data quality and standards, as part of the 'usable data' workstream, include recommendations for data standards[16] and the 'data utility framework'[17].

Diversity in data is a cross-cutting theme that aligns with HDR UK and many Alliance partners' policies on diversity and inclusion[18] [19]. Reaching the potential of making discoveries using health data requires an inclusive community and data representing all segments of society, diseases and conditions.

The working group on ethnicity coding standards is a sub-group of the Alliance Special Interest Group on Diversity in Data. It brings together both the research and custodian communities to address ethnicity coding issues in data management for health research and aligns with the aims of the 'usable data' workstream. The group aims to:

- Facilitate sharing of best practices and knowledge around ethnicity coding standards;
- Identify gaps and challenges in coding and collecting high-quality and complete ethnicity data;
- Advise the Alliance about ethnicity coding data and their management;
- Take forward and identify development activity on behalf of the Alliance about standards for ethnicity recording, including via research, scoping working practices, organising events or through other appropriate means;
- Create outputs, such as guidelines, recommendations and publications related to the remit of the working group;
- Promote good practices in handling ethnicity health data for scientific research, via relevant advocacy and engagement.

Development of recommendations

The initial three meetings of this Special Interest Group, which drew 68, 39, and 29 participants respectively, took place virtually on Zoom. The concluding meeting, attended by 60 participants, was conducted in person

at the University of Leicester's College Court Conference Centre. Members of the Special Interest Group included academic researchers, public contributors, healthcare professionals, national data custodians and NHS. A list of participating organisations is included in Appendix 1.

All meetings were chaired by Prof. Ashley Akbari (Professor of Population Data Science Research, Swansea University) and Prof. Kamlesh Khunti (Professor of Primary Care Diabetes and Vascular Medicine, University of Leicester).

All meeting discussions were captured and published in Zenodo:

- Meeting Report 13 January 2022: https://zenodo.org/record/8138582
- Meeting Report 6 April 2022: https://zenodo.org/record/8138561
- Meeting Report 18 January 2023: https://zenodo.org/record/8138530
- Meeting Report 24 May 2023: https://zenodo.org/record/8138229

During the in-person event in the University of Leicester, there were presentations[20] from keynote speakers who led conversations on specific areas of the recommendations proposed.

Insights from discussions at these meetings are summarised in Appendix 2. Below we present a set of recommendations proposed by participants and refined by the authors of this paper.

Recommendations towards an ethnicity data collection framework

Community and expert input received through our working sessions indicated that complete, detailed and accurate ethnicity data are necessary for clinicians, researchers and policymakers to understand disease outcomes and causes and to provide insights into the management of care in minority populations. However, the completeness and accuracy of ethnicity data within health and social care and routine data sources varies across settings. This could be due to inconsistent practices in data collection, missing information in EHR data sources and partly to a lack of understanding of the importance of data collection or reluctance for staff to ask for data, or for people to provide sensitive information.

While the use of categorisation to describe ethnicity, among other demographic characteristics, might be a controversial concept as definitions of individual characteristics do not adequately describe the cultural, social, and religious aspects, it should be acknowledged that categorisations are needed to enable comparable and scientific studies to address health inequalities and deliver care.

Based on these discussions, the Alliance Special Interest group on ethnicity coding standards recommends that a national framework and relevant guidelines should be implemented to ensure consistent capture of ethnicity records and use of ethnicity coding standards as recommended by the ONS.

Below we outline a proposed set of five main recommendations (highlighted in blue).

UK Health Data Research Alliance

Recommendation 1: Improve consistency and completeness in data collection

The need for consistent collection and recording of complete, accurate and high-quality data around ethnicity and other determinants of health across different healthcare settings is critical to effectively addressing health and social inequalities. Inconsistent data practices can lead to gaps in health outcomes and disparities in care provision. Standardising data collection processes and protocols across healthcare systems ensures that data is comparable and reliable. This standardisation also helps in minimising errors and discrepancies, which are often prevalent when different systems use varying data collection and recording standards.

In addition, integrating various data sources within healthcare systems is important for ensuring comprehensive data completeness and enhancing patient care quality. This could involve using examples of health systems where integration has led to better health outcomes through a more comprehensive understanding of patient demographics and needs.[21]

By focusing on these areas, healthcare systems can significantly improve the quality of data and, consequently, the quality of care and health outcomes for patients. This approach is essential in mitigating health and social inequalities, leading to a more equitable healthcare system.

Recommendation 1a: ***Collection of ethnicity data as well as wider determinants such as socio-economic status, language or marital status, and individual characteristics should be sought consistently across the health and social care sectors to enrich the completeness of information to tackle health and social inequalities.*** To address health and social inequalities effectively, it is essential to consistently collect and report details of ethnicity data along with wider determinants such as socio-economic status, language, marital status, and individual characteristics across all health and social care sectors. By enriching the completeness of information collected, healthcare providers can gain a deeper, more holistic understanding of the factors that impact health outcomes. Accurate and complete ethnicity coding is essential not only for researchers but also for the implementation of research findings and recommendations for public health policies. Such examples were evident during the COVID-19 pandemic. Viewing ethnicity as an intersection of various protected characteristics and socio-economic classes offers a nuanced perspective that is crucial for tailoring healthcare interventions and policies to meet the diverse needs of the population.

Recommendation 1b: ***While integrating multiple data sources and standardising collected data, organisations collecting and recording data should seek to provide the mechanisms and tools to individuals to update and declare their ethnicity over time.*** Agile systems will help record any changes, have the latest information and understand any changes in self – identification. Organisations should focus on integrating multiple data sources and standardising the data collected to ensure consistency and accuracy across systems. Implementing agile systems that allow individuals to update and declare changes in their ethnicity and other personal information over time is critical. Such systems must provide user-friendly tools that facilitate these updates, ensuring that records are current and reflective of any changes in self-identification. This approach

supports data accuracy and the adaptability of health services to meet evolving demographic and personal circumstances.

Recommendation 1c: ***Data collection protocols across health and social care settings should be standardised to ensure consistent, accurate, and reliable data capture on ethnicity and other health determinants.*** This will involve developing detailed protocols that specify methods and tools for data capture as well as explaining the rationale for chosen methods and tools, fostering cross-sector collaboration to adopt these standards universally, and utilising technology to streamline data entry. Regular audits, feedback mechanisms, and ongoing training will support compliance with these protocols, enhancing the integrity and utility of health data. This consistency is important for robust health research and effective healthcare policy and practice, as it helps to accurately study and address health disparities.

Recommendation 2: Alignment of approaches to standardisation and categorisation across the UK

Ethnicity is not a simple construct; rather, it is a complex and multifaceted concept that encompasses a wide range of characteristics, including cultural, historical, linguistic, and social elements. Ethnic groups are diverse and dynamic, often characterised by a variety of traditions, beliefs, and practices that can differ significantly even within the same broader ethnic category. Furthermore, people within a particular ethnic group may share common ancestry or cultural ties, but individual experiences, identities, and perspectives can vary widely. Despite this, a system for classification is extremely valuable for fostering clarity, consistency, and precision in communication and data analysis.

Standardised ethnic classifications provide a common language that enables accurate and comparable reporting across diverse settings, ensuring that information is collected uniformly and can be reliably interpreted. Standardised terminology facilitates the aggregation of data from different sources, allowing for comprehensive analyses of health outcomes among specific ethnic groups. By adopting standardised ethnic classifications, healthcare systems can enhance their ability to deliver equitable and culturally competent care, ultimately contributing to improved health outcomes for individuals from diverse backgrounds. We have identified three key recommendations to improve the consistency and quality of research.

Recommendation 2a: There should be UK-wide adoption of a standardised coding framework aligned to the one presented by the Office for National Statistics guidance[22] [2324], which includes 19 core categories.

These categories can be naturally grouped in a higher-level categorisation (6-level, such as ‘Asian background’), and may include additional categories (such as other specific Asian backgrounds like ‘Thai’) or lower-level categories (such as Uyghur Chinese). This allows all codes to be mapped to a wider hierarchy and allows translation between granularity levels, for aggregation over data sources using multiple code lists. Note that in all categorisation systems, we recommend that ‘Not stated/prefer not to say’ is not aggregated

into ‘any other’ to preserve the information about settings in which people are less willing to disclose their ethnicity.

Recommendation 2b: Data collection should be based on self-reporting wherever possible. The source of the data should be contemporaneously recorded (such as ‘reported by carer’), in order to identify processes in which self-reporting is sub-optimal and identify where ethnicity data may be less dependable. Additionally, the setting for the recording should be reported where possible, such as ‘at GP registration’, ‘at GP consultation’, and ‘during emergency care encounter’. This allows context in the recording (or lack thereof) to be examined; for example, there might be differences in self-reporting when it is perceived as related to healthcare provision rather than purely administrative purposes.

Recommendation 2c: All research and new routine data should carefully consider the specific levels that they should present as options. The most granular (specific) levels of data are the most valuable for research purposes, especially when these levels are clearly mapped to a hierarchy such that higher-level categories can be imputed where necessary for analysis. However, presenting over 100 options in a list is not appropriate in most circumstances. Where possible, presenting the 19-level code lists and then relevant sub-levels based on this selection, provides a balance of granularity and optimal user-experience. However, ***we recommend at minimum the 19-level categorisation in static data collection forms, and the 6-level categorisation in verbal data collection.*** As a guideline, if over 50% of a surveyed population self-report as a specific category, one might consider further sub-categorising this group in the future, if appropriate. Similarly, if less than 0.1% of a surveyed population self-report as a category, there may be limited value in presenting this category as an option, and a higher-level category will suffice.

Recommendation 3. Improve transparency and communication about the reason for collecting ethnicity data

High-quality and reliable research based on ethnicity data is dependent on the information available and how it is collected. There are human interactions before, during, and after the data is used by researchers, and within these interactions more communication about data use needs to take place so that it becomes a more trustworthy practice.

It isn’t enough to simply extoll the benefits of collecting and using ethnicity data, and provide one-way cascaded information about its importance. Many people who experience health inequalities have been, or perceive they have been, discriminated against because of who they are, so when they are providing personal information, like their ethnicity, they are weighing up the risks as well as the benefits.

Recommendations for communication and transparency include:

Recommendation 3a: ***Collection of ethnicity data locally should be encouraged.***

There is usually stronger support for sharing identifiable data at a local level as opposed to regionally or nationally, for example at the GP surgeries or clinics, where the relationship between the individual and the healthcare service is the strongest. This is also an opportunity to work with particular communities in an area.

Recommendation 3b: ***Those collecting ethnicity data should work with trusted partners***.

Within local areas, there are often groups that are willing to help facilitate conversations with their community. For example health services or research organisations work with local, trusted organisations (such as churches and mosques, family clubs, community groups, schools, etc.) that can help with data collection. This can mean that the conversations happen in a culturally and linguistically sensitive way – not all communities will have the same experiences or concerns.

Recommendation 3c: ***All ethnicity related materials should be co-developed with relevant members of the public.*** The same materials and messages don't always work for everyone. By co-developing materials with public contributors, they are more likely to be appropriate to and understood by the relevant communities. These materials should proactively answer the questions that a particular group may have, and specifically address any real or perceived risks in the spirit of transparency. Materials may be required in different languages, easy-read formats, physically and online.

Recommendation 3d: ***Those collecting ethnicity data need to consider where and when they collect data.*** Information about an individual's ethnicity is often recorded at the point of attending a health encounter (e.g. Emergency Department or at the general practice). While this is to be encouraged, there might be a risk of this being asked for at the same time as other questions, such as those about UK residency, which in turn confer eligibility for free healthcare. Asking for someone's ethnicity at this time, even if the individual is a UK resident, could be seen as insensitive or cause concern.

Ethnicity data could be collected during the actual health consultation, however it might not be considered to be a priority by the health professional and the patient, especially since such appointments are very time-limited.

It could be collected remotely, but the mode of doing so should be considered so as to also address those who may be digitally excluded.

Recommendation 3e: ***Trust among individuals providing their ethnicity data needs to be built.*** Just because the data has been gathered, it doesn't mean the conversation is over. If relationships are built, they need to be maintained. This can be as simple as ensuring there is always someone to contact if people have questions.

Recommendation 3f: I*nclusive and diverse participatory governance should be embedded in data collection processes*. Data collection, traditionally, can be quite an extractive process, where individuals providing their data lack power and lack the visibility of what happens next. Having members of the public as part of your

governance boards, or in patient advisory groups, and ensuring diversity within these groups, is another way to improve the quality of communication and provide the feedback loops to demonstrate why the use of that data has been beneficial.

Recommendation 4. Develop standard guidance and training for data collectors in NHS and social care settings

Health and social care professionals and other staff play a key role in data collection. But there are capacity constraints around granular data collection and possibly some lack of knowledge about its importance. Concerns from patients, service users and staff that data might be used to assess eligibility to receive public services could also undermine data collection.

Recommendation 4a: ***Standard guidance is distributed across health and care settings for healthcare professionals and other staff to consider in their interactions with patients.***

Standardised guidance should encompass protocols for data collection, the standardised coding framework, and recommendations on holding culturally sensitive interactions with patients during ethnicity data collection. Addressing patient concerns, guidance should also cover how to communicate confidentiality regarding the separation of health-related data from other potential uses, such as public services. While maintaining core principles, such guidance should allow for some variation in information and protocols to suit the specific needs of each health and care setting.

Recommendation 4b: ***Staff training to support standardisation of recording and to overcome some of the perceived barriers to data collection is required.***

Common challenges in collecting and recording ethnicity data in routine health records involve health and care providers lacking understanding of its significance and application, confusion over ethnicity categories, hesitancy and a lack of confidence in requesting ethnicity data sensitively, time limitations, insufficient resources for training and data gathering materials, variations in processes among healthcare facilities, and the utilisation of improper or outdated ethnicity codes in electronic health records [25]. Training to upskill all staff involved in collecting ethnicity data within healthcare settings is required to provide the necessary underpinning knowledge and feelings of cultural competence to be able to sensitively collect such data. In addition particular focus should be given to the use of standardised data collection protocols and the implementation of the standardised coding framework. To achieve this, health and care employers will need to provide support by granting staff the necessary time to attend training sessions, whether in person or online. This training will help staff understand the optimal methods for collecting ethnicity information and its significance in service provision, thereby equipping them with the confidence to effectively communicate its importance to patients.

Recommendation 4c: ***Training material and guidance should be developed with input from ethnic minority public contributors.***

To ensure protocols are acceptable and ethnicity data collection is clear and justified for both patients and staff, training should be developed collaboratively with ethnically diverse communities of patients and public members, facilitating effective communication in multiple languages and formats. Public involvement is essential to avoid potential issues. For example, in the UK during the COVID-19 pandemic, initial data analyses used the term "Black, Asian and Minority Ethnic" (BAME), but public feedback led to its discontinuation due to its broad grouping of diverse ethnicities[26]. Community members could also assist in disseminating information on the importance of data linkage and collection, as well as explaining how health and ethnicity data are used and interpreted.[27]

Recommendation 5. Facilitate data linkage to enrich information available on ethnicity

Accurate and complete recording of ethnicity measurements contributes directly to the representativeness of all communities in research and health and social care. Higher levels of completeness in ethnicity records could achieved through data linkage across multiple data sources. It is important to set an efficient and effective linkage approach to ensure all studies can be designed with a data linkage strategy focused on capturing ethnicity records.

We use the term ethnicity focused data linkage specifically when two or more data sources holding ethnicity records are brought together to achieve a more complete percentage of ethnicity records across a study population. Data linkage presents a unique opportunity for aggregating information from diverse data sources. Nevertheless, practical experiences in this field, as documented in recent studies [1][2][3], underscore the imperative for standardised procedures in data collection, categorisation, and coding. Considerations regarding the data sources that feed into linkage algorithms play a pivotal role in mitigating systemic biases during the generation of ethnicity data through data linkage.

An efficient linkage strategy requires a thoughtful and systematic approach to optimise resource utilisation without necessitating the linkage of every available data source. The vision for data linkage recommendations is in line with the efforts of increasing representativeness of minority and underserved communities in clinical trials and health research projects [28]. After careful consideration, we present a number of recommendations.

Recommendation 5a: ***Researchers need to evaluate the ethnicity content of data sources at the feasibility stage of their studies.***

Assessment of variables holding ethnicity related data is essential prior to inclusion of any data sources in data access requests for a project with an ethnicity focused element. This evaluation should capture elements

such as completeness of records for the study population, coding of ethnicity records and categories of ethnicity provided within each data source that is considered for linkage.

Recommendation 5b: A *definition of an ethnicity data tag for each of the existing data sources is needed.*

This recommendation is specifically focused on contributions that data providers and trusted research environments can make for enabling diversity in research through data linkage. The ethnicity tag is aimed to provide an ethnicity focused content value per data source. The elements of an ethnicity data tag comprised of: Year of record capturing, Ethnicity categories, Record completeness and total number of population with an ethnicity record in each data source. These elements can enable efficient evaluation, improvement and contribution of data sources into a linkage activity for capturing more fully existing ethnicity records.

In scenarios where only clinical data, such as primary care, secondary care, and emergency department records, contribute to the data linkage algorithms, systemic bias may manifest towards a subset of the population characterised by interactions with healthcare providers, thus potentially excluding a healthier demographic. This group has identified an ideal scenario for the effective inclusion of EHR linkage in enhancing data diversity on a population scale, involving the following two essential criteria:

I) The implementation of standardised approaches for recording ethnicity data.
II) Provision of dynamic mechanisms that allow individuals to update and declare their ethnicity over time.

By meeting these foundational criteria, we can ensure that the ethnicity data derived from the linkage of multiple data sources exhibits representativeness across the entire population. Our aims should be on leveraging data linkage, so we bring together varied data sources (primary care, hospital records, national surveys) for a holistic view of ethnic diversity within the health data domain. While integrating multiple data sources and standardising collected data, we should provide the mechanisms and tools to individuals to update and declare their ethnicity over time. This will help record any changes, have the latest information and understand any changes in self – identification.

The compatibility and interoperability of systems across different UK health data sources is a crucial step and requires coordination between the NHS secure data environments (SDEs) and Trusted Research Environments (TREs), governmental bodies, and international data while adhering to legal and governance policies. The current data linkage guidelines should be updated and include a focus on ethical considerations, privacy protection and methodological robustness. Protocols that handle sensitive ethnicity data. Data linkage reduces the amount of missing information while improving the quality of ethnicity data. This needs cross-verifying information across different data sources to identify and correct inconsistencies. Resources allocation towards the development of the necessary technological infrastructure for effective data linkage ensures secure and efficient data processing. Funded research to explore innovative data linkage methodologies. With a focus on the representation of ethnic minorities and reduction of data completeness (e.g AI and missing data methodologies etc.).

Recommendation 5c: ***Reproducibility and transparency of curation of ethnicity data and derived categories should be ensured.***

With the extraction and derivation of available ethnicity information from various linkable data sources, appropriate documentation and sharing of any mapping or hierarchies implemented should be transparent and open-source accessible. This is to ensure that the information obtained from each data source can be verified and cross-referenced across the available linked data sources in order to identify and correct inconsistencies, and ensure transparency and reproducibility of any cleaning or mapping to derived ethnic groups by future research and operational system deployments. Further funded research exploring innovative data linkage methodologies will open up opportunities to access further linkable ethnicity data, so ensuring these best practices are embedded is key to ensuring reproducibility and transferability between systems, environments and projects in the future, with further focuses on the representation of ethnic minorities and increasing data completeness through natural language processing (NLP), Artificial Intelligence (AI) and machine learning methods and missing data methodologies potential keys to enhancing the completeness of ethnicity information across the whole population.

Future considerations for international ethnicity recording and global harmonisation

Creating an international standard for ethnicity reporting may at first seem desirable with a view to enabling international comparisons. There are a number of reasons however why this is unlikely to be feasible, and indeed may not be desirable. The challenges around ethnicity data collection that have been described earlier in a national context, are further magnified at an international level, and are compounded by a number of specific issues relating to differences between countries, as set out below.

Several challenges prevent the successful adoption of a single race/ethnicity standard that could be applied internationally. First, the same terms may refer to different people in different cultural contexts (e.g., differences in the use of the term 'Asian' between the US and the UK). Second, there may be significant historical and current reasons for differences in the approach to ethnicity recording undertaken by different countries. Within the UK context, the four nations have not adopted exactly the same categories. For example, Northern Ireland has elected not to record UK vs Irish as different categories in line with the Good Friday agreement. Third, ethnicity will be collected to a greater or lesser degree in different political and cultural contexts. Concerns over historical discrimination on the basis of ethnicity has led to a number of countries (e.g. France, Germany) avoiding the collection of ethnicity data under most circumstances[29 30]. Interestingly in such countries proxies are commonly used (e.g. use of 'mother tongue', 'geographic origin')[31]. Furthermore, some countries have deemed the concept of ethnicity (as understood in the UK) to be much less important than other concepts such as Tribe or Caste.

Beyond the UK, a number of international data initiatives have attempted to address these issues but whilst well-intentioned, they risk imposing a specific view of ethnicity which is not recognised or welcomed by other countries. For example the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) may be valuable for harmonising many clinical data fields but has a US-originated recording of race and ethnicity.

A landmark publication from the United Nations, 'A human rights-based approach to data', provides six core principles for the collection and use of data, of which the importance of 'self-identification' is particularly relevant here[32] . They note that 'For the purposes of data collection, populations of interest should be self-defining.' Self-definition will be reflected both within a country but also at a global level, and runs counter to efforts for global harmonisation. They also noted that 'Data collection activities should be conducted in accordance with the human right principle of 'doing no harm'. This was also echoed in the STANDING Together initiative (STANdards for data Diversity, Inclusivity and Generalisability) which gathered input from over 350 participants across 58 countries [33]. Respondents consistently highlighted the different impact that ethnicity and other relevant attributes had across nations, and the critical role of context - with the drivers and barriers to opportunity being dramatically different between nations. The approach recommended by STANDING Together is to focus on the principle of equity, and encourage those responsible for the creation and usage of datasets to consider the most relevant attributes for their specific context ('Contextualised Groups of Interest').

In summary, challenges to the global harmonisation of ethnicity coding include that: it may not be feasible due to non-collection of relevant data for historical reasons ; it may not be meaningful due to due to differences in the construct (i.e. it is not just a case of mapping terms) and how this construct interacts with the local context; and it may not be readily compatible with a human rights-based approach to data, notably the principle of self-identification which may result in variation in how different people groups think of 'ethnicity' in their context, regardless of whether the terms they use are superficially similar. While harmonisation of categories or classification schemes is therefore neither desirable nor feasible, harmonisation of methodology, which reflects principles of self-identification, self-definition and community consultation should be pursued.

Conclusion

The interactive discussions among members of the Alliance Special Interest Group on ethnicity coding standards demonstrated a broad interest in improving collection of data on ethnicity, as well as other personal demographic information. While recent articles have been published on examples of coding standards that could be used for ethnicity data in the UK, and many groups have been working to address issues around health inequalities and data, no recommended standards on ethnicity data collection, recording and coding are available across different NHS settings. The ethnicity coding standards Special

Interest Group 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 the adoption of best practices across UK data custodians and NHS sites.

The recommendations of this group highlighted key areas for improvement.

1. **Improve consistency and completeness in data collection.**

   1a: Collection of ethnicity data as well as wider determinants such as socio-economic status, language or marital status, and individual characteristics should be sought consistently across the health and social care sectors to enrich the completeness of information to tackle health and social inequalities.

   1b: While integrating multiple data sources and standardising collected data, organisations collecting and recording data should seek to provide the mechanisms and tools to individuals to update and declare their ethnicity over time.

   1c: Data collection protocols across health and social care settings should be standardised to ensure consistent, accurate, and reliable data capture on ethnicity and other health determinants.

2. **Align approaches to standardisation and categorisation across the UK.**

   2a: There should be UK-wide adoption of a standardised coding framework aligned to the one presented by the Office for National Statistics guidance [34], which includes 19 core categories

   2b: Data collection should be based on self-reporting wherever possible. The source of the data should be contemporaneously recorded, (such as 'reported by carer'), in order to identify processes in which self-reporting is sub-optimal and identify where ethnicity data may be less dependable.

   2c: All research and new routine data should carefully consider the specific levels that they should present as options. We recommend at minimum the 19-level categorisation in static data collection forms, and the 6-level categorisation in verbal data collection.

3. **Improve transparency and communication about the reason for collecting ethnicity data.**

   3a: Collection of ethnicity data locally should be encouraged.

   3b: Those collecting ethnicity data should work with trusted partners.

   3c: All ethnicity related materials should be co-developed with relevant members of the public.

   3d: Those collecting ethnicity data need to consider where and when they collect data.

   3e. Trust among individuals providing their ethnicity data needs to be built.

   3f. Inclusive and diverse participatory governance should be embedded in data collection processes.

4. **Develop national guidance and training for data collectors in NHS and social care settings.**

4a: Standard guidance is distributed across health and care settings for healthcare professionals and other staff to consider in their interactions with patients.

4b: Staff training to support standardisation of recording and to overcome some of the perceived barriers to data collection is required.

4c. Training material and guidance should be developed with input from ethnic minority public contributors.

Facilitate data linkage to enrich the information available on ethnicity.

5a: Researchers need to evaluate the ethnicity content of data sources at the feasibility stage of their studies.

5b: A definition of ethnicity data tag for each of the existing data sources is needed.

5c. Reproducibility and transparency of curation of ethnicity data and derived categories should be ensured.

The value of collecting and recording information about ethnicity to ensure reliability in research, to inform healthcare practice and to deliver equitable care was clear. However being open about why this data is being collected and how it can be used for public benefit is crucial. Working with groups such as Understanding Patient Data and useMYdata who seek to understand views across the public and can support the development of strategies for better trust, transparency and communication will be key. Through the Alliance, we hope to raise awareness of the challenges experienced by researchers when trying to analyse data and compare research internationally. We will continue to engage with organisations such as the Centre for Ethnic Health Research, the NHS Race and Health Observatory, the Office for National Statistics, and the STANDING Together initiative among others who are actively working towards tackling health inequalities, to encourage key players and decision makers to take action towards a more standardised way to collect ethnicity data as well as data relevant to protected characteristics and demographics. Moving forward, we hope to build on these recommendations and facilitate implementation across organisations to ensure everyone is included and benefits from health and social care-data research.

UK Health Data Research Alliance

Ackowledgments

We thank all the members of the Alliance Special Interest Group on Ethnicity Coding Standards for actively participating in the workshops and for providing input on this paper.

Article co-authors: Paola Quattroni (HDR UK), Kamlesh Khunti (Co-Chair, University of Leicester), Ashley Akbari (Co-Chair, Swansea University), Ash Routen (University of Leicester), Sara Khalid (University of Oxford), Alastair Denniston (University of Birmingham), Angela Wood (University of Cambridge), Alexia Sampri (University of Cambridge), Jonathan Valabhji (NHS England), Holly Tibble (University of Edinburgh), Diana Withrow (University of Oxford), Hajira Dambha-Miller (University of Southampton).

UK Health Data Research Alliance

Appendix 1

List of organisations represented in the Alliance Special Interest group on ethnicity coding standards.

1. Ada Lovelace Institute
2. AstraZeneca
3. BREATHE Hub
4. British Heart Foundation Data Science Centre
5. Cancer Research UK
6. Cardiff University
7. Clinical Informatics & Health Outcomes Research Group, University of Oxford
8. Clinical Practice Research Datalink (CPRD)
9. DATA-CAN
10. Digital Health and Care Innovation Centre
11. Fiocruz Institute
12. Genomics England
13. GOSH
14. Health Data Research UK
15. Healthcare Quality Improvement Partnership
16. Imperial College London
17. Independent Cancer Patient Voice
18. INSIGHT Hub
19. Imperial College Healthcare NHS Trust
20. King's College London
21. Lambeth Early Action Partnership
22. London Borough of Lambeth
23. Medicines and Healthcare Products Regulatory Agency
24. National Institute for Health and Care Research (NIHR) Clinical Research Network
25. National Records of Scotland
26. NHS England
27. NHS Southeast London Clinical Commissioning Group
28. Nottingham University Hospitals NHS Trust
29. North Trent Cancer Research Network Consumer Research Panel
30. Office for National statistics (ONS)
31. Our Future Health

UK Health Data Research Alliance

32. Oxford University Hospitals NHS Foundation Trust
33. Oxford Health NHS Foundation Trust
34. Oxford University Hospitals NHS Foundation Trust
35. Public Health Agency of Canada
36. Public Health Scotland
37. Public Health Wales
38. UK Longitudinal Linkage Collaboration (UK LLC)
39. QResearch
40. Queen Mary University of London
41. Research Data Scotland
42. SAIL Databank
43. Southeast London Joint Medicines Formulary
44. South London and Maudsley NHS Trust
45. Swansea University
46. The Royal Marsden NHS foundation Trust
47. The Human Genome Project (National Human Genome Research Institute)
48. University of Cambridge
49. University of Birmingham
50. University of Exeter
51. University of Glasgow
52. University of Leeds
53. University of Leicester
54. University of Liverpool
55. University Hospitals Birmingham NHS Foundation Trust
56. University Hospitals of Leicester NHS Trust
57. University of Nottingham
58. University of Oxford
59. UseMYdata

UK Health Data Research Alliance

Appendix 2

Highlights from the ethnicity coding standards working group sessions

To address challenges and opportunities in data reporting ethnicity, we convened the community to share insights from ongoing work and understand the gaps that might need attention. Representatives from the biomedical domain provided their perspectives on the topic and initiated discussions spanning from issues around data collection, the need for training data collectors and healthcare professionals, improved communication with the public and patients, the use of terminology and the importance of building public trust and reaching out to minority groups.

The main discussion points are summarised below and have directly informed the proposed recommendations for improving the collection and use of ethnicity data to drive high-quality research.

1. Ethnicity: data recording and completeness

Community discussions focussed on how ethnicity information is currently captured and how information obtained from different data sources, including birth records, primary and secondary care electronic health records, surveys, cohort studies, registers, census data, and social care can vary. Many health and social care data sources either do not mandate the inclusion of ethnicity recording, or are varied in their collection and recording, leading to data quality challenges. The group highlighted:

- Incompleteness or missingness in coding and recordings.
- Uncertainty of what coding means or changes over time and/or between data sources, possibly due to lack of documentation.
- The need for better metadata to differentiate the differences such as ethnicity, race, lineage, and country of origin.
- Lack of consistency and granularity across data capture systems and settings across different health and social care services, along with local, regional and national data sources across the UK in recording information.
- Difficulties in harmonising data due to changes to coding standards and consistent use of ethnicity data standards are needed, leading to discrepancies where an individual's ethnicity is recorded in a way in different data sources.
- Reasons for collecting this data can vary (e.g. for insurance applications, administrative, bespoke collections for surveys, trials and research studies) and may not be well understood by individuals, which impacts how an individual may declare their ethnicity.
- Ethnicity data collection practices across health services vary, and the effects of variable ethnicity data capture and completeness are felt more acutely by particular communities; for example, Black

and South Asian communities typically experience more gaps in their ethnicity coding [35]. This can affect the healthcare provided to these groups, as well as research being inclusive of these groups, and reduce trust amongst these groups, with trust being an essential part of ethnicity data collection.

- Researchers should have access to rules and data curation methodologies to make sense of the complexities involved in using this data [36]. However more work is needed to reach consensus around standards for data curation methodologies.
- Ethnic categories evolved over time.The latest ONS Census data (2021) captures new information and ethnic categories [37].
- Guidance for best practice in how to collect ethnicity data in a standardised way would be helpful for research and clinical practice and address different ways of collecting and reporting ethnicity data across all settings.
- Harmonising data standards and codes across the UK's four nations, and agreeing on a framework that can also be adopted internationally to enable comparisons is a top priority that needs addressing.

2. The importance of definitions and terminologies

The group meetings discussed the importance of agreeing on definitions and clarifying terminology used in various contexts. In particular, the group highlighted:

- The distinction between ethnicity and race is not clear. There is also some confusion as to what we mean by using terms such as lineage, ancestry, nationality, country of birth, family origins, migrant status, language and religion, among others. People might interpret this terminology differently.
- Confusion may arise from the fact that ethnicity is a social construct.
- The ONS guidance on race and ethnicity terminology can offer an opportunity to drive consistency among groups. Additional resources are provided via The Law Society website.

The discussion highlighted that the development of a list of terminologies and definitions to ensure common understanding when using ethnicity data in research would be helpful, and sharing of best practices and existing mappings to enable reproducibility. The ONS guidance on language around race and ethnicity terminology is a good start to drive consistency.

3. Training and support for data collectors

One of the main discussions among participants was on the key role of practitioners, health and social care professionals and other individuals involved in collecting information about ethnicity directly from patients and the public. There was uncertainty about the processes used to collect information across various settings. Training, guidance and improved awareness of standard recommended processes could help reduce inconsistencies. The group reported:

- Improvements in the conversation between data collectors and participants are needed. We should consider how to make participants feel at ease, so they feel comfortable and confident providing this information. This could lead to better quality data.
- Data collectors could offer support when participants provide this information. Helping the public to understand the categorisations could be useful. Some people may not know their ethnicity or may feel restricted by the categories that are available.
- The right for participants to opt-out of providing and sharing their data should be honoured. Such data should be recognised as a separate category, rather than considered missing or unknown data.
- Ensure the data collectors understand the importance of recording ethnicity data. A lack of understanding may lead to reluctance in collecting ethnicity data.

The group suggested that the development of standard guidance for data collectors, practitioners and health and social care professionals to ensure information is requested consistently across sectors is needed.

UK Health Data Research Alliance

Appendix 3

Office for National Statistics ethnicity categories

Ethnic group variable: Census 2021

Definition of ethnic group, categories, and changes since the 2011 Census for use with research and analysis using Census 2021 data.

Definition

The ethnic group that the person completing the census feels they belong to. This could be based on their culture, family background, identity or physical appearance.

Respondents could choose one out of 19 tick-box response categories, including write-in response options.

Classification

Total number of categories: 20

CodeName

1 Asian, Asian British or Asian Welsh: Bangladeshi

2 Asian, Asian British or Asian Welsh: Chinese

3 Asian, Asian British or Asian Welsh: Indian

4 Asian, Asian British or Asian Welsh: Pakistani

5 Asian, Asian British or Asian Welsh: Other Asian

6 Black, Black British, Black Welsh, Caribbean or African: African

7 Black, Black British, Black Welsh, Caribbean or African: Caribbean

8 Black, Black British, Black Welsh, Caribbean or African: Other Black

9 Mixed or Multiple ethnic groups: White and Asian

10 Mixed or Multiple ethnic groups: White and Black African

11 Mixed or Multiple ethnic groups: White and Black Caribbean

12 Mixed or Multiple ethnic groups: Other Mixed or Multiple ethnic groups

13 White: English, Welsh, Scottish, Northern Irish or British

UK Health Data Research Alliance

CodeName

14	White: Irish
15	White: Gypsy or Irish Traveller
16	White: Roma
17	White: Other White
18	Other ethnic group: Arab
19	Other ethnic group: Any other ethnic group
-8	Does not apply*

*Students and schoolchildren living away during term-time.

Source

https://www.ons.gov.uk/census/census2021dictionary/variablesbytopic/ethnicgroupnationalidentitylanguageandreligionvariablescensus2021/ethnicgroup/classifications

UK Health Data Research Alliance

References

[1]https://style.ons.gov.uk/house-style/race-and-ethnicity/

[2]https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908434/Disparities_in_the_risk_and_outcomes_of_COVID_August_2020_update.pdf

[3] https://www.gov.uk/government/publications/final-report-on-progress-to-address-covid-19-health-inequalities/final-report-on-progress-to-address-covid-19-health-inequalities

[4] https://bmjopen.bmj.com/content/bmjopen/12/8/e064276.full.pdf

[5] https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/articles/understandingconsistencyofethnicitydatarecordedinhealthrelatedadministrativedatasetsinengland2011to2021/november2023

[6]Probabilistic Approaches for Data Integration in Biomedical Research. Sampri, A. (Author). 31 Dec 2022. Student thesis: Phd

[7] https://ebooks.iospress.nl/publication/54190

[8] https://analysisfunction.civilservice.gov.uk/policy-store/ethnicity-harmonised-standard/

[9] https://www.gov.uk/government/collections/ethnicity-data-methods-and-quality-reports

[10] https://ukhealthdata.org/wp-content/uploads/2022/07/HDRUK_ALLIANCE_BROCHURE.pdf

[11] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[12] UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems (1.0). Zenodo. https://doi.org/10.5281/zenodo.5767586

[13] Nada Karrar, Shahriar Kabir Khan, Sinduja Manohar, Paola Quattroni, David Seymour, & Susheel Varma. (2021). Improving transparency in the use of health data for research: Draft recommendations for a data use registers standard. https://doi.org/10.5281/zenodo.5084761

[14] Ben Gordon, Jake Barrett, Clara Fenessy, Caroline Cake, Adam Milward, Courtney Irwin, Monica Jones, & Neil Sebire. (2020). Data Utility Framework. https://doi.org/10.5281/zenodo.4594783

[15] Alex Bailey, Garry Coleman, Alan Harbinson, Varsha Khodiyar, Naomi Mill, Carole Morris, Chris Orton, Paola Quattroni, David Seymour, Peter Stokes, & Susheel Varma. (2022). Five Safe Data Access Request application form. Zenodo. https://doi.org/10.5281/zenodo.5946892

[16]https://ukhealthdata.org/wp-content/uploads/2021/12/211124-White-Paper-Recommendations-of-Data-Standards-v2-1.pdf

[17] https://www.hdruk.ac.uk/wp-content/uploads/2021/08/Data-Utility-Framework_withlink_updatedDC100821.pdf

[18] https://www.hdruk.ac.uk/about-us/policies/diversity-and-inclusion-2/

[19] https://www.datadiversity.org/

[20] https://ukhealthdata.org/projects/diversity-in-data/

[21] Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, van der Meulen JH. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017 Oct 1;46(5):1699-1710. doi: 10.1093/ije/dyx177. PMID: 29025131; PMCID: PMC5837697

[22] https://analysisfunction.civilservice.gov.uk/policy-store/ethnicity-harmonised-standard/

[23]https://service-manual.ons.gov.uk/content/language/ethnicity-and-race

[24] https://www.ons.gov.uk/census/census2021dictionary/variablesbytopic/ethnicgroupnationalidentitylanguageandreligionvariablescensus2021/ethnicgroupdetailed

[25] https://pubmed.ncbi.nlm.nih.gov/35641824/

[26] https://www.bmj.com/content/371/bmj.m4493.long

[27] https://pubmed.ncbi.nlm.nih.gov/35641824/

[28] https://www.gov.uk/government/publications/life-sciences-vision

[29] https://dataingovernment.blog.gov.uk/2022/01/25/comparing-ethnicity-data-for-different-countries/

[30] https://one.oecd.org/document/SDD/DOC(2018)9/En/pdf

[31] https://commission.europa.eu/system/files/2021-09/data_collection_in_the_field_of_ethnicity.pdf

[32]https://www.ohchr.org/sites/default/files/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf

[33] www.datadiversity.org/recommendations

[34]https://service-manual.ons.gov.uk/content/language/ethnicity-and-race

[35] Scobie S, Spencer, J, Raleigh V. Ethnicity coding in English health service datasets. London: Nuffield Trust ; [Internet]. 2021 [cited 2022 Nov 28].Available from: https://www. nuffieldtrust.org.uk/files/2021-06/1622731816_nuffield-trust-ethnicity-coding-web.pdf

[36] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf

[37] https://census.gov.uk/census-2021-results