|  |  |
| --- | --- |
| **Code List Checklist\*: Ethnicity** | |
| **Metadata** | |
| Title | Ethnicity |
| Name | Ethnicity |
| Author | Helen Strongman |
| Target data source | Clinical Practice Research Datalink Aurum and GOLD (March 2023) |
| Terminology | Read codes (GOLD) SNOMED-CT (Aurum) and EMIS codes (Aurum) |
| **Definition of clinical concept** | |
| Concept | All codes indicating an individual's ethnicity as conceptualised and defined by the Office of National Statistics (ONS).    The conceptualisation of ethnicity evolves over time and so should be reevaluated periodically. |
| Timeframe | No restrictions |
| Accuracy | See “synonyms” |
| Setting | Ethnicity recorded in primary care. |
| **Identify and evaluating existing code lists** | |
| Source searched | Contacted LSHTM researchers to find latest version of ethnicity code list developed by Mathur et al. |
| Existing code lists found | **Original**  Date created: 2012  Author: Rohini Mathur (Epidemiologist, LSHTM)  Clinical advisors: Liam Smeeth & CPRD team  **Updates**  Date updated: multiple updates including adaptation to Aurum and new  categorisations of the 2011 and 2021 census  Author: Rohini Mathur (Epidemiologist, LSHTM)  Most recent update: 05/06/2023  Updated by: Helen Strongman (Epidemiologist, LSHTM)  Advisor: Rohini Mathur  Changes: created Stata do files to update the original code list and explain decision making steps. Minor updates made to categorisation. |
| **Create a new code list (rerun of existing algorithm)** | |
| Synonyms | Codes with the following concepts in the Snomed or Read term were included:   * codes that specified that they referred to ethnicity/ethnic groups or categories * mixed and non-mixed ethnicity codes * codes that included terms such as white, black, Caucasian, Asian * codes in ethnicity Read chapters that are synonymous with, or represent countries with a largely homogonous ethnic population (e.g. Japanese, Indian) |
| Exceptions | Codes that described other countries or regions, nationalities or religions  were not included (e.g. England, Europe, British, Christian). |
| Methods used | A script of lower-case search terms was used to identify relevant READ, medcodeid and snomedctconceptid fields in the CPRD medical dictionary (March 2023). Terms anywhere in the string were identified. An additional search of Read Chapters was undertaken. |
| Search terms | **Inclusion:** ethn, british, english, welsh, scottish, irish, asian, indian, pakistani, bangladeshi, african and caribbean  **Exclusions:** adverse, eye, flower, urine, blood, blackout, adenoma, adore, cell, country, head, adverse, syndrome, veno, sigvaris, fruit, trypanosomiasis, poison, kit, altiven, sick, religion, algi, juzo, jobst, haddenham, gloria, read, language, ovcc, exam, vari, porphyria, assoc, carer, speak, test, encep, trust, contracept, interpret, health, operation, body, church, obscure, joint, patient, refer, activity, episco, type, child, member, fever, skin, virus, father, mother, gfr, myocard, sis, risk, society, orthodox, tick, buddhis, benign, born, lewis, traga, citizen, cattle, africanum, diet, counties, stew, hindu, jewish, muslim, sikh. |
| Hierarchy used to extend search | Inclusion Read chapters: 9S 9i 9t 9T |
| Decisions made while iterating | No additional inclusion or exclusion terms were found. |
| (Optional) Categories | Ethnicity codes were grouped as follows using the eth5 and eth16 variables:  **eth5:**  0. White  1. South Asian  2. Black  3. Other  4. Mixed  5. Not Stated    **eth16:**  1. British  2. Irish  3. Other White  4. White and Black Caribbean  5. White and Black African  6. White and Asian  7. Other Mixed  8. Indian  9. Pakistani  10. Bangladeshi  11. Other Asian  12. Caribbean  13. African  14. Other Black  15. Chinese  16. Other ethnic group  17. Not Stated  88. Unknown |
| **Review, finalise and publish** | |
| Reviewers | Rohini Mathur |
| Scope of review | The draft code list and both the search and exclusion terms were reviewed. |
| Evidence of review | The process is documented in the following HTML files:  codelist\_ethnicity\_derivation\_aurum.html  codelist\_ethnicity\_derivation\_gold.html  (See “Resources published” section for location). |
| Internal checks | A new codelist was created and compared to the previous version. |
| External checks | Rohini Mathur and colleagues developed an algorithm to derive ethnicity using earlier versions of these code lists. Where >1 ethnicity code was recorded in a patient’s primary care records, if discordance was found, the algorithm selected the most frequently recorded ethnicity. If discordant entries were equally common, the most recently recorded ethnicity was used. Where ethnicity was unknown or missing, it was replaced with that recorded in the Hospital Episode Statistics data, where available (using the same discordance logic as the primary care data).  The algorithm was validated by comparing the ethnicity distribution generated by it, to that of the 2001 UK Census. The distribution in the primary care data was found to be comparable to that of the Census when it was collapsed into the five higher-level ethnicity categories (eth5). For CPRD data after 2006, 78% of patients had ethnicity recorded. |
| Code list(s) published | The code list is published on the project’s LSHTM Data Compass: <https://doi.org/10.17037/DATA.00004742>.  And in the study’s Github repository at:  <https://github.com/hstrongman/OSA-narc-CPRD-chronology> (Aurum/GOLD March 2023)  Files include:  codelist\_ethnicity\_aurum.txt  codelist\_ethnicity\_aurum.dta  codelist\_ethnicity\_gold.txt  codelist\_ethnicity\_gold.dta |
| Resources published | Original Study:  Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, VanStaa T, Grundy E, Smeeth L. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. Journal of public health (Oxford, England). 2014;36(4):684–92.  <https://pubmed.ncbi.nlm.nih.gov/24323951>  Guidance flow diagram summarising how to generate the ethnicity variable from the data (Mather, Chapter 5, P74):  Mathur R. Ethnic inequalities in health and use of healthcare in the UK: how computerised health records can contribute substantively to the knowledge base [Internet] [doctoral]. Available from: <https://researchonline.lshtm.ac.uk/id/eprint/2478832/>  \*\* Note CPRD have validated an ethnicity code list and algorithm.  The code lists are available in the supplementary appendix to Shiekh et al 2023 but they are not matched to categories\*\*  Shiekh, S.I., Harley, M., Ghosh, R.E. et al. Completeness, agreement, and representativeness of ethnicity recording in the United Kingdom's Clinical Practice Research Datalink (CPRD) and linked Hospital Episode Statistics (HES). Popul Health Metrics 21, 3 (2023). <https://doi.org/10.1186/s12963-023-00302-0>  The Do files and HMTL documents describing the code list derivation and search strategy are in the study’s GitHub repository.  Files include:  codelist\_ethnicity.do (refers to associated text files)  codelist\_ethnicity\_description.html (early version of this checklist)  codelist\_ethnicity\_derivation\_aurum.html  codelist\_ethnicity\_derivation\_gold.html |

\* \*Matthewman J, Andresen K, Suffel A, Lin LY, Schultze A, Tazare J, Bhaskaran K, Williamson E, Costello R, Quint J, Strongman H. Checklist and guidance on creating codelists for routinely collected health data research [version 2; peer review: 3 approved]. NIHR Open Res. 2024 Sep 18;4:20.