The OpenSAFELY Collaborative *et al.*
BMC Medicine (2024) 22:288
https://doi.org/10.1186/s12916-024-03499-5

BMC Medicine

RESEARCH ARTICLE

Open Access

Consistency, completeness and external validity of ethnicity recording in NHS primary care records: a cohort study in 25 million patients' records at source using OpenSAFELY

The OpenSAFELY Collaborative, Colm D. Andrews[1*], Rohini Mathur[2,4], Jon Massey[1], Robin Park[1], Helen J. Curtis[1], Lisa Hopcroft[1], Amir Mehrkar[1], Seb Bacon[1], George Hickman[1], Rebecca Smith[1], David Evans[1], Tom Ward[1], Simon Davy[1], Peter Inglesby[1], Iain Dillingham[1], Steven Maude[1], Thomas O'Dwyer[1], Ben F. C. Butler-Cole[1], Lucy Bridges[1], Chris Bates[3], John Parry[3], Frank Hester[3], Sam Harper[3], Jonathan Cockburn[3], Ben Goldacre[1], Brian MacKenna[1], Laurie A. Tomlinson[2], Alex J. Walker[1] and William J. Hulme[1]

Abstract

Background Ethnicity is known to be an important correlate of health outcomes, particularly during the COVID-19 pandemic, where some ethnic groups were shown to be at higher risk of infection and adverse outcomes. The recording of patients' ethnic groups in primary care can support research and efforts to achieve equity in service provision and outcomes; however, the coding of ethnicity is known to present complex challenges. We therefore set out to describe ethnicity coding in detail with a view to supporting the use of this data in a wide range of settings, as part of wider efforts to robustly describe and define methods of using administrative data.

Methods We describe the completeness and consistency of primary care ethnicity recording in the OpenSAFELY-TPP database, containing linked primary care and hospital records in > 25 million patients in England. We also compared the ethnic breakdown in OpenSAFELY-TPP with that of the 2021 UK census.

Results 78.2% of patients registered in OpenSAFELY-TPP on 1 January 2022 had their ethnicity recorded in primary care records, rising to 92.5% when supplemented with hospital data. The completeness of ethnicity recording was higher for women than for men. The rate of primary care ethnicity recording ranged from 77% in the South East of England to 82.2% in the West Midlands. Ethnicity recording rates were higher in patients with chronic or other serious health conditions. For each of the five broad ethnicity groups, primary care recorded ethnicity was within 2.9 percentage points of the population rate as recorded in the 2021 Census for England as a whole. For patients with multiple ethnicity records, 98.7% of the latest recorded ethnicities matched the most frequently coded ethnicity. Patients whose latest recorded ethnicity was categorised as Other were most likely to have a discordant ethnicity recording (32.2%).

Conclusions Primary care ethnicity data in OpenSAFELY is present for over three quarters of all patients, and combined with data from other sources can achieve a high level of completeness. The overall distribution of ethnicities

*Correspondence:
Colm D. Andrews
colm.andrews@phc.ox.ac.uk
Full list of author information is available at the end of the article

across all English OpenSAFELY-TPP practices was similar to the 2021 Census, with some regional variation. This report identifies the best available codelist for use in OpenSAFELY and similar electronic health record data.

Keywords Primary care health sciences, Electronic health records, Ethnicity, Data curation

Background

Ethnicity is known to be an important determinant of health inequalities, particularly during the COVID-19 outbreak where a complex interplay of social and biological factors resulted in increased exposure, reduced protection and increased severity of illness in particular ethnic groups [1, 2]. The UK has a diverse ethnic population (The 2021 Office for National Statistics (ONS) Census estimated 9.6% Asian, 4.2% Black, 3.0% Mixed, 81.0% White, 2.2% Other [3]), which can make health research conducted in the UK generalisable to countries. Complete and consistent recording of patients' ethnic group in primary care can support efforts to achieve equity in service provision and reduces bias in research [4, 5]. Ethnicity recording for new patients registering with general practice across the UK has improved following Quality and Outcomes Framework (QOF) financial incentivisation between 2006/07 and 2011/12 [6, 7]. As a result, ethnicity is now being captured for the majority of the population in routine electronic healthcare records and is comparable to the general population [6]. The uptake and utilisation of healthcare services still varies across ethnic groups, and the recently established NHS Race and Health Observatory have led calls for a dedicated drive by NHS England and NHS Digital to emphasise the importance of collecting and reporting ethnicity data [8].

OpenSAFELY is a secure health analytics platform created by our team on behalf of NHS England. OpenSAFELY provides a secure software interface allowing analysis of pseudonymised primary care patient records from England in near real-time within highly secure data environments.

In primary care data, patient ethnicity is recorded via clinical codes, similar to how any other clinical condition or event is recorded. In OpenSAFELY-TPP, both Clinical Terms Version 3 (CTV3 (Read)) codes and Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) codes are used. SNOMED CT is an NHS standard, widely used across England.

Ethnicity is also recorded in secondary care, when patients attend emergency care, inpatient or outpatient services, independently of ethnicity in the primary care record. This is available via NHS England's Secondary Uses Service (SUS) [9]. It is common practice in OpenSAFELY to supplement primary care ethnicity, where missing, with ethnicity data from SUS [10, 11]. Throughout this paper, we refer to ethnicity rather than race as recommended by the ONS: 'The word "race" places people into categories based on physical characteristics, whilst ethnicity is self-defined and includes aspects such as culture, heritage, religion and identity'. However, we recognise that the distinction between and use of these terms may differ in different settings.

In this paper, we study the completeness, consistency and representativeness of routinely collected ethnicity data in primary care.

Methods

Study design

Retrospective cohort study across 25 million patients registered with English general practices in OpenSAFELY-TPP.

Data sources

This study uses data from the OpenSAFELY-TPP database, covering around 40% of the English population. The database includes primary care records of patients in practices using the TPP SystmOne patient information system and is linked to other NHS data sources, including in-patient hospital records from NHS England's Secondary Use Service (SUS), where ethnicity is also recorded independently of ethnicity in the primary care record.

All data were linked, stored and analysed securely within the OpenSAFELY platform https://opensafely.org/. Data include pseudonymized data such as coded diagnoses, medications and physiological parameters. No free text data are included. All code is shared openly for review and re-use under MIT open licence (opensafely/ethnicity-short-data-report at notebook). Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared.

Study population

Patients were included in the study if they were registered at an English general practice using TPP on 1 January 2022.

Ethnicity ascertainment

In primary care data, there is no categorical 'ethnicity' variable to record this information. Rather, ethnicity is recorded using clinical codes—entered by a clinician or administrator with a location and date—like any other clinical or administrative event, with specific codes relating to each ethnic group [12–14]. This means ethnicity

can be recorded by the practice in multiple, potentially conflicting, ways over time.

We created a new codelist, SNOMED:2022 [13], by identifying relevant ethnicity SNOMED CT codes and ensuring completeness by comparing the codelist to the following: another OpenSAFELY created codelist (CTV3:2020) [13], a combined ethnicity codelist from SARS-CoV2 COVID19 Vaccination Uptake Reporting Codes published by Primary Care Information Services (PRIMIS) [12, 15] and a codelist from General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) [16]. Codes which relate to religion rather than ethnicity (e.g. 'Muslim—ethnic category 2001 census') and codes which do not specify a specific ethnicity (e.g. 'Ethnic group not recorded') were excluded. In total, 258 relevant ethnicity codes were identified. We then created a codelist categorisation based on the 2001 UK Census categories, which are the NHS standard for ethnicity [17], and cross referenced it against the CTV3, PRIMIS and GDPPR codelists. The 'Gypsy or Irish Traveller' and 'Arab' groups were not specifically listed in 2001 however we categorised them as `White` and `Other` respectively as per the 2011 Census grouping [18]).

The codelist categorisation consists of two ethnicity groupings based on the 2001 census (Table 1): all analyses used the 5-group categorisation unless otherwise stated.

If a SNOMED:2022 ethnicity code appeared in the primary care record on multiple dates, the latest entry was used unless otherwise stated.

In OpenSAFELY, the function ethnicity_from_sus combines SUS ethnicity data from admitted patient care statistics (APCS), emergency care (EC) and outpatient attendance (OPA) and selects the most frequently used ethnicity code for each patient. In hospital records from SUS, recorded ethnicity is categorised as one of the 16 categories on the 2001 UK census. This accords with the 16-level grouping described above.

Subgroups

We looked at the completeness of ethnicity coding in the whole population and across each of the following demographic and clinical subgroups:

Age

Patient age was calculated as of 1 January 2022 and grouped into 5-year bands, to match the ONS age bands.

Table 1 2001 ONS Census ethnicity groupings

5-level group:
Asian or Asian British
Black or Black British,
Mixed,
White
Chinese or other ethnic groups
16-level group:
Indian
Pakistani
Bangladeshi
Any other Asian background
Caribbean, African
Any other Black background
White and Black
Caribbean
White and Black African
White and Asian
Any other Mixed background
British
Irish
Any other White background
Chinese
Any other

Sex

We used categories 'male' and 'female', matching the ONS recorded categories; patients with any other/unknown sex were excluded.

Deprivation

Overall deprivation was measured by the 2019 Index of Multiple Deprivation (IMD) [19] derived from the patient's postcode at lower super output area level. IMD was divided by quintile, with 1 representing the most deprived areas and 5 representing least deprived areas. Where a patient's postcode cannot be determined the IMD is recorded as unknown.

Region

Region was defined as the Nomenclature of Territorial Units for Statistics (NUTS 1) region derived from the patient's practice postcode.

As the rate of ethnicity recording would be expected to be lower in patients with fewer clinical interactions, and therefore fewer opportunities for ethnicity to be recorded, completeness was also compared in the clinical subgroups of dementia, diabetes, hypertension and learning disability which are more likely to require additional clinical interactions. Clinical subgroups were defined as the presence or absence of relevant SNOMED CT codes in the GP records for dementia [20], diabetes [21], hypertension [22] and learning disabilities [23] as of 1 January 2022.

Statistical methods

Completeness and distribution of ethnicity recording

The proportion of patients with either (i) primary care ethnicity recorded (that is, the presence of any code in the SNOMED:2022 codelist in the patient record) or (ii) primary care ethnicity supplemented, where missing, with ethnicity data from secondary care [24] was calculated. Completeness was reported overall and within clinical and demographic subgroups.

Amongst those patients where ethnicity was recorded, the proportion of patients within each of the 5 groups was calculated, within each clinical and demographic subgroup. We also calculated the distribution of complete ethnicity recording across practices with at least 1000 registered patients.

Consistency of ethnicity recording within patients over time

Discrepancies may arise due to errors whilst entering the data or if a patient self-reports a different ethnic group from their previously recorded ethnic group. We calculated the proportion of patients with any ethnicity recorded which did not match their 'latest' recorded grouped ethnicity for each of the five ethnic groups.

We also calculated the proportion of patients whose latest recorded ethnicity did not match their most frequently recorded ethnicity for each of the five ethnic groups.

Consistency of ethnicity recording across data sources (primary care versus secondary care)

We calculated the proportion of patients whose latest recorded ethnicity in primary care matched their ethnicity as recorded in secondary care for each of the five ethnic groups, where both primary and secondary care are recorded.

External validation against the 2021 UK census population

The UK Census collects individual and household-level demographic data every 10 years for the whole UK population. Data on ethnicity were obtained from the 2021 UK Census for England. The most recent census across the UK was undertaken on 27 March 2021. Ethnic breakdowns for the population of England were obtained via NOMIS [25].

The ethnic breakdown of the census population was compared with our OpenSAFELY-TPP population and the relative difference was calculated using the ONS value as the baseline proportion and OpenSAFELY as the comparator. In the 2021 UK Census, the Chinese ethnic group was included in the Asian ethnic group, whereas in the 2001 census, it was included in the Other ethnic group [26]. In order to provide a suitable comparison with primary care data, we regrouped the 2021 census data as per the 2001 groups. As an additional analysis, we also compared the primary care data with the census data using the 2021 census categories.

Results

Completeness of ethnicity data

19,618,135 of the 25,102,210 patients (78.2%) registered in OpenSAFELY-TPP on 1 January 2022 had a recorded ethnicity, rising to 92.5% when supplemented with secondary care data (Fig. 1, Additional file 1: Table S1).

Primary care ethnicity recording completeness was lowest for patients aged over 80 years (80.1%) and under 30, whereas ethnicity recording was highest in those over 80 when supplemented with secondary care data (97.1%). Women had a higher proportion of recorded ethnicities than men (79.8% and 76.5% respectively, 94% and 91.1% when supplemented with secondary care data). The completeness of primary care ethnicity recording ranged from 77% in the South East of England to 82.2% in the West Midlands. IMD was within 1.2 percentage points for known values (77.7% in the least deprived group 5 to 78.9% in group 3) and was lowest for the unknown group (71.6%). Primary care ethnicity recording was at least 4

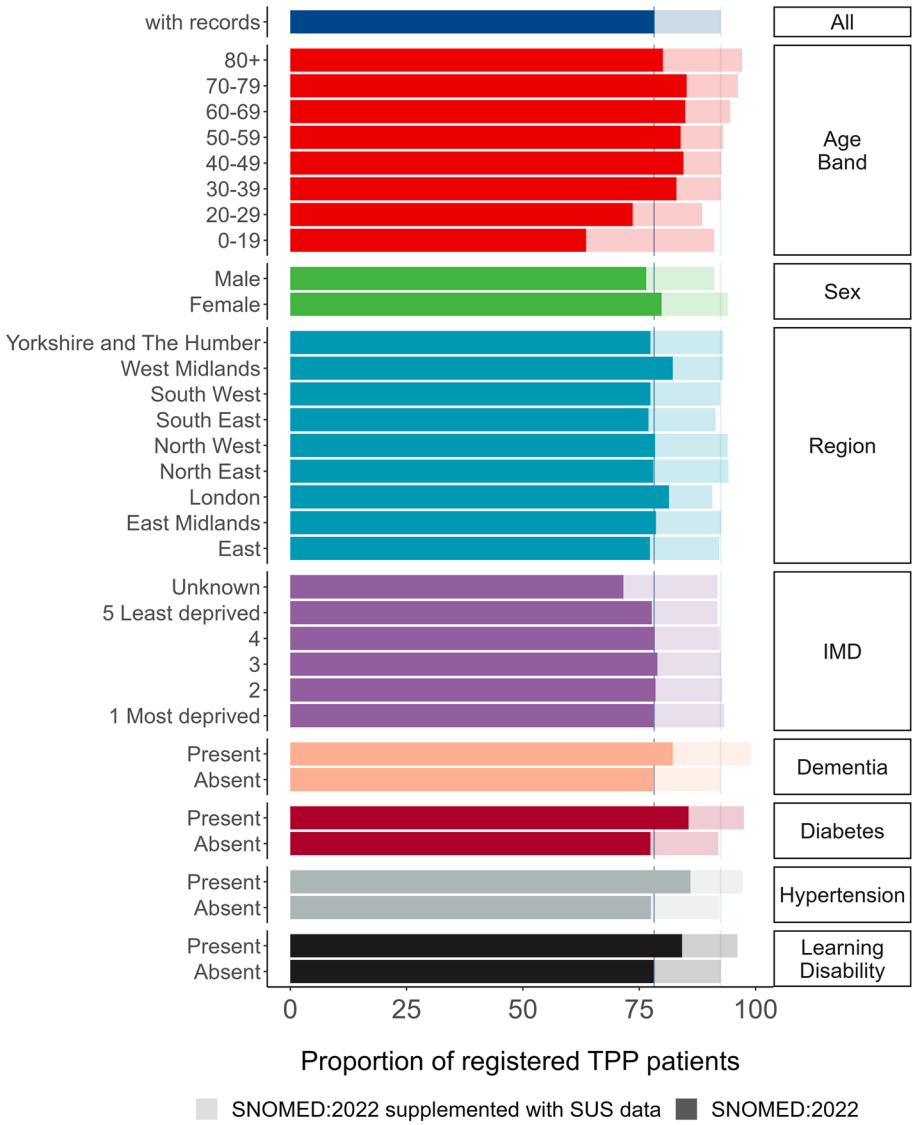

Fig. 1 Bar plot showing proportion of registered TPP population with a recorded ethnicity by clinical and demographic subgroups, based on primary care records (solid bars) and when supplemented with secondary care data (pale bars)

percentage points higher in all of the clinical subgroups compared to the general population.

Distribution of ethnicity

Using ethnicity recorded in primary care only, 6.8% of the population were recorded as Asian, 2.3% Black, 1.5% Mixed, 65.6% White and 1.9% Other, and ethnicity was not recorded for 21.8%. When supplementing with hospital-recorded ethnicity data, corresponding percentages were 7.8% Asian, 2.6% Black, 1.9% Mixed, 77.9% White, 2.3% Other and 7.5% not recorded, representing a percentage point increase ranging from 0.3% in the Black group to 12.3% in the White group.

Older patients tended to have a higher rate of recorded White ethnicity (e.g. 76.3% in the 80+group vs 50.0% in the 0–19 group), whereas younger patients had a higher rate of recording for Asian, Black, Mixed and Other groups. The higher proportion of women with recorded ethnicity was reversed in the Asian group where men (7.0% and 8.0% with secondary care data) had a higher proportion of recording than women (6.6% and 7.6% with secondary care data). The proportion of ethnicity reporting was lower for patients with dementia, hypertension or learning disabilities in every ethnic group other than White (Fig. 2/Additional file 1: Table S2). The breakdown by 16 group ethnicity is shown in Additional file 1:

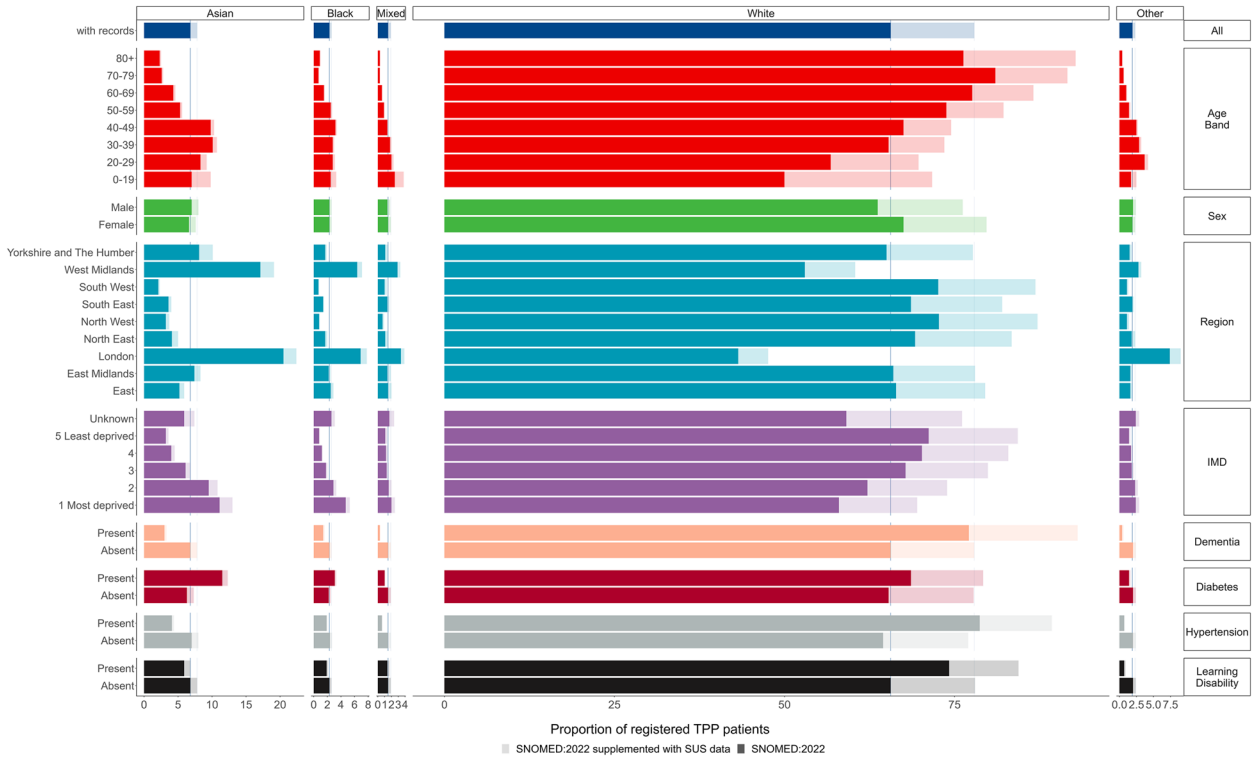

Fig. 2 Bar plot showing proportion of registered TPP population with a recorded ethnicity by clinical and demographic subgroups, based on primary care records (solid bars) and when supplemented with secondary care data (pale bars)

Table S3. There was considerable variation in the completeness of ethnicity recording across practices with at least 1000 registered patients (Fig. 3).

Consistency of ethnicity recording within patients

3.1% [260, 611] of the 19,618,135 patients with a recorded ethnicity had at least one ethnicity record that was discordant with the latest recorded ethnicity (Table 3). Patients whose latest recorded ethnicity was categorised as Mixed were most likely to have a discordant ethnicity recording (32.2%, 118,560), of whom 17.0% (62,565) also had a recorded ethnicity of White. 5.7% (33,205) of the 583,770 patients with the latest recorded ethnicity of Black also had a recorded ethnicity of White (Table 2).

Overall, for 19,364,120 (98.7%) of patients, their latest recorded ethnicity in primary care matched their most frequently recorded ethnicity in primary care (Table 3). 16,390,425 (99.5%) patients with the most recent ethnicity 'White' had matching most frequently recorded ethnicity. Other was the least concordant group, just 81.6% (399,440) of patients with the most recent ethnicity 'Mixed' had matching most frequently recorded ethnicity.

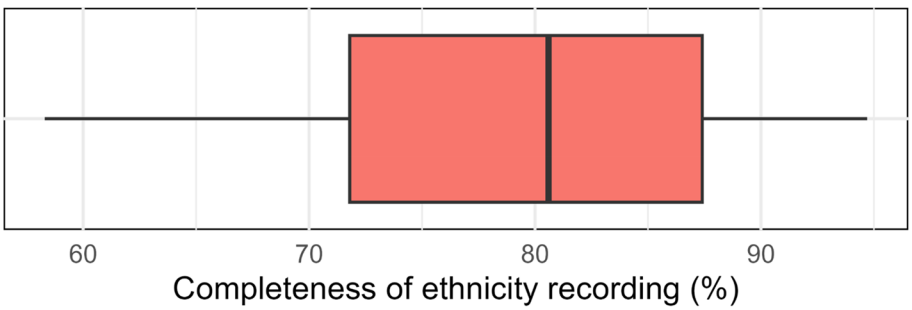

Fig. 3 Boxplot showing the 5th, 25th, 50th, 75th and 95th percentiles of completeness of ethnicity recording across practices with at least 1000 registered patients

Table 2 Count of patients with at least one recording of each ethnicity (proportion of latest ethnicity)

Latest ethnicity	Any recorded ethnicity					
	Asian	Black	Mixed	White	Other	Any discordant ethnicity
Asian: 1,708,430	1,708,430 (100.0)	8640 (0.5)	25,955 (1.5)	42,760 (2.5)	41,175 (2.4)	109,060 (6.4)
Black: 583,770	6680 (1.1)	583,770 (100.0)	41,245 (7.1)	33,205 (5.7)	11,495 (2.0)	85,075 (14.6)
Mixed: 367,980	18,400 (5.0)	32,990 (9.0)	367,980 (100.0)	62,565 (17.0)	15,920 (4.3)	118,560 (32.2)
White: 16,468,610	31,635 (0.2)	25,115 (0.2)	62,030 (0.4)	16,468,610 (100.0)	81,875 (0.5)	189,020 (1.1)
Other: 489,350	32,875 (6.7)	9430 (1.9)	16,795 (3.4)	60,865 (12.4)	489,350 (100.0)	109,545 (22.4)

Table 3 Count of patients with any recorded discordant ethnicity and a discordant 'most frequently recorded' ethnicity in primary care, according to latest ethnicity

Latest ethnicity	Total patients	Any discordant ethnicity	Discordant with most frequent ethnicity
Asian	1,708,430	109,060 (6.4)	12,685 (0.7)
Black	583,770	85,075 (14.6)	14,480 (2.5)
Mixed	367,980	118,560 (32.2)	58,760 (16.0)
White	16,468,610	189,020 (1.1)	78,185 (0.5)
Other	489,350	109,545 (22.4)	89,915 (18.4)
Overall	19,618,135	611,260 (3.1)	254,025 (1.3)

0.9% (5450) of patients with latest ethnicity 'Black' had the most frequently recorded ethnicity 'White' (Additional file 1: Table S4).

Consistency of ethnicity recording across data sources (primary care versus secondary care)

Of the 19.6 million total patients with a primary care ethnicity record, 12.9 million (66.0%) also had a secondary care ethnicity record. The proportion of patients with no secondary care coded ethnicity ranged from 31.9% in the White group to 58.6% in the Other group (Additional file 1: Table S5). SNOMED:2022 and secondary care coded ethnicity matched for 93.5% of patients with both coded ethnicities, ranging from 34.8% in the Mixed group to 96.9% in the White group (Fig. 4, Additional file 1: Table S6).

Comparison with the 2021 UK census population

The proportion of patients in each ethnicity group based on primary care records as of January 2022 was within 2.9 percentage points of the 2021 Census estimate (amended to the 2001 grouping) for the same ethnicity group across England as a whole (Asian: 8.7% primary care, 8.8% Census, relative difference (RD) −1.5; Black: 3.0%, 4.2%, RD −29.4; Mixed: 1.9%, 3.0% RD −36.5; White: 84.0%, 81.0% RD 3.6; Other: 2.5%, 2.9%, RD −15.1). When supplemented secondary care data, this increased to 3.2% (Fig. 5, Additional file 1: Table S7). In primary care records, the White population was underrepresented in all regions other than the North West (7.1% percentage points higher than Census estimates), South East (2.8%) and South West (0.6%) and was most severely underestimated in the West Midlands (−12.5%). The Asian population was overrepresented in all regions other than the North West (−3.6%) and South East (−1.6%) (Fig. 6, Additional file 1: Table S8). We also compared the primary care data to the 2021 Census estimates using 2021 rather than 2001 ethnicity groups (Additional file 1: Figs. S1 and S2 and Additional file 1: Table S9).

Discussion

Summary

This study reported ethnicity recording quality in around 25 million patients registered with a general practice in England and available for analysis in the OpenSAFELY-TPP database. Over three quarters of all patients had at least one ethnicity record in primary care data. When supplemented with hospital records, ethnicity recording was 92.5% complete, which is consistent with previously reported England-wide primary care data sources [27, 28]. 98.7% of patients' latest and most frequently recorded ethnicity matched. As the latest recorded ethnicity is computationally more efficient within OpenSAFELY, we recommend the use of the latest recorded ethnicity. The reported concordance of primary and secondary care records of 93.5% is consistent with those

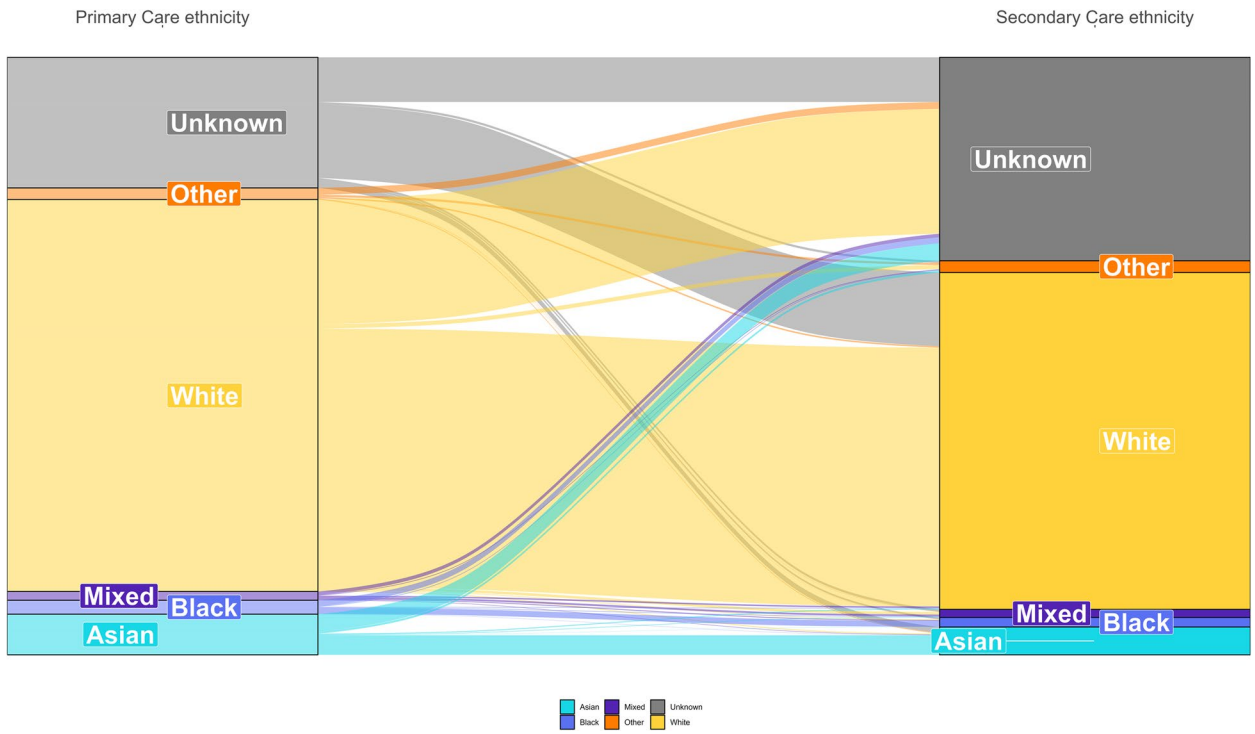

Fig. 4 Sankey plot comparing the categorisation of ethnicity in primary care and secondary care

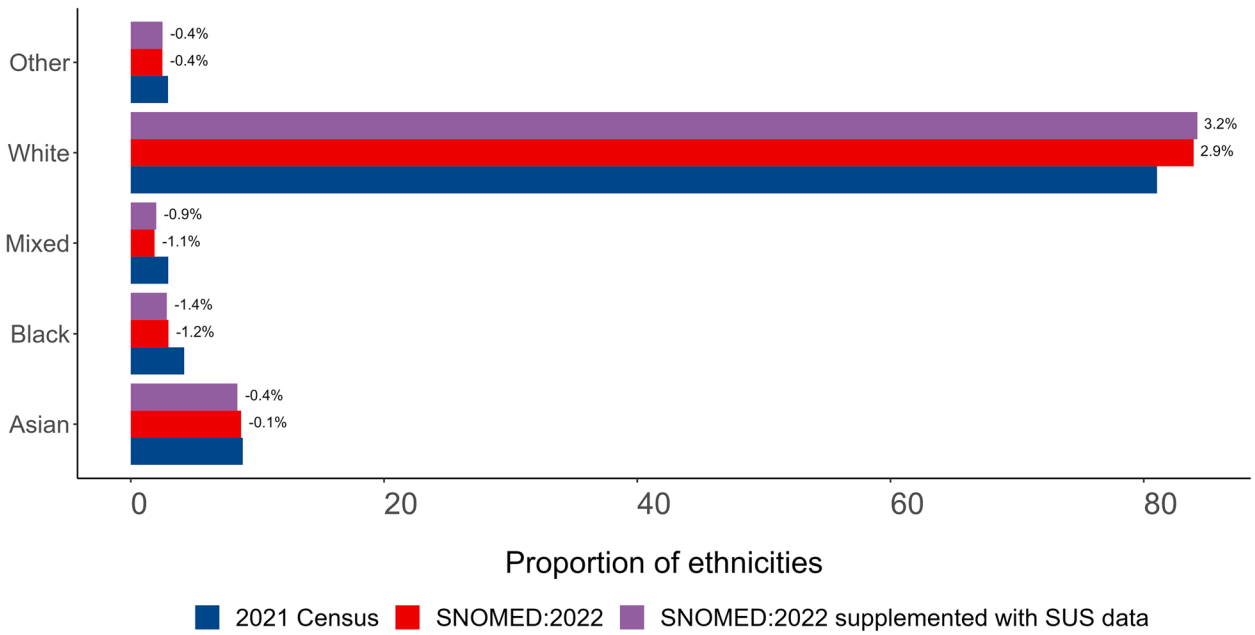

Fig. 5 Bar plot showing the proportion of 2021 Census and primary care populations per ethnicity grouped into 5 groups (excluding those without a recorded ethnicity (21.8% SNOMED:2020 and 7.5% supplemented with ethnicity data from secondary care)). Data labels indicate the percentage point difference between 2021 Census and TPP populations

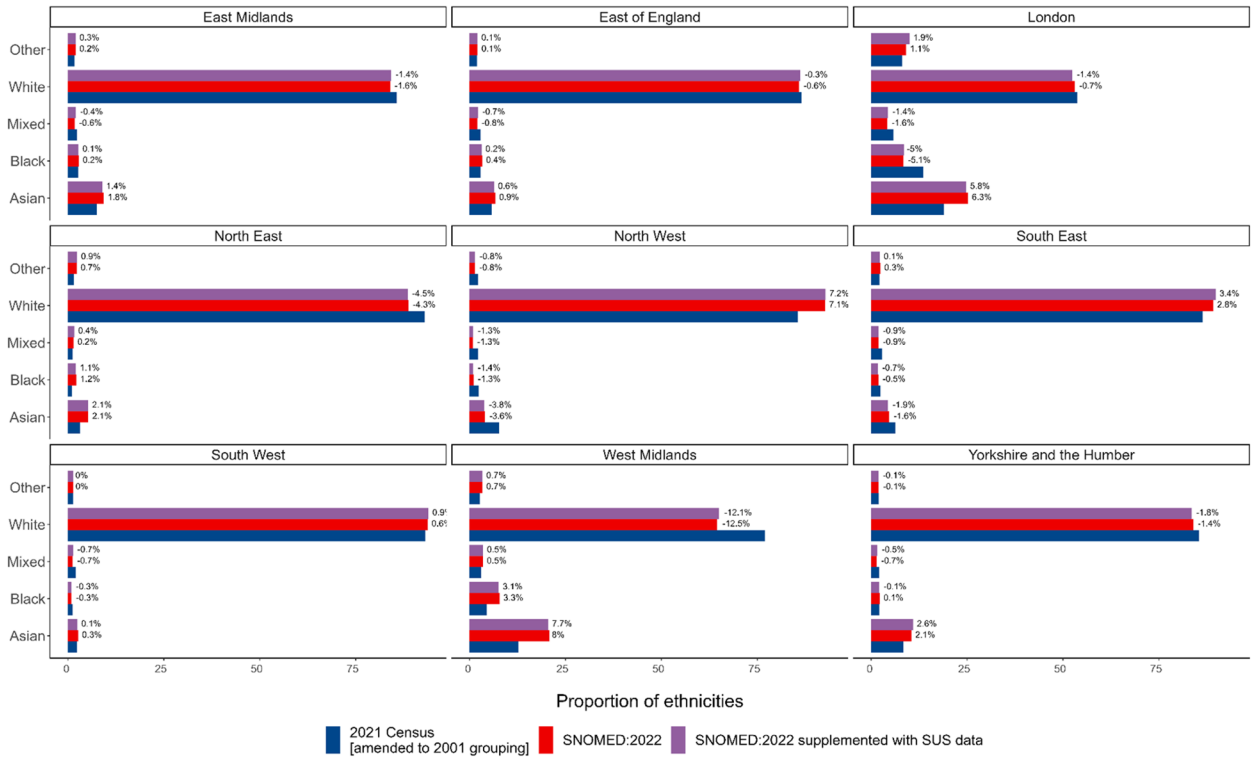

Fig. 6 Bar plot showing the proportion of 2021 Census and TPP populations in each ethnicity group by region (excluding those without a recorded ethnicity (21.8% in primary care and 7.5% supplemented with ethnicity data from secondary care)). Data labels indicate percentage point difference between 2021 Census and TPP populations

previously reported [29]. Despite regional variations, the overall ethnicity breakdown across all English OpenSAFELY-TPP practices was similar to the 2021 Census; however, larger relative differences were observed, in particular for the Mixed and Black groups. Therefore, relative to the size of certain ethnic groups, discrepant ethnicity recording practices may be a concern.

Strengths and weaknesses

This study provides a breakdown of primary care coding in OpenSAFELY-TPP by key clinical and demographic characteristics. The key strengths of this study are the use of large Electronic Health Record (EHR) datasets representing roughly 40% of the population of England registered with a GP, which enabled us to assess the quality of ethnicity data against a variety of important clinical characteristics.

Practices may utilise differing strategies for collecting ethnicity information from patients. Typically ethnicity is self-reported by the patient at registration or during consultation [30] but may not always be self-reported and may reflect an assumption made by the person entering the data. OpenSAFELY-TPP was missing ethnicity for 21.8% of patients, and the missingness of ethnicity data in EHRs may not be random [6].

This study focussed on the 5 Group ethnicity of the SNOMED:2022 codelists categorisation. However, there can be important variations in clinical care within these broad categories, as seen in COVID vaccine uptake [31, 32]. More detailed categorisations, alternative coding systems and codelists have been further explored in the OpenSAFELY-TPP Ethnicity short data report.

It is common for OpenSAFELY-TPP studies to supplement the primary care recorded ethnicity, where missing, with ethnicity data from secondary care [10, 11, 33]. The representativeness of the CTV3:2020 coded ethnicity supplemented with SUS data has been reported previously [33]. However, secondary care data is only available for people attending hospital within the time period that data were available (currently April 2019 onwards in OpenSAFELY). The population who still have no ethnicity record after supplementation are likely very different to the wider population, for example having a much lower chance of having been admitted to hospital, or interacting with healthcare services generally.

This study represents a snapshot of ethnicity recording as of 1 January 2022 and does not provide insights into

temporal trends in ethnicity recording. Trends in ethnicity recording over time are difficult to investigate due to loss of record date during transfer of clinical records when patients register with a new practice (Additional file 1: Fig. S4). Therefore, we are unable to assess the impact of QOF financial incentives being rescinded in 2011/12.

The most up-to-date formal estimates of England's population by ethnic group currently available are from the 2021 Census. Accuracy of the 2021 Census ethnicity estimates may vary by region. The 2021 census response rate was not even between regions, ranging from 95% in London to 98% in the South East, South West and East of England [34]. The 2021 census used multiple imputation to account for missing ethnicity [35]; the percentage of eligible persons who had an ethnicity value imputed or edited was not even between regions. Imputation rate was highest in London (2.0%) and lowest in the North East (1.0%) [34].

There are limitations in comparing the GP-registered population with the census population as differences naturally arise. For example, patients registered with a GP may have left the country some years ago and hence not be counted in the census; certain populations are less likely to be registered with a GP (such as Gypsy, Roma and Traveller communities [36] and migrants [37, 38]); not everyone responds to the census but some may be registered with a GP; and regional differences occur, for example due to students moving to cities during termtime. We looked at the GP-registered population in January 2022, whereas the census was taken in March 2021; therefore, some small changes in population also may have occurred during this time.

Findings in context

Over 20 studies have been conducted using the OpenSAFELY framework. It is important to understand the data issues with using ethnicity in OpenSAFELY. Whilst ethnicity data has been shown to be more complete for the CTV3:2020 codelist than the SNOMED:2022 codelist [13], the CTV3:2020 codelist included codes such as 'XaJSe: Muslim—ethnic category 2001 census' which relate to religion rather than ethnicity and were, therefore, excluded from the SNOMED:2022 codelist. The common practice of supplementing CTV3:2020 coded ethnicity with either secondary care data or the PRIMIS codelists could lead to inconsistent classification as both secondary care data and PRIMIS codelists follow the 2001 census categories.

Recording ethnicity is not straightforward. Indeed, despite often being used as a key variable to describe health, the idea of 'ethnicity' has been disputed [39]. Ethnicity is a complex mixture of social constructs, genetic make-up and cultural identity [40]. Self-identified ethnicity is not a fixed concept and evolving socio-cultural trends could contribute to changes in a person's self-identified ethnic group, particularly for those with mixed heritage [41]. It is therefore perhaps not surprising to see lower levels of concordance between latest ethnicity and most common ethnicity in those with latest ethnicity coded as 'mixed'. It is not clear to what extent this would explain all the discordance we identified or whether other factors such as data error are involved. Our findings agree with previous literature, both from the US and UK [5, 41], which suggest that the consistency of ethnicity information tends to be highest for white populations, and lowest for Mixed or Other racial/ethnic groups [42].

The 2001 census categories are the NHS standard for ethnicity [17], but we have not been able to find any explanation for the continued use of the 2001 census categories as the standard.

Due to the significant differences experienced by ethnic groups in terms of health outcomes, accurate ethnicity coding to the most granular code possible is crucial. Although we have focussed on codelist categorisations based on the 2001 census categories, ethnicity can be extracted for each of the component codes (Additional file 1: Table S8), so researchers have the option to use custom categorisations as required.

We believe that the SNOMED:2022 codelist and codelist categorisation provides a more consistent representation of ethnicity as defined by the 2001 census categories than the CTV3:2020 codelist and should be the preferred codelist and categorisation for primary care ethnicity.

Policy implications and interpretation

This paper is principally to inform interpretation of the numerous current and future analyses completed and published using OpenSAFELY-TPP and similar UK electronic healthcare databases. The practice of supplementing primary care ethnicity with secondary care ethnicity from SUS can, depending on the study design, introduce bias and should be used with caution. For example, patients who have more clinical interactions are more likely to have a recorded ethnicity and therefore patients with a recorded ethnicity in secondary care data may tend to be sicker than the general population. Ethnicity recording has been found to be more complete for patients who died in hospital compared with those discharged [5].

Conclusions

This report describes the completeness and consistency of primary care ethnicity in OpenSAFELY-TPP and suggests the adoption of the SNOMED:2022 codelist and codelist categorisation as the best standard method.

Abbreviations

APCS	Admitted patient care statistics
CTV3	Clinical Terms Version 3
EC	Emergency care
EHR	Electronic health record
GDPPR	General Practice Extraction Service Data for Pandemic Planning and Research
GP	General practitioner
GPES	General Practice Extraction Service
IMD	Index of Multiple Deprivation
NUTS 1	Nomenclature of Territorial Units for Statistics
ONS	Office for National Statistics
OPA	Outpatient attendance
PRIMIS	Primary Care Information Services
QOF	Quality and Outcomes Framework
SNOMED CT	Systematised Nomenclature of Medicine Clinical Terms
SUS	Secondary Uses Service

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-024-03499-5.

Additional file 1: Fig. S1. Bar plot showing the proportion of 2021 Census and TPP populations (amended to 2021 grouping) per ethnicity grouped into 5 groups (excluding those without a recorded ethnicity). Annotated with percentage point difference between 2021 Census and TPP populations. Fig S2. Bar plot showing the proportion of 2021 Census and TPP populations (amended to 2021 grouping) per ethnicity grouped into 5 groups per NUTS – 1 region (excluding those without a recorded ethnicity). Annotated with percentage point difference between 2021 Census and TPP populations. Fig. S3. Recording of ethnicity over time for latest and first recorded ethnicity. Unknown dates of recording may be stored as '1900 – 01 – 01'. Table S1. Count of patients with a recorded ethnicity in OpenSAFELY-TPP (proportion of registered TPP population) by clinical and demographic subgroups. All counts are rounded to the nearest 5. Table S2. Count of patients with a recorded ethnicity in OpenSAFELY TPP by ethnicity group (proportion of registered TPP population) and clinical and demographic subgroups. All counts are rounded to the nearest 5. Table S3. Count of patients with a recorded ethnicity in OpenSAFELY TPP by ethnicity group (proportion of registered TPP population) and clinical and demographic subgroups. All counts are rounded to the nearest 5. Table S4. Count of patients' most frequently recorded ethnicity (proportion of latest ethnicity). Table S6. Count of patients with a recorded ethnicity in Secondary Care by ethnicity group excluding Unknown ethnicites (proportion of Primary Care population). All counts are rounded to the nearest 5. Table S7. Count of patients with a recorded ethnicity in OpenSAFELY TPP by ethnicity group (proportion of registered TPP population) and 2021 ONS Census counts [amended to 2001 grouping] (proportion of 2021 ONS Census population). All counts are rounded to the nearest 5. Table S8. Count of patients with a recorded ethnicity in OpenSAFELY TPP [amended to the 2021 ethnicity grouping] (proportion of registered TPP population) and 2021 ONS Census counts (proportion of 2021 ONS Census population). All counts are rounded to the nearest 5. Table S9. Count of individual ethnicity code use.

Acknowledgements

We are very grateful for all the support received from the TPP Technical Operations team throughout this work and for generous assistance from the information governance and database teams at NHS England and the NHS England Transformation Directorate.

Guarantor

BG is guarantor.

Software and reproducibility

Data management was performed using Python 3.8, with analysis carried out using Python and R. Code for data management and analysis, as well as codelists are archived online https://github.com/opensafely/ethnicity-short-data-report/.

Patient and public involvement

This analysis relies on the use of large volumes of patient data. Ensuring patient, professional and public trust is therefore of critical importance. Maintaining trust requires being transparent about the way OpenSAFELY works, and ensuring patient and public voices are represented in the design and use of the platform. Between February and July 2022, we ran a 6-month pilot of Patient and Public Involvement and Engagement activity designed to be aligned with the principles set out in the Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research [43]. Our engagement focused on the broader OpenSAFELY platform and comprised three sets of activities: explain and engage, involve and iterate and participate and promote. To engage and explain, we have developed a public website at opensafely.org that provides a detailed description of the OpenSAFELY platform in language suitable for a lay audience and are co-developing an accompanying explainer video. To involve and iterate, we have created the OpenSAFELY 'Digital Critical Friends' Group, comprised of approximately 12 members representative in terms of ethnicity, gender and educational background; this group has met every 2 weeks to engage with and review the OpenSAFELY website, governance process, principles for researchers and FAQs. To participate and promote, we are conducting a systematic review of the key enablers of public trust in data-intensive research and have participated in the stakeholder group overseeing NHS England's 'data stewardship public dialogue'.

Authors' contributions

Conceptualisation: CDA, BM, RP, RM, JM and WJH. Data curation: CDA, RP, RM and JM. Formal analysis: CDA, RP, RM, JM and WJH. Funding acquisition: BG. Methodology: CDA, BM, RP, RM, JM and WJH. Project administration: CDA, RP, RM and JM. Resources: CDA, RM, JM, RP, HJC, LH, LAT and BG. Software: CDA, RM, JM, RP, HJC, LH, AM, SB, GH, RS, DE, TW, SD, PI, ID, SM, TO'D, BFCBC, LB, CB, JP, FH, SH, JC, BG, BM, AJW and WJH. Supervision: AJW, LAT and WJH. Validation: CDA, BM, RP, RM, JM and WJH. Visualisation: CDA, RP, BM, BG, AJW and WJH. Writing—original draft: CDA. Writing—review and editing: CDA, AJW, BM, HJC and WJH.

Authors' Twitter handles

Colm D Andrews-@colmresearcher.

Funding

The OpenSAFELY platform is principally funded by grants from:
NHS England [2023–2025];
The Wellcome Trust (222,097/Z/20/Z) [2020–2024];
MRC (MR/V015737/1) [2020–2021].
Additional contributions to OpenSAFELY have been funded by grants from:
MRC via the National Core Study programme, Longitudinal Health and Wellbeing strand (MC_PC_20030, MC_PC_20059) [2020–2022] and the Data and Connectivity strand (MC_PC_20029, MC_PC_20058) [2020–2022];
NIHR and MRC via the CONVALESCENCE programme (COV-LT-0009, MC_PC_20051) [2021–2024];
NHS England via the Primary Care Medicines Analytics Unit [2021–2024].
The views expressed are those of the authors and not necessarily those of the NIHR, NHS England, UK Health Security Agency (UKHSA), the Department of Health and Social Care or other funders. Funders had no role in the study design, collection, analysis and interpretation of data; in the writing of the report and in the decision to submit the article for publication.

Availability of data and materials

Access to the underlying identifiable and potentially re-identifiable pseudonymised electronic health record data is tightly governed by various legislative and regulatory frameworks, and restricted by best practice. The data in OpenSAFELY is drawn from General Practice data across England where TPP is the Data Processor. TPP developers (CB, JC, JP, FH and SH) initiate an automated process to create pseudonymised records in the core OpenSAFELY database, which are copies of key structured data tables in the identifiable records. These are linked onto key external data resources that have also been pseudonymised via SHA-512 one-way hashing of NHS numbers using a shared salt. Bennett Institute for Applied Data Science developers and PIs (BG, CEM, SB, AJW, KW, WJH, HJC, DE, PI, SD, GH, BBC, RMS, ID, KB, EJW and CTR)

holding contracts with NHS England have access to the OpenSAFELY pseudonymised data tables as needed to develop the OpenSAFELY tools. These tools in turn enable researchers with OpenSAFELY Data Access Agreements to write and execute code for data management and data analysis without direct access to the underlying raw pseudonymised patient data and to review the outputs of this code. All code for the full data management pipeline—from raw data to completed results for this analysis—and for the OpenSAFELY platform as a whole is available for review at github.com/OpenSAFELY.

Declarations

Ethics approval and consent to participate

NHS England is the data controller; TPP is the data processor; and the researchers on OpenSAFELY are acting with the approval of NHS England. This implementation of OpenSAFELY is hosted within the TPP environment which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant [44, 45]; patient data has been pseudonymised for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymised datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is via a virtual private network (VPN) connection, restricted to a small group of researchers; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; only aggregate statistical outputs leave the platform environment following best practice for anonymisation of results such as statistical disclosure control for low cell counts [46]. The OpenSAFELY research platform adheres to the obligations of the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018. In March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI) to require organisations to process confidential patient information for the purposes of protecting public health, providing healthcare services to the public and monitoring and managing the COVID-19 outbreak and incidents of exposure; this sets aside the requirement for patient consent [47]. Taken together, these provide the legal bases to link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required to share relevant health information to support the public health response to the pandemic and have been informed of the OpenSAFELY analytics platform.
This study was approved by the Health Research Authority (REC reference 20/LO/0651) and by the LSHTM Ethics Board (reference 21863).

Consent for publication

Not applicable.

Competing interests

All authors declare the following: BG has received research funding from the Bennett Foundation, the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, NHS England, the NIHR Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK, the Health Foundation, the World Health Organisation, UKRI MRC, Asthma UK, the British Lung Foundation, and the Longitudinal Health and Wellbeing strand of the National Core Studies programme; he is a Non-Executive Director at NHS Digital; he also receives personal income from speaking and writing for lay audiences on the misuse of science. BMK is also employed by NHS England working on medicines policy and clinical lead for primary care medicines data.

Author details

[1]Nuffield Department of Primary Care Health Sciences, Bennett Institute for Applied Data Science, Oxford University, Oxford OX2 6GG, UK. [2]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [3]TPP, TPP House, 129 Low Lane, Horsforth, Leeds LS18 5PX, UK. [4]Wolfson Institute for Population Health, University of London, London, Queen Mary E1 2AT, UK.

Received: 31 January 2024 Accepted: 24 June 2024
Published online: 10 July 2024

References

1. Irizar P, Pan D, Kapadia D, Bécares L, Sze S, Taylor H, et al. Ethnic inequalities in COVID-19 infection, hospitalisation, intensive care admission, and death: a global systematic review and meta-analysis of over 200 million study participants. EClinicalMedicine. 2023;57:101877.
2. Mathur R, Rentsch CT, Morton CE, Hulme WJ, Schultze A, MacKenna B, et al. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform. Lancet. 2021;397(10286):1711–24.
3. Garlick S. Ethnic group, England and Wales - Office for National Statistics. Office for National Statistics; 2022. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021. Cited 2023 May 24.
4. Knox S, Bhopal RS, Thomson CS, Millard A, Fraser A, Gruer L, et al. The challenge of using routinely collected data to compare hospital admission rates by ethnic group: a demonstration project in Scotland. J Public Health. 2020;42(4):748–55.
5. Scobie S, Spencer J, Raleigh V. Ethnicity coding in English health service datasets. Available from: https://www.nuffieldtrust.org.uk/files/2021-06/1622731816_nuffield-trust-ethnicity-coding-web.pdf. Cited 2023 Feb 12.
6. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health. 2014;36(4):684–92.
7. Contract changes 2011/12. Available from: https://web.archive.org/web/20110504084616/http://www.nhsemployers.org/PayAndContracts/GeneralMedicalServicesContract/GMSContractChanges/Pages/Contract-changes-2011-12.aspx. Cited 2023 May 24.
8. Kapadia, Zhang, Salway, Nazroo, Booth. Ethnic inequalities in healthcare: a rapid evidence review. NHS Race and Health. 2022. Available from: https://www.nhsrho.org/research/ethnic-inequalities-in-healthcare-a-rapid-evidence-review-3/. Cited 2024 June 27
9. NHS Digital. Secondary Uses Service (SUS). Available from: https://digital.nhs.uk/services/secondary-uses-service-sus. Cited 2023 May 16.
10. Fisher L, Hopcroft LEM, Rodgers S, Barrett J, Oliver K, Avery AJ, et al. Changes in English medication safety indicators throughout the COVID-19 pandemic: a federated analysis of 57 million patients' primary care records in situ using OpenSAFELY. BMJ Med. 2023;2(1):e000392. https://doi.org/10.1136/bmjmed-2022-000392.
11. Nab L, Parker EP, Andrews CD, Hulme WJ, Fisher L, Morley J, Mehrkar A, et al. Changes in COVID-19-Related Mortality across Key Demographic and Clinical Subgroups in England from 2020 to 2022: A Retrospective Cohort Study Using the OpenSAFELY Platform. Lancet Public Health. 2023;8(5):e364–77.
12. OpenCodelists: Ethnicity codes. Available from: https://www.opencodelists.org/codelist/primis-covid19-vacc-uptake/eth2001/v1/. Cited 2022 Sep 13.
13. OpenCodelists: ethnicity (SNOMED). Available from: https://www.opencodelists.org/codelist/opensafely/ethnicity-snomed-0removed/2e641f61/. Cited 2022 Sep 13.
14. OpenCodelists: Ethnicity. Available from: https://www.opencodelists.org/codelist/opensafely/ethnicity/2020-04-27/. Cited 2022 Sep 13.
15. PRIMIS develops the national Covid-19 Vaccination Uptake Reporting Specification. Available from: https://www.nottingham.ac.uk/primis/about/news/newslisting/primis-develops-the-national-covid-19-vaccination-uptake-reporting-specification.aspx. Cited 2022 Aug 19.
16. NHS Digital. General Practice Extraction Service (GPES) Data for pandemic planning and research: a guide for analysts and users of the data. Available from: https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data. Cited 2022 Aug 19.
17. Ethnic Category. Available from: https://www.datadictionary.nhs.uk/data_elements/ethnic_category.html?hl=ethnic. Cited 2022 Aug 22.
18. Gypsy, Roma and Irish Traveller ethnicity summary. Available from: https://web.archive.org/web/20220213182343/https://www.ethnicity-facts-figures.service.gov.uk/summaries/gypsy-roma-irish-traveller. Cited 2023 Jun 6.
19. McLennan D, Noble S, Noble M, Plunkett E, Wright G, Gutacker N. The English Indices of Deprivation 2019 : technical report. 2019. Available from: https://dera.ioe.ac.uk/id/eprint/34259. Cited 2022 Aug 4.

20. OpenCodelists: Dementia (SNOMED). Available from: https://www.opencodelists.org/codelist/opensafely/dementia-snomed/2020-04-22/. Cited 2022 Sep 13.
21. OpenCodelists: Diabetes (SNOMED). Available from: https://www.opencodelists.org/codelist/opensafely/diabetes-snomed/2020-04-15/. Cited 2022 Sep 13.
22. OpenCodelists: Hypertension (SNOMED). Available from: https://www.opencodelists.org/codelist/opensafely/hypertension-snomed/2020-04-28/. Cited 2022 Sep 13.
23. OpenCodelists: Wider learning disability. Available from: https://www.opencodelists.org/codelist/primis-covid19-vacc-uptake/learndis/v1/. Cited 2022 Sep 13.
24. Variable reference. Available from: https://docs.opensafely.org/study-def-variables/. Cited 2022 Nov 18.
25. Mortality statistics - underlying cause, sex and age - Nomis - Official Labour Market Statistics. Available from: https://www.nomisweb.co.uk/datasets/mortsa. Cited 2022 Jan 28.
26. List of ethnic groups. Available from: https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups. Cited 2023 Apr 17.
27. Wood A, Denholm R, Hollings S, Cooper J, Ip S, Walker V, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. BMJ. 2021;7(373):n826.
28. Pineda-Moncusí M, Allery F, Delmestri A, Bolton T, Nolan J, Thygesen JH, Handy A, et al. Ethnicity Data Resource in Population-Wide Health Records: Completeness, Coverage and Granularity of Diversity. Sci Data. 2024;11(1):221.
29. 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;21(1):3.
30. Hull SA, Mathur R, Badrick E, Robson J, Boomla K. Recording ethnicity in primary care: assessing the methods and impact. Br J Gen Pract. 2011;61(586):e290–4.
31. Watkinson RE, Williams R, Gillibrand S, Sanders C, Sutton M. Ethnic inequalities in COVID-19 vaccine uptake and comparison to seasonal influenza vaccine uptake in greater Manchester, UK: a cohort study. PLoS Med. 2022;19(3):e1003932.
32. Curtis HJ, Inglesby P, Morton CE, MacKenna B, Green A, Hulme W, et al. Trends and clinical characteristics of COVID-19 vaccine recipients: a federated analysis of 57.9 million patients' primary care records in situ using OpenSAFELY. Br J Gen Pract. 2022;72(714):e51–62.
33. Andrews C, Schultze A, Curtis H, Hulme W, Tazare J, Evans S, et al. OpenSAFELY: representativeness of electronic health record platform OpenSAFELY-TPP data compared to the population of England. Wellcome Open Res. 2022;18(7):191.
34. Measures showing the quality of Census 2021 estimates. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/measuresshowingthequalityofcensus2021estimates. Cited 2023 Feb 16.
35. Wardman L, Aldrich S, Rogers S. Census item edit and imputation process. Disponible en ligne:[06/01/2015]. http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-userguide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-andimputation-process.pdf. 2011. Cited 2022 Feb 3.
36. Tackling inequalities faced by Gypsy, roma and traveller communities. Available from: https://publications.parliament.uk/pa/cm201719/cmselect/cmwomeq/360/full-report.html. Cited 2023 May 25.
37. Kang C, Tomkow L, Farrington R. Access to primary health care for asylum seekers and refugees: a qualitative study of service user experiences in the UK. Br J Gen Pract. 2019;69(685):e537–45.
38. Knights F, Carter J, Deal A, Crawshaw AF, Hayward SE, Jones L, et al. Impact of COVID-19 on migrants' access to primary care and implications for vaccine roll-out: a national qualitative study. Br J Gen Pract. 2021;71(709):e583–95.
39. Bradby H. Ethnicity: not a black and white issue. A research note. Sociol Health Illn. 1995;17(3):405–17.
40. Lee C. "Race" and "ethnicity" in biomedical research: how do scientists construct and explain differences in health? Soc Sci Med. 2009;68(6):1183–90.
41. Saunders CL, Abel GA, El Turabi A, Ahmed F, Lyratzopoulos G. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey. BMJ Open. 2013;3(6):e002882.
42. Arday SL, Arday DR, Monroe S, Zhang J. HCFA's racial and ethnic data: current accuracy and recent improvements. Health Care Financ Rev. 2000;21(4):107–16.
43. Aitken M, Tully MP, Porteous C, Denegri S, Cunningham-Burley S, Banner N, et al. Consensus statement on public involvement and engagement with data intensive health research. Int J Popul Data Sci. 2019;4(1):586.
44. NHS Digital. BETA – Data Security Standards - NHS Digital. Available from: https://digital.nhs.uk/about-nhs-digital/our-work/nhs-digital-data-and-technology-standards/framework/beta---data-security-standards. Cited 2020 Apr 30.
45. NHS Digital. Data Security and Protection Toolkit - NHS Digital. Available from: https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/data-security-and-protection-toolkit. Cited 2020 Apr 30.
46. NHS Digital. ISB1523: Anonymisation Standard for Publishing Health and Social Care Data. Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data. Cited 2023 Jul 20.
47. Secretary of State for Health and Social Care - UK Government. Coronavirus (COVID-19): notification to organisations to share information. 2020. Available from: https://web.archive.org/web/20200421171727/https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-shareinformation. Cited 2022 Nov 3.

Publisher's Note