Office for National Statistics

Article

Refugee integration outcomes data-linkage pilot : Census 2021 linkage methodology update

Methods used to link Vulnerable Person and Vulnerable Children's Resettlement Scheme data to Census 2021 in a collaborative study by the Office for National Statistics and the Home Office.

Table of contents

1 . Foreword

In support of its work, the Home Office has long been interested in improving the evidence base around integration outcomes for refugees in the UK. The Refugee Integration Outcomes (RIO) cohort study is a novel collaboration between the Home Office and the Office for National Statistics (ONS) which aims to address this need.

Following the successful conclusion of the pilot in June 2022, we extended the study to link both refugees granted asylum and individuals who came to the UK under one of our resettlement schemes with data from Census 2021 and now, in this latest study, with administrative data from other government systems.

Refugee integration is a rich area for research, but finding robust quantitative data on outcomes, particularly over the longer term, is much more challenging. RIO aims to provide data which will begin to unlock unique and unprecedented insights into the longer-term integration outcomes for refugees.

The data sharing between Home Office and ONS, and the collaboration between their researchers, has already begun to improve our understanding and help to ensure better outcomes for refugees as they attempt to integrate into the UK. This report, built from this cross-departmental working and data sharing, demonstrates how this experimental statistical work can generate real insight.

Alongside this methodology paper, we are publishing the first analysis report of the RIO data on refugees resettled through the Vulnerable Persons Resettlement Scheme (VPRS) and Vulnerable Children's Resettlement Scheme (VCRS). This is the second report of the RIO dataset and uses data from both asylum refugees and the mainly Syrian resettled refugees who arrived in England and Wales between 2015 and 2020. It reveals very clearly both differences and similarities between these groups, and implications for further research, which should help guide other researchers on this topic.

The Home Office and ONS plan to expand the project to incorporate more recent cohorts of refugees and link to a wider range of economic, health and education data. This will test the credibility of these findings when applied to other groups of refugees and help expand the range of outcomes we can measure. As the project develops, we will publish further analysis and plan to make the dataset available to other researchers. It is complex research and results require careful interpretation, reflecting the distinct circumstances and experiences of people who are offered protection by the UK.

Jon Simmons

Deputy Director of Immigration System Statistics and Refugees Analysis and Insight (ISSRAI)

Home Office Analysis and Insight (HOAI)

2 . Main points

- Quantitative data on refugees' long-term integration outcomes in the UK is lacking, largely attributed to an absence of datasets which permit refugees to be identified.

- Linking Home Office Vulnerable Persons Resettlement Scheme (VPRS) and Vulnerable Children's Resettlement Scheme (VCRS) data to administrative data collected by other government departments and Census 2021 will help fill this gap and provide answers to questions on this hard-to-reach population.

- The Refugee Integration Outcomes (RIO) cohort study was set up to fill this evidence gap with an early pilot linking administrative data to cohorts of VPRS and VCRS refugees (a cohort is defined as a group of people with a shared characteristic moving forward through time. In this case we use year of arrival to define cohorts).

- The Office for National Statistics (ONS) and the Home Office have continued to work collaboratively on RIO, further linking cohorts of refugees granted asylum and VPRS and VCRS refugees to Census 2021.

- This second iteration of RIO, building on the RIO data linkage pilot, uses a combination of exact, deterministic, and associative matching methods, with clerical review to resolve conflicts to link Census 2021.

- When linking to Census 2021, we achieved linkage rates of 69.6% and 91.2% for refugees granted asylum and VPRS and VCRS refugees respectively; this is in part because the asylum refugees have a greater proportion of younger males who are transient and less likely to engage with services.

- The difference in linkage rates between the two cohorts greatly emphasises that these are two distinct cohorts with different characteristics and behaviours.

- Comparisons of linked and unlinked records show that our linked data are representative of the VPRS and VCRS refugee cohorts.

- We linked slightly more females, younger and older refugees granted asylum, but fewer of Sudanese and Eritrean nationality; young adult males are less likely to engage with services, but also patriarchal and African naming conventions make linkage challenging.

- We will continue to improve our linkage of refugees granted asylum to Census 2021 by adapting our linkage algorithms to take account of geographical dispersion and African naming conventions.

- The analysis presented in this article is experimental and may change as we continue to update linkages.

- This methodology article supports analysis published on 27 June 2023 on early integration outcomes for refugees resettled under the VPRS and VCRS.

- The future aims of the project are to improve linkage rates, most notably for the remaining asylum refugee residuals, explore probabilistic linkage methods, link new administrative datasets to the cohorts and produce and publish analysis for the asylum route cohort.

3 . Overview

One of the [Home Office's (HO) priorities](#) is to "protect vulnerable people and communities". To meet this goal and to develop and evaluate relevant policies, the HO is interested in improving the evidence base around integration outcomes for refugees in the UK.

The Refugee Integration Outcomes (RIO) cohort study is a collaboration between the HO and the Office for National Statistics (ONS). Our [(RIO) data linkage pilot project](#) aimed to link NHS Personal Demographics Service (PDS) data and HO border systems data to cohorts of refugees resettled under the Vulnerable Persons Resettlement Scheme (VPRS) and Vulnerable Children's Resettlement Scheme (VCRS).

Following the successful conclusion of the (RIO) data linkage pilot in June 2022, we extended the study to include refugees granted asylum between 2015 and 2020 and have linked Census 2021 for England and Wales to the VPRS and VCRS, and refugees granted asylum.

The (RIO) cohort study will provide unique insights into the integration outcomes for approximately 113,000 refugees resettled under the VPRS and VCRS or granted asylum in England and Wales between 2015 and 2020.

The RIO cohort study covers England and Wales currently, but there are plans to expand this study to Scotland and Northern Ireland and other humanitarian and protection routes in the future subject to data availability, data supplier agreement, data quality and funding.

ONS are exploring the feasibility of an anonymised person-level longitudinal data source for England and Wales, based on Census 2021 and then updated each year to reflect population change (births, deaths and migration). This new dataset is referred to as the Longitudinal Population Dataset (LPD), formerly known as the Census Data Asset (CDA) (see our [Census 2021 Data Asset longitudinal data source for population in England and Wales methodology](#)). The vision is for RIO to become integrated into the LPD as a satellite cohort study, allowing comparisons with other population groups, such as the general migrant population.

4 . Data sources

There are two main sources of data on refugees which include refugees who arrive in the UK via resettlement schemes overseen by the government, and who have been granted a protection status prior to arrival. The Refugee Integration Outcomes (RIO) cohort study focusses on those resettled under the Vulnerable Persons Resettlement Scheme (VPRS) and Vulnerable Children's Resettlement Scheme (VCRS). The second are those refugees who arrive here in other disorganised or irregular ways, and who subsequently apply for and are granted refugee status through the Asylum Refugee Route (ARR).

VPRS and VCRS data

The RIO cohort study uses data for refugees resettled in England and Wales under the VPRS and VCRS between 2015 and 2020.

This includes 16,350 people resettled under the VPRS and VCRS. Further data for resettled refugees is available in regularly published [Home Office (HO) Asylum and resettlement datasets](#). We refer to these refugees as Resettled Refugees (RR) in this report.

ARR data

The ARR data in RIO contain approximately 97,000 individuals who were granted asylum between 2015 and 2020 in England and Wales. This sample excludes those still awaiting a decision on their asylum claim, or those who were denied asylum. ARR are more challenging to link to other data owing to their more dynamic geographic mobility and greater dispersion across England and Wales. Many asylum refugees are young adults, with approximately one-third being female and two-thirds male. The majority of the ARR population in this study are from Iran, Eritrea, Sudan, Syria, and Afghanistan. Further data for asylum refugees is available in regularly published [Home Office Asylum and resettlement datasets](#).

Census 2021

The census is undertaken by the Office for National Statistics (ONS) every 10 years and is the largest statistical exercise that ONS undertakes, producing statistics that inform all areas of public life and underpin social and economic policy. It asks questions about a person, their household, and their home. In doing so, it provides a wealth of information at small geographies to inform local planning and decision-making. You can find more information on our About census page on the ONS website. The census is a rich data source that includes numerous integration indicators for a point in time, including families and households, housing, education, English and Welsh language proficiency, employment, and health.

HO border systems data

The Home Office border systems data programme was introduced in April 2015 and the (HO) has produced a series of statistical reports on the coverage of these data and its use. It was designed primarily for operational (immigration control) purposes and initially collected data on non-EU nationals departing from and arriving in the UK. As we linked to these datasets previously, more detailed information on how we use the Home Office border systems data can be found in our Refugee Integration Outcomes (RIO) data linkage pilot methodology.
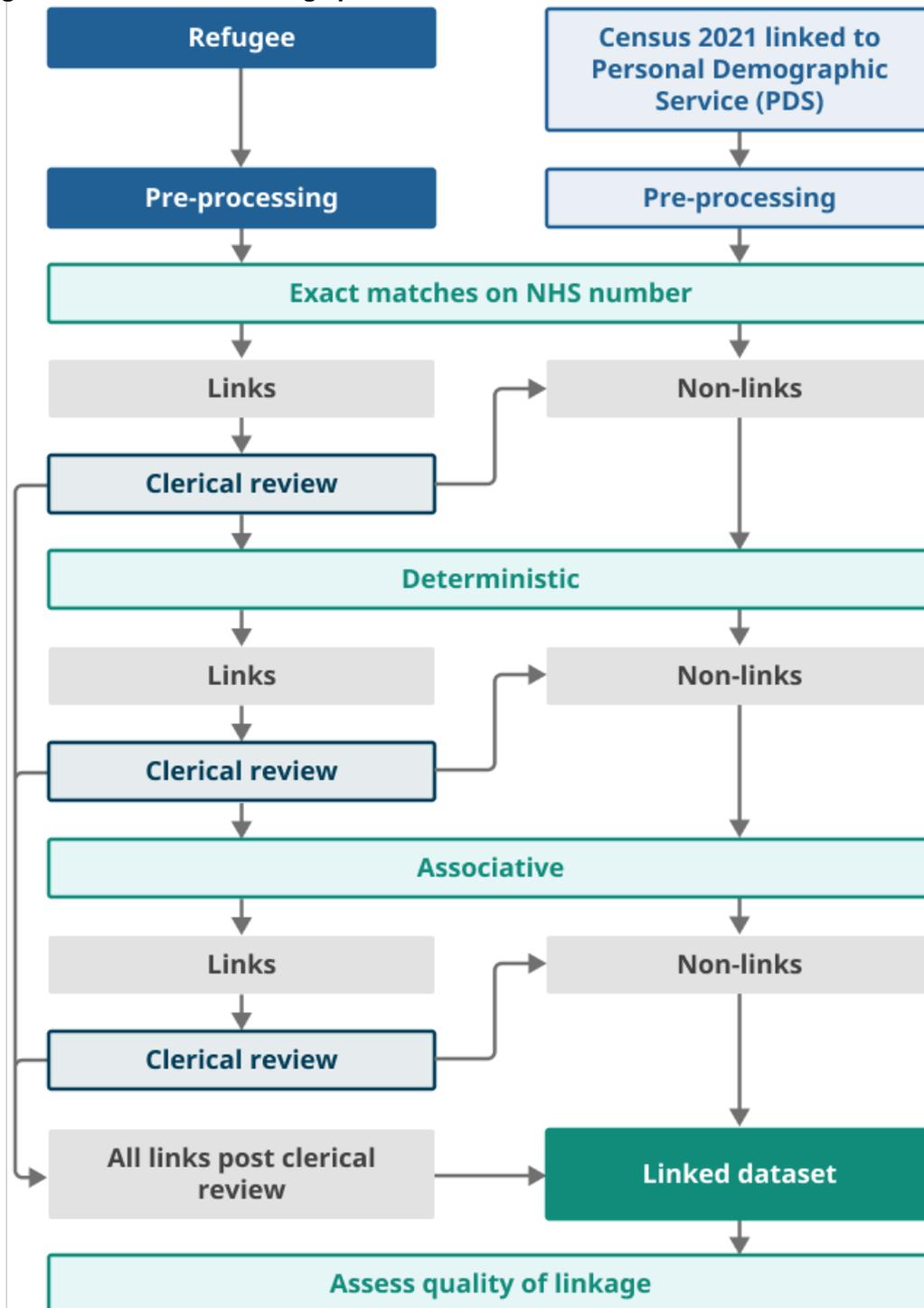
NHS Personal Demographics Service (PDS) data

The NHS Personal Demographics Service (PDS) data hold demographic details of users of health and patient care services in England and Wales. As we linked to these datasets previously, more information on the data we use from the PDS can be found in our Refugee Integration Outcomes (RIO) data linkage pilot methodology.

5 . Linkage methods

Data linkage involves multiple stages. Linkage of Census 2021 to resettled and asylum refugees started with exact matching using a census to NHS Personal Demographics Service (PDS) look-up and linking on NHS number. All non-linked records then went through deterministic and associative linking and all stages were clerically reviewed, building on the methods designed for our Refugee Integration Outcomes (RIO) data linkage pilot, as shown in Figure 1.

Figure 1. Flow chart of linkage process

Matchkeys

Deterministic linkage uses pre-determined rules to decide whether two records belong to the same individual. Combinations of identifying variables such as name, sex, date of birth, postcode and nationality are combined in different ways to create a series of "matchkeys" and used to identify matching records between the datasets. More complex deterministic methods include using partial identifiers within the matchkey series, for example, postcode sector or two or more common names. We also use the Levenshtein edit distance, which allows us to adjust the number of edits needed to be made to a name to match another record. The Levenshtein distance measures the difference between two words. These differences can be insertions, deletions or substitutions required to change one word into the other.

Matchkeys used in our Refugee Integration Outcomes (RIO) data linkage pilot study were developed specifically to link refugees resettled under the Vulnerable Person Resettlement Scheme (VPRS) and Vulnerable Children's Resettlement Scheme (VCRS) to two administrative datasets. We refer to these as Resettled Refugees (RR) in this report. These needed to be adapted to link the RR and Asylum Refugee Route (ARR) data to Census 2021. Further detail on deterministic and associative matching methods can be found in our Refugee Integration Outcomes (RIO) data linkage pilot methodology.

A substantial proportion of links were made at the exact matching stage using NHS number (92.6% for RR and 83.7% for ARR). See Section 6: Data linkage results and quality.

A full set of matchkeys used are available upon request at our email ONS.RIO.Cohort.Study@ons.gov.uk.

Clerical review

Clerical review is an additional approach which can supplement automated matching processes. It uses human decision-making to determine if the link between two records is a true match. Clerical review is time and resource intensive and can also be costly. It should therefore be targeted where it is most needed and used with other person attributes to determine a match. For example, where names follow patriarchal naming conventions, common names, and transliteration issues (in Arabic, the name Mohammed can potentially be spelt in various ways).

Limited clerical review was undertaken during our [Refugee Integration Outcomes (RIO) data linkage pilot](#). However, in the pilot, automated data matching processes were found to be insufficient on their own in providing satisfactory levels of data matching. As a result, a clerical matching team was recruited to review clerical results manually from all linkage stages. Either all records were clerically reviewed or stratified samples were taken if a large number of potential matches were made. Clerical matchers were given additional variables and extra resources on naming conventions to inform decisions.

We treat the data that we hold with respect, keeping it secure and confidential, and ensure that we comply with all relevant legislation, including the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018. Personal data are only processed when necessary for statistical purposes. Only the minimum amount of personal data required for data linkage are used, and we only link the necessary data and variables needed to answer research questions.

Table 1a and Table 1b show the number of potential links sent for clerical review and the number and proportion accepted as true matches for ARR and RRs, respectively. For ARR, 52.8% of potential links were accepted as true matches and for RR, this was 76.2%.

Samples of exact matches made on NHS number are clerically reviewed to ensure we were linking the same person. The Office for National Statistics (ONS) Data Linkage Hub linked Census 2021 and NHS Personal Demographics Service (PDS) data. This was done separately from the linkage of refugee and PDS data.

Deterministic and associative linkage use clerical matching to establish whether a potential match is in fact a true match. Associative matching includes linking individuals by collectively resolving matched records within a household. This is done by first matching households based on household-level variables (for example, postcode) before matching individuals within households. Where there is agreement between two records on a matchkey, a link is established. Matches made on matchkey one with no conflicts are considered "exact matches". However, if a refugee record linked to more than one census record, it is considered a "conflict". These conflicts are then sent for clerical review.

Insights from clerical matching are used to train and refine automatching to improve match rates. Clerical matching is critical to maintaining a rigorous approach to data linkage that supports the longitudinal integrity of RIO (for instance, we are following the same individual over time). Further detail on the clerical review process is available in our [Refugee Integration Outcomes (RIO) data linkage pilot methodology](#).

Table 1a: Clerical review results for asylum route refugees, England and Wales, 2015 to 2020

ARR

Linkage Stage	Sent to Clerical (Number)	Accepted Matches (Number)	Accepted Matches (Percentage)
Deterministic	18,476	10,153	55.0
Associative	1,664	487	29.3
Total	20,140	10,640	52.8

Source: Analysis of linked Census 2021 to ARR data, 2015 to 2020 from the Office for National Statistics

Notes

1. Deterministic includes both exact matches and conflicts.

2. Associative matching does not produce "exact matches", only conflicts.

Table 1b: Clerical review results for resettled refugees, England and Wales, 2015 to 2020

RR

Linkage Stage	Sent to Clerical (Number)	Accepted Matches (Number)	Accepted Matches (Percentage)
Deterministic	1,285	1,027	79.9
Associative	137	57	41.6
Total	1,422	1,084	76.2

Source: Analysis of linked Census 2021 to VPRS and VCRS data, 2015 to 2020 from the Office for National Statistics

Notes

1. Deterministic includes both exact matches and conflicts.

2. Associative matching does not produce "exact matches", only conflicts.

6 . Data linkage results and quality

Overview of linkage results

We present summarised results for linking refugee data to Census 2021 in Table 2. These are the overall linkage rates after accounting for deaths and other reasons for removal from the dataset.

Figure 2 shows that high linkage rates (91.2%) are achieved for Resettled Refugees (RR) but are lower for the Asylum Refugee Route (ARR) at 69.6%. Differences in linkage rates between the two cohorts emphasise how dissimilar these cohorts are, demographically and behaviourally. Therefore, these two groups should be analysed separately in research studies.

For instance, prior to arrival in the UK, and through the process of arriving in the UK, support arrangements are in place during resettlement, such as providing accommodation and registration for health services. There is, therefore, a large degree of interaction with public services and official records, meaning the majority of links can be made from exact matching on NHS number using Census 2021 linked to NHS Personal Demographics Service (PDS) data.

Another example is that engagement with health services is lower for asylum cohorts, particularly so amongst younger men. As a cohort, they are more likely to move around the country, meaning that NHS registrations may not be kept up to date or even additional new registrations are made.

Figure 2: The RR had notably higher linkage rates to Census 2021 for both males and females, ARR males had the lowest linkage rate

Linked versus unlinked comparisons of resettled and asylum refugees by sex, England and Wales, Census 2021

Figure 2: The RR had notably higher linkage rates to Census 2021 for both males and females, ARR males had the lowest linkage rate

Linked versus unlinked comparisons of resettled and asylum refugees by sex, England and Wales, Census 2021

Source: Analysis of linked and unlinked Census 2021 data from the Office for National Statistics

We set out our plans to improve linkage rates for linking Census 2021 to ARR cohorts in .

Table 2: Summary of linkage rates, overall, England and Wales, Census 2021

RIO Entry Route	Number of refugee records linked (Number)	Linkage rate (Percentage)	Residuals accounted for (Number)	Remaining unlinked records (Number)
Resettled Refugees	14,674	91.2	169	1,421
Asylum Refugees	65,137	69.6	3,508	28,429

Source: Analysis of linked VPRS, VCRS and ARR data to Census 2021 from the Office for National Statistics

Notes

1. Residuals accounted for include records linked to death registrations or flagged as emigrated moved from England and Wales to Scotland or Northern Ireland, or left for another reason.

2. These records are not included in the linkage rate calculations.

Summary of linkage rates by linkage stage

Table 3 shows that a substantial proportion of links were made at the exact matching stage using NHS number. This suggests that most of the records with high quality name, date of birth and postcode variables were previously linked to NHS data during our Refugee Integration Outcomes (RIO) data linkage pilot for RR and linkage of NHS Personal Demographic Service (PDS) data to ARR. The linkage rates exclude "residuals accounted for" because a refugee had either died, emigrated from the UK or was known to have gone away. Table 4 provides further analysis of residuals.

Table 3: Linkage rates to Census 2021 by linkage stage and residuals explained, England and Wales, 2015 to 2020

	Linkage stage	Count of Records (Number)	Linkage Rate (percentages)
Resettled refugees (RR)			
	Exact using NHS number	13,590	84.4
	Deterministic	1,027	6.4
	Associative	57	0.4
	Total linked	14,674	91.2
Asylum refugees (ARR)			
	Exact using NHS number	54,498	58.2
	Deterministic	10,152	10.9
	Associative	487	0.5
	Total linked	65,137	69.6

Source: Analysis of linked Census 2021 to VPRS and VCRS and ARR data, 2015 to 2020 from the Office for National Statistics

Notes

1. Some numbers have been removed rounded or aggregated to prevent disclosure of individuals.

Analysis of residuals

There are several reasons why some refugee records did not link to Census 2021, outlined in our Refugee Integration Outcomes (RIO) data linkage pilot project methodology.

For example:

- missed matches in the linkage (records that did not link but should have)

- moved from England and Wales to another UK country before or soon after 21 March 2021 (Census Day): a cross-border move

- left UK before or soon after 21 March 2021: an emigration event

- died before or soon after 21 March 2021

- did not complete a census form (census non-response)

Regarding the final point, there was no actual census record to link to. However, final Census 2021 published outputs will include coverage estimation to account for non-response.

We can explain census linkage failure for 10.6% (169) of RR and 11.0% (3,508) of ARR residual records, shown in Table 3. This analysis is based on information from linked death registrations and NHS Personal Demographics Service (PDS) data. PDS data record emigrations, moves to Scotland and Northern Ireland, deaths, and other reasons for flagging an NHS record for removal. PDS may record lower numbers of emigrations and cross border moves as an individual would have to notify a health care professional that they are moving country.

Further analysis of linked Home Office (HO) border systems data will give us more information on how many have left and returned or how many have left but not since returned. Resettled refugees are granted indefinite leave to remain and refugee status on arrival in the UK. They are able to travel outside the UK but may lose their indefinite leave to remain if they travel back to the country they sought asylum from or stay outside the UK for two or more continuous years.

Table 4: Reason for removal from NHS Personal Demographics Service (PDS) data, England and Wales, 2015 to 2020

Reason for removal	RR Count (Number)	ARR Count (Number)
PDS Death and linked to registered deaths	22	50
Linked to registered deaths only	49	243
PDS Death only	23	60
Emigrated from UK	14	127
Cross border move to Scotland or Northern Ireland	73	273
Removed/Left for other reason	21	2,855
Residual explained	169	3,508

Source: Analysis of linked death registrations and NHS Personal Demographics Service (PDS) data to VPRS, VCRS and ARR data, 2015 to 2020 from the Office for National Statistics

Notes

1. Left for other reason includes practice request, gone away, address not known, cancelled, address out of practice area, practice request immediate removal and other reason.

2. Cross border moves contain moves to Northern Ireland or Scotland.

3. Emigrated is defined as leaving the United Kingdom.

Comparison of unlinked and linked records

To explore biases in the linked data, we compare the characteristics of linked and unlinked records in Figure 2 and Figures 3 to 5.

RRs have equally high linkage rates for males and females and negligible bias by age, sex and nationality. For ARR, males and people aged 18 to 44 years have lower linkage rates. Considering the transient nature of ARR and, as reported in our Quality assurance of administrative data used in population statistics methodology, younger men are less likely to engage with NHS services, which will have contributed to lower linkage rates.
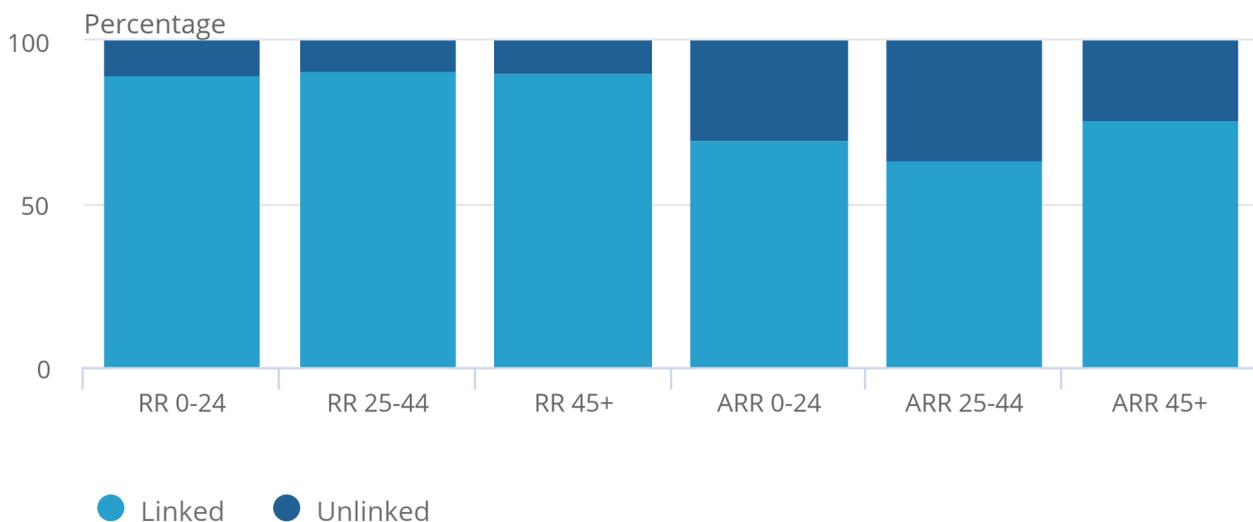
Eritrean and Sudanese refugees have lower linkage rates than other nationalities. These countries' cultures typically follow patriarchal naming conventions, with second name tending to be father's forename. The Eritrean main language, Tigrinya, is a Semitic language which may lead to more mistakes when recording information in English, lowering the chances of making a successful link. Further development of matching algorithms to support African naming conventions as well as more resources, around languages and naming conventions given to our clerical team may improve linkage results.

Figure 3: Asylum refugees aged 18 to 44 years are the hardest to link

Linked versus unlinked comparisons of resettled and asylum refugees by broad age group, England and Wales, Census 2021

Figure 3: Asylum refugees aged 18 to 44 years are the hardest to link

Linked versus unlinked comparisons of resettled and asylum refugees by broad age group, England and Wales, Census 2021
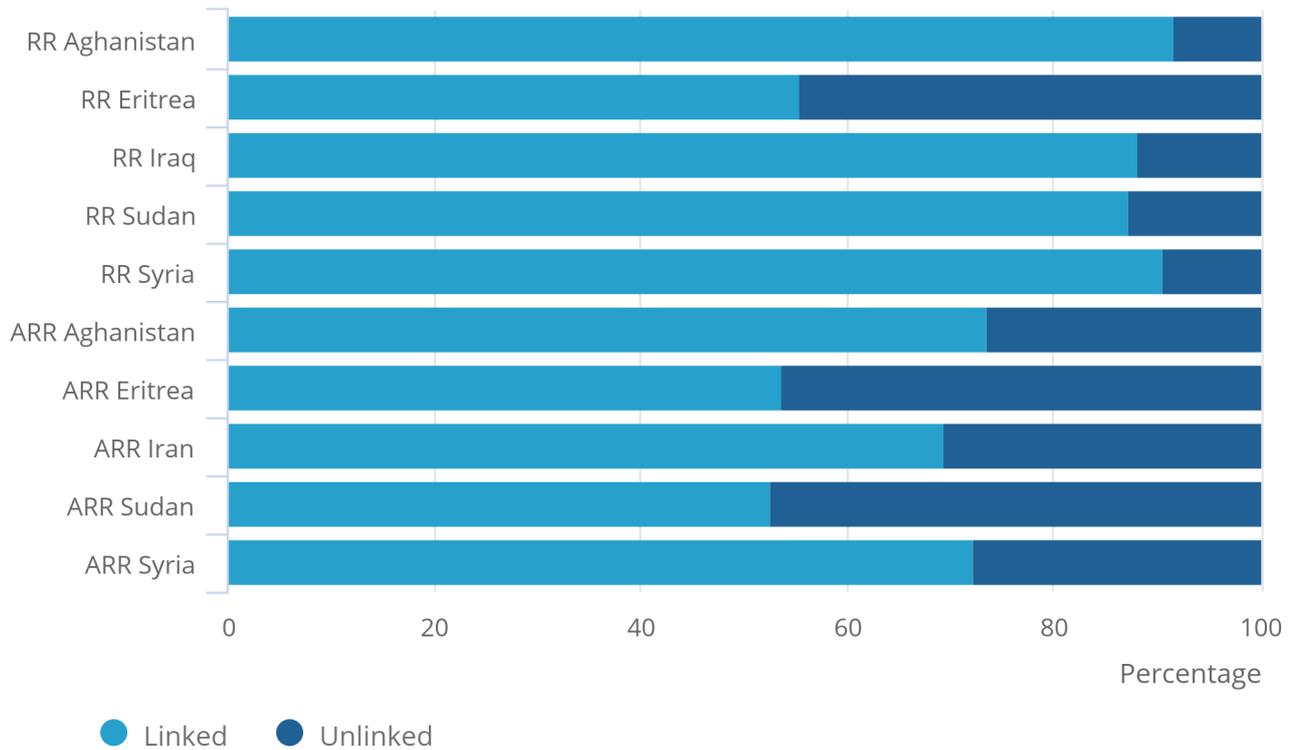
Source: Analysis of linked and unlinked Census 2021 data from the Office for National Statistics

Figure 4: Lower linkage rates achieved for Eritrean and Sudanese asylum refugees

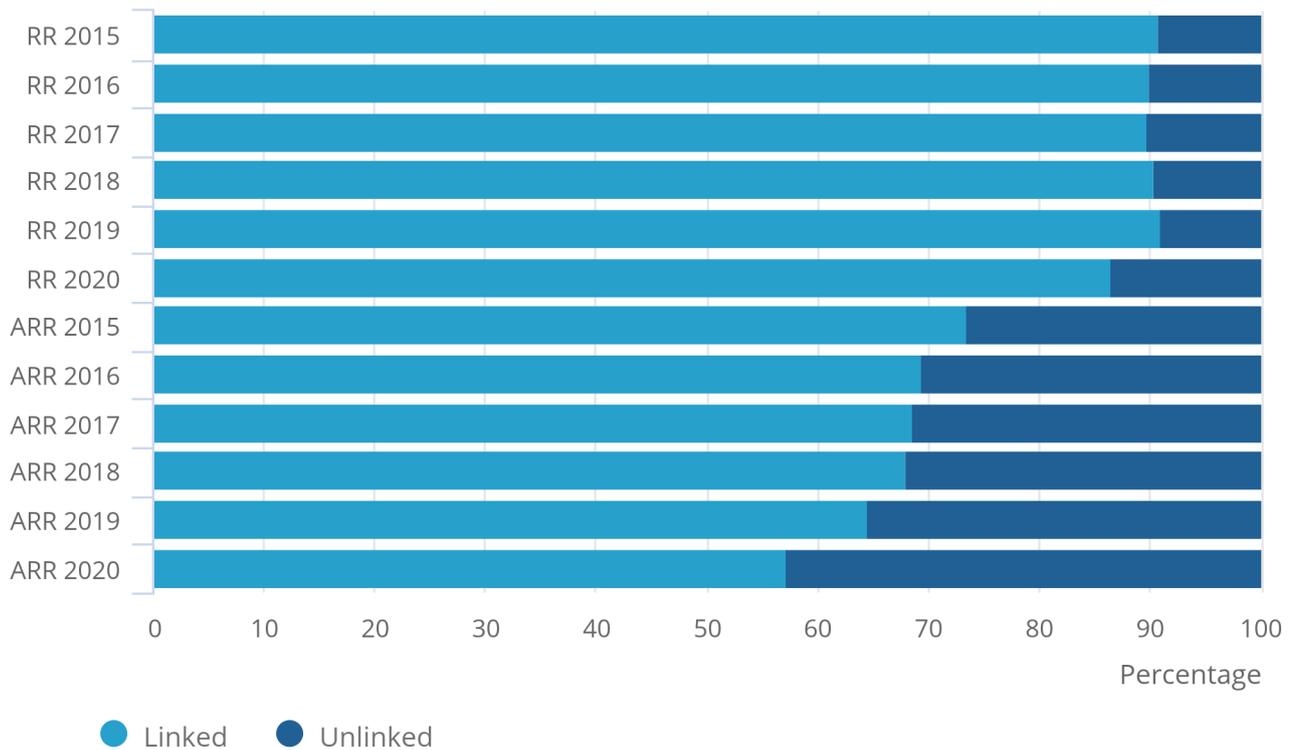Linked versus unlinked comparisons of resettled and asylum refugees by nationality, England and Wales, Census 2021

Source: Analysis of linked and unlinked Census 2021 data from the Office for National Statistics

Figure 5: Lower linkage rates achieved for later cohorts may be due to lags in registering with health services

Linked versus unlinked comparisons of resettled and asylum refugees by year of arrival or grant of asylum, Census 2021

Figure 5: Lower linkage rates achieved for later cohorts may be due to lags in registering with health services

Linked versus unlinked comparisons of resettled and asylum refugees by year of arrival or grant of asylum, Census 2021

Source: Office for National Statistics - analysis of linked and unlinked Census 2021 data

7 . Discussion and next steps

Following the success of our Refugee Integration Outcomes (RIO) data linkage pilot, we have extended study coverage to include cohorts of asylum refugees and linked Census 2021. High-quality linkage of resettled refugees to Census 2021 has enabled Home Office (HO) and the Office for National Statistics (ONS) to publish, for the first-time, analysis based on experimental data on the early outcomes of refugees resettled under the Vulnerable Persons Resettlement Scheme (VPRS) and Vulnerable Children's Resettlement Scheme (VCRS) who arrived in England and Wales between 2015 and 2020.

Asylum refugee data are more challenging to link because of the diversity in this group and being predominantly younger males, more transient, and also less likely to engage with official surveys and government services, as reported in our Quality assurance of administrative data used in population statistics methodology. Lower achieved linkage rates for asylum refugees in this study have strengthened this hypothesis. Initial analysis of residual records (not linked to Census 2021) shows that 3.6% of asylum refugees left or died. Further analysis of HO border systems data may shed further light on refugees' emigration from the UK since grant of asylum. Other possible reasons for not linking administrative data are outlined in our Refugee Integration Outcomes (RIO) data linkage pilot methodology.

We will look to improve linkage rates for the remaining asylum refugee residuals that we could not link to NHS Personal Demographics Service (PDS) data, Census 2021 and HO border systems data. Our focus will be improving linkage on Sudanese and Eritrean names. We will also explore probabilistic linkage methods to expand the number of possible matches sent for clerical review and to expand clerical search to a wider geographical area. This is currently limited to within-postcode.

Future updates on RIO linkage will be published as we link additional administrative data. Our ambition is to also extend the analysis for refugees resettled under the VPRS and VCRS to refugees granted asylum.

8 . Related links

Refugee Integration Outcomes (RIO) data linkage pilot
Methodology | Revised 15 June 2022
Methods used to link refugee data to various administrative data in a pilot study led by the Office for National Statistics (ONS) and Home Office.

Qualitative evaluation of the Vulnerable Persons Resettlement Scheme (VPRS) and the Vulnerable Children's Resettlement Scheme (VCRS)
Web page | Released 30 March 2023
Outputs from a 3-year qualitative research project conducted by Ipsos to evaluate the Vulnerable Persons and Vulnerable Children's Resettlement Schemes.

Asylum and resettlement datasets
Dataset | Last updated 25 May 2023
Listing of the detailed asylum and resettlement datasets.

Early integration outcomes for refugees resettled in England and Wales: 2015 to 2021
Article | Released 27 June 2023
A collaboration between the Home Office and the Office for National Statistics providing new analysis on refugees resettled between 2015 and 2020 under the Vulnerable Persons and Vulnerable Children's Resettlement Schemes in England and Wales.

9 . Cite this methodology

Office for National Statistics (ONS), released 27 June 2023, ONS website, methodology article, Refugee integration outcomes data-linkage pilot: Census 2021 linkage methodology