STUDY PROTOCOL

REVISED 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) [version 2; peer review: 1 approved with reservations, 1 not approved]

Neha Pathak[1-4], Parth Patel[1,2], Rachel Burns[1,2], Lucinda Haim[1,2], Claire X. Zhang[1,3], Yamina Boukari[1], Arturo Gonzales-Izquierdo[1,2], Rohini Mathur[5], Caroline Minassian[5], Alexandra Pitman[6], Spiros Denaxas[1,2,7], Harry Hemingway[1,2], Andrew Hayward[3], Pam Sonnenberg[8], Robert W. Aldridge[1-3]

[1]Institute of Health Informatics, University College London, London, NW1 2DA, UK
[2]Health Data Research UK, London, UK
[3]Institute of Epidemiology & Healthcare,, University College London, London, WC1E 6BT, UK
[4]Guy's & St Thomas's NHS Foundation Trust, London, SE1 9RT, UK
[5]London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
[6]Division of Psychiatry, University College London, London, W1T 7BN, UK
[7]The Alan Turing Institute, London, NW1 2DB, UK
[8]Institute of Global Health, University College London, London, WC1E 6JB, UK

v2 **First published:** 03 Jul 2020, **5**:156
https://doi.org/10.12688/wellcomeopenres.15931.1
Latest published: 24 May 2021, **5**:156
https://doi.org/10.12688/wellcomeopenres.15931.2

Open Peer Review

Reviewer Status ? ✗

	Invited Reviewers	
	1	**2**
version 2 (revision) 24 May 2021		
version 1 03 Jul 2020	? report	✗ report

Abstract

An estimated 14.2% (9.34 million people) of people living in the UK in 2019 were international migrants. Despite this, there are no large-scale national studies of their healthcare resource utilisation and little is known about how migrants access and use healthcare services. One ongoing study of migration health in the UK, the Million Migrants study, links electronic health records (EHRs) from hospital-based data, national death records and Public Health England migrant and refugee data. However, the Million Migrants study cannot provide a complete picture of migration health resource utilisation as it lacks data on migrants from Europe and utilisation of primary care for all international migrants. Our study seeks to address this limitation by using primary care EHR data linked to hospital-based EHRs and national death records.

Our study is split into a feasibility study and a main study. The feasibility study will assess the validity of a migration phenotype, a transparent reproducible algorithm using clinical terminology codes to determine migration status in Clinical Practice Research Datalink (CPRD), the largest UK primary care EHR. If the migration phenotype is found to be valid, the main study will involve using the phenotype in the linked dataset to describe primary care and hospital-based healthcare resource utilisation and mortality in migrants compared to non-migrants. All outcomes will be explored according to sub-conditions identified as research priorities through patient and public involvement, including preventable causes of inpatient admission, sexual and reproductive health conditions/interventions and mental health conditions. The results will generate evidence to inform policies that aim to improve migration health and universal health coverage.

Keywords

migration, migrant, primary care, healthcare usage, mortality, electronic health records, inclusion health

1. **Laurence Lessard-Phillips**, University of Birmingham, Birmingham, UK

2. **Laurence Gruer**, University of Edinburgh, Edinburgh, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Neha Pathak (neha.pathak.09@ucl.ac.uk)

Author roles: Pathak N: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Patel P**: Conceptualization, Methodology, Writing – Review & Editing; **Burns R**: Conceptualization, Methodology, Writing – Review & Editing; **Haim L**: Conceptualization, Methodology, Writing – Review & Editing; **Zhang CX**: Conceptualization, Methodology, Writing – Review & Editing; **Boukari Y**: Conceptualization, Methodology, Writing – Review & Editing; **Gonzales-Izquierdo A**: Conceptualization, Methodology, Writing – Review & Editing; **Mathur R**: Conceptualization, Methodology, Writing – Review & Editing; **Minassian C**: Conceptualization, Methodology, Writing – Review & Editing; **Pitman A**: Conceptualization, Methodology, Writing – Review & Editing; **Denaxas S**: Conceptualization, Methodology, Writing – Review & Editing; **Hemingway H**: Conceptualization, Methodology, Writing – Review & Editing; **Hayward A**: Conceptualization, Methodology, Writing – Review & Editing; **Sonnenberg P**: Conceptualization, Methodology, Writing – Review & Editing; **Aldridge RW**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: RWA received consultancy fees from Doctors of the World, outside the submitted work to complete the 2017 Médecins du Monde Observatory Report.

Grant information: This work was supported the Wellcome Trust through a Clinical Research Career Development Fellowship to RWA [206602], a Clinical Research Training Fellowship to NP [211162] and a Sir Henry Wellcome Postdoctoral Fellowship to RM [201375]. ACH's salary is provided by Central and North West London NHS Community Trust. HH is a National Institute for Health Research (NIHR) Senior Investigator. His work is supported by: Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA; it is chaired, by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC. The National Institute for Health Research University College London Hospitals Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the Wellcome Trust, the NIHR or the Department of Health, International Organisation for Migration or Public Health England. SD is supported by an Alan Turing Fellowship and the UCL Hospitals Biomedical Research Centre. (UCLH BRC).AP is supported by the NIHR UCLH Biomedical Research Centre (BRC).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: 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) [version 2; peer review: 1 approved with reservations, 1 not approved]** Wellcome Open Research 2021, **5**:156 https://doi.org/10.12688/wellcomeopenres.15931.2

First published: 03 Jul 2020, **5**:156 https://doi.org/10.12688/wellcomeopenres.15931.1

REVISED Amendments from Version 1

The revised protocol clarifies the definition of international migrants that will be used in this study. Edits have been made to address minor errors with referencing and the reporting of findings from previous studies. The planned sensitivity analyses have also been updated to reflect the consideration of additional methods to assess comparability between migrant and non-migrant groups. The strengths and limitations of the study have been revised to include addition considerations raised by reviewers.

Any further responses from the reviewers can be found at the end of the article

Introduction

An estimated 14.3% (9.4 million people) of people living in the UK in 2019 were international migrants[1]. Despite this, little is known about how migrants access and use healthcare services in the UK. A systematic review of migrant healthcare in Europe showed high emergency care service use but low uptake of preventive services including outpatient care and screening[2]. One study in Scotland also showed that people of South Asian ethnicities, including those born outside of the UK, had higher rates of avoidable hospital admissions compared to the white Scottish population[3]. However, existing studies of migrant healthcare utilisation in the UK are mostly limited to outpatient, hospital and emergency care. In addition, some have used proxy measures of migration which are unable to provide a true estimate of the impact of migration. For example, a study in England using registration with a GP after the age of 15 as a proxy for migration estimated hospital admission rates to be half the rate of the general population[4].

The Million Migrants study is an ongoing population-based linked cohort study examining secondary healthcare utilisation and mortality in 1.5 million non-European Union (EU) migrants to England[5]. It will link Public Health England (PHE) records of non-EU migrants and refugees to secondary care electronic health records (EHRs) and death registration records. The novel record linkage and cohort size means the Million Migrants will be able to examine in detail the health needs of migrants in England in all hospital-based services (emergency, inpatient and outpatient care) without relying on proxy measures of migration. However, information governance restrictions prevent linkage of PHE migrant and refugee records to UK EHRs from primary care, often the first point of contact in the UK health system and a central part of the NHS Long Term Plan for preventive care[6]. The Million Migrants study is also limited to individuals migrating from outside of the EU. These two factors mean it cannot provide a complete picture of migration health.

To use UK primary care EHR to study migration health without linking to PHE records, a valid migration phenotype is necessary: a transparent reproducible algorithm using clinical terminology codes to determine migration status[7]. A valid migration phenotype is one that determines the migration status for a large number of individuals with high certainty and who are representative of migrants in the general population. A phenotype that is poorly defined or lacks comprehensiveness leads to selection bias and reduces the validity of any findings[7].

A recent study using Clinical Practice Research Datalink (CPRD), one of the largest UK primary care EHRs, described phenotypes for social factors amongst older individuals including migration status[9]. The study estimated that 1.3% of individuals aged ≥ 65 years in CPRD GOLD were international migrants. However, the study did not evaluate the migration phenotype in CPRD Aurum. As 81.3% of migrants in England are aged between 16 and 64 years old[1], it is likely that applying a migration phenotype to individuals of any age in CPRD GOLD and Aurum will identify a higher proportion of international migrants. If this phenotype is then found to be broadly representative of the UK migrant population, it will be possible to use CPRD and datasets linked to CPRD to describe primary care and hospital-based healthcare resource utilisation and mortality in migrants from EU and non-EU countries compared to non-migrants across the UK.

This protocol describes the planned methods of a feasibility study and a main study to describe healthcare resource utilisation and mortality for migrants in the UK using CPRD. This will generate evidence to address the gaps outlined in migration health research and inform policy aimed at increasing equitable healthcare for international migrants attending UK primary care. The definition of migrants used in this study will reflect that of the International Organization for Migration, where a migrant is an individual who “moves away from his or her place of usual residence, whether within a country or across an international border, temporarily or permanently, and for a variety of reasons”[10].

Aims and objectives

The feasibility study aims to assess the validity of a migration phenotype in CPRD. Specific objectives are:

1. To develop a migration phenotype.
2. To assess the completeness of recording of migration status using the migration phenotype.
3. To assess the representativeness of recording of migration status using the migration phenotype.

The main study will be completed if the phenotype is found to be valid and aims to describe healthcare resource utilisation and mortality in migrants to the UK who have registered with primary care. Specific objectives are:

1. To describe patterns of primary care and hospital-based healthcare resource utilisation by migrants compared to non-migrants.
2. To describe the costs of primary care and hospital-based healthcare resource utilisation by migrants compared to non-migrants.

3. To estimate total healthcare resource utilisation patterns across primary and secondary care and investigate whether distinct groups of patients exist based on degree of utilisation.

4. To describe mortality outcomes in migrants compared to non-migrants.

Methods

Ethical approvals

The feasibility and main study were approved by the MHRA (UK) Independent Scientific Advisory Committee (ISAC protocol 19_062R), under Section 251 (NHS Social Care Act 2006). This study will be carried out as part of the CALIBER programme. CALIBER, led from the UCL Institute of Health Informatics, is a research resource consisting of anonymised, coded variables extracted from linked electronic health records, methods and tools, specialised infrastructure, and training and support[11,12].

Feasibility study

Study design. An observational, retrospective longitudinal population-based cohort study.

Data resource and processing. Data will be extracted from CPRD using the CALIBER resource. CPRD collects de-identified data of patients registered with a network of GP practices across the UK. The data encompass 45 million patients, including 13 million currently registered patients, across two datasets: CPRD GOLD and CPRD Aurum[13]. CPRD GOLD contains data contributed by practices using Vision® electronic patient record system software and is broadly representative of the UK general population with respect to age, sex and ethnicity[14]. CPRD Aurum contains data from practices using EMIS Web® electronic patient record system software and is broadly representative of the UK general population with respect to age, sex, geographical spread and deprivation[16].

Study population. Individuals of all ages listed in CPRD where the individual record was of 'acceptable' research quality as verified by the CPRD and the GP that the patient is registered to has been deemed to be contributing 'up-to-standard' (UTS) data at the study start date[14].

The study start date is 1st January 1997. The end of the study period is limited by the most recent data available: December 2018 for CPRD GOLD and September 2018 for CPRD Aurum. An individual will stop contributing to active follow up at the earliest of: the date a patient's care was transferred out of a CPRD practice, the practice's last collection date, patients' date of death or the last date of the study.

Comparator population. The comparators for validation of this cohort are published aggregate Office of National Statistics (ONS) data on the population of the UK by country of birth[1] and aggregate ONS 2011 English Census data on country of birth[17].

Outcomes

1. A consensus list of diagnostic terms indicating migration to the UK (a migration phenotype).

2. Overall and annual percentage of individuals recorded as international migrants in a UK primary care sample (completeness).

3. Percentage of individuals recorded as international migrants in a UK primary care sample compared to published aggregate ONS statistics: by year, age, sex and country of birth (representativeness).

Development of phenotype. Previously established methods by CALIBER will be used for the development of a migration phenotype[12]. The CPRD code browsers will be searched for diagnostic terms relating to migration using the following search terms: *migrant*, *migrat*, *countr*, *asylum*, *refugee*, *visa*, *abroad*, *born in*, *origin*, *illegal*, *language*. This initial phenotype will then be reviewed and refined by migration health experts and experts in using CPRD from the CALIBER team. Finally, each diagnostic term will be assigned a category based on the type of term (visa status, language, country of birth, origin) and a category based on the certainty of migration status ("definite", "probable", "possible"). We have found 434 diagnostic terms in an initial search (see Extended data[15]).

Analysis plan. Previously developed methodology to assess the validity of phenotypes in CPRD will be used to achieve outcomes 2 and 3 including:

Completeness: we will examine the percentage of recorded migrants in CPRD throughout the study period, per year and at the time of the 2011 English census will be calculated by dividing the number of individuals identified as migrants by our phenotype by the total number of individuals in the CPRD dataset. This will be done for all migrants and sub-groups according to type of migration term and certainty of migration status. Distribution by sex, age and geographical region of birth will be estimated.

Representativeness: we will undertake a comparison of recorded migrants in CPRD with the percentage of migrants in ONS country of birth statistics per year (examined visually and using chi-squared test of proportions; calculating ratio of proportion in CPRD compared to proportion in ONS)[17]. Comparison of recorded migrants in CPRD living in England on the date of the 2011 English census to 2011 English Census data on country of birth stratified by sex, age and geographical region of origin (examined visually and using chi-squared test of proportions; calculating ratio of proportion in CPRD compared to proportion in ONS).

Main study

Study design. An observational, retrospective longitudinal population-based cohort record linkage study.

Data resources, processing and linkage. Data will be extracted from the CPRD GOLD and Aurum datasets and linked to Hospital Episodes Statistics (HES) datasets, death registration data and Index of Multiple Deprivation records obtained through the CALIBER resource[11,18]. CPRD GOLD and Aurum have been described earlier in this paper in the feasibility study section of the methods. For patients in English practices that have consented to take part in the CPRD linkage schemes, a subset of CPRD data is linked to HES, ONS mortality records and patient and practice-level IMD records. We describe the linked records that will be used for our study below. Data linkage in England is carried out by the Trusted Third Party NHS Digital[19].

HES Admitted Patient Care data (HES APC): records for all admissions to, or attendances at English NHS healthcare providers including private patients treated in NHS hospitals, patients resident outside of England and care delivered by treatment centres funded by the NHS. All NHS healthcare providers in England, including acute hospital trusts, primary care trusts and mental health trusts provide data. HES APC data includes the complete set of hospital episode information (admission and discharge dates, diagnoses (identifying primary diagnosis), specialists seen under and procedures undertaken) for each linked patient with a hospitalisation record.

HES Outpatient (HES OP): records for all outpatient appointments occurring in England only including information on the type of consultation, appointment dates, hospital specialty, referral source, waiting times, clinical diagnosis and procedures performed.

HES Accident and Emergency (HES A&E): records for all patient care administered in the accident and emergency setting in England. These data are a subset of national A&E data collected by NHS England to monitor the national standard that 95% of patients attending A&E should wait no longer than 4 hours from arrival to admission, transfer or discharge. A&E data is submitted by A&E providers of all types in England. Data collected includes details about patients' attendance, outcomes of attendance, waiting times, referral source, A&E diagnosis, A&E treatment (drugs prescribed not recorded), A&E investigations and Health Resource Group.

Death Registration data: records from the ONS including information on the official date and causes of death using ICD-10 codes.

Patient-level IMD 2015: The latest available patient postcode of residence in CPRD for English practices in the linkage scheme is mapped to a Lower Layer Super Output Area (LSOA) boundary. The LSOA of residence then allows linkage to 2015 English Index of Multiple Deprivation (composite and individual domains). Data are provided as quintiles, deciles or twentiles of the deprivation score to prevent disclosure of patient location.

Practice-level IMD (Standard): The general practice postcode linkages are available for all practices in CPRD GOLD and CPRD Aurum and are linked to 2015 English Index of Multiple Deprivation (composite and individual domains), 2016 Scottish Index of Multiple Deprivation (composite and individual domains), 2017 Northern Ireland Index of Multiple Deprivation (composite and individual domains), 2014 Welsh Index of Multiple Deprivation (composite and individual domains). The most recent national Indices of Deprivation are provided for each country. Data is provided as quintiles or deciles of the deprivation score to prevent disclosure of patient location. Access is provided by CPRD subject to ISAC approval. This dataset will only be used if patient-level IMD data is not available for an individual.

Study population. Individuals of all ages listed in CPRD where the individual record was of 'acceptable' research quality as verified by the CPRD and the GP that the patient is registered to has been deemed to be contributing 'up-to-standard' (UTS) data at the study start date.

The study start date is 1st January 1997, although the exact start date will be informed by the feasibility study taking representativeness of migrant phenotype over time into account. For primary care analyses, the end of the study period is limited by the most recent data available: December 2018 for CPRD GOLD and September 2018 for CPRD Aurum. For hospital-based care analyses, the study start and end dates will be limited by the coverage of the latest releases of linked data: HES APC (April 1997 to November 2019), HES OP (Apr 2003 to November 2018), HES A&E (April 2007 to November 2018).

An individual will stop contributing to active follow up at the earliest of: the date a patient's care was transferred out of a CPRD practice, the practice's last collection date for GOLD/Aurum data extraction, patients' date of death or the last date of the study.

Exposure. Migration to the UK is the exposure of interest. This will be defined using the migration phenotype developed and validated as outlined previously in the feasibility study section.

Comparator population. The non-exposed cohort: individuals with no evidence of migration to the UK as defined by the migration phenotype.

Outcomes. We have selected outcomes that are important to researchers and policy-makers as well as migrants and refugees who have attended our public engagement workshops. Where possible, outcomes are in alignment with the Million Migrants study to facilitate triangulation of results[5]. Outcomes fall into one of three categories: primary care, hospital-based care and mortality. Table 1 summarises the clinical and statistical definition of these outcomes. All outcomes will be explored by subgroup conditions where appropriate. Table 2 summarises clinical definitions of subgroups of conditions which have also been aligned with the Million Migrants study. Details of diagnostic terms for conditions within each sub-group can be found in the Extended Data file[15].

Table 1. Outcomes by category with clinical and statistical definitions.

Outcome	Clinical definition	Statistical definition	Likely statistical modelling approach
Primary care outcomes			
Consultations	Any type of consultation with primary care with any member of staff.	Numerical indicator for number of consultations.	Poisson regression
Prescriptions	Prescription for any medication issued in primary care.	Numerical indicator for number of prescriptions.	Poisson regression
Referrals to secondary care	Referral made from primary care to hospital-based services.	Numerical indicator for number of referrals.	Poisson regression
Missed appointments	Appointments in primary care that were not attended.	Numerical indicator for number of appointments coded as did not attend.	Poisson regression
Diagnosis of existing health conditions	Presence of a health condition from one of the sub-groups outlined in Table 2.	Binary indicator for presence of health condition (yes/no) from which a numerical indicator for number of people with a condition can be estimated.	Poisson regression
Hospital-based outcomes			
Hospital attendances	Hospital attendances in inpatient, outpatient, or A&E.	Numerical indicator for number of attendances.	Poisson regression
Hospital admissions	Admission into the hospital as an inpatient.	Numerical indicator for number of admissions.	Poisson regression
Duration of hospital admission	Days spent in hospital as an inpatient.	Numerical indicator for number of days.	Poisson regression
30 day emergency readmissions	Emergency admissions to any hospital in England occurring within 30 days of the last, previous discharge from Hospital.	Numerical indicator for number of emergency readmissions recorded within 30 days of the index admission discharge date.	Poisson regression
Missed outpatient appointments	Outpatient appointments that were not attended.	Numerical indicator for number of outpatients appointments coded as did not attend.	Poisson regression
Missed procedures	Procedures that were not attended.	Numerical indicator for number of appointments for procedures coded as did not attend.	Poisson regression
Diagnosis of existing health conditions	Presence of health conditions by sub-groups of conditions outlined in Table 2.	Binary indicator for presence of health condition (yes/no) from which a numerical indicator for number of people with a condition can be estimated.	Poisson regression
Mortality outcomes			
Death from all causes	Deaths in England from any cause	Binary indicator for presence of death due to any cause (yes/no).	Standardised mortality ratio (SMR).
Death from specific conditions	Deaths in England from conditions within sub-groups outlined in Table 2.	Binary indicator for presence of death due to any cause (yes/no).	Cox proportional hazards model.

Sample size. Based on a feasibility count in 2019, there are 416,353 events with a diagnostic term indicating migration to the UK in CPRD GOLD records of acceptability research quality between 2007 and 2016. We have based our sample size calculation on the full study primary outcome of primary care consultations. We estimate a general population (e.g. migrants and non-migrants combined) primary care consultation rate of 1800 per 100 person years over the study period. Based on our feasibility counts of diagnostic terms indicating migration, the study has sufficient statistical power (80%) to detect a Hazard Ratio of 0.99 for this outcome when comparing all migrants compared to all non-migrants at the 5% significance level. The study also has sufficient statistical power (80%) to detect a Hazard Ratio of 0.90 for this outcome when comparing migrant

Table 2. Clinical definition for primary care, hospital-based and mortality subgroup outcomes.

Outcome subgroups	Clinical definition
Ambulatory care sensitive (ACS) conditions	Conditions where effective community care can prevent inpatient hospital admission or death[20].
Amenable conditions	Conditions where hospital admissions or death could be avoided through high- quality preventative healthcare[21].
Preventable conditions	Conditions where all or most hospital admissions or deaths from a specific cause could be avoided by established medical or public health interventions[21].
Avoidable conditions	Conditions that are considered preventable, amenable or both, where each admission or death is only counted once. When cause of admissions or death falls within both the preventable and amenable definition, all admissions or deaths from that cause are counted in both categories when they are presented separately[21].
Sexual and reproductive health conditions and treatments	Conditions and treatments related to sexual and reproductive health. These are defined using the seven domains from the Guttmacher-Lancet commission on sexual and reproductive healthcare and rights[22]: abortion, contraception, gender-based violence, HIV and sexually transmitted infections, infertility, maternal and newborn health, and reproductive cancers.
Mental health outcomes	Psychiatric disorders including severe mental illness (psychotic disorders), common mental disorders (mixed anxiety and depression, depressive episode, phobias, obsessive compulsive disorder, panic disorder, eating disorders, post-traumatic stress disorder, perinatal mental health conditions), and personality disorders. Suicide attempt/self-harm.
All causes	Death due to any cause.
ICD-10 chapter	Death due to specific conditions such as infectious disease, disease of the blood, cardiovascular diseases, digestive disease, genitourinary disease, musculoskeletal disease, nervous disease, respiratory disease, endocrine disease, injury or external causes, mental and behavioural, or Neoplasms[20].

subgroups (e.g. international migrants from Poland or India) to all non-migrants at the 5% significance level.

After completion of the feasibility study, we will use the results to update our sample size calculation with the number of individuals with diagnostic terms indicating migration. We will use the results of this updated sample size calculation to assess whether to proceed to the full study or not in conjunction with the overall representativeness compared to aggregate ONS data on migration as demonstrated by the feasibility study. If the feasibility study finds completeness or representativeness is worse than the 2017 study of social factors including migration in older people[9] or the updated sample size calculation means that the study does not have the level of statistical power required, we will not proceed with the main study.

Analysis plan. All statistical analyses will be carried out using the latest available versions of R software.

Patterns of healthcare resource utilisation: Annual incidence rates and incidence rate ratios will be calculated for all primary and hospital-based care outcomes presented in Table 1 and subgrouped by outcomes in Table 2. Poisson regression will be used to generate rate ratios, with robust standard errors to produce 95% confidence intervals.

Costs of healthcare resource utilisation: Methods previously used to study this in patients with irritable bowel syndrome in linked CPRD and HES data[23] will be replicated. Absolute costs will be calculated as total mean individual annual costs with 95% confidence intervals. The costs of health services in primary care will be obtained from nationally calculated unit costs as NHS reference costs[24] and costs of medications from the British National Formulary[25]. The cost of secondary healthcare utilisation will be calculated according to national tariff prices based on the national average unit costs of providing each service; this is published as the National Schedule of Reference Costs[24].

Total healthcare utilisation patterns: Markers of total healthcare utilisation within primary and secondary care will be identified and patients will be classified according to total healthcare utilisation defined by their chronological sequence of clinical events in all healthcare settings. An exploratory multivariate statistical technique such as Cluster Analysis (K-mean clustering or hierarchical clustering) will be applied to determine whether separable groups of patients who have missed opportunities for preventive healthcare exist.

Mortality outcomes: Standardised mortality ratios (SMR) using ONS death data will be summarised by age and gender. For deaths due to specific conditions, an appropriate regression model will be used. Suicide rates will be based on the ONS definition of suicide, which includes deaths with an underlying cause of intentional self-harm, as well as those with an underlying cause of undetermined intent.

Covariates. The following covariates will be included in the analysis model for all outcomes and sub-conditions: age, sex, deprivation level (Index of Multiple Deprivation quintile), and ethnicity. Additional lists of covariates will be developed where relevant to specific conditions in the sub-groups outlined in Table 2.

Sensitivity analyses. Where possible, stratified measures will be calculated according to: sex, age, socioeconomic status, ethnicity, migrant visa type, geographical region of birth, general practice consultation type (e.g. face to face versus telephone-based), staff type (e.g. role, gender), method of hospital admission and hospital specialty.

CPRD practices may not be representative of all practices in the UK or of practices serving international migrants to the UK. To mitigate this, proportions of migrants will be described regionally - if there is a large amount of variation, analyses will be weighted to account for this using previously described methods by Aldridge *et al.*[26].

The distribution of covariates across migrant and non-migrant groups will be assessed. Additional methods to achieve comparability between groups will be considered in sensitivity analysis where the uneven distribution of covariates is likely to introduce significant bias.

Information governance

All analyses will be completed on the UCL Data Safe Haven (DSH), an information technology infrastructure certified to national and international information governance standards. The dataset will be securely destroyed after 20 years, in line with UCL's record retention policy. There may be small numbers with specific outcomes or of specific migrant types and in line with CPRD policy, we will not report any data with a cell containing <5 events and, where necessary, we will 'protect' these counts with secondary suppression.

Dissemination of results

We will disseminate research findings to a variety of stakeholders, including patients, healthcare professionals, voluntary organisations, policy-makers, politicians and the public. We will achieve this through the co-creation of research dissemination materials (e.g. lay reports and videos) as well as research engagement stands and workshops in patient and public settings.

Study status

At the time of submission, CPRD GOLD data has been extracted for analysis, cleaned and prepared for validation and validation started with ongoing refinements. Data has been prepared and explored for subsequent analyses in GOLD. A linkage request for linkage to IMD data has been completed and the data provided by CPRD. A linkage request for HES and ONS data is being prepared. Analyses using Aurum data have not yet started.

Discussion

This protocol describes a method of creating and validating and EHR phenotype to describe the healthcare utilisation, morbidity and mortality of international migrants to the UK across primary and secondary care.

Many of the strengths of this study are shared with the Million Migrants study[5]. These include the large size of the cohort and extensive stakeholder engagement. We have collaborated with migrants, refugees and advocacy groups as well as a range of clinical, research and policy stakeholders to ensure ethical and efficient data use and optimise the impact of our research findings. It will also be possible to triangulate secondary care and mortality outcomes for non-EU migrants in the present study with the results of the Million Migrants study.

Unique strengths of the present study include the methods used to develop the migration phenotype, specifically the involvement of migration health experts and clinicians. The study includes primary care data and imposes no restrictions on country of birth or visa types. This means that our study addresses important limitations of the Million Migrants study and profiles a larger part of the patient journey. Another unique strength is the cluster analyses: these will focus on identifying clusters of patients attending GP services that have missed opportunities for care/less resource utilisation so may not be benefiting from preventive services largely delivered in primary care. These findings can then be used to inform development and evaluation of interventions to improve care for underserved groups.

Nonetheless, there are some important sources of bias that must be considered when interpreting any results relating to the fact that determining migration status is dependent on clinician coding. First, clinician coding may be incorrect resulting in misclassification bias. Second, clinician coding may be incomplete resulting in missing data, and therefore, there may be under-recording of migration and the presence of migrants in the comparator population. Third, language coding was incentivised between 2008 to 2011 so representativeness may be better during that period and the cohort may be skewed towards non-English speaking migrants (selection bias)[27]. This could also be a unique strength of the study as the cohort could be particularly useful for understanding healthcare access and use by non-English speaking migrants who may face additional barriers to care. Fourth, this study only captures the healthcare utilisation of migrants who are known to the NHS, rather than those who do not use healthcare or face significant access barriers that prevent them from accessing care. Findings are unlikely to be representative of migrant subgroups like asylum seekers and undocumented migrants and others who are unable to access without fear of being charged for NHS services[28]. Fifth, CPRD does not provide routine linkages to data on individual-level deprivation, and the study's use of area-level deprivation does not account for the individual-level measures of socio-economic position that play a role in the association between migration and healthcare utilisation.

Conclusion

In summary, this study has been designed as a novel linkage study to complement the Million Migrants study by including data from primary care and EU migrants. The findings of this study will address important gaps in migration health research and inform policy aimed to increase equitable healthcare for international migrants attending UK primary care.

Data availability

Underlying data

No data are associated with this article.

Extended data

Open Science Framework: 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). https://doi.org/10.17605/OSF.IO/NPA5W[15]

This projections the following extended data:

- Extended data.pdf (PDF contains additional materials listed below)
  - Provisional Read code list for international migration
  - Revised definition of avoidable conditions
  - Avoidable conditions definition for children and young people
  - Definition of ambulatory care sensitive conditions
  - Definition of sexual and reproductive health outcomes
  - Definition of mental and behavioural disorders

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Reporting guidelines

Open Science Framework: RECORD checklist for '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)' https://doi.org/10.17605/OSF.IO/NPA5W[15]

Acknowledgements

We are extremely grateful for the time and views of all the people born outside of the UK that attended our public engagement workshops when developing this study.

References

1. **Population of the UK by country of birth and nationality: individual country data**. Office for National Statistics. [cited 2020 Mar 10].
Reference Source
2. Graetz V, Rechel B, Groot W, *et al.*: **Utilization of health care services by migrants in Europe-a systematic literature review.** *Br Med Bull.* 2017; **121**(1): 5–18.
PubMed Abstract | Publisher Full Text
3. 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): e226–36.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Steventon A, Bardsley M: **Use of secondary care in England by international immigrants.** *J Health Serv Res Policy.* 2011; **16**(2): 90–4.
PubMed Abstract | Publisher Full Text
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 [version 1; peer review: 2 approved, 2 approved with reservations].** *Wellcome Open Res.* 2019; **4**: 4.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Plan NLT: **Online version of the NHS Long Term Plan**. NHS Long Term Plan, 2019; [cited 2020 Mar 10].
Reference Source
7. Hripcsak G, Albers DJ: **Next-generation phenotyping of electronic health records.** *J Am Med Inform Assoc.* 2013; **20**(1): 117–21.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Herrett E, Thomas SL, Schoonen WM, *et al.*: **Validation and validity of diagnoses in the General Practice Research Database: a systematic review.** *Br J Clin Pharmacol.* 2010; **69**(1): 4–14.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Jain A, van Hoek AJ, Walker JL, *et al.*: **Identifying social factors amongst older individuals in linked electronic health records: An assessment in a population based study.** *PLoS One.* 2017; **12**(11): e0189038.
PubMed Abstract | Publisher Full Text | Free Full Text
10. International Organization for Migration: **Who is a migrant?** [cited 2021 May 13].
Reference Source
11. 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.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Denaxas S, Gonzalez-Izquierdo A, Direk K, *et al.*: **UK phenomics platform for developing and validating EHR phenotypes: CALIBER.** *Epidemiology.* 2019; [cited 2019 Apr 8].
Publisher Full Text
13. **Clinical Practice Research Datalink CPRD**. [cited 2020 Mar 10].
Reference Source
14. Herrett E, Gallagher AM, Bhaskaran K, *et al.*: **Data Resource Profile: Clinical Practice Research Datalink (CPRD).** *Int J Epidemiol.* 2015; **44**(3): 827–36.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Pathak N, Patel P: **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).** 2020.
http://www.doi.org/10.17605/OSF.IO/NPA5W
16. Wolf A, Dedman D, Campbell J, *et al.*: **Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum.** *Int J Epidemiol.* 2019; **48**(6): 1740–1740g.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Census: **LC2103EW - Country of birth by sex by age**. Data Viewer - Nomis - Official Labour Market Statistics, 2011; [cited 2019 May 20].
Reference Source
18. **CPRD linked data**. CPRD. [cited 2020 Mar 10].
Reference Source

19. **NHS Digital** [cited 2020 Mar 10].
Reference Source

20. **2.6 Unplanned hospitalisation for chronic ambulatory care sensitive conditions**. NHS Digital. [cited 2019 Feb 21].
Reference Source

21. **Avoidable mortality in the UK QMI.** Office for National Statistics. [cited 2019 Feb 21].
Reference Source

22. Starrs AM, Ezeh AC, Barker G, *et al.*: **Accelerate progress—sexual and reproductive health and rights for all: report of the Guttmacher–Lancet Commission.** *Lancet.* 2018; **391**(10140): 2642–2692.
PubMed Abstract | Publisher Full Text

23. Canavan C, West J, Card T: **Calculating Total Health Service Utilisation and Costs from Routinely Collected Electronic Health Records Using the Example of Patients with Irritable Bowel Syndrome Before and After Their First Gastroenterology Appointment.** *Pharmacoeconomics.* 2016; **34**(2): 181–94.
PubMed Abstract | Publisher Full Text | Free Full Text

24. **Reference costs.** NHS Improvement. [cited 2019 Feb 21].
Reference Source

25. BNF: **British National Formulary - NICE**. [cited 2019 Feb 21].
Reference Source

26. Aldridge RW, Zenner D, White PJ, *et al.*: **Tuberculosis in migrants moving from high-incidence to low-incidence countries: a population-based cohort study of 519 955 migrants screened before entry to England, Wales, and Northern Ireland.** *Lancet.* 2016; **388**(10059): 2510–8.
PubMed Abstract | Publisher Full Text | Free Full Text

27. British Medical Association and NHS employers: **Clinical directed enhanced services (DES) guidance for GMS contract 2008/2009**. 2008; [cited 2020 Mar 10].
Reference Source

28. Weller SJ, Crosby LJ, Turnbull ER, *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.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ✕

Version 1

Reviewer Report 18 February 2021

https://doi.org/10.21956/wellcomeopenres.17475.r42303

✕ **Laurence Gruer**
Usher Institute, University of Edinburgh, Edinburgh, UK

This paper sets out the rationale and method for what could be a large and important record linkage study of the comparative health of and health service utilisation by international migrants resident in the UK. It describes an initial feasibility study using the concept of a "migration phenotype". If this was successful, a main study would be done.

The viability of the study therefore hinges on being able to create a reliable "migration phenotype". This phenotype would be based on relevant data items in individuals' electronic health records in the Clinical Practice Research Datalink (CPRD) which could clearly differentiate "migrants" from "non-migrants". If reliable, the phenotype could then be used as the basis for the main study, in which migrants could be compared with non-migrants across a wide range of health and health service indicators.

The "migration phenotype" would be generated by developing a "*a transparent reproducible algorithm using clinical terminology codes to determine migration status*". This would entail the following: "*The CPRD code browsers will be searched for diagnostic terms relating to migration using the following search terms: *migrant*, *migrat*, *countr*, *asylum*, *refugee*,*visa*, *abroad*, *born in*, *origin*, *illegal*, *language*. This initial phenotype will then be reviewed and refined by migration health experts and experts in using CPRD from the CALIBER team. Finally, each diagnostic term will be assigned a category based on the type of term (visa status, language, country of birth, origin) and a category based on the certainty of migration status ("definite", "probable", "possible")."*

This could work if the relevant terms are recorded in the CPRD with a high degree of completeness and accuracy. Are they?

In the introduction, it is stated *"An estimated **14.3%** (9.4 million people) of people living in the UK in 2019 were international migrants." It later states: "A recent study using Clinical Practice Research Datalink (CPRD), the largest UK primary care EHR, described phenotypes for social factors amongst older individuals including migration status. The study estimated that **1.6%** of individuals aged* ≥ 65 years in CPRD were international migrants." As 1.6% seemed very low, I read the paper about this study

(Jain *et al.* 2017). I discovered the **above statement was incorrect**. In fact, the study found that **data completeness** for immigration status was only 1.6%. This comprised 0.7% where country of birth was recorded and 0.9% where a "first language" code was recorded. Indeed, the paper explicitly said***: the most incompletely recorded social factor was immigration status.*** It added*, "among those with data on immigrant status, there was marked over-representation of immigrants (n = 7,866, ~81% of the total) among those with recorded data but under-representation when immigrant status was considered as a binary variable (1.3% of the total study population compared to 9.9% non-UK born individuals in the English Census)."* This means that **country of birth,** the main indicator of immigration status, **was not recorded at all for 99.3%** of individuals in the CRPD database to be used to develop the phenotype. Even if the less reliable term "first language' is added, the two key terms for determining immigrant status were **missing** for **98.4%** of the individuals in the database **.** None of the other search terms proposed for the phenotype algorithm could compensate for this vast amount of missing data. While completeness of recording of "ethnicity" was much higher at about 80%, this is not an adequate proxy for "migrant" as many people with a non-White British ethnicity were born in the UK and are thus non-immigrants. Whilst the study by Jain *et al.* was limited to people over 65, there is no reason to believe the recording of immigrant status in the CRPD database would be any better for younger people.

In the Discussion, the study team acknowledge the risk of missing data, stating *"clinician coding may be incomplete resulting in missing data, and therefore, there may be under-recording of migration and the presence of migrants in the comparator population".* What they don't seem to have appreciated is that over 98% of the relevant data are missing!

Assuming the findings in the paper by Jain *et al.* are correct, they indicate that the CPRD cannot be used to develop a reliable immigrant phenotype.

From our experience in Scotland, the only reliable source of country of birth data is the Census. It was for this reason that the Scottish Health and Ethnicity Study was made possible through the successful linkage of the Census, with self-reported ethnic group and country of birth, to health and death records.

I thus respectfully invite the authors to reconsider whether either the feasibility study or the main study are viable.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

No

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: For the past 15 years I have specialised in research on ethnicity and health. From 2014-20 I was a co-investigator on the Scottish Health and Ethnicity Linkage Study, Phase 4, and am a co-author of an analysis of completeness of ethnic coding for hospital admissions in Scotland.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 13 May 2021

Neha Pathak, University College London, London, UK

Thanks very much for your helpful comments. We appreciate the time you have taken to read and provide constructive feedback on our protocol. We have addressed all of your points below, and included your comments alongside our responses in bold italics. We have numbered our responses in order to cross-reference them with the review.

******

Reviewer comment 1:
The viability of the study therefore hinges on being able to create a reliable “migration phenotype”. This phenotype would be based on relevant data items in individuals’ electronic health records in the Clinical Practice Research Datalink (CPRD) which could clearly differentiate “migrants” from “non-migrants”. If reliable, the phenotype could then be used as the basis for the main study, in which migrants could be compared with non-migrants across a wide range of health and health service indicators. The “migration phenotype” would be generated by developing a “a transparent reproducible algorithm using clinical terminology codes to determine migration status”. This would entail the following: “The CPRD code browsers will be searched for diagnostic terms relating to migration using the following search terms: *migrant*, *migrat*, *countr*, *asylum*, *refugee*,*visa*, *abroad*, *born in*, *origin*, *illegal*, *language*. This initial phenotype will then be reviewed and refined by migration health experts and experts in using CPRD from the CALIBER team. Finally, each diagnostic term will be assigned a category based on the type of term (visa status, language, country of birth, origin) and a category based on the certainty of migration status (“definite”, “probable”, “possible”).”
This could work if the relevant terms are recorded in the CPRD with a high degree of completeness and accuracy. Are they?

Author response 1:
We agree that this study hinges on a migration phenotype and for that reason have included a feasibility study of the migration phenotype as part of the study protocol outlining how we would assess completeness and representativeness.

Since submitting this study protocol, we have completed the feasibility study, which shows that while migrants are under-recorded in CPRD GOLD compared to ONS migrant population estimates, the cohort’s demographic characteristics are largely representative of the wider migrant population according to ONS. Sufficient power can also be achieved with the present migrant cohort to examine a variety of primary care outcomes. However, as this is a protocol paper we do not believe that these results should be included in this paper.

Publishing study protocols is important for transparency in research which is why we have written this protocol including the feasibility study.

Reviewer comment 2:
In the introduction, it is stated “An estimated 14.3% (9.4 million people) of people living in the UK in 2019 were international migrants.” It later states: “A recent study using Clinical Practice Research Datalink (CPRD), the largest UK primary care EHR, described phenotypes for social factors amongst older individuals including migration status. The study estimated that 1.6% of individuals aged ≥ 65 years in CPRD were international migrants.” As 1.6% seemed very low, I read the paper about this study (Jain et al. 2017). I discovered the above statement was incorrect. In fact, the study found that data completeness for immigration status was only 1.6%. This comprised 0.7% where country of birth was recorded and 0.9% where a “first language” code was recorded. Indeed, the paper explicitly said: the most incompletely recorded social factor was immigration status. It added, “among those with data on immigrant status, there was marked over-representation of immigrants (n = 7,866, ~81% of the total) among those with recorded data but under-representation when immigrant status was considered as a binary variable (1.3% of the total study population compared to 9.9% non-UK born individuals in the English Census).” This means that country of birth, the main indicator of immigration status, was not recorded at all for 99.3% of individuals in the CRPD database to be used to develop the phenotype. Even if the less reliable term “first language’ is added, the two key terms for determining immigrant status were missing for 98.4% of the individuals in the database . None of the other search terms proposed for the phenotype algorithm could compensate for this vast amount of missing data. While completeness of recording of “ethnicity” was much higher at about 80%, this is not an adequate proxy for “migrant” as many people with a non-White British ethnicity were born in the UK and are thus non-immigrants. Whilst the study by Jain et al. was limited to people over 65, there is no reason to believe the recording of immigrant status in the CRPD database would be any better for younger people.
In the Discussion, the study team acknowledge the risk of missing data, stating “clinician coding may be incomplete resulting in missing data, and therefore, there may be under-recording of migration and the presence of migrants in the comparator population”. What they don’t seem to have appreciated is that over 98% of the relevant data are missing!

Assuming the findings in the paper by Jain et al. are correct, they indicate that the CPRD cannot be used to develop a reliable immigrant phenotype.

Author response 2:
Thank you for highlighting that the figure 1.6% should read 1.3% in the introduction. We have corrected the figure to 1.3% in the introduction, in line with Jain et al’s calculation that approx. 81% of the 1.6% of individuals with immigration status codes (country of birth and language) were international migrants.

We do not agree that Jain et al’s paper indicates that the CPRD cannot be used to develop and test a reliable phenotype. Rather, Jain et al’s study shows that a phenotype can be developed, but that it has not been evaluated in the whole CPRD population, i.e. it did not complete evaluation of the phenotype in under 65 year olds in CPRD GOLD, and did not evaluate the phenotype at all in CPRD Aurum. According to ONS 2011 census estimates, those aged 65 years and over only make up 27% of the migrant population, so Jain et al’s study hasn’t evaluated a migration phenotype in approximately three-quarters of the migrant population. We have outlined why we think it is reasonable to pursue completing a feasibility study of the phenotype across all age ranges in paragraph four of the introduction and, in giving us approval to complete this study, the Independent Scientific Advisory Committee for CPRD agree that it is reasonable to proceed with the feasibility study.

Reviewer comment 3:
From our experience in Scotland, the only reliable source of country of birth data is the Census. It was for this reason that the Scottish Health and Ethnicity Study was made possible through the successful linkage of the Census, with self-reported ethnic group and country of birth, to health and death records.

Author response 3:
We agree that country of birth data collected in the Census is an excellent source of data . However, it is not standard to create bespoke linkages between CPRD and other datasets, and furthermore, such linkages have now been suspended by CPRD during the pandemic. We acknowledge that the inability to undertake bespoke linkage of census data to CPRD records limits the completeness of country of birth data. We are completing this study in recognition of its limitations as it would otherwise not be possible to study primary care outcomes at all in electronic health records for this population. Census data linked to primary care records would be limited by the fact that it is collected every 10 years, so even with a bespoke linkage this will not enable the identification of migration status for large numbers of recent migrants.

Reviewer comment 4:
I thus respectfully invite the authors to reconsider whether either the feasibility study or the main study are viable.

Author response 4:
The feasibility study has been reviewed and approved by the Independent Scientific Advisory Committee for CPRD. This committee reviews the scientific appropriateness of any study before it is conducted. We will not proceed with the main study if our feasibility study demonstrates that the use of the migration phenotype is not viable for the study of migrant health outcomes in CPRD. The feasibility study is being undertaken to determine whether the main study will be feasible or not, but we do not believe any of the points raised by the reviewer suggest that a feasibility study is not viable.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 September 2020

https://doi.org/10.21956/wellcomeopenres.17475.r40474

? **Laurence Lessard-Phillips**
Institute for Research into Superdiversity, University of Birmingham, Birmingham, UK

This is a fascinating research protocol for looking into migrants' healthcare utilisation and mortality, with the aims to develop - and apply- a migration phenotype in two stages: first through a feasilbility study, and then through a main study. The main data source is the CPRD.

This is highly relevant and timely, with the potential to explore a relatively unexplored/limited area of research, often due to the lack of available data.

In order to make the research protocol clearer, I would suggest that the authors consider the following points:

- It is unclear what migration proxies are being criticised; some of those proxies are also used in the data that will be used to assess the phenotype. It would be interesting to hear a bit more about how this will be dealt with and what types of assessment criteria will be used in the feasibility study, as these were not entirely clear to me.

- On the point above: would it be worth looking into getting access to the APS (or LFS), which are used to produce the ONS 'estimates' (with their own source of biases) so that there could be a more detailed comparison?

- Is there a possibility to also consider the different between country of birth and nationality? Could there be instances where individuals are misclassified if nationality/citizenship is not looked into (nationality may also grant specific rights to individuals, regardless of where they, or their parents, were born).

- Out of curiosity (and linked to issues of healthcare charges): will the data risk including visitors as well as migrants?

- Is the comparator population in the main study too heterogenous? This stems from my interest in migrant generations, but I assume that heterogeneity within the non-migration could be established from some of the included covariates.

- It would be good to emphasise the fact that expert input has been used for the development of the phenotype (as it seems to be alluded to toward the end of the protocol).

- It was good to mention that clinician coding could lead to bias. Could another source of bias also be that the data source, as relevant as it might be, may also exclude those who 1) do not use healthcare and/or 2) are being barred from using healthcare? Given the link to the

DOTW/MdM reports, it may be relevant to mention.

- What is the level of certainty that EU 'migrants' will be captured within the CPRD data? Could there be different terms being used? What if they are not? This is probably where expert input will be useful in the assessment of the phenotype.
- The covariates included are important, but is there a danger of associating area-level IMD to individual deprivation/SES? Is there a way to capture this?
- On the IMD: given the longitudinal aspect of the study, would it be worth also using earlier measures (if available)?
- One minor point: is the reference in the third introductory sentence the correct one, as it seems to deal with Scotland when the text refers to a European systematic review... Maybe use the original reference within that article?

Of course, the points above are quite minor, and most are meant to be points to reflect on as the study develops and be dealt with relatively quickly. As mentioned at the start of this review, this study, and the data that it will generate, has great potential for our understanding of migrant health in the UK.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: One of the authors (RW Aldridge) is sitting on the advisory board of a project on which I am PI.

Reviewer Expertise: (Sociology of) migration; migrant and ethnic inequalities; social inclusion; 'migrant integration'; quantitative methods; social science.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 May 2021

Neha Pathak, University College London, London, UK

Thanks very much for your helpful comments. We appreciate the time you have taken to read and provide constructive feedback on our protocol. We have addressed all of your points below, and included your comments alongside our responses in bold italics. We have numbered our responses in order to cross-reference them with the review.

******

Reviewer comment 1:
It is unclear what migration proxies are being criticised; some of those proxies are also used in the data that will be used to assess the phenotype. It would be interesting to hear a bit more about how this will be dealt with and what types of assessment criteria will be used in the feasibility study, as these were not entirely clear to me.
On the point above: would it be worth looking into getting access to the APS (or LFS), which are used to produce the ONS 'estimates' (with their own source of biases) so that there could be a more detailed comparison?

Author response 1:
We have separated out "proxies" from the original sentence to make the examples easier to follow: "In addition, some have used proxy measures of migration which are unable to provide a true estimate of the impact of migration. For example, a study in England using registration with a GP after the age of 15 as a proxy for migration estimated hospital admission rates to be half the rate of the general population."
The assessment criteria for the feasibility study includes completeness and representativeness of the resultant CPRD GOLD migrant cohort, compared to ONS aggregate data. We have chosen not to compare results to the Labour Force Survey because the ONS states "The Labour Force Survey (LFS) is not designed to measure changes in the levels of population or long-term international migration… levels and changes in levels should be used with caution".

Reviewer comment 2:
Is there a possibility to also consider the different between country of birth and nationality? Could there be instances where individuals are misclassified if nationality/citizenship is not looked into (nationality may also grant specific rights to individuals, regardless of where they, or their parents, were born).

Author response 2:
We agree that the conflation of country of birth and nationality through miscoding by clinicians is likely to affect the summary outputs by region of birth provided as part of the feasibility study and any sensitivity analysis conducted using region of birth. However, it is not possible to study nationality using CPRD electronic health record codes as it is difficult to ascertain whether non-UK origin terms definitively indicate nationality. The potential for misclassification bias, and other sources of bias, has therefore been addressed in the Discussion: "Nonetheless, there are some important sources of bias that must be considered when interpreting any results relating to the fact that determining migration status is dependent on clinician coding. First, clinician coding may be incorrect resulting in misclassification bias. Second, clinician coding may be incomplete resulting in missing data, and therefore, there may be under-recording of migration and the presence of migrants in the comparator population. Third, language coding was incentivised between 2008 to 2011

so representativeness may be better during that period and the cohort may be skewed towards non-English speaking migrants (selection bias)."
Discussion of misclassification bias will be expanded on in publications related to this study.

Reviewer comment 3:
Out of curiosity (and linked to issues of healthcare charges): will the data risk including visitors as well as migrants?

Author response 3:
Our migration phenotype is designed to incorporate codes that would include overseas visitors. This is an active choice that we have made to capture the breadth of migrant typologies (available in the Extended data accompanying this protocol) and to fulfil the definition of international migration that we have used, which has been added to the Introduction: "The definition of migrants used in this study will reflect that of the International Organization for Migration, where a migrant is an individual who "moves away from his or her place of usual residence, whether within a country or across an international border, temporarily or permanently, and for a variety of reasons".

Reviewer comment 4:
Is the comparator population in the main study too heterogenous? This stems from my interest in migrant generations, but I assume that heterogeneity within the non-migration could be established from some of the included covariates.

Author response 4:
We recognise that the comparator population will be heterogeneous. Actions we will take have been added to the 'Sensitivity analysis' section: "The distribution of covariates across migrant and non-migrant groups will be assessed. Additional methods to achieve comparability between groups will be considered in sensitivity analysis where the uneven distribution of covariates is likely to introduce significant bias."

Reviewer comment 5:
It would be good to emphasise the fact that expert input has been used for the development of the phenotype (as it seems to be alluded to toward the end of the protocol).

Author response 5:
We agree that expert input is a strength of the development of the phenotype. We have already included a sentence explaining this in the Methods section: "This initial phenotype will then be reviewed and refined by migration health experts and experts in using CPRD from the CALIBER team." To highlight this further, we have included a new sentence in the third paragraph of the "Discussion" section: "Unique strengths of the present study include the methods used to develop the migration phenotype, specifically the involvement of migration health experts and clinicians."

Reviewer comment 6:
It was good to mention that clinician coding could lead to bias. Could another source of bias also be that the data source, as relevant as it might be, may also exclude those who 1) do not use healthcare and/or 2) are being barred from using healthcare? Given the link to the

DOTW/MdM reports, it may be relevant to mention.

Author response 6:
Thank you, we agree that these are important considerations and we have updated the Discussion section to reflect this: “Fourth, this study only captures the healthcare utilisation of migrants who are known to the NHS, rather than those who do not use healthcare or face significant access barriers that prevent them from accessing care. Findings are unlikely to be representative of migrant subgroups like asylum seekers and undocumented migrants and others who are unable to access without fear of being charged for NHS services.”

Reviewer comment 7:
What is the level of certainty that EU 'migrants' will be captured within the CPRD data? Could there be different terms being used? What if they are not? This is probably where expert input will be useful in the assessment of the phenotype.

Author response 7:
The breakdown of migrants by WHO region of birth and ONS Nomis continent of birth as part of feasibility study will provide insight into this.

Reviewer comment 8:
The covariates included are important, but is there a danger of associating area-level IMD to individual deprivation/SES? Is there a way to capture this?

Author response 8:
‘Patient level IMD' is still intended as an area-level measure as it provides the area-level deprivation based on the home postcode of the patient. We will use it as a person-based location reference for determining area-level IMD, which is preferable to using a GP practice-based location reference (i.e. practice-level IMD). CPRD unfortunately does not provide routine linkages to individual deprivation/SES data. We have updated the Discussion section to acknowledge the lack of individual deprivation/SES data as a limitation to the study: “Fifth, CPRD does not provide routine linkages to data on individual-level deprivation, and the study’s use of area-level deprivation does not account for the individual-level measures of socioeconomic position that play a role in the association between migration and healthcare utilisation.”

Reviewer comment 9:
On the IMD: given the longitudinal aspect of the study, would it be worth also using earlier measures (if available)?

Author response 9:
From our preliminary results, the completeness and representativeness of the phenotype is poorer in earlier years of the study. The feasibility study will report on changes to completeness and representativeness over the years, and the main study will likely be limited to later years as completeness and representativeness improve over time. This means that earlier measures of deprivation will not be relevant to the main study.

Reviewer comment 10:

One minor point: is the reference in the third introductory sentence the correct one, as it seems to deal with Scotland when the text refers to a European systematic review... Maybe use the original reference within that article?

Author response 10:
Thank you for making us aware. We have now corrected this citation.

Competing Interests: No competing interests were disclosed.