CPRD GOLD Ethnicity Record Documentation

22nd May 2023

Documentation Control Sheet

Over time, it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made and should be filed inside the front cover of the new or amended document.

Version	Affected Areas of Summary of Change	Prepared by	Date	Reviewed by	Date
1.0	Initial Draft	Eleanor Axson	25/04/2023	Suhail Shiekh	16/05/2023

Summary of changes

Version 1.0 Initial Draft

CPRD GOLD Ethnicity Record data

Introduction

Ethnic inequalities in health have been widely documented and remain a priority for epidemiological and health services research. Reliable and accurate ethnicity data is essential to further understand ethnic inequalities in health and adapt health services to address the needs of underserved ethnic groups. The CPRD databases have been shown to be representative of the ethnic distributions of the UK and the devolved nations [1].

Ethnicity records for patients in CPRD can be found in several locations, including in their primary care record in CPRD GOLD and in linked secondary care datasets [2], including Hospital Episode Statistics (HES) Admitted Patient Care (HES APC) data, HES Outpatient (HES OP) data, HES Accident and Emergency (HES A&E) data, and the HES Diagnostic Imaging Dataset (HES DID).

An algorithm has been developed to make best use of the available ethnicity data, to maximise the number of patients with an ethnicity categorisation, and to provide a standardised procedure for ethnicity categorisation using CPRD databases. The CPRD Ethnicity Record lists categorised ethnicity for all patients in CPRD GOLD for whom ethnicity data is available in either primary care or secondary care records.

Methodology

The CPRD Ethnicity Record dataset is comprised of a single derived ethnicity category for each patient in CPRD GOLD as categorised into various ethnic categorisations by the most recent version of the CPRD Ethnicity Algorithm.

1. Ethnicity Categories

The CPRD Ethnicity Record provides ethnicity categorised in six categories (Asian, black, mixed/multiple, white, other unknown).

2. Selection of Ethnicity Records

A list of ethnicity-related SNOMED, Read, and local EMIS® codes was used to extract all records from the Observation file relating to ethnicity in CPRD GOLD.

Where available, ethnicity from linked HES APC, HES OP, and HES A&E datasets [2] was obtained from the CPRD-recoded ‘ethnos’ variable (see below) and ethnicity from HES DID was obtained from the ‘did_ethcat’ variable.

Ethnicity recording in HES APC, HES OP, and HES A&E

Ethnicity (ethnos) is recorded in each episode of the original HES APC, HES OP, and HES A&E datasets and these are recoded by CPRD (see Table 1 below) and provided in the HES data tables.

Original 'ethnos' value	CPRD-recoded 'ethnos' value
0 = White A = British (White) B = Irish (White) C = Any other White background	White
1 = Black – Caribbean M = Caribbean (Black or Black British)	Black_Caribbean
2 = Black – African N = African (Black or Black British)	Black_African
3 = Black – Other P = Any other Black background	Black_Other
4 = Indian H = Indian (Asian or Asian British)	Indian
5 = Pakistani J = Pakistani (Asian or Asian British)	Pakistani
6 = Bangladeshi K = Bangladeshi (Asian or Asian British)	Bangladeshi
L = Any other Asian background	Other_Asian
7 = Chinese R = Chinese (other ethnic group)	Chinese
D = White and Black Caribbean (Mixed) E = White and Black African (Mixed) F = White and Asian (Mixed) G = Any other Mixed background	Mixed
8 = Any other ethnic group S = Any other ethnic group	Other
9 = Not given X = Not known Z = Not stated	Unknown

Table 1: Ethnicity recoding by CPRD

3. Ethnicity Categorisation

Patients with a single ethnicity recording in the data are assigned to the appropriate ethnic category based on the value of that single record.

Patients with multiple and/or conflicting ethnicity recordings in the data are assigned an ethnic category based on the following procedure:

1) The most frequently recorded ethnicity code in all available data:
    - CPRD GOLD
    - HES APC (2003-onwards)
    - HES A&E (2007-2020)
    - HES OP (2003-onwards)
    - HES DID (2012-onwards)

2) IF there are multiple ethnicities with the same frequency, the most recent is chosen

3) IF there are multiple ethnicities with the same frequency and most recent date, precedence is given in the following order:
    - CPRD GOLD value
    - HES APC value
    - HES A&E value
    - HES OP value
    - HES DID value

4) IF there are multiple ethnicities with the same frequency, most recent date, and data source, the ethnicity that occurs most frequently in the 2021 England and Wales census [3].

5) IF 'other' is the most frequently recorded ethnic group, then the second most frequently recorded ethnic group is assigned instead
    - Application of steps 1-4 for the second most frequently recorded ethnic group
    - A value of 'other' ethnicity will only be assigned if there are no other useable ethnic groups coded

The CPRD Ethnicity Algorithm does not impute missing ethnicity and a value of 'unknown' ethnicity will be assigned if there are no known ethnicities recorded in any available dataset.

Additionally, the CPRD Ethnicity Algorithm will assign a value of 'unknown' ethnicity if there is a code indicating the patient has declined to provide ethnicity data at any time in their record.

HES Linkage Eligibility

Availability of ethnicity records from the linked HES datasets is limited to those patients meeting the linkage eligibility for the relevant HES dataset per the Linkage Eligibility dataset specified in the relevant CPRD Ethnicity Record Release Notes. If a patient was not eligible for linkage to HES datasets, only primary care records were available to the algorithm. The availability of HES linkage is for England-based patients only. Users should be aware of the implications of data availability on the chances of a patient having a known versus unknown ethnicity value and consider these in relation to their work.

Backwards Compatibility

Because CPRD primary care data and linked data are dynamic and updated data sources, all new issues of the CPRD Ethnicity Record cannot guarantee to contain the same categorisation of ethnicity for patients. Furthermore, it is likely that additional data sources will be added and refinements to the algorithm/underlying code lists will be developed over time which may render the CPRD Ethnicity Record non-backwards compatible.

Using the CPRD Ethnicity Record in Publications

When reporting and writing about ethnicity, CPRD recommends following the UK Government style guide for writing about ethnicity [4], where appropriate.

The following statement is required to appear in publications that have utilised the CPRD Ethnicity Record:

> The CPRD Ethnicity Record sources underlying data from Hospital Episode Statistics (HES) Copyright © [CURRENT YEAR], re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

Caveats and Notes

No restrictions on the research standard 'acceptable' flag were included in the CPRD Ethnicity Record. Restrictions on acceptability can be applied retrospectively, by the user, if needed, using the 'acceptable' flag in the Patient table in CPRD GOLD.

No restrictions on date of ethnicity recording were included in the CPRD Ethnicity Record; therefore, any ethnicity recording occurring at any time in a patient's record in primary and secondary care was available for the algorithm.

If you need any further information on the CPRD Ethnicity Record, please contact CPRD Enquiries (enquiries@cprd.com).

Relevant Publications

Shiekh SI 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). *Population Health Metrics*. 2023 Mar 14;21(1). https://doi.org/10.1186/s12963-023-00302-0

CPRD Ethnicity Record: Data dictionary

Column name	Description	Type	Format
patid	CPRD GOLD patient identifier	INTEGER	20
ethnic_6	Six category ethnicity values: 1 = Asian 2 = black 3 = mixed/multiple 4 = white 5 = other 6 = unknown	INTEGER	1

References

1. Shiekh SI, Harley M, Ghosh RE, Ashworth M, Myles P, Booth HP, 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 Metr. 2023 Mar 14;21(1).
2. Clinical Practice Research Datalink. CPRD linked data [Internet]. 2022 [cited 2022 Mar 4]. Available from: https://www.cprd.com/linked-data
3. Office for National Statistics. Ethnic group, England and Wales: Census 2021 [Internet]. 2022 [cited 2023 Apr 25]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021#ethnic-groups-in-england-and-wales
4. UK Government. Writing about ethnicity [Internet]. Style Guide. 2021 [cited 2022 May 11]. Available from: https://www.ethnicity-facts-figures.service.gov.uk/style-guide/writing-about-ethnicity