International Journal of Population Data Science

Journal Website: www.ijpds.org

Linking migration and hospital data in England: linkage process and evaluation of bias

Rachel Burns[1,*], Sacha Wyke[2], Yamina Boukari[1], Sirinivasa Vittal Katikireddi[3], Dominik Zenner[4,5], Ines Campos-Matos[2,6], Katie Harron[7], and Robert W. Aldridge[1]

Submission History

Submitted:	07/07/2023
Accepted:	14/11/2023
Published:	12/02/2024

[1]Centre for Public Health Data Science, Institute of Health Informatics, University College London, 222 Euston Road, London NW1 2DA, United Kingdom
[2]UK Health Security Agency, 61 Colindale Ave, London NW9 5EQ United Kingdom
[3]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Berkeley Square, 99 Berkeley Street, Glasgow, G3 7HR, United Kingdom
[4]Global Public Health Unit, Wolfson Institute of Population Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Yvonne Carter Building, 58 Turner Street, London E1 2AB, United Kingdom
[5]Infection and Population Health Department, Institute of Global Health, University College London
[6]Office for Health Improvement and Disparities, Department of Health841 and Social Care, 39 Victoria Street, London SW1H 0EU, United Kingdom
[7]UCL Great Ormond Street, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, United Kingdom

Abstract

Introduction
Difficulties ascertaining migrant status in national data sources such as hospital records have limited large-scale evaluation of migrant healthcare needs in many countries, including England. Linkage of immigration data for migrants and refugees, with National Health Service (NHS) hospital care data enables research into the relationship between migration and health for a large cohort of international migrants.

Objectives
We aimed to describe the linkage process and compare linkage rates between migrant sub-groups to evaluate for potential bias for data on non-EU migrants and resettled refugees linked to Hospital Episode Statistics (HES) in England.

Methods
We used stepwise deterministic linkage to match records from migrants and refugees to a unique healthcare identifier indicating interaction with the NHS (linkage stage 1 to NHS Personal Demographic Services, PDS), and then to hospital records (linkage stage 2 to HES). We calculated linkage rates and compared linked and unlinked migrant characteristics for each linkage stage.

Results
Of the 1,799,307 unique migrant records, 1,134,007 (63%) linked to PDS and 451,689 (25%) linked to at least one hospital record between 01/01/2005 and 23/03/2020. Individuals on work, student, or working holiday visas were less likely to link to a hospital record than those on settlement and dependent visas and refugees. Migrants from the Middle East and North Africa and South Asia were four times more likely to link to at least one hospital record, compared to those from East Asia and the Pacific. Differences in age, sex, visa type, and region of origin between linked and unlinked samples were small to moderate.

Conclusion
This linked dataset represents a unique opportunity to explore healthcare use in migrants. However, lower linkage rates disproportionately affected individuals on shorter-term visas so future studies of these groups may be more biased as a result. Increasing the quality and completeness of identifiers recorded in administrative data could improve data linkage quality.

Keywords
record linkage; data linkage; administrative data; hospital records; migrant; refugee

*Corresponding Author:
Email Address: r.burns@ucl.ac.uk (Rachel Burns)

https://doi.org/10.23889/ijpds.v9i1.2181

Introduction

Despite migrants comprising almost 17% of people living in England in 2021, large-scale evaluation of their access to and use of healthcare has been limited [1, 2]. Previous research suggests that healthcare utilisation may vary by country of origin, reason for migration, and length of stay in England [3, 4]. However, this information is not routinely collected in administrative data sources such as hospital records, leading to challenges in the identification of migrants and the assessment of their healthcare needs. Linking administrative data sources could become an important tool to help harness existing data to uncover migrant health inequalities and inform service provision and policy.

Routinely collected administrative data have already been used in some countries to study the relationship between migration and health. Countries such as Sweden and Norway have unique identifiers shared across datasets that can ascertain migration status within routine datasets through exact linkage [5, 6]. These countries and others such as Canada also have migration registers that document entry into the country. For example, Canada's federal database with all applications for permanent immigration has been used to identify migrants in provincial-level routine health data to explore a range of outcomes including maternal health and migrant mortality [7–9]. Another possible migration specific data source are pre-migration health assessments coordinated by the International Organization for Migration (IOM). Since England lacks both a shared unique identifier and migration-specific registries, linkage of its pre-migration health assessment datasets to health data offers a unique opportunity to quantify population-based migrant health patterns.

Pre-migration health assessments have been coordinated by the International Organization for Migration (IOM) for member states over the last two decades. IOM has coordinated the health checks for refugees prior to their arrival to a host country such as the UK. Additionally, IOM organises the pre-entry tuberculosis (TB) screening programme for visa applicants from high-incidence TB countries migrating to the UK. These datasets contain personal identifiers that would allow for the accurate identification of a large cohort of resettled refugees and migrants within administrative data in England. To date, the pre-entry TB screening dataset has been linked to England's TB registry data and to National Health Service (NHS) numbers, the unique health identifier held in the NHS Personal Demographic Service (PDS), but has not been used to assess healthcare utilisation more broadly [10, 11].

In this study, we describe the methods used to link the refugee health check and pre-entry TB screening datasets to hospital records in England. We compare linkage rates between different migrant sub-groups to explore the potential for bias and the implications of this for epidemiological analyses. Bias in linked data can arise from inaccurate registration and recording of individuals, imperfect identifiers that can lead to linkage errors, and cleaning and coding decisions [12]. Moreover, changes in data collection and policy could impact the recording of individuals, especially given the large period over which administrative datasets are collected.

Comparing characteristics of linked and unlinked groups is important to identify which groups are most likely to be affected by bias. For example, research in the UK has shown that individuals with foreign name structures are less likely to link than others [13]. This can result in missed matches, where records do not link due to a lack of complete or discriminative identifiers, and could lead to an underestimation of the true health needs of certain groups [12, 14]. Evidence on the characteristics of linked and unlinked groups can be used by researchers to consider linkage bias in future analyses using the linked data [15].

The two objectives of this study were: first, to describe the linkage process of the migrant pre-entry TB screening and the refugee health check datasets to hospital data for migrants and refugees in England and second, to evaluate which migrant sub-groups were less likely to link and therefore understand the representativeness of the linked cohort [16]. Our goal was to establish the Million Migrant cohort and enable large-scale longitudinal research on migrants' hospital-based healthcare utilisation.

Methods

Study design and population

The study population consisted of non-European Union (EU) migrants undergoing a pre-entry TB screening between 1 January 2005 and 31 December 2020 as part of the UK visa application process and refugees enrolled in a UK refugee resettlement programme and receiving a pre-arrival health check between 1 January 2013 and 31 December 2020.

Data sources

Non-EU migrant pre-entry TB screening data (migrant cohort) - As part of the UK visa application process, individuals from countries where tuberculosis incidence is high (>40 cases per 100,000 people or more) and who are planning to come and live in the UK for more than 6 months receive a pre-entry TB screening by IOM or by international clinics recognised by the UK Home Office and quality assessed by United Kingdom Health Security Agency (UKHSA) (see Supplementary Table 1.1 in Appendix 1 for list of countries). The migrant cohort includes records of the international non-EU migrants who were screened by the UK pre-entry TB screening programme between 1 January 2005 and 31 December 2020.

Refugee health check data (refugee cohort) - The refugee cohort includes records of all refugees enrolled in UK refugee resettlement programmes who had a completed pre-arrival health assessment conducted by IOM between 1 January 2013 and 31 December 2020.

Personal Demographic Service (PDS) - PDS is the national database of patient demographic information that can be used to facilitate linkage of health datasets. It contains the most up to date identifying information for individuals who interact with primary or secondary care services in England, Wales, and the Isle of Man. This database is held by NHS England [17].

Hospital Episode Statistics (HES) - HES contains patient and clinical information for all inpatient (including day cases), outpatient (including all hospital-based speciality clinics) and accident and emergency episodes in England [18]. Diagnoses associated with admission are coded using ICD-10. HES is

linked with Office of National Statistics mortality records and includes all reported deaths in England since 1937. HES data from 1 January 2005 through 23 March 2020 were included in this study. The HESID, a pseudonymised number used to uniquely identify each patient record, was used to identify all HES records associated with each patient.

The data sources are described in more detail in the Supplementary Tables 2.1 and 3.1 in Appendices 2, 3.

Data flows

A two-stage linkage process was used to link the non-EU migrant pre-entry TB screening and the refugee health check datasets to HES (Figure 1). Stage 1 combined migrant and refugee datasets to form the Million Migrant cohort and linked the Million Migrant cohort to the NHS Personal Demographic Service (PDS) in order to obtain an NHS number and postcode. Stage 2 linked Million Migrant-PDS linked data to HES.

The migrant and refugee datasets from IOM were sent to the UK Health Security Agency (UKHSA, previously PHE, Public Health England) Data Lake, a protective storage centre holding large amounts of data. In step A, data were deduplicated (e.g. removal of exact duplicate records) and individuals with any missing personal identifiers (forename, surname, sex and date of birth) were removed as a prerequisite for linkage through the Demographic Batch Service. The two datasets were then combined into the Million Migrant cohort, containing information on personal identifiers, visa category, region of origin, and date of health check or screening. UKHSA researchers then sent the Million Migrant linkage file containing identifiers to the UKHSA's Demographics Batch Service to link to PDS and acquire NHS number and UK postcode (step B). The Demographic Batch Service is an offline service used by UKHSA that allows for exact matches only being returned [19]. Next, UKHSA researchers used NHS number, date of birth, sex, and postcode to link the Million Migrant-PDS linkage file to HES data and obtain the pseudonymised HESID (step C). Finally, the anonymised Million Migrant cohort was merged with the extract of HES within the UKHSA Data Lake, using the pseudonymised HESID (step D).

Linkage

Linkage process

To link the Million Migrant cohort and HES, multi-step deterministic record linkage was employed in a two-stage process using available personal identifiers (Table 1). The data linkage was done on Microsoft SQL Server Management Studio v18.1 and analysis on R Studio 4.1.2.

Stage 1: Linking the million migrant cohort to the PDS

The refugee and migrant cohorts were deduplicated and individuals with any missing personal identifiers were removed. For the refugee cohort, when an individual had more than one refugee health check, only the most recent record was retained for linkage as it was assumed that the individual would have entered the UK for resettlement on that date and remained until the end of the study period. For the migrant cohort, individuals screened for TB may have multiple records on the IOM screening database, due to multiple visa applications each requiring a new TB screening or multiple screenings resultant from a positive TB test. Someone with a positive TB test would need to undergo TB treatment and then repeat screening(s) for visa clearance. Where migrants had more than one pre-entry screening record, only the first valid record was retained for the linkage. The first valid record was defined as the first recorded TB screening date with at least 6 months in between it and the second recorded screening date (e.g., 01/01/2013 and 07/01/2013). This amount of time was used as 6 months is the minimum visa length requiring an individual to undergo a TB screening.

The refugee and migrant cohorts were then combined into the Million Migrant cohort and linkage variables cleaned (for example, through the removal of erroneous punctuation and the transformation of accented forenames and surnames into their base unaccented version). The Million Migrant cohort was deterministically linked by UKHSA's Demographic Batch Service to the PDS to obtain NHS number and postcode. Only exact matches between all linkage variables were able to be retrieved due to the infrastructure of the Demographic Batch Service, limiting the linkage to a deterministic approach. To maximise match levels, several tracing rounds were attempted with modifications to specific linkage variables to account for discrepancies in name structure or errors in recorded date of birth (Table 2).

Stage 2: Linking the million migrant-PDS linked data to HES data

The Million Migrant-PDS linked data, enriched with NHS numbers and postcodes, were linked to HES following the stepwise deterministic matching procedure developed by NHS Digital (Table 3). NHS number, date of birth, sex, and postcode were used as the linkage variables to match records. Match rank corresponded to the step at which the match was determined and represents the quality of the matching. A lower value match rank indicated a match based on a greater number of restrictions and was therefore considered to be stronger evidence of a true positive match. Individuals in the Million Migrant-PDS linked data who linked to the same HESID or had linked to multiple different HESIDs were excluded as the researchers were unable to determine the correct HESID from the available information.

Evaluation of potential bias through linkage error

Migrant characteristics were compared in the linked and unlinked cohorts at each step for each linkage stage across age, sex, visa type, region of origin, and length of time in England (estimated as years since pre-entry screening or health check). The entire Million Migrant cohort was used as the inception cohort but we did not expect all migrants to link to either PDS or HES for a number of reasons. The individual might never have arrived in England, may have migrated to

Figure 1: Data flow and linkage process for the refugee health check and the non-EU migrant TB screening databases, NHS personal demographic service, and hospital episode statistics

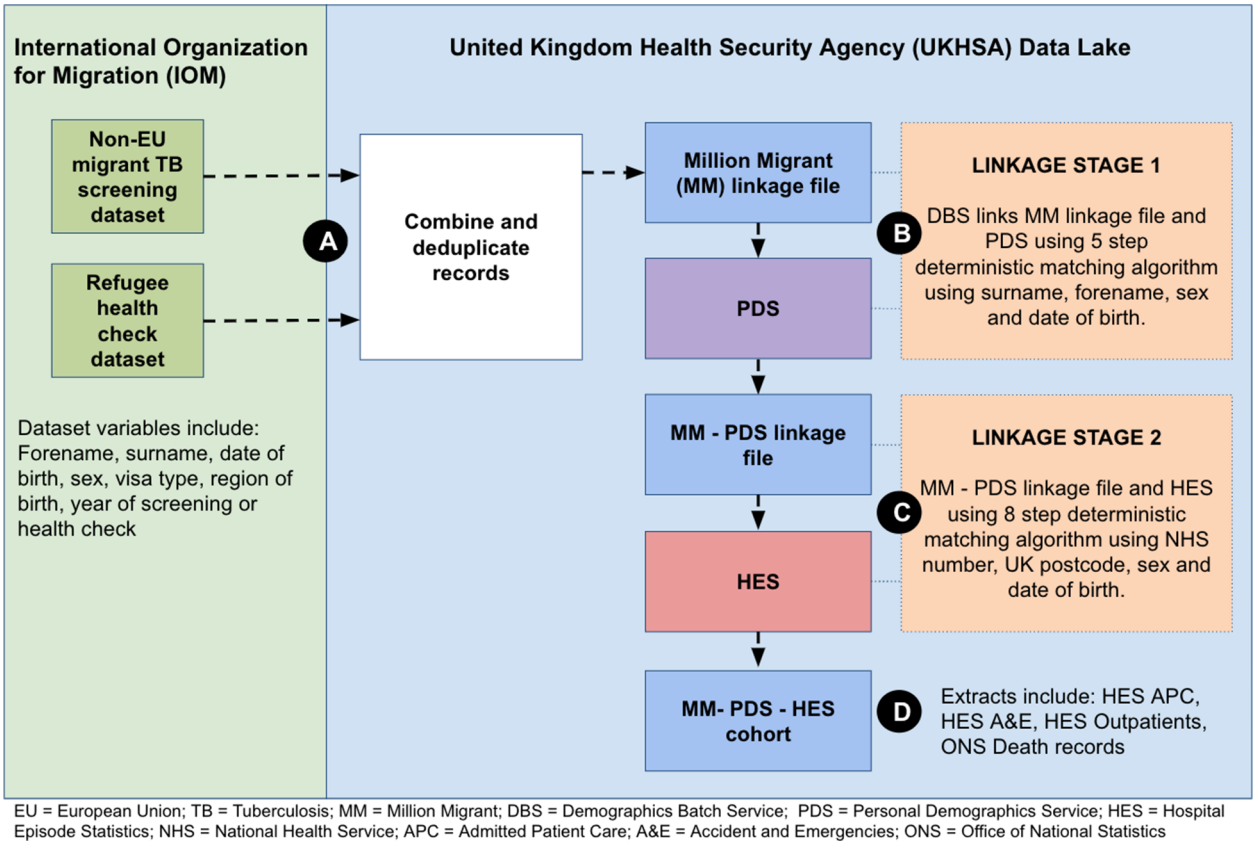


Table 1: Availability of personal identifiers in the refugee cohort, migrant cohort, PDS, and HES

Linkage identifiers	Data sources			
	Million Migrant cohort		PDS	HES
	Migrant	Refugee		
Forename(s)	√	√	√	
Surname(s)	√	√	√	
Date of birth	√	√	√	√
Sex	√	√	√	√
Postcode			√	√
NHS number			√	√

Scotland, Northern Ireland, or Wales (thus be unidentifiable in PDS or HES), or may not have registered with or utilised the NHS.

Because the second linkage stage depended upon successful linkage in the first stage (i.e., NHS numbers and postcodes acquired from linkage stage 1 were used as personal identifiers in linkage stage 2), the linkage rates were estimated for both the entire Million Migrant cohort and the Million Migrant-PDS linked cohort (i.e., the Million Migrant cohort with NHS numbers linked after stage 1). Overall linkage rates for HES were calculated as the percentage of individuals in the Million Migrant cohort and Million Migrant-PDS cohort who linked to any HES record. To evaluate potential bias arising from missed matches, characteristics of HES linked and unlinked individuals were compared for those in the Million Migrant cohort and in the Million Migrant-PDS cohort.

Standardised differences (mean difference in standard deviation units) when comparing linked to unlinked HES records were used to detect potential biases according to the following variables: age, sex, visa type, region of origin, and length of time in England. We used standardised differences as these are thought to be more informative when comparing linked and unlinked records than P-values in large samples. Standardised mean differences of 0.2 or less were considered as small, greater than 0.2 to 0.5 as moderate, and greater than 0.5 to over 0.8 as large, as previously defined in the literature [20]. Assessing the standardised mean differences can help identify variables that may be more affected by linkage error and are therefore potential sources of bias (See Appendix 4 for more details).

To identify characteristics that were predictive of linkage in HES, three separate multivariable logistic regression models

Table 2: Modifications to linkage variables in the Million Migrant cohort by round in exact deterministic linkage to PDS

Round	Forename	Surname	Sex	Date of birth
1	Accented characters transformed into base unaccented version; hyphens and apostrophes replaced with a wildcard (*)	Accented characters transformed into base unaccented version; hyphens and apostrophes replaced with a wildcard (*)	Exact	Exact
2	Unaccented characters; hyphens and apostrophes replaced with a wildcard (*) Placed a wildcard (*) after the first forename (e.g., Anna* if Anna Marie).	Unaccented characters were retained; hyphens and apostrophes replaced with a wildcard (*) If more than one surname, placed a wildcard (*) in front of the last surname (e.g., *Smith if Roberts Smith).	Exact	Exact
3	Placed a wildcard (*) after all forenames (e.g., Anna* if Anna Marie or Anna).	Unaccented characters were retained	Exact	Exact
4	Unaccented characters; hyphens and apostrophes replaced with a wildcard (*) Placed a wildcard (*) after all forenames (e.g., Anna* if Anna Marie or Anna).	Unaccented characters; hyphens and apostrophes replaced with a wildcard (*)	Exact	Exact
5	Unaccented characters	Unaccented characters	Exact	Exchanged month and day

Table 3: NHS Digital stepwise deterministic matching algorithm according to detail of identifying variables

Match rank	NHS Number	Date of birth	Sex	Postcode
1	Exact	Exact	Exact	Exact
2	Exact	Exact	Exact	
3	Exact	Partial	Exact	Exact
4	Exact	Partial	Exact	
5	Exact			Exact
6*		Exact	Exact	Exact
7**		Exact	Exact	Exact
8	Exact			

*Where NHS number does not contradict the match and date of birth is not 1 January and the postcode is in the 'ignore' list such as postcodes for communal establishments such as hospitals, care homes, prisons and boarding schools.
**Where NHS number does not contradict the match and the date of birth is not 1 January.

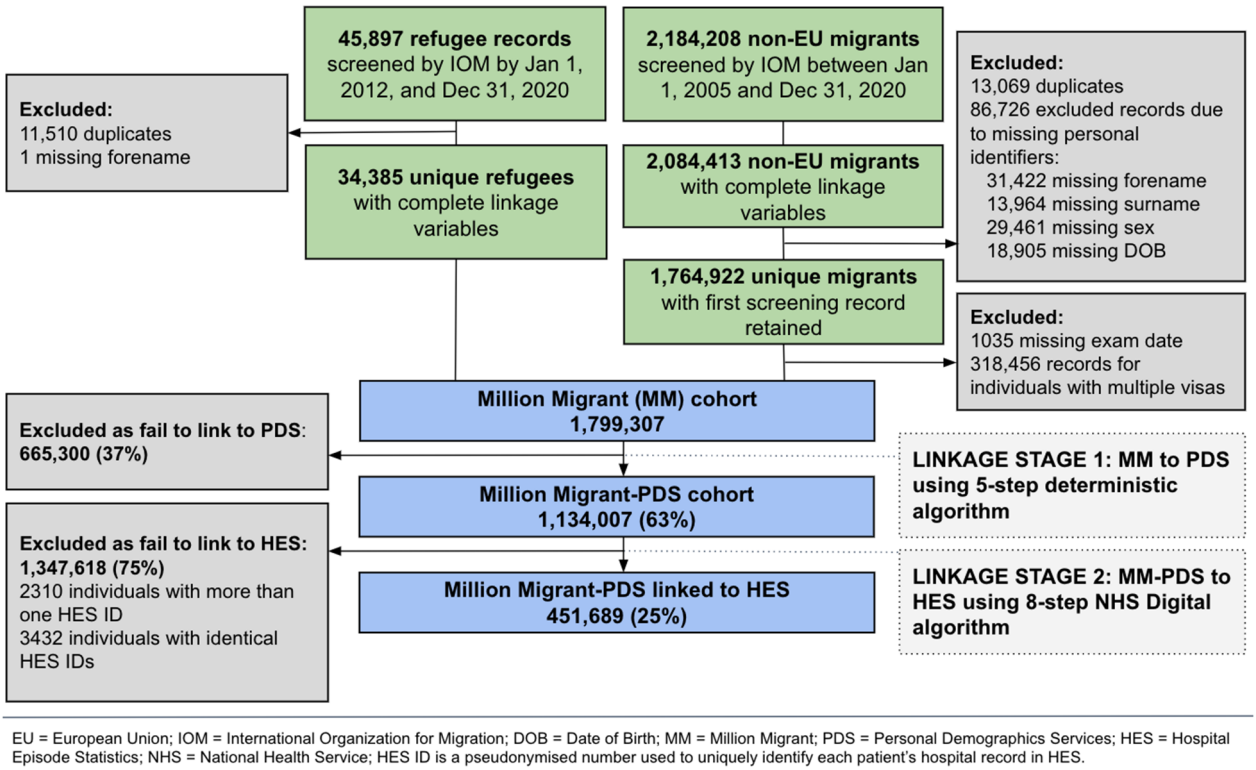
were used to evaluate linkage with HES for both the Million Migrant cohort and Million Migrant-PDS cohort controlling for age and sex, using the following variables as the exposure: visa type, region of origin, and length of time in England. Given the importance of age and sex on healthcare utilisation and hospitalisation, these two variables were used as controls when calculating the relative risk for each exposure separately (See Appendix 4 for more details on statistical analyses).

Higher linkage rates were expected for both PDS and HES for individuals arriving on more permanent visas such as settlement and dependent, family, and refugee and for individuals with a longer length of time since migration to England (i.e., years since pre-entry TB screening or refugee health check). We assumed that individuals on more permanent visas and longer term migrants would have a higher likelihood of interaction with primary or secondary care (and would therefore be more likely to link to PDS). Furthermore, we expected higher linkage rates to PDS in migrants who arrived after the introduction of the NHS Immigration Health Surcharge in 2015 due to the immediate allocation of NHS numbers by the Home Office upon payment of the fee. However, this study cannot distinguish between higher linkage rates and higher NHS contact rates. Higher linkage rates could be due to higher healthcare utilisation and/or better recording of identifiers.

Results

The Million Migrant cohort consisted of 34,385 (1.9%) resettled refugees and 1,764,922 (98.1%) non-EU migrants, totalling 1,799,307 unique individuals. Within the Million Migrant cohort, 1,134,007 (63.0%) linked to PDS allowing ascertainment of NHS number and UK postcode (linkage stage 1) and 451,689 (25.0%) linked to a HES record (linkage stage 2) (Figure 2; Supplementary Table 5.1 in Appendix 5). Overall, 40% of migrants with an NHS number linked to a HES record, indicating that they had interacted with hospital services at

Figure 2: Results of linkage at stage 1 (Million Migrant to PDS) and stage 2 (Million Migrant-PDS to HES) and final linkage rates



some point during the study period (Supplementary Table 6.5 in Appendix 6).

Distribution of migrant characteristics in linked records

Linkage stage 1: Million migrant to PDS

At linkage stage 1, 880,344 (78.0%) of individuals who linked to PDS did so at the first step of the 5-step deterministic method (Table 2), i.e., exact linkage by sex, date of birth, and unaccented forename and surname with hyphens and apostrophes replaced with a wildcard to allow for flexibility in their recording (Supplementary Table 5.2. and 5.3). The second step linked an additional 191,352 (17.0%) individuals to PDS, where a wildcard was placed after the first forename and before the last surname to account for variations in recorded names. Differences appeared by visa type, region of origin, and length of time in England. Almost 1 in 4 refugees linked at step 3 (25.4% of refugees versus less than 3% for other visa categories), where a wildcard was placed after all first forenames and hyphens and apostrophes were retained. Evaluation by region of origin showed that additional steps in this methodology captured a greater percentage of individuals from South Asia (27.5%) in step 2 and individuals from Latin America and the Caribbean (48.8%) and Middle East and North Africa (20.9%) in step 3.

Linkage stage 2: Million migrant-PDS to HES

Linkage at stage 2, from Million Migrant-PDS to HES using the NHS Digital 8-step algorithm, demonstrated a similar pattern to linkage stage 1 with 310,118 (69.0%) linking at step 1 (exact NHS number, date of birth, sex, and postcode) and 117,328 (26.0%) at step 2 (exact NHS number, date of birth, and sex) (Supplementary Table 5.4 and 5.5). Younger migrants (aged 0-34) were more likely to link at step 2 than older migrants (aged 50+). Individuals from East Asia and the Pacific and Europe and Central Asia were disproportionately linked at steps 2-5 compared to those from the Middle East and North Africa, South Asia, and Sub-Saharan Africa. Evaluation by visa type showed a greater proportion of individuals on more permanent visas such as family reunion, refugee, and settlement and dependent visas linked at step 1. Conversely, students and workers (individuals on work and working holiday visas) were more likely to link at step 2 where the postcode was omitted (30.3% of students and 32.8% of work visa holders compared to 16.3% of refugees).

Linkage rates by demographic characteristics of migrants

In linkage stage 1, linkage rates varied according to all demographic characteristics (Figure 3; Supplementary Table 6.1 and 6.2 in Appendix 6). Linkage rates to PDS were higher for younger (<35 years) migrants (at least 64.5%) compared to older (50+ years) migrants (from 42.4%) and for female (67.7%) migrants than male (58.4%) migrants. Migrants from Europe and Central Asia (74.2%) and North America (78.9%) had higher linkage rates than those from the Middle East and North Africa (59.2%) and South Asia (59.1%). Individuals on family (66.9%), settlement and dependent (70.1%), and work (69.8%) visas had higher linkage rates than those on student (56.8%) and working holiday (47.9%) visas. Migrants who migrated less than 5 years from the end of the study

Figure 3: Percent of individuals who were linked to PDS by age, sex, visa type, region of origin, and length of time in England

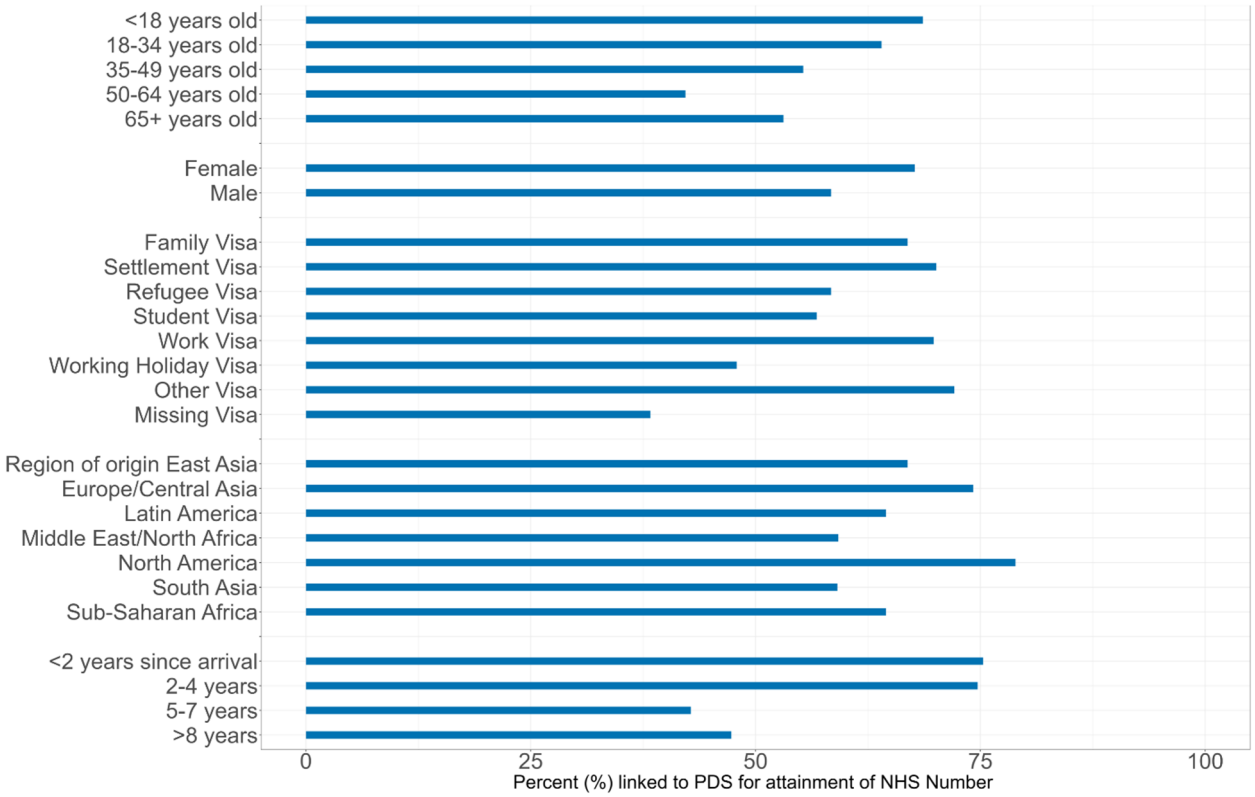
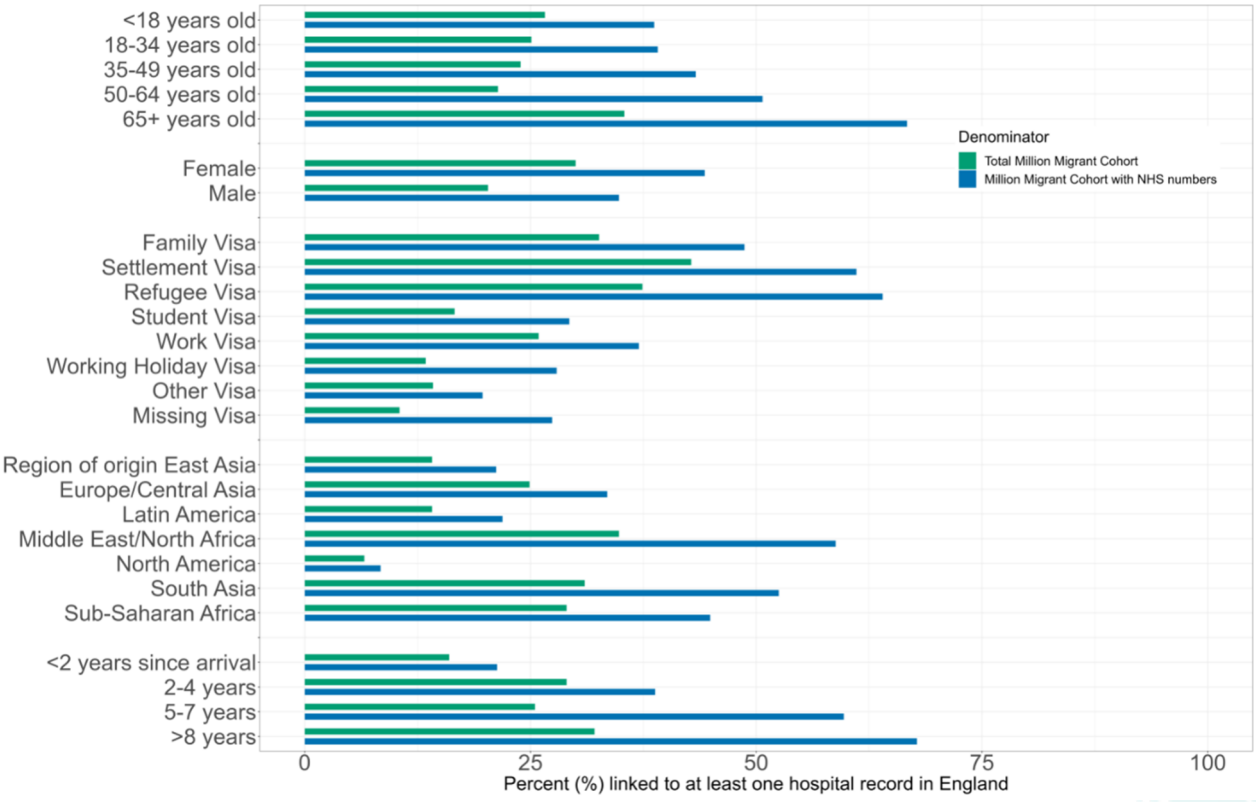


Figure 4: Percent of individuals who were linked to HES out of the total Million Migrant cohort (green) and the Million Migrant-PDS cohort (blue) by age, sex, visa type, region of origin, and length of time in England



period in 2020 were almost 2 times more likely to link than those who had been in England for more than 5 years (74.7% versus 42.8%).

For linkage stage two, HES linkages patterns were relatively similar by sex, visa type, and region of origin for both denominators, the total Million Migrant cohort and the Million

Table 4: Sociodemographic characteristics of the migrant sample from the Million Migrant cohort linked and unlinked to Hospital Episode Statistics (row percentages) and standardised mean differences

Variable	Overall* N = 1,799,307 (100%)	Linked N = 451,689 (25%)	Unlinked N = 1,347,618 (75%)	SMD	95% CI
Sex				0.26	0.26, 0.26
Female	886,791 (49.3)	266,309 (30.0)	620,482 (70.0)		
Male	912,516 (50.7)	185,380 (20.3)	727,136 (79.7)		
Age Group				0.06	0.06, 0.07
0 to 17	235,577 (13.1)	62,598 (26.6)	172,979 (73.4)		
18 to 34	1,297,617 (72.1)	325,118 (25.1)	972,499 (74.9)		
35 to 49	213,349 (11.9)	51,086 (23.9)	162,263 (76.1)		
50 to 64	41,282 (2.3)	8,821 (21.4)	32,461 (78.6)		
65+	11,458 (0.6)	4,061 (35.4)	7,397 (64.6)		
Missing	24 (0.0)	5 (20.8)	19 (79.2)		
Region of origin				0.44	0.43, 0.44
East Asia & Pacific	582,906 (32.4)	82,467 (14.1)	500,439 (85.9)		
Europe & Central Asia	49,012 (2.7)	12,196 (24.9)	36,816 (75.1)		
Latin America & Caribbean	948 (0.1)	134 (14.1)	814 (85.9)		
Middle East & North Africa	45,300 (2.5)	15,768 (34.8)	29,532 (65.2)		
North America	437 (0.0)	29 (6.6)	408 (93.4)		
South Asia	790,340 (43.9)	245,384 (31.0)	544,956 (69.0)		
Sub-Saharan Africa	330,303 (18.4)	95,688 (29.0)	234,615 (71.0)		
Missing	61 (0.0)	23 (37.7)	38 (62.3)		
Visa Type				0.63	0.62, 0.63
Family Reunion	94,209 (5.2)	30,691 (32.6)	63,518 (67.4)		
Settlement and Dependents	467,571 (26.0)	199,976 (42.8)	267,595 (57.2)		
Refugee	33,978 (1.9)	12,692 (37.4)	21,286 (62.6)		
Students	828,790 (46.1)	137,827 (16.6)	690,963 (83.4)		
Work	155,635 (8.6)	40,259 (25.9)	115,376 (74.1)		
Working Holiday Maker	35,915 (2.0)	4,803 (13.4)	31,112 (86.6)		
Other	167,918 (9.3)	23,836 (14.2)	144,082 (85.8)		
Missing	15,291 (0.8)	1,605 (10.5)	13,686 (89.5)		
Length of time in England (in years)				0.36	0.36, 0.37
<2	572,985 (31.8)	91,826 (16.0)	481,159 (84.0)		
2 to 4	494,524 (27.5)	143,285 (29.0)	351,239 (71.0)		
5 to 7	276,943 (15.4)	70,701 (25.5)	206,242 (74.5)		
>8	454,855 (25.3)	145,877 (32.1)	308,978 (67.9)		

SMD = Standardised Mean Difference; *Overall percentages are column percentages.

Migrant-PDS cohort (Figure 4; Supplementary Tables 6.3, 6.4 and 6.5 in Appendix 6). More females (59.0%) than males (41.0%) linked to HES. For the total Million Migrant cohort, individuals entering on more permanent visas were more likely to link than those on more temporary visas: a third of settlement and dependent (32.6%), family (42.8%), and refugee (37.4%) visa holders linked to HES compared to students (16.6%), workers (25.9%), working holiday (13.4%) and other (14.2%). When restricting to Million Migrant-PDS cohort, linkage rates for almost all visa types double. A higher proportion of individuals in the total Million Migrant cohort from the Middle East and North Africa (34.8%), South Asia (31.0%) and Sub-Saharan Africa (29.0%) linked to HES compared to those from East Asia and the Pacific (14.1%) and Europe and Central Asia (23.9%). This pattern was consistent within the Million Migrant-PDS cohort.

With the total Million Migrant cohort, children and older migrants were more likely to link compared to other age groups, with 26.6% of those under 18 and 35.4% of those over 65 years old linking to HES. When limiting to just the Million Migrant-PDS cohort, HES linkage rates instead linearly increase by age group. HES linkage rates were higher for individuals who had been in England longer within the total Million Migrant cohort (32.4% for individuals with more than eight years versus 15.2% for those with less than two years in England). These differences are more pronounced within the Million Migrant-PDS cohort, with 67.8% of migrants with more than 8 years in England linked compared to 21.3% of newly arrived migrants.

Comparing characteristics of linked and unlinked migrants

Differences were detected in the distribution of demographic characteristics of HES linked and unlinked migrants. In the Million Migrant cohort, the largest standardised mean

Table 5: Adjusted odds ratios for a link between Million Migrant cohort and Hospital Episode Statistic records for visa type, region of origin, and length of time in England controlling for age and sex

Variable*	Total	aOR	95% CI
Visa type			
Family Reunion	94,240	—	—
Settlement and Dependents	467,492	1.58	1.55, 1.60
Refugee	33,913	1.5	1.46, 1.54
Students	828,862	0.39	0.38, 0.39
Work	155,616	0.73	0.72, 0.75
Working Holiday Maker	35,926	0.31	0.30, 0.32
Other	167,893	0.32	0.31, 0.33
Missing	15,280	0.26	0.25, 0.28
Region of origin			
East Asia & Pacific	582,896	—	—
Europe & Central Asia	49,010	2.11	2.07, 2.16
Latin America & Caribbean	947	1.06	0.88, 1.27
Middle East & North Africa	45,300	3.62	3.54, 3.70
North America	437	0.46	0.31, 0.66
South Asia	790,327	3.33	3.30, 3.36
Sub-Saharan Africa	330,305	2.75	2.72, 2.78
Length of time in England (in years)			
<2	539,438	—	—
2 to 4	499,624	2.26	2.24, 2.28
5 to 7	294,064	1.95	1.93, 1.97
>8	466,096	3.21	3.18, 3.24

aOR = adjusted odds ratio. CI = confidence interval. *Note: Visa type, region of origin, and length of time in England were all separate exposures in separate logistic regression models.

differences were seen by region of origin and visa type (Table 4). For the Million Migrant-PDS cohort, the differences increased for region of origin (0.63; 0.63,0.63) and length of time in England (0.81; 0.81,0.82) (Supplementary Table 6.4 and 6.5).

Evaluation of linkage from million migrant to HES

Consistent with the linkage rate estimates, we found differences in the odds of linkage to HES across visa type, region of origin, and length of time in England when controlling for age and sex (Table 5). Individuals on more permanent visas such as settlement and dependent and refugee visas had higher odds of linkage to a hospital record. Compared to those on family visas, individuals on student, working holiday and other visas had less than half the odds of linkage to a hospital record. Individuals from the Middle East and North Africa and South Asia were more than three times more likely to link to a hospital record than individuals from East Asia, followed by those from Sub-Saharan Africa and Europe and Central Asia. Individuals who had been in England more than eight years had three times the odds of linkage than those who had been here less than two years. Similar but more pronounced trends were seen when restricting the cohort to the Million Migrant-PDS cohort (Supplementary Table 6.6). Specifically, refugees had 40% higher odds of HES linkage than those on settlement and dependent visas and longer term migrants (arrived more than eight years ago) had almost 10 times the rate of hospital linkage than those who had arrived less than two years ago.

Discussion

This study is the first to link administrative records from pre-migration health screening and hospital data for non-EU migrants and resettled refugees in England. In this two-stage linkage process, we evaluated two deterministic algorithms and found that linkage to a unique health identifier (stage 1) improved over the 15-year study period and older migrants and migrants arriving on more permanent visas were more likely to link to a hospital record (stage 2).

Key findings

Over two-thirds of the Million Migrant cohort were linked to PDS for attainment of NHS numbers, and a quarter linked to HES. Almost half of migrants with NHS numbers linked to at least one hospital record. Multi-stage linkage methods allowed for the relaxation of record requirements in each dataset to agree on all personal identifiers. This was particularly important for individuals with multiple forenames and surnames or several postcodes that could be recorded differently across datasets. We identified important variability by key sociodemographic characteristics between linked and unlinked migrants. More recent migrants (>5 years residing in England) were more likely to link to PDS than migrants

who had their screening or health check more than five years before the end of the study period. Conversely, linkage rates to hospital records increased with length of stay in England. Similarly, women, children, and older migrants were more likely to link to a hospital record. Individuals on more permanent visas such as settlement and dependent, family, and refugee visas were more likely to link to hospital records.

Interpretation and contextualisation of findings

There are several explanations for why an individual may not have been successfully linked in either PDS or HES. First, we could not accurately quantify the denominator of individuals eligible for linkage with HES. The total denominator included varying proportions of individuals who never arrived in England (e.g., someone who underwent a pre-entry TB screening for their UK visa but never actually migrated), settled in Scotland, Northern Ireland or Wales and thus were unidentifiable in the hospital data from England, or never utilised any healthcare services within the NHS and were subsequently never allocated an NHS number. On average, 7% of non-EU migrants migrate to Scotland, Northern Ireland or Wales [21]. Removing this proportion from the total denominator would increase the PDS linkage rate to 68% from 62%. A comparable study in Canada showed linkage rates of the federal immigration database on permanent residents to provincial health care registries ranging from 74-86% for four Canadian provinces.[7] However, these linkages included only permanent residents to Canada who were found to be more likely to link in this study and thus could potentially explain the different linkage rates. Second, individuals with missing identifiers (<4.0% of records) were excluded from the linkage algorithm and removed from the denominator. Incorrect or inconsistent recording of personal identifiers between datasets could have resulted in missed matches in either dataset. Third, it is likely that a proportion of individuals never attended hospital whilst residing in England, especially given the age distribution of the cohort (75% arrived between the ages of 18–35 years old) and the majority reside in England between 1-4 years [22]. Moreover, migrants are on average healthier than the general population, resulting in less healthcare interactions which would mean lower linkage rates [23].

Lower PDS linkage rates of individuals from South Asia, Latin America and the Caribbean, and the Middle East and North Africa for round 1 (exact name agreement) could be explained by challenges in linking individuals with multiple forenames and surnames and the subsequent inconsistencies in recorded names. Similar trends were seen in a study linking school and hospital records, with ethnic minorities from Asian, Black, Chinese, and Mixed backgrounds less likely to link to PDS and HES in England [13]. Similarly, higher HES linkage rates were found when postcode was removed as a linkage variable for younger migrants and individuals arriving on student and work visas. As both are highly mobile populations, mismatch between postcodes held within the two datasets (PDS and HES) would be likely. To improve linkage rates for these groups, additional steps in the deterministic algorithm would be needed such as methods that identify partial agreements in string comparisons, inclusion of soundex for non-English names, or the application of probabilistic linkage methods [24].

While we would expect linkage rates to increase with length of time in England (e.g., more time to engage with the health system), more recent migrants (>5 years residing in England) linked at a higher rate to PDS. This trend could be driven by a complex range of factors including use of NHS prior to their first pre-entry TB screening test (i.e., someone who had already lived in England and departed prior to the start of the pre-entry TB screening programme), reason for migration and respective intended duration of stay, and in part from the introduction of the NHS Immigration Health Surcharge in 2015 [25]. The policy automatically allocates an NHS number and adds the individual to PDS prior to arrival to England once the payment is processed [26]. This is encouraging for future linkage work with these cohorts in England, but does not solve the issues with uncertainty around the total denominator. The study's PDS linkage rates for individuals who migrated more than eight years ago were 15% higher than a study conducted using data from 2009 and 2010 that found only 32.5% of individuals who had a pre-entry TB screening linked to PDS, suggesting length of stay does increase linkage rate [11].

Certain demographic groups had higher hospital linkage rates than others. Women, children, and older migrants linked at a higher rate to at least one hospital record. This mirrors evidence from the general population in England that these population groups tend to utilise hospital services at a higher rate [27]. Similarly, more individuals on more permanent visas and individuals with a longer length of stay in England had higher hospital linkage rates, reflecting the likelihood that needing to use a hospital would increase with time spent in England. Individuals entering on student, work, or working holiday visas are often shorter term migrants who would have less time to need to go to the hospital for care. Moreover, individuals on shorter term visas might engage with other health systems outside of England [28]. A record-linkage study in Norway found hospitalisation varies with length of stay, with longer stays being associated with higher admission rates [6]. However, it is important to note that linkage rates and NHS contact rates (whether primary or second care) cannot be distinguished in this study. Therefore, a higher linkage rate for a particular group could be due to either higher healthcare use or linkage error.

Strength and limitations

This study demonstrated novel multistep deterministic linkage methods to link immigration and hospital data for non-EU migrants and refugees in England. The Million Migrant study is a large population-based cohort with low levels of missing data on linkage variables and created of a highly powered longitudinal cohort of migrants and their data in England. The cohort is highly representative of migrants from non-EU countries intending to stay for at least six months and refugees resettled on a UK government resettlement scheme, but it is not representative of the entire migrant population in the UK. Importantly, our results will not be generalisable to migrants arriving on tourist visas or through irregular routes, asylum seekers, and migrants from countries who are not part of the TB pre-entry screening programme. However, individuals who underwent a TB pre-entry screening could overstay their initial visa and become undocumented and thus could theoretically be included in this cohort.

However, there are several potential sources of bias in this linkage study. Linkage error due to missing, mis-recorded, or dynamic identifiers could introduce bias in analyses of the linked dataset. This is particularly important when inclusion or exclusion from the analysis cohort relies on accurate linkage and could lead to a selection bias if the missed matches are not missing completely at random. More generally, estimating the linkage rates within the Million Migrant cohort was challenging. Without a known denominator, it was difficult to identify missed matches or false matches. Furthermore, as there was no comparable gold standard dataset, false matches (or two records incorrectly identified as a match) could not be identified. This limited measuring linkage quality in a meaningful way.

Uncertainty around the total denominator of the cohort could have resulted in an underestimation of the true linkage rate. We explored this uncertainty by determining linkage rates with the denominator as the total Million Migrant cohort and as those who had an NHS number. The standardised differences using both denominators were minimal or moderate for most sociodemographic variables except visa type, indicating that the linked Million Migrant-HES cohort was largely representative of the original unlinked cohort in relation to the variables observed. Researchers using this linked dataset will need to conduct sensitivity analyses around the denominator to determine any impact on effect estimates.

Implications

We described a novel record linkage study that links together data from non-EU migrant pre-entry screening and resettled refugee health check and hospitalisation data to generate the Million Migrant cohort. This data resource will allow for the opportunity to accurately identify migrants in routine health data using NHS numbers and improve evidence on the hospital-based events (e.g., readmissions) and mortality of migrants in England. Moreover, these data could be used to explore both population-based maternal health outcomes and the health of the children of migrants through analysis of maternal delivery and infant birth records. However, as there were lower linkage rates of individuals on shorter term visas, men and older migrants, future studies using these linked data have a greater risk of bias for these groups. We hope to make this data resource available to approved researchers in 2024 to assess the health needs and secondary care use of migrants, filling an important research gap and providing policymakers with detailed information about the health needs of this population. Lastly, the multistep deterministic data linkage methodologies should be expanded upon and replicated in other administrative data sources including primary care data, education, housing, or other social welfare programmes datasets. These would allow for a more in depth exploration of the relationship of migration to broader social and structural determinants of health.

Conclusion

Data linkage of migrant and hospital data offers an efficient way to study access to secondary care and create longitudinal cohorts for research, service evaluation and monitoring. Combining disparate data sources can help with the identification of migrants within electronic health records, and the approaches described here could be replicated in other administrative data sources. Hospital records disaggregated by migrant status and length of stay will provide an important resource to explore access to healthcare and the impact of migration as a core determinant of health.

Acknowledgments

The research costs for the study have been supported by MRC Grant Ref: MR/V028375/1 and by a Wellcome Clinical Research Career Development Fellowship (206602). SVK acknowledges funding from the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17).

Contributors

Conceptualization: RB, IC-M, RWA, KH. Methodology: RB, SW, KH, IC-M, RWA. Formal analysis: RB. Data curation: RB, IC-M, MM. Writing – original draft preparation: RB. Writing – review and editing: RB, YB, DZ, SW, SVK, IC-M, KH, RA. Visualisations: RB. Supervision: SW, KH, IC-M, RWA. Project administration: RB. Funding acquisition: RWA.

Statement on conflicts of interest

RB has received funding from Doctors of the World and is the chair of the UK Refugee and Migrant Health Research Consortium.

Ethics statement

The study has received sponsorship and approval from UCL's Joint Research Office. NHS Research Ethics Committee approval was received on 3 June 2019 (19/LO/086). We received Public Health England Caldicott panel approval (application entitled: "Hepatitis outcomes and healthcare quality and access in international migrants and refugees"), providing the legal basis under regulation 3 and 7 of the Health Service (Control of Patient Information) Regulations 2002 for the access to and use of patient identifiable data without individual consent for the data linkage of the datasets.

Code availability

Our scripts for data processing and transformation is available for inspection in our public GitHub repository: https://github.com/rburns520/Million_Migrant_Linkage.

References

1. International migration, England and Wales: Census 2021. Office for National Statistics https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/

internationalmigrationenglandandwales/census2021#:...:text=Out%20of%20the%2059.6%20million,were%20born%20outside%20the%20UK.

2. Burns R, Zhang CX, Patel P, Eley I, Campos-Matos I, Aldridge RW. Migration health research in the United Kingdom: A scoping review. *J Migr Health* 2021; 4: 100061. https://doi.org/10.1016/j.jmh.2021.100061

3. Saunders CL, Steventon A, Janta B, *et al.* Healthcare utilization among migrants to the UK: cross-sectional analysis of two national surveys. *J Health Serv Res Policy* 2021; 26: 54–61. https://doi.org/10.1177/1355819620911392

4. Zhang CX, Boukari Y, Pathak N, *et al.* Migrants' primary care utilisation before and during the COVID-19 pandemic in England: An interrupted time series analysis. *Lancet Reg Health - Eur* 2022; 20: 100455. https://doi.org/10.1016/j.lanepe.2022.100455

5. Dunlavy A, Cederström A, Katikireddi SV, Rostila M, Juárez SP. Investigating the salmon bias effect among international immigrants in Sweden: a register-based open cohort study. *Eur J Public Health* 2022; 32: 226–32. https://doi.org/10.1093/eurpub/ckab222

6. Elstad JI. Register study of migrants' hospitalization in Norway: world region origin, reason for migration, and length of stay. *BMC Health Serv Res* 2016; 16: 1–12. https://doi.org/10.1186/s12913-016-1561-9

7. Urquia ML, Walld R, Wanigaratne S, *et al.* Linking national immigration data to provincial repositories: the case of Canada. *Int J Popul Data Sci* 2021; 6. https://doi.org/10.23889/ijpds.v6i1.1412

8. Debbarman S, Prior H, Walld R, Urquia ML. Assessing the migrant mortality advantage among foreign-born and interprovincial migrants in Manitoba, Canada. *Can J Public Health* 2022; : 1–12. https://doi.org/10.17269/s41997-022-00727-4

9. Ewesesan R, Chartier MJ, Nickel NC, Wall-Wieler E, Urquia ML. Psychosocial and behavioral health indicators among immigrant and non-immigrant recent mothers. *BMC Pregnancy Childbirth* 2022; 22: 612. https://doi.org/10.1186/s12884-022-04937-z

10. 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. *The Lancet* 2016; 388: 2510–8. https://doi.org/10.1016/S0140-6736(16)31008-X

11. Stagg HR, Jones J, Bickler G, Abubakar I. Poor uptake of primary healthcare registration among recent entrants to the UK: a retrospective cohort study. *BMJ Open* 2012; 2: e001453. http://dx.doi.org/10.1136/bmjopen-2012-001453

12. Shaw RJ, Harron KL, Pescarini JM, *et al.* Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses. *Eur J Epidemiol* 2022; 37: 1215–24. https://doi.org/10.1007/s10654-022-00934-w

13. Libuy N, Harron K, Gilbert R, Caulton R, Cameron E, Blackburn R. Linking education and hospital data in England: linkage process and quality. *Int J Popul Data Sci* 2021; 6. https://doi.org/10.23889/ijpds.v6i1.1671

14. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. *Int J Epidemiol* 2019; 48: 2050–60. https://doi.org/10.1093/ije/dyaa028

15. Gilbert R, Lafferty R, Hagger-Johnson G, *et al.* GUILD: guidance for information about linking data sets. *J Public Health* 2018; 40: 191–8. https://doi.org/10.1093/pubmed/fdx037

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

17. NHS Digital. Demographics. https://digital.nhs.uk/services/demographics

18. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES APC). *Int J Epidemiol* 2017; 46: 1093–1093i. https://doi.org/10.1093/ije/dyx015

19. Demographics Batch Service. https://digital.nhs.uk/developer/api-catalogue/demographics-batch-service.

20. Harron KL, Doidge JC, Knight HE, *et al.* A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol* 2017; 46: 1699–710. https://doi.org/10.1093/ije/dyx015

21. Where do migrants live in the UK? 2022; published online March 24. https://migrationobservatory.ox.ac.uk/resources/briefings/where-do-migrants-live-in-the-uk/.

22. Permanent or Temporary: How Long do Migrants stay in the UK? Migraiton Obs. Univ. Oxf. 2020; published online Dec 16. https://migrationobservatory.ox.ac.uk/resources/briefings/permanent-or-temporary-how-long-do-migrants-stay-in-the-uk/#:...:text=leave%20to%20remain).-,The%20majority%20of%20foreign%2Dborn%20people%20living%20in%20the%20UK,in%20the%20UK%20for%20longer.

23. Fernández-Reino M. The health of migrants in the UK. Migration Observatory, 2021 https://migrationobservatory.ox.ac.uk/resources/briefings/the-health-of-migrants-in-the-uk/.

24. Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linking to enhance deterministic algorithms and reduce linkage errors in hospital administrative data. *J Innov Health Inform* 2017; 24: 891. http://dx.doi.org/10.14236/jhi.v24i2.891

25. Foreign & Commonwealth Office. UK introduces Health Surcharge. UK Gov. 2015; published online March 20. https://www.gov.uk/government/news/uk-introduces-health-surcharge.

26. Management of NHS Numbers and PDS Records. NHS Digit. https://digital.nhs.uk/services/national-back-office-for-the-personal-demographics-service/management-of-nhs-numbers-and-pds-records.

27. Hospital Admitted Patient Care Activity 2018-19. NHS Digital, 2019 https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2018-19.

28. Gorman D, Porteous L. Influences on Polish migrants' breast screening uptake in Lothian, Scotland. *Public Health* 2018; 158: 86–92. https://doi.org/10.1016/j.puhe.2017.11.026

Abbreviations

aOR:	Adjusted Odds Ratio
CI:	Confidence Interval
EU:	European Union
HES:	Hospital Episodes Statistics
NHS:	National Health Service
PDS:	Personal Demographic Service
IOM:	International Organization for Migration
SMD:	Standardised Mean Difference
TB:	Tuberculosis
UCL:	University College London
UK:	United Kingdom
UKHSA:	United Kingdom Health Security Agency

Supplementary Appendix

Appendix 1: UK tuberculosis pre-entry screening programme and refugee resettlement schemes

UK Tuberculosis pre-entry screening programme

All individual's resident in a non-European Country where tuberculosis is common (40 cases per 100,000 people) and planning to come and live in the UK for more than 6 months are screened for tuberculosis as part of the UK visa application process. The UK tuberculosis pre-entry screening programme was first piloted in 15 countries from 2005 and then subsequently rolled out to 101 countries through four phases between May 2012 and March 2014. The non-EU migrant pre-entry tuberculosis dataset included individuals screened between 1 January 2005 and 31 December 2020.

UK refugee resettlement schemes

The UK government accepts refugees identified by the United Nations High Commissioner for Refugees (UNHCR) under a series of resettlement schemes. These schemes include: the Gateway Protection Programme, the Mandate Resettlement Scheme, the Syrian Vulnerable Persons Resettlement Scheme, and the Vulnerable Children Resettlement Scheme. These individuals receive a pre-entry health assessment prior to arrival in the UK that aims to identify health and social care needs to be addressed and accommodated for once in the UK.[3] The refugee pre-arrival dataset used includes individuals who were given a health assessment between 1 March 2013 and 31 December 2022.

The cohort does not include individuals on Afghan Citizens Resettlement Scheme (ACRS), Afghan Relocations and Assistance Policy (ARAP) or Ukrainain resettlement schemes.

Supplementary Table 1.1: List of countries in the UK tuberculosis pre-entry screening programme and rollout date

Country	Pre-entry screening rollout phase
Cambodia	Pre-entry pilot scheme
United Republic of Tanzania	Pre-entry pilot scheme
Bangladesh	Pre-entry pilot scheme
Kenya	Pre-entry pilot scheme
Eritrea	Pre-entry pilot scheme
Pakistan	Pre-entry pilot scheme
Sudan	Pre-entry pilot scheme
Ivory Coast	Pre-entry pilot scheme
Niger	Pre-entry pilot scheme
Laos	Pre-entry pilot scheme
Thailand	Pre-entry pilot scheme
Burkina Faso	Pre-entry pilot scheme
Ghana	Pre-entry pilot scheme
Somalia	Pre-entry pilot scheme
Togo	Pre-entry pilot scheme
Malaysia	Phase 1 - completed 31/12/12
India	Phase 1 - completed 31/12/12
South Africa	Phase 1 - completed 31/12/12
Philippines	Phase 1 - completed 31/12/12
Mali	Phase 1 - completed 31/12/12
Lesotho	Phase 1 - completed 31/12/12
Uganda	Phase 2 - completed 31/10/13
Morocco	Phase 2 - completed 31/10/13
Indonesia	Phase 2 - completed 31/10/13
Zambia	Phase 2 - completed 31/10/13
Zimbabwe	Phase 2 - completed 31/10/1
Malawi	Phase 2 - completed 31/10/13
Vietnam	Phase 2 - completed 31/10/13
Sierra Leone	Phase 2 - completed 31/1/13
Gambia	Phase 2 - completed 31/10/13
Ethiopia	Phase 2 - completed 31/10/1
Russia	Phase 3 - completed 31/12/13
Senegal	Phase 3 - completed 31/12/13
Mauritania	Phase 3 - completed 31/12/13
Mongolia	Phase 3 - completed 31/12/13
Mozambique	Phase 3 - completed 31/12/13

Continued

Supplementary Table 1.1: Continued

Country	Pre-entry screening rollout phase
East Timor	Phase 3 - completed 31/12/13
Central African Republic	Phase 3 - completed 31/12/13
Tajikistan	Phase 3 - completed 31/12/13
Republic of Congo	Phase 3 - completed 31/12/13
Moldova	Phase 3 - completed 31/12/13
Namibia	Phase 3 - completed 31/12/13
Botswana	Phase 3 - completed 31/12/13
Solomon Islands	Phase 3 - completed 31/12/13
Ukraine	Phase 3 - completed 31/12/13
Angola	Phase 3 - completed 31/12/13
Uzbekistan	Phase 3 - completed 31/12/13
South Korea	Phase 3 - completed 31/12/13
Burundi	Phase 3 - completed 31/12/13
Liberia	Phase 3 - completed 31/12/13
Papua New Guinea	Phase 3 - completed 31/12/13
Nepal	Phase 3 - completed 31/12/13
Haiti	Phase 3 - completed 31/12/13
Peru	Phase 3 - completed 31/12/13
Nigeria	Phase 3 - completed 31/12/13
Myanmar	Phase 3 - completed 31/12/13
North Korea	Phase 3 - completed 31/12/13
Afghanistan	Phase 3 - completed 31/12/13
Suriname	Phase 3 - completed 31/12/13
Democratic Republic of the Congo	Phase 3 - completed 31/12/13
Ecuador	Phase 3 - completed 31/12/13
Equatorial Guinea	Phase 3 - completed 31/12/13
Kazakhstan	Phase 3 - completed 31/12/13
Gabon	Phase 3 - completed 31/12/13
Bhutan	Phase 3 - completed 31/12/13
Madagascar	Phase 3 - completed 31/12/13
Guinea Bissau	Phase 3 - completed 31/12/13
Chad	Phase 3 - completed 31/12/13
Cameroon	Phase 3 - completed 31/12/13
China	Phase 3 - completed 31/12/13
Guinea	Phase 3 - completed 31/12/13
Rwanda	Phase 3 - completed 31/12/13
Djibouti	Phase 3 - completed 31/12/13
Guyana	Phase 3 - completed 31/12/13
Kyrgyzstan	Phase 3 - completed 31/12/13
Swaziland	Phase 3 - completed 31/12/13
Bolivia	Phase 3 - completed 31/12/13
Iraq	Phase 4 - completed 31/03/14
Guatemala	Phase 4 - completed 31/03/14
Benin	Phase 4 - completed 31/03/14
Dominican Republic	Phase 4 - completed 31/03/14
Georgia	Phase 4 - completed 31/03/14
Azerbaijan	Phase 4 - completed 31/03/14
Brunei	Phase 4 - completed 31/03/14
Belarus	Phase 4 - completed 31/03/14
South Sudan	Phase 4 - completed 31/03/14
Sri Lanka	Phase 4 - completed 31/03/14
Panama	Phase 4 - completed 31/03/14
Vanuatu	Phase 4 - completed 31/03/14
Paraguay	Phase 4 - completed 31/03/14
Armenia	Phase 4 - completed 31/03/14
Turkmenistan	Phase 4 - completed 31/03/14
Algeria	Phase 4 - completed 31/03/14

Appendix 2: Description of data resources used in the linkage

Supplementary Table 2.1: Description of data resources used in the linkage

Data	Population	Years collected	Identifiers
Non-EU migrant pre-entry tuberculosis screening dataset	Derived from records on non-EU migrants' pre-entry tuberculosis screening. Part of the Million Migrant cohort.	1 January 2005 and 31 December 2020	Forename, surname, sex, date of birth
Refugee pre-arrival health check dataset	Derived from records on refugees' pre-arrival health check. Part of the Million Migrant cohort.	1 January 2013 and 31 December 2020	Forename, surname, sex, date of birth
NHS Personal Demographic Service	National database of patient demographic information in England used to identify the NHS numbers and postcode in the Million Migrant cohort.	2004 onwards	Forename, surname, sex, date of birth
Hospital Episode Statistics	All activity in English hospitals with information on inpatient admissions, outpatient appointments, deaths and A&E attendances.	APC (inpatients): April 1997 onwards* HES-ONS link mortality: 1997 onwards Outpatient: 2003 onwards A&E: 2001 onwards	NHS number, sex, date of birth, UK postcode

Appendix 3: Description of demographic variables in million migrant cohort

Supplementary Table 3.1: Description of demographic variables in million migrant cohort

Demographic variable	Description
Sex	Sex was recorded as Male or Female during an individual's TB pre-entry screening or refugee health check.
Age groups	Age groups were calculated by subtracting the date of the linkage (XXXXX) with an individual's date of birth. The following age groups were used: 0–17 years old, 18–34, 35–49, 50–64, 65+.
Region of origin	Region of origin was derived from country of origin recorded from the IOM datasets and categorised into the World Bank's 7-group classification of geographical regions.
Visa type	Visa type was recorded at the non-EU migrant pre-entry tuberculous screening or refugee pre-arrival health check and included student, work, working holiday, family, settlement and dependent, refugee, other and unknown categories.
Length of time in England	Length of time in England was calculated from the date of either pre-entry TB screening or refugee health assessment to date of linkage (01/01/2022)

Appendix 4: Statistical analysis

Standardised mean difference

Definition

Standardised mean difference (SMD) is mean difference between linked and unlinked records divided by their standard deviation.

Equation

The *smd* function in R Studio 4.1.2 calculates the standardised mean difference for each level k of a grouping variable (e.g., age group) compared to a reference r level[1]:

$$d_k = \sqrt{(\bar{x}_r - \bar{x}_k)^T S_{rk}^{-1} (\bar{x}_r - \bar{x}_k)}$$

d_k Cohen's d or SMD
$\bar{x}$ Sample mean
S_{rk} Covariances for reference group r and group k

If x is categorical, then $\bar{x}$ is the vector of proportions of each category level within a specific group, and S_{rk} is the multinomial covariance matrix.

Interpretation

We used standardised mean differences as these are thought to be more informative when comparing linked and unlinked records than P-values in large samples. Standardised mean differences of 0.2 or less were considered as small, greater than 0.2 to 0.5 as moderate, and greater than 0.5 to over 0.8 as large, as previously defined in the literature.[2--4] Assessing the standardised mean differences can help identify variables that may be more affected by linkage error and are therefore potential sources of bias.

References

1. Saul, Bradly. Using smd. 2020. https://cran.r-project.org/web/packages/smd/vignettes/smd_usage.html#:...:text= The%20smd%20package%20provides%20the,matrix%20%2C%20list%20%2C%20and%20data.

2. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699–710. https://doi.org/10.1093/ije/dyx177

3. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28:3083–107.

4. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

Explaining the model and confounder selection

The multivariable logistic regression models were performed to describe the association between each explanatory variable (either visa type, region of origin, or length of time in the UK) and the outcome of having a HES record (e.g., linked in the deterministic linkage procedure or did not link). Logistic regression was chosen because the outcome was binary (linked to HES or unlinked to HES). These analyses were not performed to describe a causal relationship between the explanatory variable and outcome but were done to be purely descriptive in nature.

I conducted model checks for multicollinearity between region of origin, visa type, and length of stay by calculating the generalised variance inflation factor (GVIF) using the vif commend in the car package. Evidence of multicollinearity was assumed if the adjusted GVIF exceeded 10.

Three separate logistic regression models were used for each explanatory variable (visa type, region of origin, and length of time). For each explanatory variable, we adjusted the final models by age and sex, given the importance of both age and sex on healthcare utilisation and hospitalisation. These two variables were used as controls when calculating the odds for each exposure separately.

Equation

$$\text{Log}\,[Y/(1-Y)] = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Log (Likelihood) — Explanatory variable — Age group — Sex

Explanatory variable: visa type, region of origin, length of stay.

Appendix 5: Performance of linkage stages

Supplementary Table 5.1: Migrant records by linkage stage (stage 1 and 2)

	Linked MM-PDS-HES	Excluded as a fail to link PDS (stage 1)	Excluded as fail to link HES (stage 2)	Overall
Records	451,689	665,300	1,347,618	1,799,307
Percentage	25	37	75	100

Supplementary Table 5.2: Linkage million migrant cohort to PDS by linkage step and sociodemographic characteristics (row percentages)

Variable	PDS Linkage step					Overall N (%)
	1 N (%)	2 N (%)	3 N (%)	4 N (%)	5 N (%)	
Individuals	880344 (78)	191352 (17)	24406 (2.2)	2777 (0.2)	35128 (3.1)	1,134,007 (100)
Sex						
Female	492,990 (82.1)	73,196 (12.2)	12,054 (2.0)	1,603 (0.3)	20,919 (3.5)	600,762 (53.0)
Male	387,354 (72.6)	118,156 (22.2)	12,352 (2.3)	1,174 (0.2)	14,209 (2.7)	533,245 (47.0)
Age	24 (20, 29)	24 (20, 30)	24 (17, 31)	24 (16, 32)	22 (20, 25)	24 (20, 29)
Age Group						
0 to 17	122,316 (75.6)	28,764 (17.8)	6,169 (3.8)	862 (0.5)	3,605 (2.2)	161,716 (14.3)
18 to 34	647,072 (77.9)	139,085 (16.7)	14,016 (1.7)	1,374 (0.2)	29,372 (3.5)	830,919 (73.3)
35 to 49	91,914 (78.0)	20,090 (17.0)	3,404 (2.9)	474 (0.4)	1,994 (1.7)	117,876 (10.4)
50 to 64	13,964 (80.2)	2,616 (15.0)	624 (3.6)	59 (0.3)	143 (0.8)	17,406 (1.5)
65+	5,076 (83.4)	797 (13.1)	193 (3.2)	#	#	6,085 (0.5)
Region of origin						
East Asia & Pacific	307,009 (78.8)	46,618 (12.0)	5,060 (1.3)	134 (0.0)	30,901 (7.9)	389,722 (34.4)
Europe & Central Asia	34,253 (94.2)	1,512 (4.2)	539 (1.5)	25 (0.1)	32 (0.1)	36,361 (3.2)
Latin America & Caribbean	135 (22.1)	175 (28.6)	298 (48.8)	#	#	611 (0.1)
Middle East & North Africa	19,040 (71.0)	1,437 (5.4)	5,591 (20.9)	713 (2.7)	23 (0.1)	26,804 (2.4)
North America	246 (71.3)	76 (22.0)	#	#	#	345 (0.0)
South Asia	328,987 (70.4)	128,429 (27.5)	5,410 (1.2)	62 (0.0)	4,099 (0.9)	466,987 (41.2)
Sub-Saharan Africa	190,654 (89.4)	13,103 (6.1)	7,484 (3.5)	1,835 (0.9)	70 (0.0)	213,146 (18.8)
Visa Type						
Family Reunion	50,176 (79.6)	10,071 (16.0)	1,454 (2.3)	292 (0.5)	1,045 (1.7)	63,038 (5.6)
Settlement and Dependents	264,551 (80.8)	55,056 (16.8)	4,642 (1.4)	754 (0.2)	2,558 (0.8)	327,561 (28.9)
Refugee	13,843 (69.8)	809 (4.1)	5,045 (25.4)	128 (0.6)	14 (0.1)	19,839 (1.7)
Students	358,784 (76.2)	73,724 (15.7)	9,258 (2.0)	1,245 (0.3)	27,731 (5.9)	470,742 (41.5)
Work	74,879 (68.9)	29,366 (27.0)	2,235 (2.1)	209 (0.2)	2,000 (1.8)	108,689 (9.6)
Working Holiday Maker	11,438 (66.5)	4,354 (25.3)	#	#	1,116 (6.5)	17,197 (1.5)
Other	101,913 (84.2)	17,068 (14.1)	1,361 (1.1)	107 (0.1)	631 (0.5)	121,080 (10.7)
Missing	4,760 (81.2)	904 (15.4)	131 (2.2)	#	#	5,861 (0.5)
Length of time in England (in years)						
<2	320,650 (74.3)	77,178 (17.9)	9,430 (2.2)	1,509 (0.3)	22,615 (5.2)	431,382 (38.0)
2 to 4	295,229 (80.0)	50,762 (13.8)	10,826 (2.9)	862 (0.2)	11,499 (3.1)	369,178 (32.6)
5 to 7	98,931 (83.6)	15,886 (13.4)	2,668 (2.3)	255 (0.2)	660 (0.6)	118,400 (10.4)
>8	165,534 (77.0)	47,526 (22.1)	1,482 (0.7)	151 (0.1)	354 (0.2)	215,047 (19.0)

Suppressed due to low cell count. Age and region of origin missing categories removed.
Note: The table only includes successfully linked records.

Supplementary Table 5.3: Linkage million migrant cohort to PDS by linkage step and sociodemographic characteristics (column percentages)

Variable	PDS Linkage step					Overall N (%)
	1 N (%)	2 N (%)	3 N (%)	4 N (%)	5 N (%)	
Individuals	880344 (78)	191352 (17)	24406 (2.2)	2777 (0.2)	35128 (3.1)	1,134,007 (100)
Sex						
Female	492,990 (56.0)	73,196 (38.3)	12,054 (49.4)	1,603 (57.7)	20,919 (59.6)	600,762 (53.0)
Male	387,354 (44.0)	118,156 (61.7)	12,352 (50.6)	1,174 (42.3)	14,209 (40.4)	533,245 (47.0)
Age	24 (20, 29)	24 (20, 30)	24 (17, 31)	24 (16, 32)	22 (20, 25)	24 (20, 29)
Age Group						
0 to 17	122,316 (13.9)	28,764 (15.0)	6,169 (25.3)	862 (31.0)	3,605 (10.3)	161,716 (14.3)
18 to 34	647,072 (73.5)	139,085 (72.7)	14,016 (57.4)	1,374 (49.5)	29,372 (83.6)	830,919 (73.3)
35 to 49	91,914 (10.4)	20,090 (10.5)	3,404 (13.9)	474 (17.1)	1,994 (5.7)	117,876 (10.4)
50 to 64	13,964 (1.6)	2,616 (1.4)	624 (2.6)	59 (2.1)	143 (0.4)	17,406 (1.5)
65+	5,076 (0.6)	797 (0.4)	193 (0.8)	#	#	6,085 (0.5)
Region of origin						
East Asia & Pacific	307,009 (34.9)	46,618 (24.4)	5,060 (20.7)	134 (4.8)	30,901 (88.0)	389,722 (34.4)
Europe & Central Asia	34,253 (3.9)	1,512 (0.8)	539 (2.2)	25 (0.9)	32 (0.1)	36,361 (3.2)
Latin America & Caribbean	135 (0.0)	175 (0.1)	298 (1.2)	#	#	611 (0.1)
Middle East & North Africa	19,040 (2.2)	1,437 (0.8)	5,591 (22.9)	713 (25.7)	23 (0.1)	26,804 (2.4)
North America	246 (0.0)	76 (0.0)	15 (0.1)	#	#	345 (0.0)
South Asia	328,987 (37.4)	128,429 (67.1)	5,410 (22.2)	62 (2.2)	4,099 (11.7)	466,987 (41.2)
Sub-Saharan Africa	190,654 (21.7)	13,103 (6.8)	7,484 (30.7)	1,835 (66.1)	70 (0.2)	213,146 (18.8)
Visa Type						
Family Reunion	50,176 (5.7)	10,071 (5.3)	1,454 (6.0)	292 (10.5)	1,045 (3.0)	63,038 (5.6)
Settlement and Dependents	264,551 (30.1)	55,056 (28.8)	4,642 (19.0)	754 (27.2)	2,558 (7.3)	327,561 (28.9)
Refugee	13,843 (1.6)	809 (0.4)	5,045 (20.7)	128 (4.6)	14 (0.0)	19,839 (1.7)
Students	358,784 (40.8)	73,724 (38.5)	9,258 (37.9)	1,245 (44.8)	27,731 (78.9)	470,742 (41.5)
Work	74,879 (8.5)	29,366 (15.3)	2,235 (9.2)	209 (7.5)	2,000 (5.7)	108,689 (9.6)
Working Holiday Maker	11,438 (1.3)	4,354 (2.3)	#	#	1,116 (3.2)	17,197 (1.5)
Other	101,913 (11.6)	17,068 (8.9)	1,361 (5.6)	107 (3.9)	631 (1.8)	121,080 (10.7)
Missing	4,760 (0.5)	904 (0.5)	131 (0.5)	33 (1.2)	33 (0.1)	5,861 (0.5)
Length of time in England (in years)						
<2	320,650 (36.4)	77,178 (40.3)	9,430 (38.6)	1,509 (54.3)	22,615 (64.4)	431,382 (38.0)
2 to 4	295,229 (33.5)	50,762 (26.5)	10,826 (44.4)	862 (31.0)	11,499 (32.7)	369,178 (32.6)
5 to 7	98,931 (11.2)	15,886 (8.3)	2,668 (10.9)	255 (9.2)	660 (1.9)	118,400 (10.4)
>8	165,534 (18.8)	47,526 (24.8)	1,482 (6.1)	151 (5.4)	354 (1.0)	215,047 (19.0)

Suppressed due to low cell count. Age and region of origin missing categories removed.
Note: The table only includes successfully linked records.

Supplementary Table 5.4: Linkage million migrant cohort to HES by linkage step and sociodemographic characteristics (row percentages)

Variable	HES Linkage step						Overall N (%)
	1 N (%)	2 N (%)	3 N (%)	4 N (%)	5 N (%)	6 N (%)	
Individuals	310,118 (69)	117,328 (26)	4,381 (1.0)	2,440 (0.5)	744 (0.2)	16,678 (3.7)	451,689 (100)
Sex							
Female	190,332 (71.5)	62,756 (23.6)	2,711 (1.0)	1,410 (0.5)	369 (0.1)	8,731 (3.3)	266,309 (59.0)
Male	119,786 (64.6)	54,572 (29.4)	1,670 (0.9)	1,030 (0.6)	375 (0.2)	7,947 (4.3)	185,380 (41.0)
Age	25 (21, 31)	25 (20, 30)	23 (20, 29)	22 (19, 27)	25 (19, 31)	23 (20, 26)	25 (21, 30)
Age Group							
0 to 17	40,058 (64.0)	19,647 (31.4)	695 (1.1)	421 (0.7)	152 (0.2)	1,625 (2.6)	62,598 (13.9)
18 to 34	222,697 (68.5)	83,117 (25.6)	3,208 (1.0)	1,815 (0.6)	464 (0.1)	13,817 (4.2)	325,118 (72.0)
35 to 49	37,151 (72.7)	12,165 (23.8)	424 (0.8)	188 (0.4)	95 (0.2)	1,063 (2.1)	51,086 (11.3)
50 to 64	6,934 (78.6)	1,682 (19.1)	41 (0.5)	13 (0.1)	21 (0.2)	130 (1.5)	8,821 (2.0)
65+	3,276 (80.7)	717 (17.7)	#	#	12 (0.3)	43 (1.1)	4,061 (0.9)
Region of origin							
East Asia & Pacific	46,312 (56.2)	23,157 (28.1)	2,892 (3.5)	1,749 (2.1)	199 (0.2)	8,158 (9.9)	82,467 (18.3)
Europe & Central Asia	7,715 (63.3)	3,854 (31.6)	23 (0.2)	18 (0.1)	37 (0.3)	549 (4.5)	12,196 (2.7)
Latin America & Caribbean	88 (65.7)	37 (27.6)	#	#	#	#	134 (0.0)
Middle East & North Africa	12,642 (80.2)	2,900 (18.4)	25 (0.2)	9 (0.1)	28 (0.2)	164 (1.0)	15,768 (3.5)
North America	12 (41.4)	11 (37.9)	#	#	#	#	29 (0.0)
South Asia	177,523 (72.3)	60,300 (24.6)	1,305 (0.5)	591 (0.2)	302 (0.1)	5,363 (2.2)	245,384 (54.3)
Sub-Saharan Africa	65,808 (68.8)	27,064 (28.3)	136 (0.1)	72 (0.1)	177 (0.2)	2,431 (2.5)	95,688 (21.2)
Visa Type							
Family Reunion	22,267 (72.6)	7,452 (24.3)	308 (1.0)	116 (0.4)	59 (0.2)	489 (1.6)	30,691 (6.8)
Settlement and Dependents	151,324 (75.7)	44,600 (22.3)	970 (0.5)	375 (0.2)	236 (0.1)	2,471 (1.2)	199,976 (44.3)
Refugee	10,419 (82.1)	2,127 (16.8)	23 (0.2)	#	26 (0.2)	89 (0.7)	12,692 (2.8)
Students	81,441 (59.1)	41,701 (30.3)	2,371 (1.7)	1,562 (1.1)	256 (0.2)	10,496 (7.6)	137,827 (30.5)
Work	25,450 (63.2)	13,220 (32.8)	382 (0.9)	217 (0.5)	96 (0.2)	894 (2.2)	40,259 (8.9)
Working Holiday	2,959 (61.6)	1,378 (28.7)	175 (3.6)	78 (1.6)	#	204 (4.2)	4,803 (1.1)
Other	15,176 (63.7)	6,423 (26.9)	144 (0.6)	77 (0.3)	54 (0.2)	1,962 (8.2)	23,836 (5.3)
Missing	1,082 (67.4)	427 (26.6)	#	#	#	73 (4.5)	1,605 (0.4)
Length of time in England (in years)							
<2	59,861 (65.2)	22,901 (24.9)	2,378 (2.6)	1,102 (1.2)	296 (0.3)	5,288 (5.8)	91,826 (20.3)
2 to 4	94,530 (66.0)	40,303 (28.1)	1,593 (1.1)	1,070 (0.7)	254 (0.2)	5,535 (3.9)	143,285 (31.7)
5 to 7	49,736 (70.3)	18,580 (26.3)	136 (0.2)	115 (0.2)	82 (0.1)	2,052 (2.9)	70,701 (15.7)
>8	105,991 (72.7)	35,544 (24.4)	274 (0.2)	153 (0.1)	112 (0.1)	3,803 (2.6)	145,877 (32.3)

Suppressed due to low cell count. Age and region of origin missing categories removed.
Note: The table only includes successfully linked records.

Supplementary Table 5.5: 5Linkage million migrant cohort to HES by linkage step and sociodemographic characteristics (column percentages)

Variable	HES Linkage step						Overall N (%)
	1 N (%)	2 N (%)	3 N (%)	4 N (%)	5 N (%)	6 N (%)	
Individuals	310,118 (69)	117,328 (26)	4,381 (1.0)	2,440 (0.5)	744 (0.2)	16,678 (3.7)	451,689 (100)
Sex							
Female	190,332 (61.4)	62,756 (53.5)	2,711 (61.9)	1,410 (57.8)	369 (49.6)	8,731 (52.4)	266,309 (59.0)
Male	119,786 (38.6)	54,572 (46.5)	1,670 (38.1)	1,030 (42.2)	375 (50.4)	7,947 (47.6)	185,380 (41.0)
Age	25 (21, 31)	25 (20, 30)	23 (20, 29)	22 (19, 27)	25 (19, 31)	23 (20, 26)	25 (21, 30)
Age Group							
0 to 17	40,058 (12.9)	19,647 (16.7)	695 (15.9)	421 (17.3)	152 (20.4)	1,625 (9.7)	62,598 (13.9)
18 to 34	222,697 (71.8)	83,117 (70.8)	3,208 (73.2)	1,815 (74.4)	464 (62.4)	13,817 (82.8)	325,118 (72.0)
35 to 49	37,151 (12.0)	12,165 (10.4)	424 (9.7)	188 (7.7)	95 (12.8)	1,063 (6.4)	51,086 (11.3)
50 to 64	6,934 (2.2)	1,682 (1.4)	41 (0.9)	13 (0.5)	21 (2.8)	130 (0.8)	8,821 (2.0)
65+	3,276 (1.1)	717 (0.6)	#	#	#	43 (0.3)	4,061 (0.9)
Region of origin							
East Asia & Pacific	46,312 (14.9)	23,157 (19.7)	2,892 (66.0)	1,749 (71.7)	199 (26.7)	8,158 (48.9)	82,467 (18.3)
Europe & Central Asia	7,715 (2.5)	3,854 (3.3)	23 (0.5)	18 (0.7)	37 (5.0)	549 (3.3)	12,196 (2.7)
Latin America & Caribbean	88 (0.0)	37 (0.0)	#	#	#	#	134 (0.0)
Middle East & North Africa	12,642 (4.1)	2,900 (2.5)	25 (0.6)	#	28 (3.8)	164 (1.0)	15,768 (3.5)
North America	12 (0.0)	11 (0.0)	#	#	#	#	29 (0.0)
South Asia	177,523 (57.2)	60,300 (51.4)	1,305 (29.8)	591 (24.2)	302 (40.6)	5,363 (32.2)	245,384 (54.3)
Sub-Saharan Africa	65,808 (21.2)	27,064 (23.1)	136 (3.1)	72 (3.0)	177 (23.8)	2,431 (14.6)	95,688 (21.2)
Visa Type							
Family Reunion	22,267 (7.2)	7,452 (6.4)	308 (7.0)	116 (4.8)	59 (7.9)	489 (2.9)	30,691 (6.8)
Settlement and Dependents	151,324 (48.8)	44,600 (38.0)	970 (22.1)	375 (15.4)	236 (31.7)	2,471 (14.8)	199,976 (44.3)
Refugee	10,419 (3.4)	2,127 (1.8)	#	#	26 (3.5)	89 (0.5)	12,692 (2.8)
Students	81,441 (26.3)	41,701 (35.5)	2,371 (54.1)	1,562 (64.0)	256 (34.4)	10,496 (62.9)	137,827 (30.5)
Work	25,450 (8.2)	13,220 (11.3)	382 (8.7)	217 (8.9)	96 (12.9)	894 (5.4)	40,259 (8.9)
Working Holiday	2,959 (1.0)	1,378 (1.2)	175 (4.0)	#	#	204 (1.2)	4,803 (1.1)
Other	15,176 (4.9)	6,423 (5.5)	144 (3.3)	77 (3.2)	54 (7.3)	1,962 (11.8)	23,836 (5.3)
Missing	1,082 (0.3)	427 (0.4)	#	#	#	73 (0.4)	1,605 (0.4)
Length of time in England (in years)							
<2	59,861 (19.3)	22,901 (19.5)	2,378 (54.3)	1,102 (45.2)	296 (39.8)	5,288 (31.7)	91,826 (20.3)
2 to 4	94,530 (30.5)	40,303 (34.4)	1,593 (36.4)	1,070 (43.9)	254 (34.1)	5,535 (33.2)	143,285 (31.7)
5 to 7	49,736 (16.0)	18,580 (15.8)	136 (3.1)	115 (4.7)	82 (11.0)	2,052 (12.3)	70,701 (15.7)
>8	105,991 (34.2)	35,544 (30.3)	274 (6.3)	153 (6.3)	112 (15.1)	3,803 (22.8)	145,877 (32.3)

Suppressed due to low cell count. Age and region of origin missing categories removed.
Note: The table only includes successfully linked records.

Appendix 6: Linking rates

Supplementary Table 6.1: Linking rate for PDS linkage (stage 1) by sociodemographic characteristics (row percentages)

Variable	Overall N = 1,799,307 (100%)	PDS Linkage	
		Linked N = 1,134,007 (63%)	Un-linked N = 665,300 (37%)
Sex			
Female	886,791 (49.3)	600,762 (67.7)	286,029 (32.3)
Male	912,516 (50.7)	533,245 (58.4)	379,271 (41.6)
Age	24 (20, 30)	24 (20, 29)	25 (21, 31)
Age Group*			
0 to 17	235,577 (13.1)	161,716 (68.6)	73,861 (31.4)
18 to 34	1,297,617 (72.1)	830,919 (64.0)	466,698 (36.0)
35 to 49	213,349 (11.9)	117,876 (55.3)	95,473 (44.7)
50 to 64	41,282 (2.3)	17,406 (42.2)	23,876 (57.8)
65+	11,458 (0.6)	6,085 (53.1)	5,373 (46.9)
Region of origin			
East Asia & Pacific	582,906 (32.4)	389,722 (66.9)	193,184 (33.1)
Europe & Central Asia	49,012 (2.7)	36,361 (74.2)	12,651 (25.8)
Latin America & Caribbean	948 (0.1)	611 (64.5)	337 (35.5)
Middle East & North Africa	45,300 (2.5)	26,804 (59.2)	18,496 (40.8)
North America	437 (0.0)	345 (78.9)	92 (21.1)
South Asia	790,340 (43.9)	466,987 (59.1)	323,353 (40.9)
Sub-Saharan Africa	330,303 (18.4)	213,146 (64.5)	117,157 (35.5)
Missing	61 (0.0)	31 (50.8)	30 (49.2)
Visa Type			
Family Reunion	94,209 (5.2)	63,038 (66.9)	31,171 (33.1)
Settlement and Dependents	467,571 (26.0)	327,561 (70.1)	140,010 (29.9)
Refugee	33,978 (1.9)	19,839 (58.4)	14,139 (41.6)
Students	828,790 (46.1)	470,742 (56.8)	358,048 (43.2)
Work	155,635 (8.6)	108,689 (69.8)	46,946 (30.2)
Working Holiday	35,915 (2.0)	17,197 (47.9)	18,718 (52.1)
Other	167,918 (9.3)	121,080 (72.1)	46,838 (27.9)
Missing	15,291 (0.8)	5,861 (38.3)	9,430 (61.7)
Length of time in England (in years)			
<2	572,985 (31.8)	431,382 (75.3)	141,603 (24.7)
2 to 4	494,524 (27.5)	369,178 (74.7)	125,346 (25.3)
5 to 7	276,943 (15.4)	118,400 (42.8)	158,543 (57.2)
>8	454,855 (25.3)	215,047 (47.3)	239,808 (52.7)

*Age group missing category removed due to low cell numbers.

Supplementary Table 6.2: Linking rate for PDS linkage (stage 1) by sociodemographic characteristics (column percentages)

Variable	Overall N = 1,799,307 (100%)	PDS Linkage	
		Linked N = 1,134,007 (63%)	Un-linked N = 665,300 (37%)
Sex			
Female	886,791 (49.3)	600,762 (53.0)	286,029 (43.0)
Male	912,516 (50.7)	533,245 (47.0)	379,271 (57.0)
Age	24 (20, 30)	24 (20, 29)	25 (21, 31)
Age Group*			
0 to 17	235,577 (13.1)	161,716 (14.3)	73,861 (11.1)
18 to 34	1,297,617 (72.1)	830,919 (73.3)	466,698 (70.1)
35 to 49	213,349 (11.9)	117,876 (10.4)	95,473 (14.4)
50 to 64	41,282 (2.3)	17,406 (1.5)	23,876 (3.6)
65+	11,458 (0.6)	6,085 (0.5)	5,373 (0.8)
Region of origin			
East Asia & Pacific	582,906 (32.4)	389,722 (34.4)	193,184 (29.0)
Europe & Central Asia	49,012 (2.7)	36,361 (3.2)	12,651 (1.9)
Latin America & Caribbean	948 (0.1)	611 (0.1)	337 (0.1)
Middle East & North Africa	45,300 (2.5)	26,804 (2.4)	18,496 (2.8)
North America	437 (0.0)	345 (0.0)	92 (0.0)
South Asia	790,340 (43.9)	466,987 (41.2)	323,353 (48.6)
Sub-Saharan Africa	330,303 (18.4)	213,146 (18.8)	117,157 (17.6)
Missing	61 (0.0)	31 (0.0)	30 (0.0)
Visa Type			
Family Reunion	94,209 (5.2)	63,038 (5.6)	31,171 (4.7)
Settlement and Dependents	467,571 (26.0)	327,561 (28.9)	140,010 (21.0)
Refugee	33,978 (1.9)	19,839 (1.7)	14,139 (2.1)
Students	828,790 (46.1)	470,742 (41.5)	358,048 (53.8)
Work	155,635 (8.6)	108,689 (9.6)	46,946 (7.1)
Working Holiday	35,915 (2.0)	17,197 (1.5)	18,718 (2.8)
Other	167,918 (9.3)	121,080 (10.7)	46,838 (7.0)
Missing	15,291 (0.8)	5,861 (0.5)	9,430 (1.4)
Length of time in England (in years)			
<2	572,985 (31.8)	431,382 (38.0)	141,603 (21.3)
2 to 4	494,524 (27.5)	369,178 (32.6)	125,346 (18.8)
5 to 7	276,943 (15.4)	118,400 (10.4)	158,543 (23.8)
>8	454,855 (25.3)	215,047 (19.0)	239,808 (36.0)

*Age group missing category removed due to low cell numbers.

Supplementary Table 6.3: Sociodemographic characteristics and standardised mean differences (SMD) of the migrant sample from the Million Migrant cohort linked and non-linked to HES (row percentages)

Variable	Overall N = 1,799,307 (100%)	HES Linked N = 451,689 (25%)	HES Un-linked N = 1,347,618 (75%)	SMD	95% CI
Sex				0.26	0.26, 0.26
Female	886,791 (49.3)	266,309 (59.0)	620,482 (46.0)		
Male	912,516 (50.7)	185,380 (41.0)	727,136 (54.0)		
Age Group*				0.06	0.06, 0.07
0 to 17	235,577 (13.1)	62,598 (13.9)	172,979 (12.8)		
18 to 34	1,297,617 (72.1)	325,118 (72.0)	972,499 (72.2)		
35 to 49	213,349 (11.9)	51,086 (11.3)	162,263 (12.0)		
50 to 64	41,282 (2.3)	8,821 (2.0)	32,461 (2.4)		
65+	11,458 (0.6)	4,061 (0.9)	7,397 (0.5)		
Region of origin				0.44	0.43, 0.44
East Asia & Pacific	582,906 (32.4)	82,467 (18.3)	500,439 (37.1)		
Europe & Central Asia	49,012 (2.7)	12,196 (2.7)	36,816 (2.7)		
Latin America & Caribbean	948 (0.1)	134 (0.0)	814 (0.1)		
Middle East & North Africa	45,300 (2.5)	15,768 (3.5)	29,532 (2.2)		
North America	437 (0.0)	29 (0.0)	408 (0.0)		
South Asia	790,340 (43.9)	245,384 (54.3)	544,956 (40.4)		
Sub-Saharan Africa	330,303 (18.4)	95,688 (21.2)	234,615 (17.4)		
Missing	61 (0.0)	23 (0.0)	38 (0.0)		
Visa Type				0.63	0.62, 0.63
Family Reunion	94,209 (5.2)	30,691 (6.8)	63,518 (4.7)		
Settlement and Dependents	467,571 (26.0)	199,976 (44.3)	267,595 (19.9)		
Refugee	33,978 (1.9)	12,692 (2.8)	21,286 (1.6)		
Students	828,790 (46.1)	137,827 (30.5)	690,963 (51.3)		
Work	155,635 (8.6)	40,259 (8.9)	115,376 (8.6)		
Working Holiday	35,915 (2.0)	4,803 (1.1)	31,112 (2.3)		
Other	167,918 (9.3)	23,836 (5.3)	144,082 (10.7)		
Missing	15,291 (0.8)	1,605 (0.4)	13,686 (1.0)		
Length of time in England (in years)				0.36	0.36, 0.37
<2	572,985 (31.8)	91,826 (20.3)	481,159 (35.7)		
2 to 4	494,524 (27.5)	143,285 (31.7)	351,239 (26.1)		
5 to 7	276,943 (15.4)	70,701 (15.7)	206,242 (15.3)		
>8	454,855 (25.3)	145,877 (32.3)	308,978 (22.9)		

*Age group missing category removed due to low cell numbers.

Supplementary Table 6.4: Sociodemographic characteristics and standardised mean differences (SMD) of the migrant sample from the million migrant-PDS cohort linked and non-linked to HES (row percentages)

Variable	Overall N = 1,134,007 (100%)	HES Linked N = 451,689 (40.0%)	HES Un-linked N = 682,318 (60%)	SMD	95% CI
Sex				0.20	0.20, 0.20
Female	600,762 (53.0)	266,309 (44.3)	334,453 (55.7)		
Male	533,245 (47.0)	185,380 (34.8)	347,865 (65.2)		
Age Group*				0.11	0.11, 0.11
0 to 17	161,716 (14.3)	62,598 (38.7)	99,118 (61.3)		
18 to 34	830,919 (73.3)	325,118 (39.1)	505,801 (60.9)		
35 to 49	117,876 (10.4)	51,086 (43.3)	66,790 (56.7)		
50 to 64	17,406 (1.5)	8,821 (50.7)	8,585 (49.3)		
65+	6,085 (0.5)	4,061 (66.7)	2,024 (33.3)		
Region				0.63	0.63, 0.63
East Asia & Pacific	389,722 (34.4)	82,467 (21.2)	307,255 (78.8)		
Europe & Central Asia	36,361 (3.2)	12,196 (33.5)	24,165 (66.5)		
Latin America & Caribbean	611 (0.1)	134 (21.9)	477 (78.1)		
Middle East & North Africa	26,804 (2.4)	15,768 (58.8)	11,036 (41.2)		
North America	345 (0.0)	29 (8.4)	316 (91.6)		
South Asia	466,987 (41.2)	245,384 (52.5)	221,603 (47.5)		
Sub-Saharan Africa	213,146 (18.8)	95,688 (44.9)	117,458 (55.1)		
Missing	31 (0.0)	23 (74.2)	8 (25.8)		
Visa Type				0.67	0.67, 0.67
Family Reunion	63,038 (5.6)	30,691 (48.7)	32,347 (51.3)		
Settlement and Dependents	327,561 (28.9)	199,976 (61.1)	127,585 (38.9)		
Refugee	19,839 (1.7)	12,692 (64.0)	7,147 (36.0)		
Students	470,742 (41.5)	137,827 (29.3)	332,915 (70.7)		
Work	108,689 (9.6)	40,259 (37.0)	68,430 (63.0)		
Working Holiday	17,197 (1.5)	4,803 (27.9)	12,394 (72.1)		
Other	121,080 (10.7)	23,836 (19.7)	97,244 (80.3)		
Missing	5,861 (0.5)	1,605 (27.4)	4,256 (72.6)		
Length of time in England (in years)				0.79	0.79, 0.80
<2	431,382 (38.0)	91,826 (21.3)	339,556 (78.7)		
2 to 4	369,178 (32.6)	143,285 (38.8)	225,893 (61.2)		
5 to 7	118,400 (10.4)	70,701 (59.7)	47,699 (40.3)		
>8	215,047 (19.0)	145,877 (67.8)	69,170 (32.2)		

*Age group missing category removed due to low cell numbers.

Supplementary Table 6.5: Sociodemographic characteristics and standardised mean differences (SMD) of the migrant sample from the Million Migrant-PDS cohort linked and non-linked to HES (column percentages)

Variable	Overall N = 1,134,007 (100.0%)	HES Linked N = 451,689 (40.0%)	HES Un-linked N =682,318 (60%)	SMD	95% CI
Sex				0.20	0.20, 0.20
Female	600,762 (53.0)	266,309 (59.0)	334,453 (49.0)		
Male	533,245 (47.0)	185,380 (41.0)	347,865 (51.0)		
Age Group*				0.11	0.11, 0.11
0 to 17	161,716 (14.3)	62,598 (13.9)	99,118 (14.5)		
18 to 34	830,919 (73.3)	325,118 (72.0)	505,801 (74.1)		
35 to 49	117,876 (10.4)	51,086 (11.3)	66,790 (9.8)		
50 to 64	17,406 (1.5)	8,821 (2.0)	8,585 (1.3)		
65+	6,085 (0.5)	4,061 (0.9)	2,024 (0.3)		
Region of origin				0.63	0.63, 0.63
East Asia & Pacific	389,722 (34.4)	82,467 (18.3)	307,255 (45.0)		
Europe & Central Asia	36,361 (3.2)	12,196 (2.7)	24,165 (3.5)		
Latin America & Caribbean	611 (0.1)	134 (0.0)	477 (0.1)		
Middle East & North Africa	26,804 (2.4)	15,768 (3.5)	11,036 (1.6)		
North America	345 (0.0)	29 (0.0)	316 (0.0)		
South Asia	466,987 (41.2)	245,384 (54.3)	221,603 (32.5)		
Sub-Saharan Africa	213,146 (18.8)	95,688 (21.2)	117,458 (17.2)		
Missing	31 (0.0)	23 (0.0)	8 (0.0)		
Visa Type				0.67	0.67, 0.67
Family Reunion	63,038 (5.6)	30,691 (6.8)	32,347 (4.7)		
Settlement and Dependents	327,561 (28.9)	199,976 (44.3)	127,585 (18.7)		
Refugee	19,839 (1.7)	12,692 (2.8)	7,147 (1.0)		
Students	470,742 (41.5)	137,827 (30.5)	332,915 (48.8)		
Work	108,689 (9.6)	40,259 (8.9)	68,430 (10.0)		
Working Holiday	17,197 (1.5)	4,803 (1.1)	12,394 (1.8)		
Other	121,080 (10.7)	23,836 (5.3)	97,244 (14.3)		
Missing	5,861 (0.5)	1,605 (0.4)	4,256 (0.6)		
Length of time in England (in years)				0.79	0.79, 0.80
<2	431,382 (38.0)	91,826 (20.3)	339,556 (49.8)		
2 to 4	369,178 (32.6)	143,285 (31.7)	225,893 (33.1)		
5 to 7	118,400 (10.4)	70,701 (15.7)	47,699 (7.0)		
>8	215,047 (19.0)	145,877 (32.3%)	69,170 (10.1)		

*Age group missing category removed due to low cell numbers.

Supplementary Table 6.6: Adjusted odds ratios for a link between million migrant-PDS cohort and HES records for visa type, region of origin, and length of time in England, controlling for age and sex

Variable*	**Total**	**aOR = Adjusted Odds Ratio**	**95% CI**
Visa Type			
Family Reunion	63065	—	—
Settlement and Dependents	327492	1.7	1.67, 1.73
Refugee	19807	2.33	2.26, 2.41
Students	470814	0.4	0.39, 0.41
Work	108675	0.59	0.57, 0.60
Working Holiday Maker	17199	0.37	0.36, 0.39
Other	121056	0.24	0.23, 0.24
Missing	5863	0.4	0.37, 0.42
Region of origin			
East Asia & Pacific	389724	—	—
Europe & Central Asia	36359	1.98	1.93, 2.02
Latin America & Caribbean	610	1.06	0.87, 1.29
Middle East & North Africa	26806	5.88	5.73, 6.04
North America	345	0.37	0.25, 0.54
South Asia	466980	4.72	4.67, 4.77
Sub-Saharan Africa	213147	3.31	3.27, 3.35
Length of time in England (years)			
< 2	407121	—	—
2 to 4	376063	2.47	2.44, 2.49
5 to 7	129218	5.49	5.42, 5.56
> 8	221569	9.93	9.81, 10.1

aOR = adjusted odds ratio. CI = confidence interval. *Note: Visa type, region of origin, and length of time in England were all separate exposures in separate GLM binomial models.