

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 31/10/13
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 31/10/13
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 level¹:

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

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* command 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

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

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.

