 doi:10.20944/preprints202110.0081.v1

Title

Development and validation of a primary care electronic health record phenotype to study migration and health in the UK

Authors

Neha Pathak, MBBS[1,2]
Claire X. Zhang, MSc[1,3]
Yamina Boukari, PhD[1]
Rachel Burns, MSc[1]
Rohini Mathur, PhD[4]
Arturo Gonzalez-Izquierdo, PhD[1,7]
Prof Spiros Denaxas, PhD[1,7]
Prof Pam Sonnenberg, PhD[5]
Prof Andrew Hayward, PhD[6]
Prof Robert W. Aldridge, PhD[1]

Affiliations & addresses

[1]Institute of Health Informatics, University College London, 222 Euston Rd, London, United Kingdom NW1 2DA
[2]Guy's & St Thomas' NHS Foundation Trust, London, United Kingdom
[3]Public Health England, Wellington House, 133-155 Waterloo Rd, London, United Kingdom SE1 8UG
[4]Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, Keppel Street, London, United Kingdom WC1E 7HT
[5]Institute for Global Health, University College London, 30 Guilford Street, London, United Kingdom WC1N 1EH
[6]Institute of Epidemiology & Health Care, University College London, 1-19 Torrington Place
London, United Kingdom WC1E 7HB
[7]Health Data Research UK, London, United Kingdom

Corresponding author

Professor Robert W. Aldridge
Public Health Data Science
Institute of Health Informatics
University College London
222 Euston Rd, London
United Kingdom NW1 2DA
r.aldridge@ucl.ac.uk
020 3549 5541

Word count

3,556

 doi:10.20944/preprints202110.0081.v1

Abstract

International migrants comprised 14% of the UK population in 2020, but migrant health in the UK has rarely been studied at a population level using primary care electronic health records (EHRs). Given the difficulty of determining migration status using EHRs, this study developed a migration phenotype and assessed its validity. We developed a phenotyping algorithm using codes for country of birth, visa status, non-English main/first language and non-UK origin. It was applied to a Clinical Practice Research Datalink (CPRD) GOLD database of 16,071,111 primary care patients between 1997 and 2018. We compared the completeness and representativeness of the identified migrant population to Office for National Statistics (ONS) country of birth and 2011 census data by year, age, sex, geographic region of birth and ethnicity. Between 1997-2018, 403,768 migrants (2.51% of the CPRD GOLD population) were identified using the phenotype. 178,749 (1.11%) of these migrants were identified by codes indicating foreign country of birth or visa status, 216,731 (1.35%) a non-English main/first language, and 8,288 (0.05%) non-UK origin. The cohort was similarly distributed compared to ONS migration statistics in terms of sex and region of birth. Recording of migration improved from identifying approximately one-tenth of the expected proportion of migrants according to the ONS in 2004 to a quarter in 2018. Younger migrants were better represented than those aged 50 and over. The migration phenotype identified a large number of migrants and can be used to undertake large-scale migration health research in CPRD GOLD to inform healthcare policy, practice and action. While the cohort was representative of the UK migrant population in terms of sex and region of birth, migration status was under-recorded in earlier years and older ages, and future studies for these groups should therefore be interpreted with caution.

Keywords

Migration, phenotype, validation, algorithm, primary care, Clinical Practice Research Datalink

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

Introduction

Background

International migrants comprising 14% of the UK population in 2020 (1). The conditions prior to, during and after migration expose individuals to a range of health risks, resulting in differences in health outcomes between migrants and non-migrants in the migrant's country of arrival (2). In the UK, there are well-established multi-generational minority ethnic communities but a history of 'hostile' migration policies (3). The study of migrant health is therefore needed to complement the study of ethnic inequalities to understand how migration intersects with ethnicity, as well as its effects over and above ethnicity to shape risk factors for health, physical and mental health outcomes, and healthcare access (4).

While migrants' hospitalisation and mortality outcomes have been studied on a population level using electronic health records (EHRs) (5, 6), primary care outcomes are scarcely investigated at this scale, despite often being the first point of contact with the UK health system and a central part of the National Health Service (NHS) strategy for preventive care (7). Most studies examining primary care outcomes in UK migrants are qualitative or employ quantitative survey methods. When EHRs have been used, primary care registration data could only be linked to disease-specific migrant health datasets like tuberculosis screening (8). Linkage of census data has only been attempted in Northern Ireland for prescriptions outcomes (9). Additionally, three studies have identified migration status without the use of data linkages (10-12), all conducted within the Clinical Practice Research Datalink (CPRD), one of the largest UK primary care EHR resources. Using predominantly country of birth and language codes, they estimated that 1.3% of individuals aged ≥ 65 years in CPRD could be identified as international migrants (11). However, with 67.7% of migrants in England aged between 16 and 64 years old at the time of the 2011 census (13), a large proportion of migrants at younger ages were not identified by these studies.

Thus, a valid migration phenotype, which is a transparent reproducible algorithm using clinical terminology codes (14), is needed to determine migration status for individuals of all ages using UK primary care EHRs in order to study a broad range of migration health outcomes. A migration phenotype should determine the migration status of a large number of individuals who use primary care and are representative of the UK migrant population. CPRD with its associated linked datasets is an ideal database to use in the development of this phenotype so that it can be used to study primary, secondary and tertiary healthcare utilisation, mortality and other health outcomes in migrants from EU and non-EU countries.

We aimed to develop a migration phenotype for UK NHS primary care EHRs and assess its validity in individuals of all ages by describing completeness of recording of migration status, as well as representativeness compared to Office for National Statistics (ONS) country of birth and 2011 census statistics.

Materials & Methods

Study design

This is a study validating a migration phenotype for a population-based cohort study of migration health in the UK using linked EHRs, with a previously published protocol (15).

Ethics & approvals

This study is based in part on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency (MHRA). It was approved by the MHRA Independent Scientific Advisory Committee (ISAC protocol 19_062R) and carried out as part of the CALIBER programme (16). The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the authors alone.

Data resource

We extracted data from the CPRD GOLD January 2019 build, which comprised approximately 16 million individuals from 761 practices covering 3.53% of the UK population (17). CPRD GOLD contains de-identified data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

of patients across a network of GP practices across the UK that use Vision® EHR software. This data source is broadly representative of the age, sex and ethnicity demographics of the UK general population (18).

Inclusion criteria

We included individuals of all ages in CPRD GOLD between 1 January 1997 and 31 December 2018 whose record was of 'acceptable' research quality. This means CPRD has verified that the individual and their GP practice were contributing 'up-to-standard' data (18). An individual was included at the latest of 1 January 1997, their current registration date or the date on which their GP practice started contributing up-to-standard data to CPRD GOLD. An individual was excluded at the earliest of 31 December 2018, the date their care was transferred out of a CPRD GOLD practice, the practice's last data collection date for CPRD, or the individual's date of death.

Development of the migration phenotype

We created the phenotype using a systematic approach previously developed from the CALIBER platform described elsewhere (19). The phenotype was created in three stages (exploration, development and implementation) with feedback at each stage from a team of clinicians, computer scientists, epidemiologists, public health practitioners, bioinformatics and migration health experts.

We searched for Read V2 terms relating to international migration using the following: *migrant*, *migrat*, *countr*, *asylum*, *refugee*, *visa*, *abroad*, *born in*, *origin*, *illegal*, *language*, with the asterisk representing a wildcard search operator. The initial list of terms was reviewed and refined by two experts in migration health research. Each term was assigned a category (Figure 1) based on the type of term ("visa status indicating migration to the UK", "main/first language not English", "country of birth outside of the UK", "non-UK origin") and a category based on the certainty of migration status ("definite", "probable", "possible"). Each individual was classified once using their highest certainty of migration category.

Outcomes

The following three outcomes were used:

1. Migration phenotype: The total number of terms used in the migration phenotype.
2. Completeness: The percentage of migrants recorded in CPRD for the whole study period, in each year and at the time of the 2011 census.
3. Representativeness: The percentage of migrants in CPRD compared with annual ONS country of birth statistics (1), and the percentage of migrants in CPRD living in England and Wales on the date of the 2011 census (27 March 2011) compared with census data (20).

Data analysis

We counted the number of different terms used in the migration phenotype, including by the category of term described in Table 1. We compared the list of terms to the Jain et al study (11).

To assess completeness, we estimated the distribution of migrants across the study period and at the time of the 2011 census by sex, year of birth, World Health Organisation (WHO) region of birth, continent of birth, 13 CPRD practice region (classified by CPRD as 10 regions in England, with Scotland, Wales and Northern Ireland as separate regions) and ethnicity (18 category groupings, then further aggregated into the 6 higher-level groups of White British, White Non-British, Mixed, Asian/Asian British, Black/Black British, Other to address small group sizes; Table S1).

To assess representativeness, we compared the percentage of migrants in CPRD with annual ONS country of birth statistics (1, 20) both visually and using the chi-squared test for proportions. Ratios were calculated of the proportion of migrants in CPRD compared to ONS country of birth statistics in each year between 2004 and 2018 (from 2004 onwards, ONS data is sectioned into periods January-December for a more consistent comparison across years) (11). We also compared, visually and using the chi-squared test, the percentage of migrants in CPRD living in England on the date of the 2011 census with 2011 census data on country of birth (13) stratified by sex, age, geographical region

of origin, and ethnicity. Ratios were calculated of the proportion of migrants in CPRD compared to ONS census data.

We conducted subgroup analyses based on certainty of migration status (i.e. definite, probably, possible).

Bias

Misclassification may lead to differential bias where migration status is more likely to be recorded for individuals experiencing a specific outcome than those who do not. This could lead to a false association between migration and outcomes studied. We assessed bias by comparing the distribution of migrants recorded in CPRD GOLD to ONS population statistics, and created categories of migration status to address differences in level of certainty of classification across codes included in the phenotype.

Tools

Data were supplied by the CALIBER research team in multiple files, and imported into R software for cleaning and analysis. All data cleaning and analysis code has been made available as open-source metadata.

Results

Migrant phenotype

434 terms indicating migration to the UK were identified from the Read Version 2 terminology system, and are listed in Table S2. The majority of terms indicated country of birth outside of the UK (51.84%; 225 out of 434 terms) or having a non-English main or first language (42.16%; 183 out of 434 terms). The remaining terms related to visa status indicating migration to the UK (3.46%; 15 out of 434 terms) or a non-UK origin (2.53%; 11 out of 434 terms).

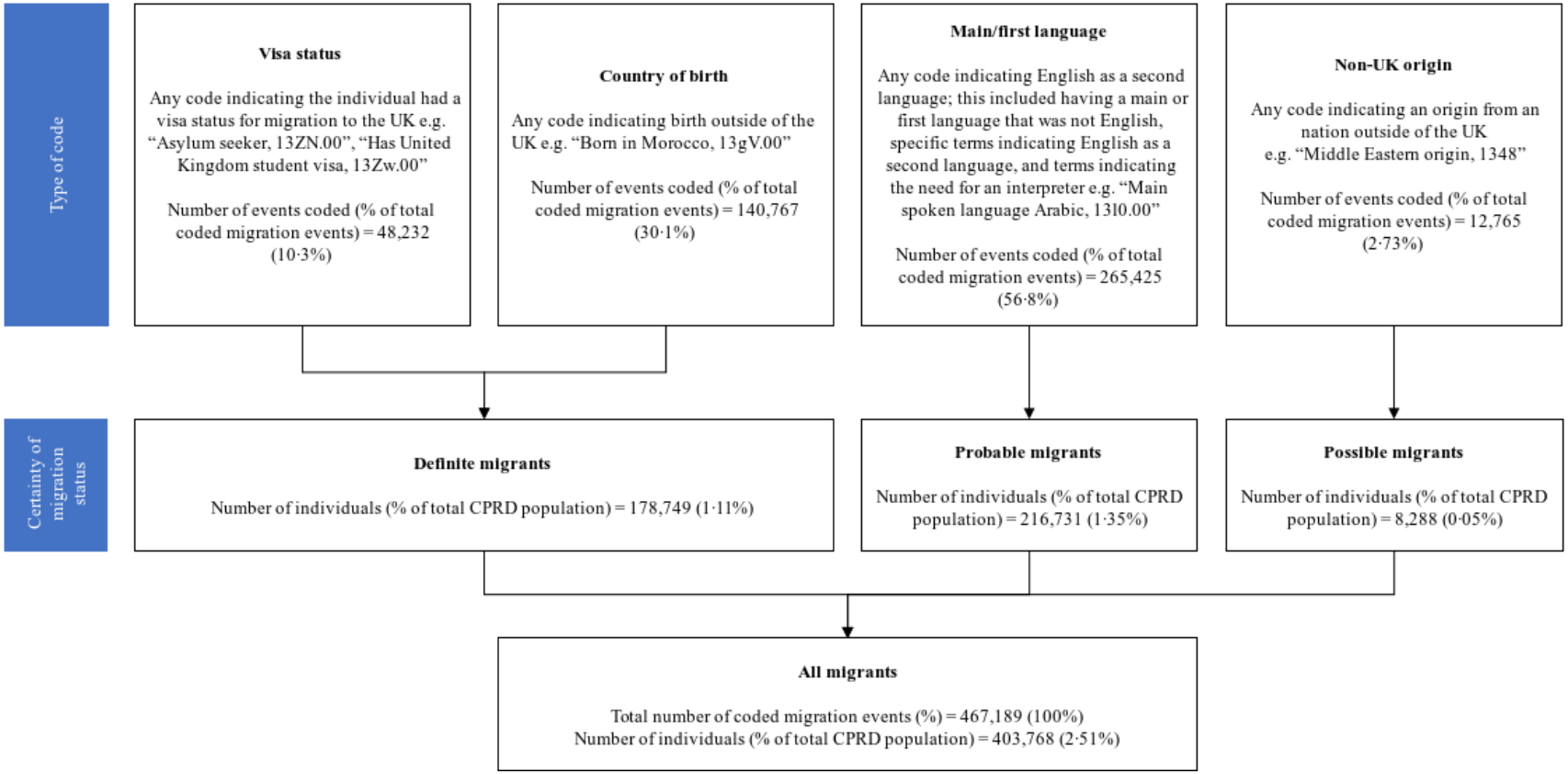
67 Read codes included by Jain et al's were excluded as they were largely related to reading other languages (11). The expert group discussed that preferred written language may not always correspond to a person's main/first language, so these terms were excluded from the present migration phenotype. A further 36 language, country of birth and origin related terms were included in the present migration phenotype that were not included by Jain et al.

Completeness

Of the patients in CPRD between January 1997 and December 2018, 2.51% (403,768/16,071,111) had at least one term indicating migration to the UK (Figure 1). 467,189 events indicating migration were coded across 403,768 individuals. 44.3% of these 403,768 individuals were classified as "definite" migrants, 53.7% as "probable" migrants, and 2.05% "possible" migrants. The most commonly coded migration-related events indicated a non-English first/main language 56.8%. The least commonly coded event was related to being of non-UK origin (2.73%). The percentage of migrants in CPRD GOLD increased from 0.20% in 1997 to 3.64% in 2018. Table S3 details the number and percentage of individuals in CPRD recorded as migrants annually between 1997 and 2018. At the time of the 2011 census, 2.52% of CPRD GOLD patients in England and Wales had at least one term indicating international migration, and their demographic characteristics are detailed in Table S4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

Figure 1. Categorisation of migrants by certainty of migration status using type of migration code



 doi:10.20944/preprints202110.0081.v1

Table 1 summarises the distribution of migrants in CPRD GOLD for the demographic factors of sex, year/decade of birth, ethnicity, region of birth, and primary care practice region. Just over half of migrants were female (53.7%) and the median year of birth was 1982 (IQR 1973-1990). The most common ethnicity amongst all migrants was White Non-British (34.3%) followed by Asian/Asian British (26.7%) and Black/Black British (9.2%). 42.4% of migrants in CPRD GOLD were registered with a London practice, and the proportion of patients in a region that were recorded as migrants was also highest in London (7.44%; Table S5).

Of the 140,423 patients with country of birth codes that aligned with a WHO region of birth, the most common was European Region (12.5%) followed by African Region (5.86%) and Western Pacific Region (4.36%). Of the 140,641 patients with country of birth codes that aligned with ONS Nomis continent of birth codes, the most common was the Middle East & Asia (12.5%) followed by Europe (12.4%) and Africa (5.86%).

Distribution of sex and year of birth was consistent across certainty of migration status categories. However, ethnicity was better recorded in “probable” migrants with only 7.71% of unknown ethnicity compared to 28.8% of “definite” migrants with unknown ethnicity.

 doi:10.20944/preprints202110.0081.v1

Table 1. Demographic characteristics of recorded migrants in CPRD GOLD by certainty of migration status (1997-2018)

Demographic characteristic		Migrants (%)	Definite migrants (%)	Probable migrants (%)	Possible migrants (%)	Definite + Probable migrants (%)
Totals*		403,768 (100%)	178,749 (44.3%)	216,731 (53.7%)	8,288 (2.05%)	395,480 (97.9%)
Sex	Male	187,057 (46.3%)	83,399 (46.7%)	99,849 (46.1%)	3,809 (46.0%)	183,248 (46.3%)
	Female	216,704 (53.7%)	95,346 (53.3%)	116,879 (53.9%)	4,479 (54.0%)	212,225 (53.7%)
Year of birth	1900-1919	456 (0.11%)	194 (0.11%)	212 (0.10%)	50 (0.60%)	406 (0.01%)
	1920-1939	9,303 (2.30%)	3,387 (1.89%)	5,584(2.58%)	332 (4.01%)	8,971(2.27%)
	1940-1959	31,169 (7.71%)	12,803 (7.16%)	17,292 (7.98%)	1,074 (13.0%)	30,095 (7.61%)
	1960-1979	130,715 (32.4%)	62,582 (35.0%)	64,325 (29.7%)	3,808 (45.9%)	126,907 (32.1%)
	1980-1999	179,702 (44.5%)	86,459 (48.4%)	90,780 (41.9%)	2,463 (29.7%)	177,239 (44.8%)
	2000-2018	52,423 (13.0%)	13,324 (7.45%)	38,538 (17.8%)	561 (6.77%)	51,862 (13.1%)
Ethnicity	White British	6,125 (1.52%)	3,519 (1.97%)	2,525 (1.17%)	81 (0.977%)	6,044 (1.53%)
	White Non-British	138,410 (34.3%)	48,554 (27.2%)	89,557 (41.3%)	299 (3.61%)	138,111 (34.9%)
	Mixed	11,008 (2.73%)	5,373(3.01%)	5,453 (2.52%)	82 (0.989%)	10,826 (2.74%)
	Asian/Asian British	107,630 (26.7%)	35,850(20.1%)	69,791 (32.2%)	1,989 (24.0%)	105,641 (26.7%)
	Black/African/Caribbean/Black British	37,101 (9.19%)	21,100 (11.8%)	14,374 (6.63%)	1,627(19.6%)	35,474 (8.99%)
	Other	31,454 (7.79%)	12,819 (7.17%)	18,314 (8.45%)	321 (3.87%)	31,133 (7.87%)
	Unknown	72,040 (17.8%)	51,534 (28.8%)	16,717 (7.71%)	3,789 (45.7%)	68,251 (17.3%)
WHO region of birth**	African Region	23,675 (5.86%)	23,675 (13.2%)	..	..	..
	European Region	50,588 (12.5%)	50,588 (28.3%)	..	..	..
	Eastern Mediterranean Region	13,701 (3.39%)	13,701 (7.66%)	..	..	..
	Region of the Americas	12,114 (3.00%)	12,114 (6.78%)	..	..	..
	South East Asian Region	14,813 (3.67%)	14,813 (8.29%)	..	..	..
	Western Pacific Region	17,621 (4.36%)	17,621 (9.86%)	..	..	..
	Unknown	263,345 (65.2%)	46,237 (25.9%)	..	..	..
Continent of birth**	Africa	23,675 (5.86%)	23,675 (5.86%)	..	..	..
	Europe	50,015 (12.4%)	50,015 (12.4%)	..	..	..
	Middle East & Asia	50,296 (12.5%)	50,296 (12.5%)	..	..	..
	The Americas & Caribbean	12,114 (3.00%)	12,114 (3.00%)	..	..	..
	Antarctica & Oceania	4,297 (1.06%)	4,297 (1.06%)	..	..	..
	Unknown	263,127 (65.2%)	38,352 (21.5%)	..	..	..

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

Practice region	England	379,844 (94.07%)	163,301 (91.35%)	208,884 (96.38%)	7,446 (92.41%)	372,185 (94.11%)
	London	171,368 (42.4%)	84,467 (47.3%)	81,530 (37.6%)	5,371 (64.8%)	165,997 (42.0%)
	South Central	48,740 (12.1%)	26,361 (14.7%)	21,716 (10.0%)	663(8.00%)	48,077 (12.2%)
	South East Coast	43,089 (10.7%)	19,468 (10.9%)	23,324 (10.8%)	297(3.58%)	42,792 (10.8%)
	North West	31,964 (7.92%)	11,666 (6.53%)	20,006 (9.23%)	292 (3.52%)	31,672 (8.01%)
	West Midlands	29,629 (7.34%)	5,756 (3.22%)	23,556 (10.9%)	317 (3.82%)	29,312 (7.41%)
	East of England	24,006 (5.95%)	5,394 (3.02%)	18,405 (8.49%)	207 (2.50%)	23,799 (6.02%)
	South West	19,734 (4.89%)	8,158 (4.56%)	11,463 (5.29%)	113 (1.36%)	19,621 (4.96%)
	North East	4,980 (1.23%)	611 (0.342%)	4,357 (2.01%)	12 (0.145%)	4,968 (1.26%)
	East Midlands	4,594 (1.14%)	1,078 (0.603%)	3,342(1.54%)	174 (2.10%)	4,420 (1.12%)
	Yorkshire & The Humber	1,740 (0.43%)	342 (0.191%)	1,185 (0.547%)	213 (2.57%)	1,527 (0.386%)
	Scotland	12,135 (3.01%)	8,090 (4.53%)	3,822 (1.76%)	223 (2.69%)	11,912 (3.01%)
	Wales	10,868 (2.69%)	6,858 (3.84%)	3,618 (1.67%)	392 (4.73%)	10,476 (2.65%)
	Northern Ireland	921 (0.23%)	500 (0.280%)	407 (0.188%)	14 (0.169%)	907 (0.229%)

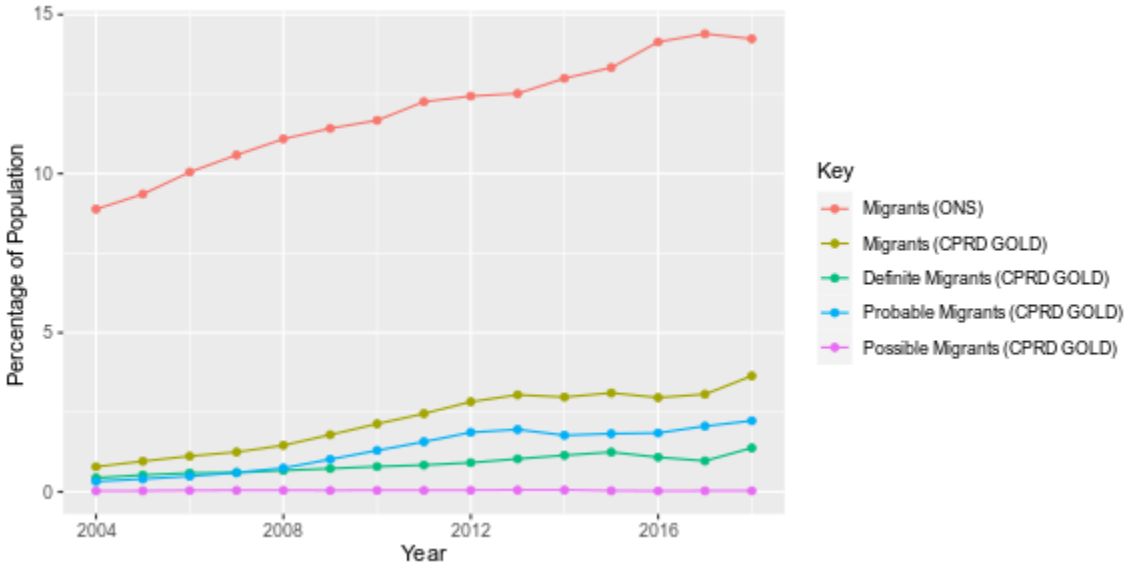
*Percentages are calculated across columns except for first row
**Country of birth codes only available for those in the ‘definite’ migration certainty category

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

Representativeness

The percentage of patients recorded as migrants increased over time in CPRD GOLD by 4.6 times between 2004 (0.79%) and 2018 (3.64%) compared to the 1.6 fold increase in migrants as per ONS data over the same period (8.89% in 2004 to 14.2% in 2018; Figure 2). "Probable" migrants increased faster than the other two certainty categories, the "possible" certainty category remained poorly recorded throughout.

Figure 2. Percentage of international migrants in CPRD and international migrants in ONS by certainty of migration status (2004-2018).

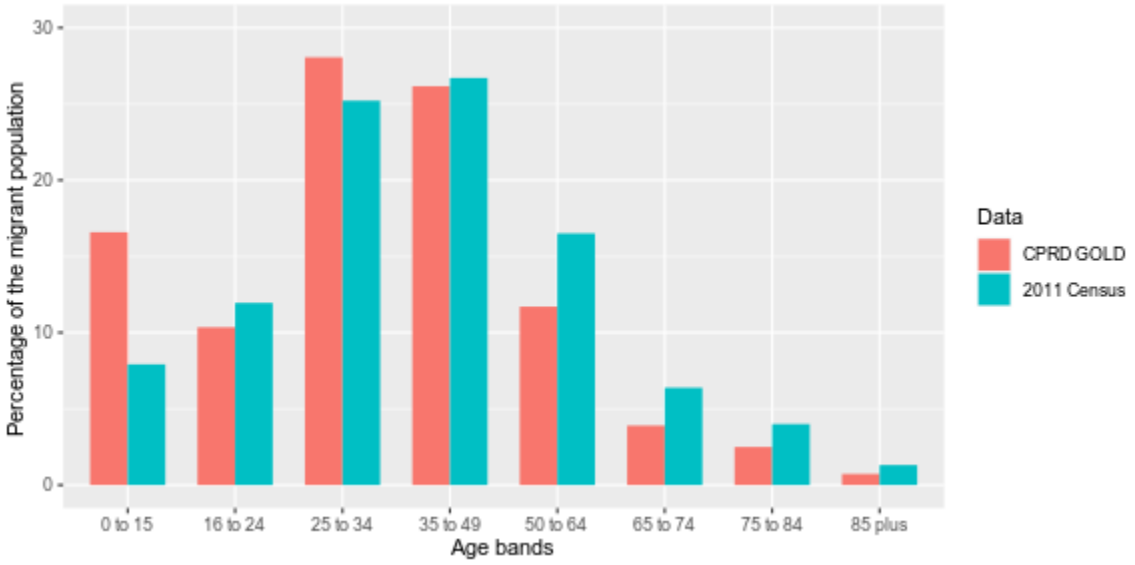


While the percentage of migrants in CPRD GOLD was consistently lower than in ONS country of birth data ($p<0.0001$), the ratio of the percentage of migrants recorded in CPRD compared ONS increased over time from 0.09 in 2004 to 0.26 in 2018 (Table S6). Migrants were under-recorded in CPRD compared to ONS 2011 census data in all age bands (Table S7), with the highest numbers recorded in age band 25-34 year olds (5.22% in CPRD and 25.2% in ONS) and lowest in the age band 85 years and older (0.64% in CPRD, 7.83% in ONS). Migrants aged 0-15 years were most well-recorded in CPRD (2.1% in CPRD, 5.8% in ONS, ratio = 0.41), while 85 years and older were the most poorly recorded group (0.64% in CPRD, 7.83% in ONS, ratio = 0.08).

Comparing the whole migrant cohort within CPRD GOLD and ONS 2011 census data (Figure 3 and Table S8), differences are smallest across age bands between 16 and 49 years old, but greatest for the 0-15 year-old band and age bands above 50 years old. The proportion of females is similarly higher than males in both datasets (52.3% in CPRD and 51.6% in ONS).

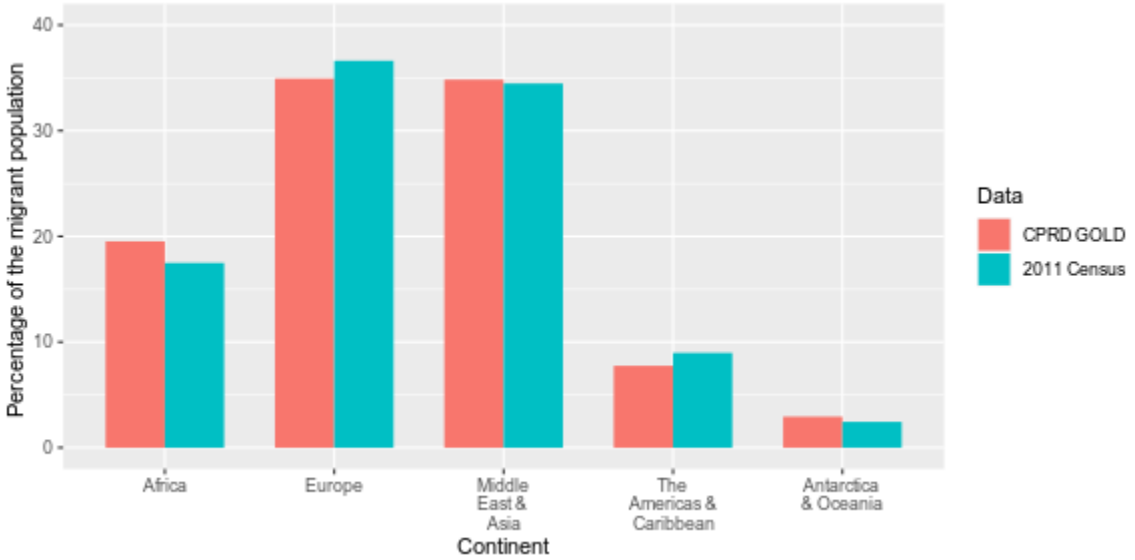
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

Figure 3. Percentage age breakdown of CPRD (2011) and ONS migrant population at the time of the 2011 census



The CPRD migrant cohort and migrants in the 2011 census are similar by continent of birth (Figure 4). Migrants were mostly born in Europe (34.9% in CPRD and 36.6% in ONS) or the Middle East and Asia (34.8% in CPRD and 34.5% in ONS).

Figure 4. Percentage of migrants in CPRD (2011) and 2011 census according to continent of birth as defined by ONS Nomis

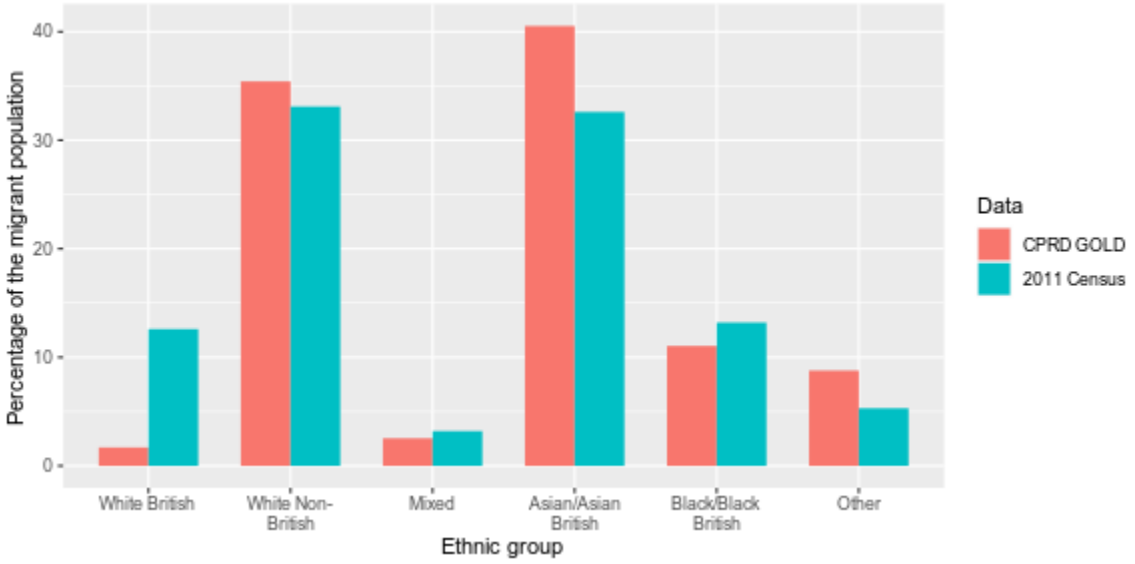


Among the CPRD migrant cohort with known ethnicity, Asian/Asian British ethnicity was more frequently recorded than amongst non-UK born individuals in ONS census data (40.5% in CPRD and 32.6% in ONS) while the White British ethnic group was recorded less frequently (1.70% in CPRD and 12.6% in ONS). White British migrants in the ONS data are likely to reflect those born to British nationals living abroad, or those who identify as White British post-arrival to the UK (21). The remaining ethnic groups had approximately similar proportions between

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0081.v1

datasets (Figure 5). A comparison of ethnicity using the more granular 18 group classification (Table S9) resulted in small numbers, limiting the ability to draw definitive conclusions.

Figure 5. Percentage of migrants in CPRD (2011) and 2011 census by 6 higher-level ethnic groups



Discussion

We developed and evaluated a phenotyping algorithm that identified over 400,000 migrants in CPRD GOLD. The vast majority of these were either "definite" migrants (codes indicating visa or a country of birth outside the UK) or "probable" migrants (codes indicating a first or main language that was not English). Migration status was under-recorded in CPRD GOLD compared to ONS data, particularly in individuals over the age of 50 years, but increased over the years to capture a quarter of the expected proportion of migrants by 2018. The distribution of sex and geographic region of birth were similar between migrants in in CPRD GOLD and ONS datasets. Ethnicity was well-recorded in migrants in CPRD, however the Asian/Asian British ethnic group was overrepresented compared to ONS data.

Several explanations may account for the lower number of migrants identified in CPRD compared with ONS data. Firstly, GPs do not routinely record migration related information in EHRs. Recording may be limited to situations where, for example, an interpreter is needed, or differential health risks in a recent migrant's country of birth/origin will affect clinical decision-making. Secondly, barriers to primary care experienced by migrants, such as language, discrimination, lack of knowledge about services (22), and fear of data sharing for the purposes of immigration enforcement (3), could affect migrants' ability or willingness to register with an NHS GP practice. This corroborates findings of lower levels of primary-care registration amongst newly-arrived migrants to the UK (8) and undocumented migrants and asylum seekers making up a large proportion of patients attending non-NHS primary care (3). The under-recording of migrants could thus represent a lower number of migrants registering with primary care services. Thirdly, barriers to primary care access could also result in lower attendance at consultations, thereby limiting the opportunity for a GP to ask questions on country of birth, language, or visa type. If there are more opportunities to code migration status with increasing time (and more appointments attended) since GP registration, migrants represented in CPRD GOLD may be those who have lived in the UK longer. As such, generalisability of the phenotype only extends to migrants who have registered with primary care, and they are less likely to be newly-arrived migrants (8).

The improved recording of migration status over time, in younger age groups, and in certain ethnic groups could also be explained by healthcare provider coding behaviours or patient healthcare utilisation patterns. Improvements in coding of migrant status over time could reflect the incentivising of GPs to record main/first language terms as

part of the Quality Outcomes Framework between 2008-2011 (23). These codes made up the majority of the migration phenotype, and the rate of increase in recording over time was faster in "probable" migrants (terms related to a non-English main/first language) than "definite" migrants. The better recording of migration in younger age groups may be explained by children having more routine contact with primary care unrelated to disease or illness, such as for childhood immunisations and developmental checks. Healthcare use at older ages related to chronic disease may not be as readily accessed by migrants. Older migrants may have migrated to the UK before EHRs existed or before clinical coding in EHRs was well-established, and their migration status may not have been coded retrospectively. As a smaller proportion of older migrants are recorded as migrants in CPRD GOLD, there may be greater bias when studying health outcomes associated with older age groups. The better representation of migrants in the Asian/Asian British ethnic group could reflect a higher rate of consultations in this ethnic group as previously described in CPRD GOLD (24). However, GPs could also deem migration to be more relevant to patients from an Asian/Asian British ethnic group, for example, due to assumptions made about language proficiency or specific health risks. Interpretation of findings should take this into account when analysing migration and ethnicity data using this phenotype.

Potential sources of bias also affect this study, with the main limitation being misclassification of migration status. Migrants make up considerably less of the general population than non-migrants, and as a result the percentage of migrants misclassified as non-migrants is likely to be low. This means that estimates of outcomes in the non-migrant group would be minimally influenced by misclassification, whereas estimates of the same outcomes in the migrant group may be influenced to a greater extent. This may occur in particular as a result of the inclusion of language terms in the phenotype. Furthermore, the representativeness of CPRD GOLD practices serving migrants compared to all UK GP practices is unknown, and may have affected the low percentage of migrants in CPRD in regions like London (7%) where ONS estimates of Londoners born abroad are much higher (35%) (1). Migrants are also likely to be more mobile than non-migrants within the UK; as CPRD cannot link an individual's record from multiple CPRD practices, migrants may be more likely than non-migrants to be incorrectly counted as more than one individual within the dataset. Significant variation exists between GP practices in their recording of patient sociodemographic indicators, and a more resource-intensive source of validation, such as a nationwide survey of GP practices, is needed to examine these issues further.

Other limitations of the phenotype include the under-identification of older migrants aged 50 years and over. First, language codes also make up the "probable" category of migrants, likely over-identifying migrants from non-English speaking countries and under-identify migrants from English-speaking countries, subsequently underrepresenting economic migrants who have good English proficiency. Second, inclusion of written language codes could also be explored in further development of phenotype certainty categories. Third, aggregation of ethnic groups into 6 higher level categories to deal with small group sizes in migrants loses granularity when comparing the CPRD migrant population with ONS statistics by ethnicity to assess representativeness.

Nevertheless, the involvement of experts in migration health and CPRD to develop the migrant phenotype was a strength of this study. Compared to previous approaches, we included a further 36 relevant diagnosis terms indicating migration to create a more comprehensive phenotype. We categorised terms according to certainty of migration status, allowing future studies to study migration health with varying degrees of certainty for how accurately the phenotype identifies migrant patients in CPRD GOLD. The specificity of the phenotype can be improved by omitting the "possible" migrant certainty category (defined by non-UK origin, making up only 2.1% of all migrants). As the proportion of migrants recorded in CPRD GOLD has improved over time, studying healthcare outcomes in more recent years may be of more value. The cohort in later years should be compared to the 2021 Census as a matter of priority when these data become available.

Conclusions

We used a migration phenotype to identify a large cohort of the UK migrant population and demonstrated the feasibility of using CPRD GOLD to undertake large-scale population-based migration health research in the UK. This will allow researchers and policy-makers to use primary care EHRs to monitor health outcomes and healthcare in migrants for evidence-based action. However, migrants were under-recorded in the CPRD GOLD database compared to ONS population estimates, particularly in older age groups who may have been in the country longer. Migrants in CPRD GOLD were largely representative of the UK migrant population in terms of sex and geographical region of birth. Improvements in recording of migration status in CPRD were also observed over time.

References

1. Office for National Statistics. Population of the UK by country of birth and nationality. 2020 [Accessed 2020 Mar 10]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/datasets/populationoftheunitedkingdombycountryofbirthandnationalityunderlyingdatasheets.

2. Abubakar I, Aldridge RW, Devakumar D, et al. The UCL-Lancet Commission on Migration and Health: the health of a world on the move. Lancet. 2018;392(10164):2606-54.

3. Weller S, Crosby L, Turnbull E, et al. The negative health effects of hostile environment policies on migrants: A cross-sectional service evaluation of humanitarian healthcare provision in the UK. Wellcome Open Research. 2019;4(109).

4. Bhopal RS. Migration, ethnicity, race, and health in multicultural societies. 2nd ed. ed. New York: Oxford University Press; 2014.

5. Burns R, Pathak N, Campos-Matos I, et al. Million Migrants study of healthcare and mortality outcomes in non-EU migrants and refugees to England: Analysis protocol for a linked population-based cohort study of 1.5 million migrants. Wellcome Open Research. 2019;4(4).

6. Katikireddi SV, Cezard G, Bhopal RS, et al. Assessment of health care, hospital admissions, and mortality by ethnicity: population-based cohort study of health-system performance in Scotland. Lancet Public Health. 2018;3(5):e226-e36.

7. National Health Service. The NHS Long Term Plan. 2019 [Accessed 2021 Jul 15]. Available from: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/.

8. Stagg HR, Jones J, Bickler G, Abubakar I. Poor uptake of primary healthcare registration among recent entrants to the UK: a retrospective cohort study. BMJ Open. 2012;2(4):e001453.

9. Bosqui T, O'Reilly D, Väänänen A, et al. First-generation migrants' use of psychotropic medication in Northern Ireland: a record linkage study. International Journal of Mental Health Systems. 2019;13(1):77.

10. Jain A, van Hoek AJ, Walker JL, et al. Inequalities in zoster disease burden: a population-based cohort study to identify social determinants using linked data from the U.K. Clinical Practice Research Datalink. Br J Dermatol. 2018;178(6):1324-30.

11. Jain A, van Hoek AJ, Walker JL, Mathur R, Smeeth L, Thomas SL. Identifying social factors amongst older individuals in linked electronic health records: An assessment in a population based study. PLoS One. 2017;12(11):e0189038.

12. Jain A, Walker JL, Mathur R, et al. Zoster vaccination inequalities: A population based cohort study using linked data from the UK Clinical Practice Research Datalink. PLoS One. 2018;13(11):e0207183.

13. Office for National Statistics. LC2103EW - Country of birth by sex by age. 2011. [Accessed 2019 May 20]. Available from: https://www.nomisweb.co.uk/census/2011/LC2103EW/view/2092957699?rows=c_cob&cols=c_age

14. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117-21.

15. Pathak N, Patel P, Burns R, et al. Healthcare resource utilisation and mortality outcomes in international migrants to the UK: analysis protocol for a linked population-based cohort study using Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES) and the Office for National Statistics (ONS). Wellcome Open Research. 2021;5(156).

16. Denaxas SC, George J, Herrett E, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012;41(6):1625-38.

17. Clinical Practice Research Datalink. Release Notes: CPRD GOLD January 2019. 2019.

18. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.

19. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. Journal of the American Medical Informatics Association. 2019;26(12):1545-59.

20. Office for National Statistics. QS203EW (Country of birth (detailed)) - Nomis - Official Labour Market Statistics. 2013. [Accessed 2021 Jan 11]. Available from: https://www.nomisweb.co.uk/census/2011/qs203ew

21. Office for National Statistics. 2011 Census analysis: Ethnicity and religion of the non-UK born population in England and Wales: 2011. 2015. [Accessed 2021 Jun 01]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/2011censusanalysisethnicityandreligionofthenonukbornpopulationinenglandandwales/2015-06-18

22. 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-e545.

23. NHS Digital. Quality Outcomes Framework (QOF). 2021. [Accessed 2021 Jun 01]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof

24. Mukhtar TK, Bankhead C, Stevens S, et al. Factors associated with consultation rates in general practice in England, 2013-2014: a cross-sectional study. Br J Gen Pract. 2018;68(670):e370-e377.

Funding

Wellcome Trust Clinical Research Career Development Fellowship [206602] and Clinical Research Training Fellowship [211162].

Author contributions

Conceptualization, NP, RWA, AH and PS.; Methodology, NP, RWA, AH, PS and SD; Validation, NP; Formal Analysis, NP; Data Curation, NP and AGI.; Writing – Original Draft Preparation, NP and CZ; Writing – Review & Editing, NP, CZ, RB, RM, RWA, SD, AGI, YB, AH and PS; Visualization, NP.; Supervision, RWA, AH and PS; Project Administration, NP and CZ; Funding Acquisition, NP and RWA.

Data availability statement

Data used in this study were provided by the Clinical Practice Research Datalink (CPRD). Researchers can only access this data through a direct research application to the CPRD Independent Scientific Advisory Committee. Open source metadata in the form of code lists and coding scripts have been made available.

Conflicts of interest

The authors declare no conflicts of interest. NP and RWA receive funding from the Wellcome Trust. RWA has undertaken paid research consulting work on migration and health for Doctors of World and International Labor Organization in the last five years. CZ is employed by Public Health England and contributes to the development of national guidance and policy in migrant health. CZ is also a Trustee for the charity Art Refuge. The views expressed are those of the authors and not necessarily those of the Wellcome Trust, UCL, London School of Hygiene and Tropical Medicine, Public Health England, Guy's & St Thomas' NHS Foundation Trust, and Health Data Research UK.