**Wellcome Open Research**

Wellcome Open Research 2020, 5:156 Last updated: 03 JUL 2020

STUDY PROTOCOL

# 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 1; peer review: awaiting peer review]

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

**v1** **First published:** 03 Jul 2020, **5**:156
https://doi.org/10.12688/wellcomeopenres.15931.1
**Latest published:** 03 Jul 2020, **5**:156
https://doi.org/10.12688/wellcomeopenres.15931.1

**Open Peer Review**

**Reviewer Status** *AWAITING PEER REVIEW*

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

## 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

**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 1; peer review: awaiting peer review]** Wellcome Open Research 2020, **5**:156 https://doi.org/10.12688/wellcomeopenres.15931.1

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

## 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]. Existing studies in the UK are limited to hospital admissions care, use proxy measures of migration and present mixed results. One study in Scotland showed that people of South Asian ethnicities, including migrants, had higher rates of avoidable hospital admissions compared to the white Scottish population[2] whilst another 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[3].

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[4]. 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[5]. 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[6]. 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), the largest UK primary care EHR, described phenotypes for social factors amongst older individuals including migration status[8]. The study estimated that 1.6% of individuals aged ≥ 65 years in CPRD were international migrants. 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 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.

## 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[9,10].

### 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[11]. 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[12]. 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[13].

***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[12].

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[14].

***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[10]. 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)[10]. 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[9,16]. 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[17].

*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[4]. 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].

***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 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[8] or the updated sample size calculation means that the study does not have the level of

**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. |

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 sub-grouped 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[18] 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[19] and costs of medications from the British

**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[21]. |
| Amenable conditions | Conditions where hospital admissions or death could be avoided through high- quality preventative healthcare[22]. |
| 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[22]. |
| 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[22]. |
| 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[23]: 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[18]. |

National Formulary[20]. 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[19].

*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.*[24].

## 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[4]. 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 are the inclusion of primary care data and the fact that there is no restriction on country of birth or visa types for inclusion. 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)[25]. 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.

## 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

### 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. 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
3. 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
4. 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
5. Plan NLT: **Online version of the NHS Long Term Plan**. NHS Long Term Plan. 2019; [cited 2020 Mar 10].
   Reference Source
6. 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
7. 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
8. 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
9. 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
10. 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
11. **Clinical Practice Research Datalink CPRD**. [cited 2020 Mar 10].
    Reference Source
12. 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
13. 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
14. Census: **LC2103EW - Country of birth by sex by age**. Data Viewer - Nomis - Official Labour Market Statistics. 2011; [cited 2019 May 20].
    Reference Source
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. **CPRD linked data**. CPRD. [cited 2020 Mar 10].
    Reference Source
17. **NHS Digital** [cited 2020 Mar 10].
    Reference Source
18. 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
19. **Reference costs.** NHS Improvement. [cited 2019 Feb 21].
    Reference Source
20. BNF: **British National Formulary - NICE**. [cited 2019 Feb 21].
    Reference Source
21. **2.6 Unplanned hospitalisation for chronic ambulatory care sensitive conditions**. NHS Digital. [cited 2019 Feb 21].
    Reference Source
22. **Avoidable mortality in the UK QMI** - Office for National Statistics. [cited 2019 Feb 21].
    Reference Source
23. Starrs AM, Ezeh AC, Barker G, *et al.*: **Accelerate progress—sexual and reproductive health and rights for all: report of the Guttmacher–Lancet Commission.** *The Lancet.* 2018; [cited 2018 May 27].
    Reference Source
24. 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
25. British Medical Association and NHS employers: **Clinical directed enhanced services (DES) guidance for GMS contract 2008/2009**. 2008; [cited 2020 Mar 10].
    Reference Source