Refugee Integration Outcomes (RIO) data linkage pilot

Methods used to link refugee data to various administrative data in a pilot study led by the Office for National Statistics (ONS) and the Home Office.

Contact:
Nicky Rogers and Gemma Hanson
demographic.methods@ons.gov.uk
+44 1329 444866 and +44 1329 447330

Release date:
15 June 2022

Next release:
To be announced

Table of contents

1 . Main points

- Quantitative data on refugees' long-term integration outcomes in the UK is lacking, and this is largely attributed to a lack of datasets which permit refugees to be identified; linking Home Office refugee data to administrative data collected by other government departments would help to fill this gap and analysis of these linked data could answer questions on this hard-to-reach population.
- The Office for National Statistics (ONS) and the Home Office set up a pilot project to link a sample of refugees who arrived via the Vulnerable Persons and Vulnerable Children's resettlement schemes.
- The pilot aimed to test the feasibility of linking refugee data to two administrative data sources: Home Office Exit Checks and NHS Personal Demographic Service (PDS) data, as we know that these data sources are likely to include refugees.
- The pilot helped us develop specific linkage algorithms based around Arabic naming conventions, as well as using associative matching methods to draw strength from data on family units.
- We achieved match rates of 96% and 97% for PDS and Exit Checks respectively.
- Allowing for variations in name spellings helped us to deal with possible transliteration issues of names that may occur between Home Office refugee data and other administrative datasets, thereby improving match rates.

2 . Overview

This methodology details new innovative data linkage methods developed through linkage of refugee data to administrative data. The Office for National Statistics (ONS) and the Home Office set up a pilot project to link a sample of resettled refugees. We have worked closely with Home Office experts to improve our understanding of these data and to adapt linkage methods to deal with different naming conventions. We report our methods and linkage results from the pilot study in this methodology.

The Home Office is the lead government department for immigration and passports, drugs policy, crime, fire, counterterrorism and police. One of the Home Office's priorities is to "protect vulnerable people and communities". To meet this goal and to develop and evaluate relevant policies, the Home Office is interested in producing the evidence base around integration outcomes for refugees in the UK.

It is widely accepted that quantitative data on refugee outcomes, particularly over the longer term, is lacking. This is largely because of a lack of datasets which permit refugees to be identified (Ruiz and Vargas-Silva, 2018). Linking Home Office refugee data to administrative data collected by other government departments would help to fill this gap and analysis of these linked data could answer questions on this hard-to-reach population.

It is the ONS' mission to provide the best insights on population and migration. We do this by working with other government departments and using a range of new and existing data sources to meet the needs of our users. This is increasingly important in a rapidly changing policy and societal context, where we know our users need better evidence to support decision making at both national and local levels.

As part of the ONS' commitments to inclusive data, we want to ensure that "our statistics reflect the experiences of everyone in our society so that everyone counts, and is counted, and no one is forgotten" (Statistics for the Public Good, 2020). Linking Home Office refugee data with other administrative data sources will ultimately help inform local authorities, government, charities and other organisations with resource allocation for these vulnerable populations. It also has the potential to increase public awareness of societal issues.

The Vulnerable Persons and Vulnerable Children's resettlement schemes (VPRS and VCRS) were established by the UK government to resettle vulnerable adults and children. The VPRS was launched in 2014 for those in greatest need, including people requiring urgent medical treatment, survivors of violence and torture, and women and children at risk. The scheme aimed to resettle 20,000 people fleeing the conflict in Syria by March 2020.

The VCRS was launched in 2016 with the aim of resettling up to 3,000 at-risk children and their families from the Middle East and North Africa (MENA) region. The region consists of 19 countries including:

- Algeria
- Bahrain
- Egypt
- Iran
- Iraq
- Israel
- Jordan
- Kuwait
- Lebanon
- Libya
- Morocco
- Oman
- Palestine
- Qatar
- Saudi Arabia
- Syria
- Tunisia
- United Arab Emirates
- Yemen

By the time the schemes closed in February 2021, the total number of individuals resettled through the VPRS was 20,319, with a further 1,838 resettled through the VCRS.

For both schemes, refugees were identified and referred to the Home Office by the United Nations Refugee Agency (UNHCR). The process involved collecting identifying information and other supporting data for the Home Office's caseworking system. As part of its evaluation of VPRS and VCRS, the Home Office also collects data relating to a range of early integration outcomes through local authorities and community sponsor groups. They do this at two timepoints, once resettled refugees have been in the UK for approximately 6 and 12 months. Data on integration outcomes have not been collected beyond this point, as intensive caseworker support and contact with refugees steps down over time. The Home Office wanted to minimise the administrative burden placed on local authorities and community sponsor groups.

For the pilot study, we linked the refugee data for those resettled in England and Wales to two administrative data sources: Home Office Exit Checks and NHS Personal Demographic Service (PDS) data.

Since the aim of the pilot was to test the feasibility of linking refugee data to administrative data, we decided to link to these two data sources as we know they are likely to include refugees. For example, upon arrival to the UK, refugees' entry is recorded in Exit Checks data. Also, as a condition of participating in the VPRS and VCRS schemes, local authorities and community sponsor groups are required to ensure refugees are registered with a GP shortly after arrival, and so should be present in the PDS.

We faced unique challenges in the use of linked administrative data in the pilot. For example, bias from the linkage errors where records cannot be linked or are linked together incorrectly (Harron et al., 2017). This can be particularly challenging where unique identifiers for linkage across data sources are not available and there is reliance on name, address, sex or gender, date of birth and even nationality to link records together.

In addition, a specific challenge we faced in this study is the treatment of different naming conventions. Names can be a highly discriminative variable in data linkage. However, the number of different ways names can be structured can be problematic and we need to understand this and develop algorithms to optimise linkage. Missed linkages can result in bias if subgroups of records are more or less likely to be linked (Bohensky et al., 2010; Ford et al., 2006; Lariscy, 2011).

Typically, algorithms around record linkage are designed primarily with English language or Western naming conventions in mind. Therefore, certain naming conventions in different languages make it harder for algorithms to correctly identify a match. For example, refugees may be addressed by informal titles that would not commonly be formalised in administrative data sources. Furthermore, there can be significant variation in the way names originally in non-Latin scripts are transliterated into Latin script. For example, in Arabic, the name Muhammad, when transliterated into Latin script, can potentially be spelt in various ways, including Mohammed, Mohamad and Mohamed. It is therefore possible that the same name may be written in different ways on various official documents, and in different administrative datasets.

In this article, we highlight the methodology used to optimise linkage algorithms for Arabic naming conventions to successfully link refugee data to administrative data.

3 . Data sources

Vulnerable Persons (VPRS) and Vulnerable Children’s (VCRS) Resettlement Scheme data

The pilot uses data for refugees resettled in the UK as of June 2020. This includes 19,755 resettled under the VPRS, and 1,826 resettled under the VCRS.

The Home Office refugee dataset consists of:

- data collected by the United Nations Refugee Agency (UNHCR) and provided to the Home Office as part of the referral process
- additional caseworking data collected by the Home Office as part of the resettlement process
- data collected from local authorities and community sponsors in the UK as part of the Home Office’s evaluation of the schemes

The UNHCR data are collected in the host country before refugees depart for the UK. Examples of the data collected, which have been used for this pilot include:

- full name
- date of birth
- nationality
- gender
- relationship between individuals within a family group

The UNHCR data are entered into the Home Office caseworking system, which is added to data collected as part of the resettlement process, such as arrival date, and the local authority or community sponsor group that the refugee has been resettled to.

Lastly, data collected by the Home Office from local authorities and community sponsors for evaluation purposes includes:

- postcode
- changes in family composition
- economic status
- benefits claimed
- enrolment in English language training (ESOL) and education
- services accessed

These variables are based on the Home Office Indicators of Integration framework 2019, which was developed as a means of conceptualising integration across a range of different domains.

For the purposes of this methodology, we refer to these data together as “Refugee” data.

Exit Checks data

The Home Office Exit Checks programme was introduced in April 2015. It was designed primarily for operational (immigration control) purposes and initially collected data on non-EU nationals departing from and arriving in the UK. The data are a linked database that combines data from Home Office systems. They build event histories that consist of an individual’s travel in and out of the country, together with data relating to immigration status (for example, type and periods of leave granted indicated on a traveller’s visa). These combined data are used by the Home Office for operational and security purposes but might also have statistical benefits.

The Exit Checks data have coverage of the UK but exclude entries and exits within the Common Travel Area (CTA). The CTA is an administrative arrangement between the UK, Ireland and the Crown Dependencies (Isle of Man, Guernsey and Jersey), which is implemented in UK domestic law in statute. Under the CTA, British and Irish citizens can move freely and reside in either jurisdiction and enjoy associated rights and privileges. These include the right to work, study and vote in certain elections, as well as to access social welfare benefits and health services.

NHS Personal Demographic Services (PDS) data

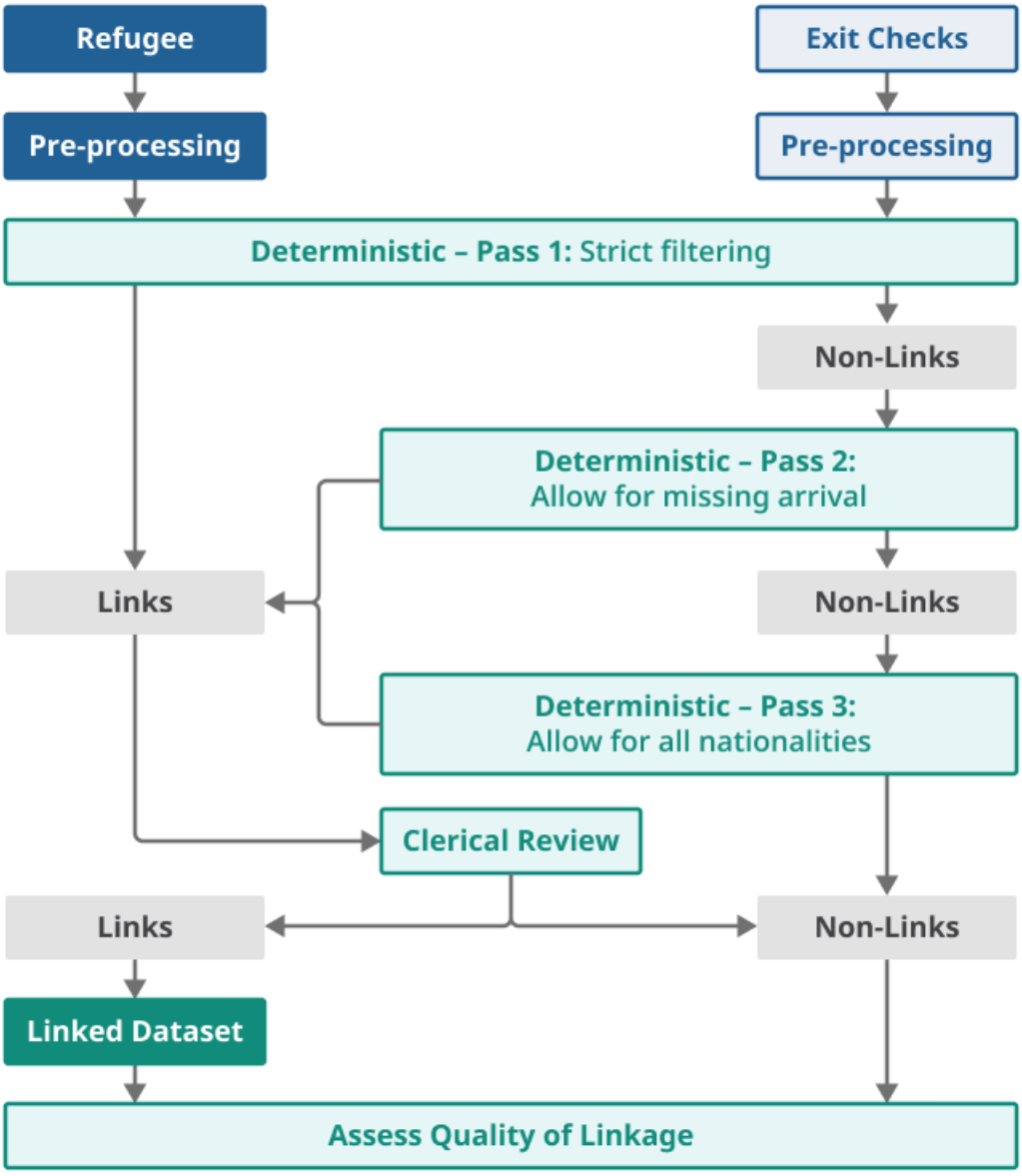
The NHS Personal Demographic Services (PDS) holds demographic details of users of health and patient care services in England and Wales. It captures data on NHS patient details such as:

- name
- address
- date of birth
- NHS number

4 . Linkage methods

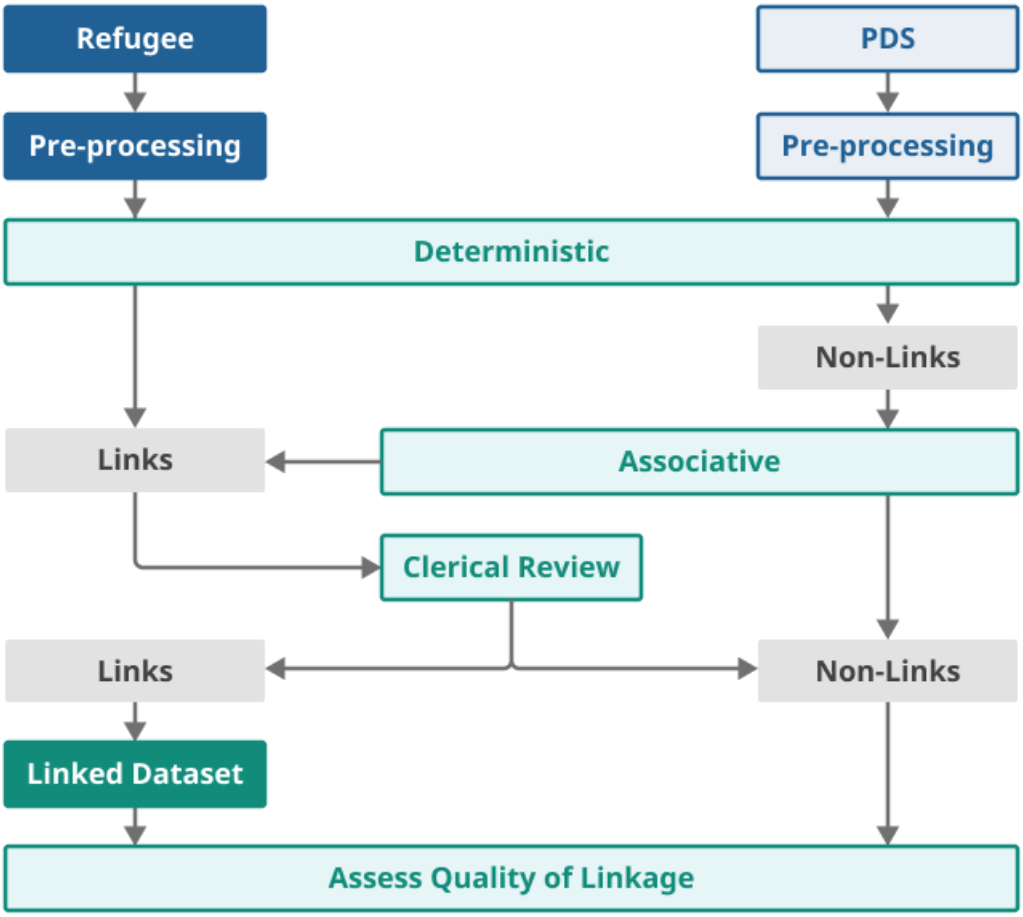
Data linkage involves multiple stages of linkage. There are several standard tools that are used for data linkage. Typically, these include deterministic, probabilistic, associative, and clerical. The aim of the pilot was to develop these traditional methods to optimise linkage for naming conventions across a wider set of cultures, as well as minority groups that have been found hard to link in administrative data. We focus on deterministic, associative, and clerical matching in our approach. Figures 1a and 1b illustrate this process for linking Exit Checks and Personal Demographic Services (PDS) data to the refugee data.

Figure 1a: Main steps in linkage to Exit Checks



Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Figure 1b: Main steps in linkage to Personal Demographics Service data



Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

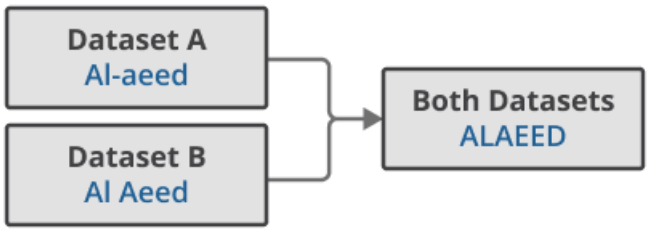
In the next sections we describe the pre-processing and linkage steps in more detail.

Data preparation

It is commonplace in data linkage for the datasets to contain inconsistent, inaccurate, or incomplete data. Often the datasets differ in their structure, format and content, and subsequently vital, but time consuming, pre-processing is required (Harron et al., 2017; Playford et al., 2016). To ensure as much consistency as possible between datasets there is some standard pre-processing that is required to be completed for all datasets. We summarise standard processes applied across all the datasets including Refugee, Exit Checks and PDS data.

Firstly, names are standardised across each dataset to optimise linkage outcomes. For example, where a name appears as “Al-aeed” on one dataset and “Al Aeed” on another, the removal of non-alphanumeric characters and full capitalisation will result in “ALAEED” on both datasets (Figure 2).

Figure 2: Name standardisation



Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Postcodes are standardised in a very similar way with spaces removed, and dates of birth are formatted DDMMYYYY. The variables recording sex on both datasets are recorded to the same coding scheme. We ensure that nationality codes are standardised to the same code on each dataset.

In addition to standardising the data, we derive variables to assist with linkage. Depending on the data, an individual’s name can be recorded as a single variable in one dataset but can be split into separate variables in another. Similarly, an individual might include all elements of their name in one dataset but omit their middle names in another. It is for this reason that we created a series of variables that separate each element of name.

Following a workshop with a Home Office linguist expert in Arabic naming conventions, we identified that it is commonplace for individuals to include informal titles in their name. These titles may not be present on formal documentation but may be included within self-reported datasets such as the PDS data. Therefore, we created a variable that removes informal titles (for example, Abu, Abou, Sheikh, Shaikh, Sharif, Sharifah, Hajj, Hajji, Hadj, Hadji, Hadja).

Other than the derived variables for name we also separate each component of birth and arrival dates, as well as creating postcode area (PO), district (PO15), and sector (PO15 5) from postcode.

Filter rules applied to Exit Checks data

As we progress through the linkage, we loosen the filtering to optimise additional linkages resulting in the Exit Checks data to be filtered in three different ways. With this approach however, potential error could be introduced through loosening the filtering. We compensate for this by quality assuring the matches through clerical review of candidate pairs. In addition to this, the Exit Checks data were filtered to align closely with the time period refugees arrived in the UK (2015 to 2020).

Filter one

The first filter removes all missing arrival dates, keeps only visa types that refer to refugee resettlement and filters on specific nationalities. We start with a more stringent filter to reduce the chance of false positives.

Filter two

The second filter allows the arrival date to be missing or recorded as “null” and includes all visa types and nationalities. This maximises matches where the arrival date is potentially missing.

Filter three

The third filter removes all missing arrival dates but allows all possible visa types and nationalities to be included. This compensates for potential data missingness or error that may prevent a match being made.

Deterministic linkage

Deterministic linkage uses pre-determined rules to decide whether two records belong to the same individual. The refugee data do not contain a universal unique identifier such as NHS number or National Insurance Number (NINo). Therefore, emphasis is placed on other identifying variables such as name, gender, date of birth, postcode and nationality.

These identifying variables are combined in different ways to create a series of “matchkeys” and used to identify matching records between the two 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.

These matchkeys are unique keys, which aim to eliminate some of the discrepancies in data that might otherwise prevent a match. Matchkeys are ordered and applied by strength to find the best quality matches first. Matchkey 1 is an exact match and therefore considered to be the best quality match. This is followed by the remaining matchkeys run in the order shown in the tables in Appendix 1. The deterministic linkage to Exit Checks was run in three separate passes based on each filter rule applied.

Where there is agreement between two records on a matchkey, a link is established. The links must be unique within a matchkey for them to be accepted. For example, if a refugee record was found to match to two different Exit Checks records it would not be linked.

The refugee record would then be compared against subsequent matchkeys until a unique match is found. If multiple matches were made with “A” matching to “B” on one matchkey, and then also “A” to “B” matching on another matchkey, then the first matchkey the match was made on will be taken as true with the later match dropped.

For example, if there is a match established on matchkeys 2 and 4, then matchkey 2 would be taken as true. The reason for taking the first matchkey as the true match, is that the matchkeys are ordered with the strictest matchkeys occurring first and therefore this will be a better-quality match. If, however, “A” matches to “B” on one matchkey but then later “A” matches to “C” on a different matchkey, the match is flagged as conflicting and requires clerical review.

Associative matching

Associative matching is 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. An important factor within the refugee cohort is that we know that resettled refugees often travel together, so to help optimise match rates we take strength from family units to produce potential candidate pairs. These potential pairs are then sent for clerical review.

We use associative matching to link any remaining residuals in the refugee data following deterministic linkage to the PDS. The purpose of using an associative matching approach is to take advantage of family units, which can help to produce potential candidate pairs that are then sent for clerical review. To generate potential pairs, matches are grouped by United Nations Refugee Agency (UNHCR) reference number (family ID number) in the refugee data and postcode in the PDS. Initially, we join refugee residuals by UNHCR reference number and then join PDS residuals by postcode, before filtering by surname and then middle names.

Clerical review

Clerical review is a gold standard approach to data linkage as it allows for human decision about the status of a link for each individual pair. However, it is resource intensive and therefore often not a practical method for linking data. Nonetheless, clerical review remains a useful tool when estimating the precision of matches. Following the deterministic methods, a clerical review of matched records allows us to check for false positive matches (when two records have incorrectly matched) or false negative matches (when two records that should have matched did not).

The matches made on matchkey 1 are considered true matches because of the full match of distinctive person attributes in each dataset. However, to increase the overall match rate, later matchkeys contain criteria where the full data do not need to match exactly. This inevitably leads to some false matches. A clerical check is conducted for matches based on matchkeys 2 onwards. Each match is judged to be either a TRUE match where it is certain the data have matched correctly, or a FALSE match where it is certain the data have matched incorrectly.

Where there are fewer than 100 matches made on a matchkey, a clerical check is carried out on all the matches found and a percentage calculated. Where there were over 100 matches made on a matchkey, a clerical check is taken on a sample of the matches made for that matchkey. The sample size was determined by how strict the matchkey was, with the looser the matchkey the more records needing to be reviewed.

Table 1: Example of paired records for clerical review

Matchkey	Name	DOB	Sex	Postcode
2	MARKJONES	27/03/1961	M	NP108XG
	MARKJONES	28/03/1961	M	NP109XG
7	JANETAYLOR	21/01/1980	F	SN15TH
	JAYNEJONES	21/01/1980	F	SN15RR

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Notes

1. These are dummy data for illustration purposes only.

5 . Results of data linkage

We present summarised results for linking refugee data to Home Office Exit Checks data and NHS Personal Demographic Service data (PDS) in Table 2. Precision rates are also shown. Precision is defined as the proportion of links made that are true matches and is used as a standardised measure of linkage quality. A precision rate of 99.9% means that 99.9% of the links made were true matches.

High linkage rates were achieved as we expect refugees to be registered with a GP and therefore present in the PDS, and also recorded in the Exit checks data. Using these datasets helped us to develop matchkeys that optimise linkage rates. We discuss this further in Section 6.

Data processing ahead of linkage removed approximately 5,300 refugees who had been resettled in Scotland and Northern Ireland. A further 170 records were removed from the refugee dataset as these represented babies born since a family arrived in the UK with no personal identifying information included for onward linkage. This left approximately 16,300 resettled refugees to be included in the pilot, covering England and Wales.

Table 2: Summary of linkage rates

	Number of refugee records linked	Linkage rate (%)	Precision rate (%)
Exit Checks	15,466	96.8	99.9
PDS	15,663	96.3	100

Source: Office for National Statistics - analysis of linked Home Office Exit Checks data and NHS Personal Demographic Service data

Notes

1. 295 refugees who arrived after 1 February 2020 are excluded from the Exit Checks linkage rate as at the time of linkage we did not have Exit Checks data that went beyond this date.

Table 3 presents results for each linkage stage. Details of each matchkey are listed in Appendix 1 and detailed linkage rates by matchkey are presented in Appendix 2.

Table 3: Summary of linkage rates by linkage stage

	Number of refugee records linked	Linkage rate (%)
Exit Checks data		
Pass 1: Exact: Matchkey 1 and filtered to include specific visa types and nationalities.	10,303	64.5
Pass 1: Matchkeys 2-14 and filtered to include specific visa types and nationalities.	1,650	10.3
Pass 2: Matchkeys 1-6 and looser filtering on Exit Checks to include 'Null' or missing arrival dates	3,456	21.6
Pass 3: Matchkeys 1-10 and filtering to allow for all nationalities	27	0.2
Clerical review of conflicting matches	30	0.2
Personal Demographic Service data		
Exact: Matchkey 1	9	0.1
Deterministic: Matchkeys 2-24	14,299	87.9
Associative Matching	1,270	7.8
Clerical review of conflicting matches	85	0.5

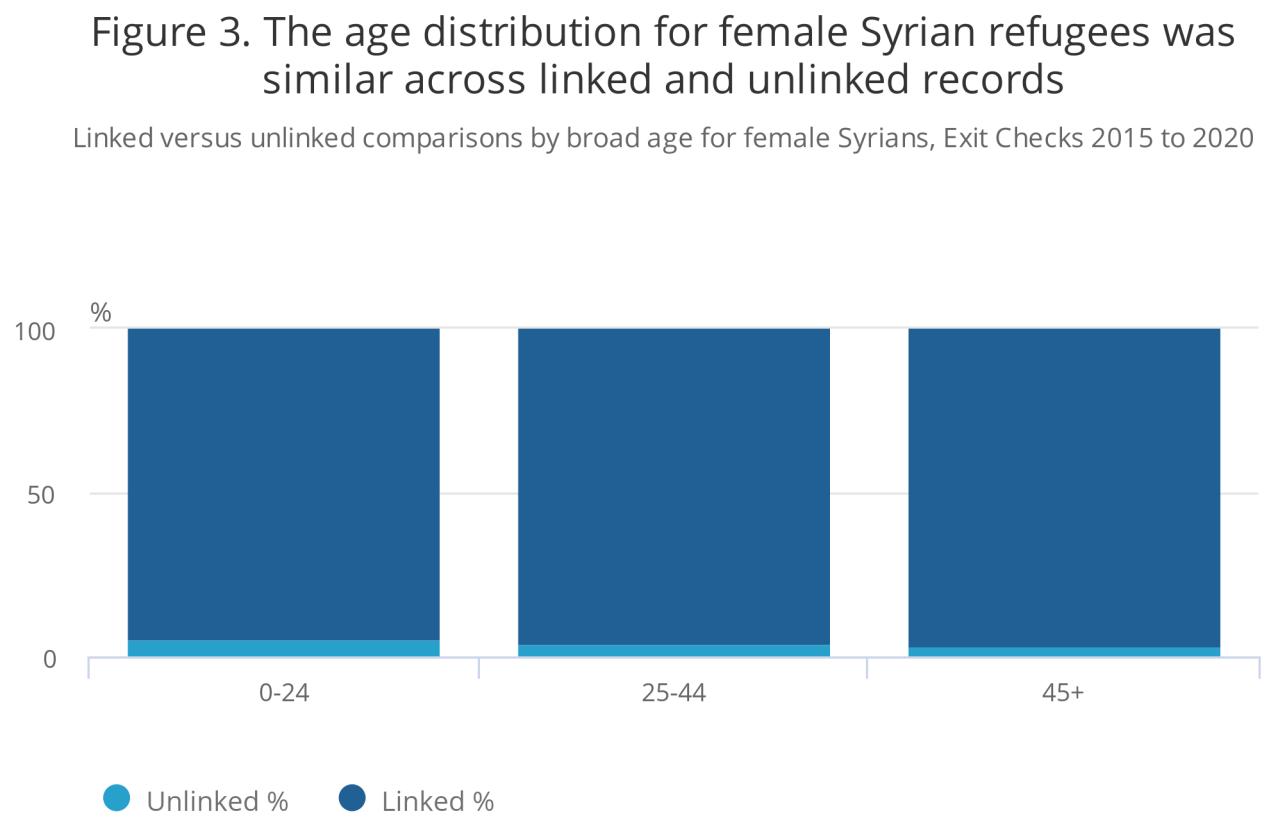
Source: Office for National Statistics - linkage rates of Home Office Exit Checks data and NHS Personal Demographic Services data

Comparison of unlinked and linked records

To ensure that there were no biases in the resulting linked dataset, these records were examined and compared with the unlinked records. Results suggest that in both data linkages there were in fact no biases present. Figures 3 to 7 highlight that the distributions of the linked and unlinked records remained consistent when examined across different characteristics, namely sex, broad age, nationality, and broad nationality.

Figure 3. The age distribution for female Syrian refugees was similar across linked and unlinked records

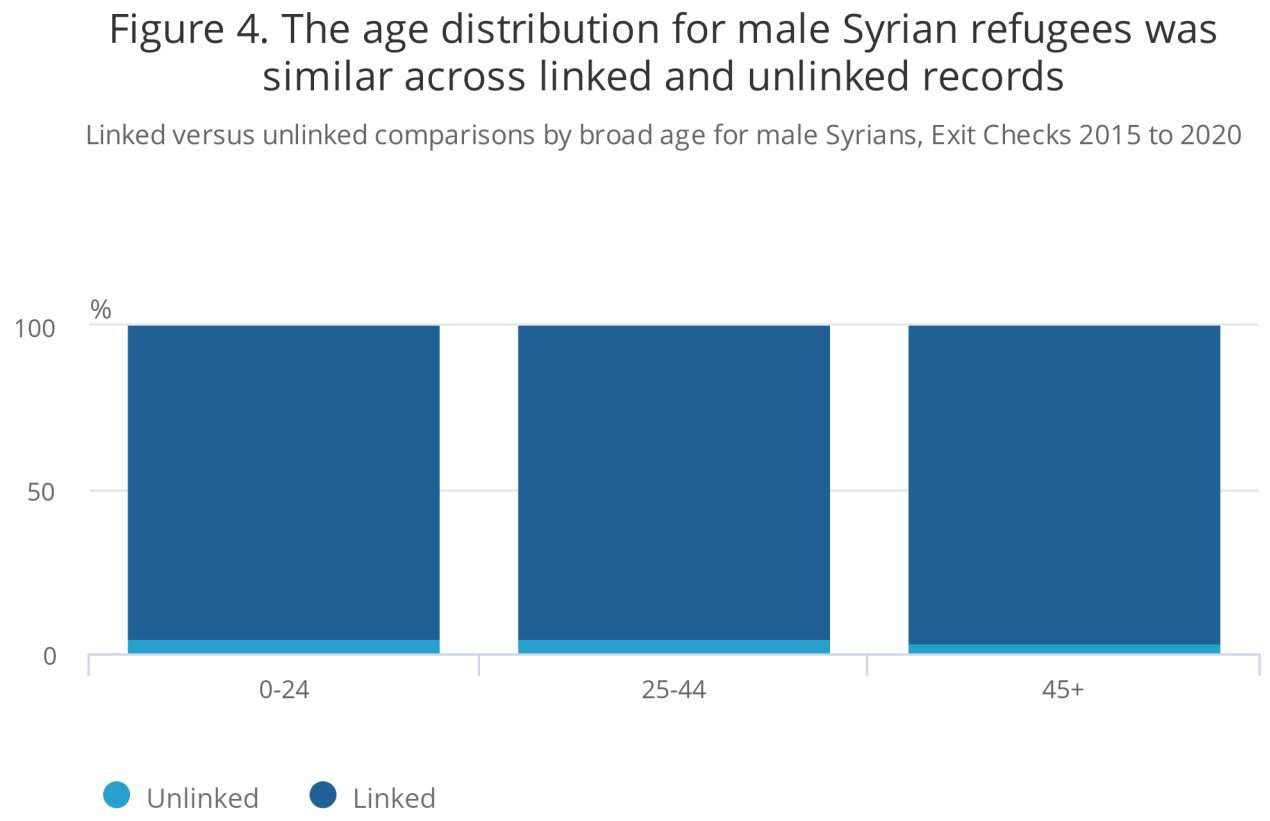
Linked versus unlinked comparisons by broad age for female Syrians, Exit Checks 2015 to 2020



Source: Office for National Statistics - analysis of linked and unlinked Home Office Exit Checks data

Figure 4. The age distribution for male Syrian refugees was similar across linked and unlinked records

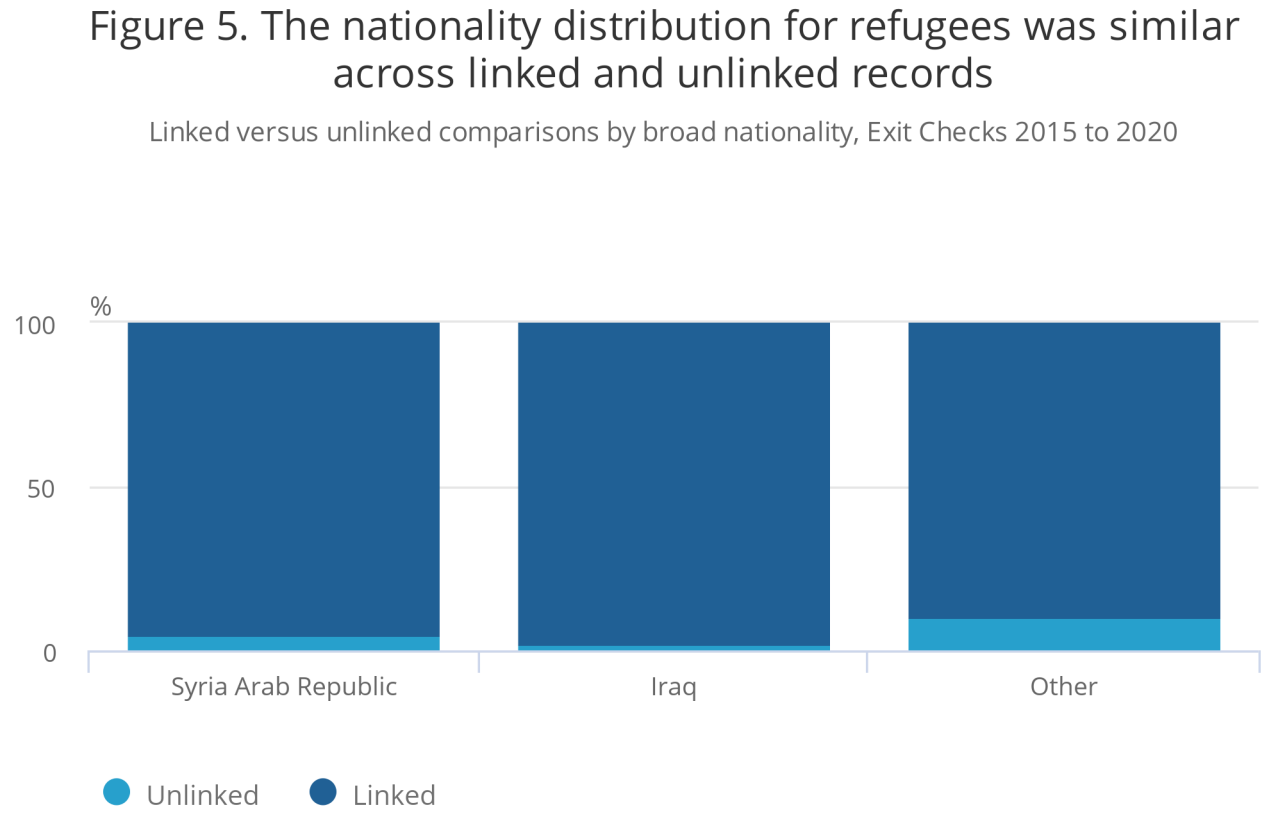
Linked versus unlinked comparisons by broad age for male Syrians, Exit Checks 2015 to 2020



Source: Office for National Statistics - analysis of linked and unlinked Home Office Exit Checks data

Figure 5. The nationality distribution for refugees was similar across linked and unlinked records

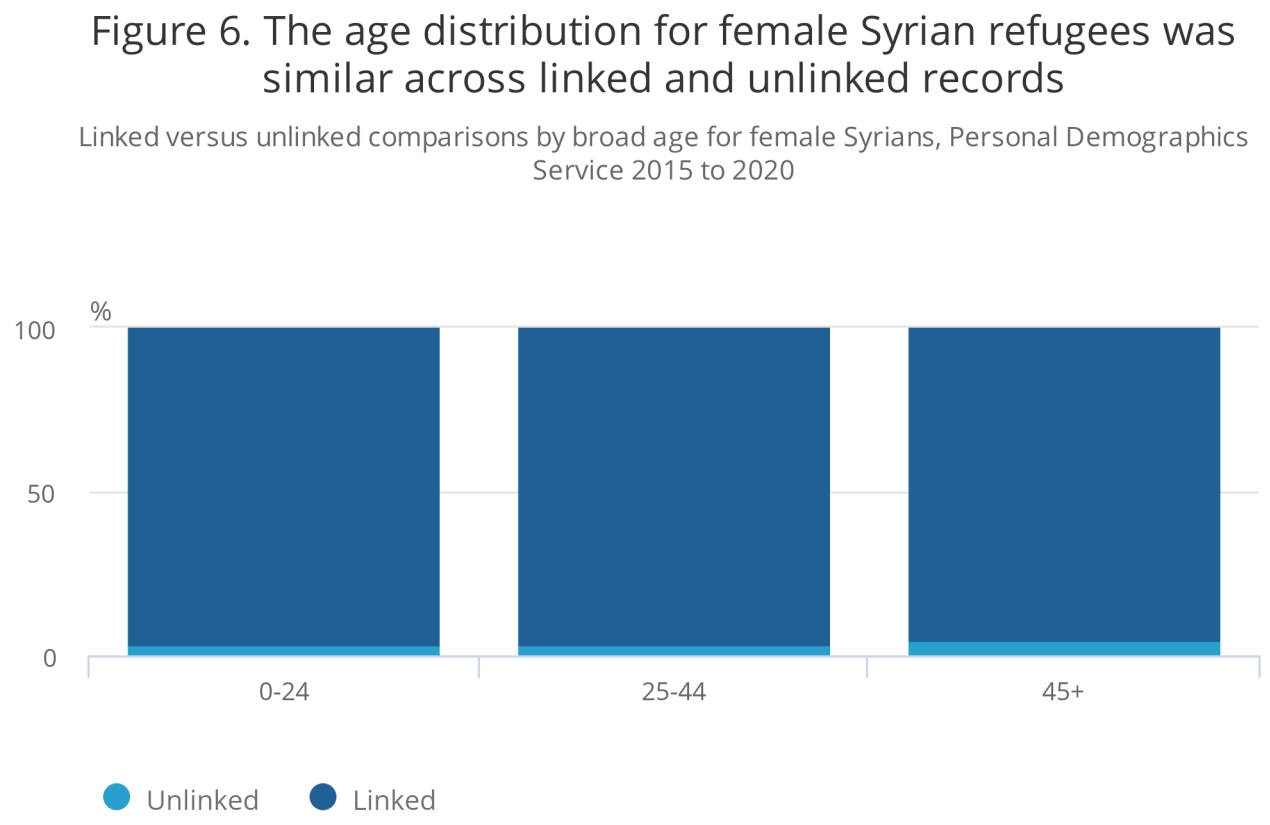
Linked versus unlinked comparisons by broad nationality, Exit Checks 2015 to 2020



Source: Office for National Statistics - analysis of linked and unlinked Home Office Exit Checks data

Figure 6. The age distribution for female Syrian refugees was similar across linked and unlinked records

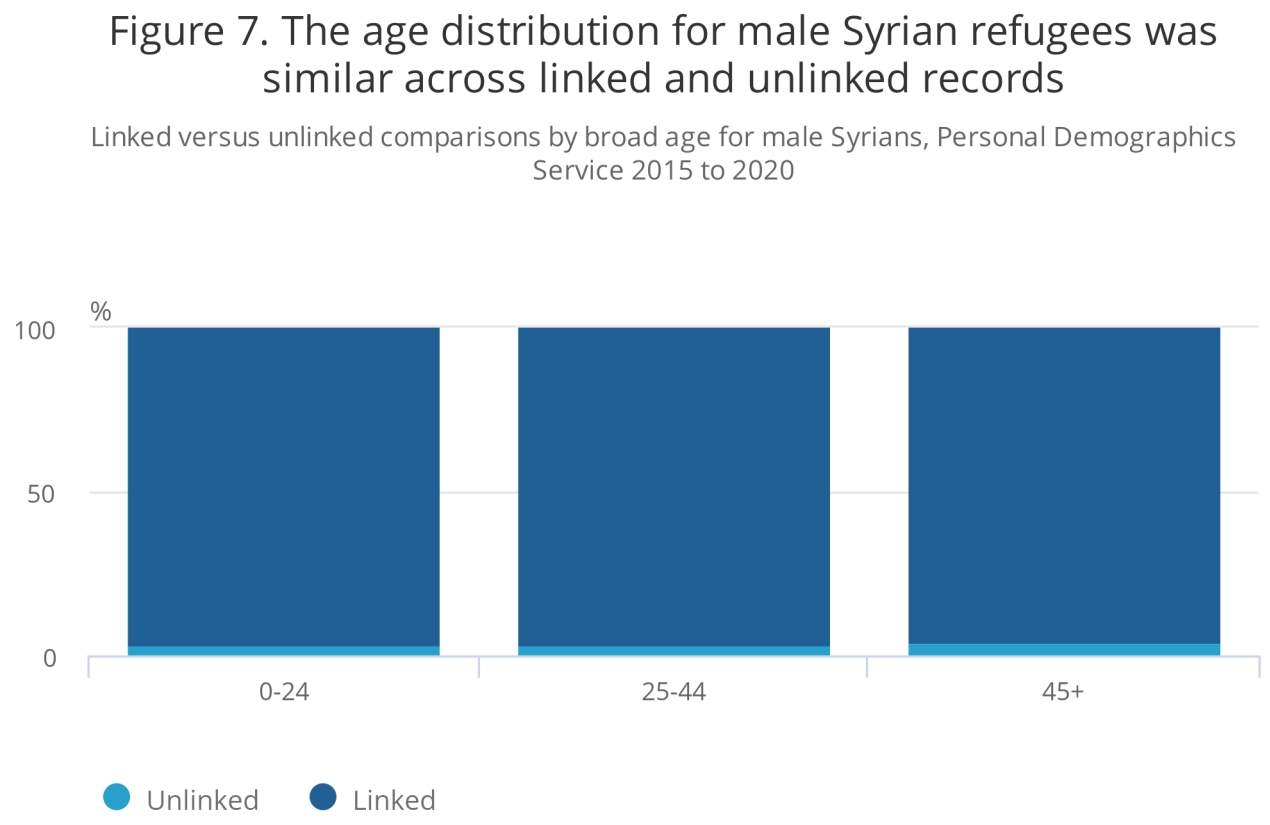
Linked versus unlinked comparisons by broad age for female Syrians, Personal Demographics Service 2015 to 2020



Source: Office for National Statistics - analysis of linked and unlinked NHS Personal Demographic Services data

Figure 7. The age distribution for male Syrian refugees was similar across linked and unlinked records

Linked versus unlinked comparisons by broad age for male Syrians, Personal Demographics Service 2015 to 2020



Source: Office for National Statistics - analysis of linked and unlinked NHS Personal Demographic Services data

6 . Approaches to linking data and future considerations

Exit Checks and Personal Demographic Service data (PDS) linkage

We approached linking the refugee data to both Exit Checks and PDS in an iterative way. This allows us to have control over the data filtering and loosening of matchkeys to maximise our match rate and achieve high quality matches.

For Exit Checks this consisted of a three-stage approach through which we achieved a total match rate of 96.8% with precision of 99.9% (ONS, 2021). The three different levels of filtering on the Exit checks data enabled us to start with a strict filtering enhancing the likelihood of high-quality matches. For PDS however, we used multiple methods of data linkage, this included deterministic and associative matching. Using this approach, we achieved a total match rate of 96.3% with precision of 100% (ONS, 2021).

There are several key reasons why we were able to achieve such high match rates. Firstly, as the aim of the pilot was to understand the feasibility of linking refugee data to administrative data, we selected the Exit Checks and PDS data sources because we were confident that they would include the refugees that we were interested in.

High match rates were also achieved by allowing for variations in name spellings. This allowed us to deal with possible transliteration issues of names that may occur in the original recording of name information in the refugee data. An example of this is the name Muhammad, which is spelt in several ways, and so a simple misspelling can cause a match not to be made (false negative) or a match to be made when it should not (false positive). We overcame this by using Levenshtein distance within the matchkey, allowing varying levels of changes between words to be classed as a match. Furthermore, by consulting with Home Office language experts we learnt that informal titles are often included within someone's Arabic names. However, these informal titles may not be present in formal identification documentation and subsequently are unlikely be present within the Exit Checks data but may be present on self-reported data such as the PDS. Removing informal titles allows for these differences in the recording of names between datasets.

While our focus for the pilot has been to adapt linkage algorithms for Arabic names, we will need to extend this to other cultural naming conventions as we extend the study to include refugees granted asylum. Asylum refugee data will potentially be more challenging to link because of the diversity in nationalities in this group, representing a wider range of cultures and naming conventions. During this experimental development phase, we will seek to understand reasons why we may not link some administrative records and develop strategies to address these. Possible reasons include:

Unobserved exit from the study

Either through death or embark (emigration) or a cross-border move from England and Wales to Scotland or Northern Ireland. There are a number of reasons for an exit being unobserved, including:

- death that occurred outside England and Wales or failure to link death
- embark not captured in administrative data source (PDS or Exit Checks), cross-border flow not flagged in PDS, embark recorded in data but failure to link this event

Not in the administrative data at all

By choice individuals may not engage with public services and "go off grid" or have yet to attend school, register with a GP, find employment and so on.

There may, therefore, be periods of latency. For some refugees there may be periods where they go unrecorded in the administrative data. For example, young healthy men not appearing in health service data, women not working, or individuals not claiming benefits or in education.

In the administrative data, but not visible through a refugee variable or other characteristics such as nationality

Here we would rely on our data linkage strategy with clerical review to achieve optimal linkage rates.

In the administrative data and visible

However, data are of poorer quality or suboptimal, therefore lower linkage rates are achieved.

Linkage failure

Factors such as missing data (refugee does not appear in the dataset either through incompleteness or under-coverage), and inaccuracies in either the refugee data or the administrative data such as transpositions, misspellings, use of formal or informal names, can all contribute to linkage failure.

Changes in key linkage variables, for example change to a surname, or change of nationality because of naturalisation, may also contribute to linkage failure if linkage strategies do not deal with this. If there is no record to link to in a dataset, we would not class this as linkage failure.

7 . Future developments

The pilot helped us develop specific linkage algorithms based around naming conventions across a wider set of cultures and using associative matching methods to draw strength from data on family units.

Following the success of this pilot, we are now collaborating with the Home Office to move forward with a full Refugee Longitudinal Cohort Study. We will extend the study to include a sample of refugees granted asylum between 2015 and 2020, and potentially beyond. We are also in the process of linking additional administrative and Census 2021 data to the study to allow a fuller picture of integration outcomes. Administrative data include education, health, income and benefit data for England and Wales. We are initially focussed on England and Wales, but plan to extend to Scotland and Northern Ireland. We have started engaging with devolved administrations on acquiring equivalent education, health and benefits data for these countries.

Such a study has the potential to contribute towards informing policy and decision makers but also all sections of society on health, social care, economic and social outcomes. The study is aligned to the Office for National Statistics (ONS) strategic objective of inclusivity and recommendations made by the National Statistician’s Inclusive Data Task Force (IDTF). This will ultimately help inform local authorities, government, charities and other organisations with resource allocation for these vulnerable populations as well as the potential to increase public awareness of societal issues.

To progress this, we are exploring strategies to reduce linkage failure. This includes how we can adapt our linkage methodology to link asylum route refugees who come from a wider range of countries than the Middle East and North Africa (MENA) region, including, for example, Eritrea, Sudan, Afghanistan and Pakistan.

8 . References

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., and Brand, C. A. (2010). 'Data linkage: a powerful research tool with potential problems'. BioMed Central (BMC) health services research, 10 (1), pages 1 to 7

Blackwell, L. and Rogers, N. 'A Longitudinal Error Framework to Support the Design and Use of Integrated Datasets' in Cernat, A. and Sakshaug, J.W. 'Measurement Error in Longitudinal Data', Oxford: Oxford University Press

Ford, J. B., Roberts, C. L., & Taylor, L. K. (2006). 'Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data'. Paediatric and perinatal epidemiology, 20(4), pages 329 to 337

'Integrated Communities Strategy', 2018

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. (2017). 'Challenges in administrative data linkage for research'. Big data and society, 4(2), 2053951717745678

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., and van der Meulen, J. H. (2017). 'A guide to evaluating linkage quality for the analysis of linked data'. International journal of epidemiology, 46(5), pages 1699 to 1710

Lariscy, J. T. (2011). 'Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox'. Journal of aging and health, 23(8), pages 1263 to 1284

Playford, C. J., Gayle, V., Connelly, R., and Gray, A. J. (2016). 'Administrative social science data: The challenge of reproducible research'. Big Data & Society, 3(2), 2053951716684143.

Ruiz, I., and Vargas-Silva, C. (2018). 'Differences in labour market outcomes between natives, refugees and other migrants in the UK'. Journal of Economic Geography, 18(4), pages 855 to 885.

9 . Acknowledgement

We wish to acknowledge our colleagues at the Home Office for making the data available to us. We greatly value their experience and knowledge of the data and their linguist expertise, which supported the development of our methodology.

Authors

Gemma Hanson, Nicky Rogers, Elzemiek Scott-Kortlever, Louisa Blackwell, Zainab Ismail and Sarah Cummins

10 . Appendices

Appendix 1

Appendix 1a: Details of each matchkey used in Exit Checks linkage
Pass 1 Matchkeys

	Matchkey	Inconsistency resolved by Matchkey
1	NAME ST | SEX | DOB | NAT CODE | ARRIVAL DATE	Full match
2	NAME ST | SEX | DOB | NAT CODE	Incorrect arrival date
3	NAME ORDERED | SEX | DOB | NAT CODE	Incorrect arrival date and allows for names to be reported in a different order
4	NAME ST | SEX | DOB | ARRIVAL DATE	Incorrect reporting nationality
5	COMMON NAME >2 | SEX | DOB | NAT CODE | ARRIVAL DATE	Allows for not all names to be the same (two or more in common)
6	NAME ST | SEX | DOB (EDIT DIS = 3) | NAT CODE | ARRIVAL DATE	Allows for date of birth to have 3 edits different between the two datasets
7	NAME ST | DOB | NAT CODE | ARRIVAL DATE	Gender missing
8	NAME (EDIT DIS >.70) | SEX | DOB | NAT CODE | ARRIVAL DATE	Allows for slight differences in the spelling of the name
9	COMMON NAME >2 | SEX | DOB | NAT CODE | ARRIVAL DATE +/- 7 DAYS	Looks for >2 common names and allows for the arrival dates to be within 7 days of each other
10	NAME (EDIT DIS >.80) | SEX | DOB <2 | NAT CODE | ARRIVAL DATE	Allows for edit differences in names and allows for DOB to have 2 differences
11	NAME ST | SEX | DOB YEAR | NAT CODE | ARRIVAL DATE	DOB year match only
12	NAME ST | SEX | DOB MONTH | NAT CODE | ARRIVAL DATE	DOB month match only
13	NAME (EDIT DIS >.65) | SEX | DOB | NAT CODE | ARRIVAL DATE	Allows for slight differences in the spelling of the name
14	NAME (EDIT DIS >.70) | SEX | DOB | NAT CODE | ARRIVAL DATE +/- 7 DAYS	Allows for edit differences in names and allows for the arrival dates to be within 7 days of each other

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 1b: Details of each matchkey used in Exit Checks linkage
Pass 2 Matchkeys

	Matchkey	Inconsistency resolved by matchkey
1	NAME ST | SEX | DOB | NAT CODE | ARRIVAL DATE IS NULL	Allows for arrival date to be null
2	NAME ST | SEX | DOB | ARRIVAL DATE IS NULL	Missing nationality and arrival date as null
3	NAME ST | DOB | NAT CODE | ARRIVAL DATE IS NULL	Incorrect arrival date and allows for names to be reported in a different order
4	NAME ST (EDIT DIS >.70) | SEX | DOB | NAT CODE | ARRIVAL DATE IS NULL	Allows for edit differences in names and missing arrival date
5	NAME ST | SEX | DOB <2 | NAT CODE | ARRIVAL DATE IS NULL	Allows for DOB to have 2 differences and missing arrival date
6	NAME ST | SEX | DOB | NAT CODE	Arrival date missing

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 1c: Details of each matchkey used in Exit Checks linkage
Pass 3 Matchkeys

	Matchkey	Inconsistency resolved by Matchkey
1	NAME ST | SEX | DOB | NAT CODE | ARRIVAL DATE	Full match
2	NAME ST | SEX | DOB | NAT CODE	Missing arrival date
3	NAME ST | SEX | DOB | ARRIVAL DATE	Missing nationality
4	COMMON NAME >2 | SEX | DOB | ARRIVAL DATE	Allows for not all names to be the same (2< in common) and missing nationality code
5	NAME ST | SEX | DOB <2 | ARRIVAL DATE	Allows for date of birth to have 2 edits different between the two datasets and missing nationality code
6	NAME ST | DOB | ARRIVAL DATE	Gender and nationality code missing
7	NAME (EDIT DIS >.70) | SEX | DOB | ARRIVAL DATE	Allows for slight differences in the spelling of the name and missing nationality code
8	NAME ST | SEX | DOB excl. one element | NAT CODE | ARRIVAL DATE	Allows for date of birth to have 1 element excluded
9	NAME ST | SEX | DOB excl. one element | ARRIVAL DATE	Allows for date of birth to have 1 element excluded and missing nationality
10	COMMON NAME >2 | SEX | DOB | ARRIVAL DATE +/- 7 days	Allows for not all names to be the same (2< in common), missing nationality and arrival date within 7 days
11	NAME (EDIT DIS >.80) | SEX | DOB<2 | ARRIVAL DATE	Allows for slight differences in the spelling of the name, DOB with 2 edit differences and missing nationality code
12	NAME ST | SEX | DOB excl. one element | ARRIVAL DATE +/- 7 days	Allows for date of birth to have 1 element excluded and missing nationality and arrival date within 7 days.
13	NAME (EDIT DIS >.70) | SEX | DOB | ARRIVAL DATE +/- 7 days	Allows for slight differences in the spelling of the name, missing nationality code and arrival date within 7 days

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 1d: Details of each matchkey used in PDS linkage
Pass 1 Matchkeys

	Matchkey	Inconsistency resolved by Matchkey
1	NAME NT | SEX | DOB | POSTCODE	Full match
2	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE	Missing middle name
3	NAME NT | DOB | POSTCODE	Missing gender
4	NAME NT | SEX | DOB | POSTCODE SECTOR	Postcode sector only needs to match
5	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE SECTOR	Missing middle name and postcode sector only needs to match
6	NAME NT | SEX | DOB | POSTCODE DISTRICT	Postcode district only needs to match
7	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE DISTRICT	Missing middle name and postcode district only needs to match
8	NAME NT | SEX | DOB | LOCAL AUTHORITY	Postcode missing but match on local authority
9	FIRSTNAME | LASTNAME | SEX | DOB | LOCAL AUTHORITY	Missing middle name and postcode missing but match on local authority
10	NAME NT | SEX | DOB | POSTCODE AREA	Postcode area only needs to match
11	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE AREA	Missing middle name and postcode area only needs to match
12	NAME NT | SEX | DOB | ARRIVAL DATE /REGISTRATION DATE	Missing postcode
13	FIRSTNAME | LASTNAME | SEX | DOB	Missing middle names
14	NAME (EDIT DIS >.70) | SEX | DOB | POSTCODE	Edit distance on full name
15	NAME (EDIT DIS >.70) | SEX | DOB | LOCAL AUTHORITY	Edit distance on full name and missing postcode
16	COMMON NAME >2 | SEX | DOB | POSTCODE	2 or more common names from full name
17	COMMON NAME >2 | SEX | DOB | LOCAL AUTHORITY	2 or more common names from full name and missing postcode
18	NAME NT | SEX | DOB <2 | POSTCODE	Edit distance of 2 or less on DOB

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 2

Appendix 2a: Details of linkage rates by matchkey for Exit Checks linkage
Pass 1 linkage rates by matchkeys

	Matchkey	Count	% Refugee (16,264)	% During time window (15,969)
1	NAME ST | SEX | DOB | NAT CODE | ARRIVAL DATE	10,303	63.1	64.2
2	NAME ST | SEX | DOB | NAT CODE	808	5	5.1
3	NAME ORDERED | SEX | DOB | NAT CODE	93	0.6	0.6
4	NAME ST | SEX | DOB | ARRIVAL DATE	16	0.1	0.1
5	COMMON NAME >2 | SEX | DOB | NAT CODE | ARRIVAL DATE	38	0.2	0.2
6	NAME ST | SEX | DOB (EDIT DIS = 3) | NAT CODE | ARRIVAL DATE	1	0	0
7	NAME ST | DOB | NAT CODE | ARRIVAL DATE	29	0.2	0.2
8	NAME (EDIT DIS >.70) | SEX | DOB | NAT CODE | ARRIVAL DATE	0	0	0
9	COMMON NAME >2 | SEX | DOB | NAT CODE | ARRIVAL DATE +/- 7 DAYS	4	0	0
10	NAME (EDIT DIS >.80) | SEX | DOB <2 | NAT CODE | ARRIVAL DATE	404	2.5	2.5
11	NAME ST | SEX | DOB YEAR | NAT CODE | ARRIVAL DATE	0	0	0
12	NAME ST | SEX | DOB MONTH | NAT CODE | ARRIVAL DATE	0	0	0
13	NAME (EDIT DIS >.65) | SEX | DOB | NAT CODE | ARRIVAL DATE	95	0.6	0.6
14	NAME (EDIT DIS >.70) | SEX | DOB | NAT CODE | ARRIVAL DATE +/- 7 DAYS	162	1	1
Total		11,953	73.5	74.6

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 2b: Details of linkage rates by matchkey for Exit Checks linkage
Pass 2 linkage rates by matchkeys

	Matchkey	Count excl. conflicting	% Refugee (16,264)	% During time window (15,969)
1	NAME ST | SEX | DOB | NAT CODE | ARRIVAL DATE IS NULL	3174	19.5	19.9
2	NAME ST | SEX | DOB | ARRIVAL DATE IS NULL	4	0	0.03
3	NAME ST | DOB | NAT CODE | ARRIVAL DATE IS NULL	0	0	0
4	NAME ST (EDIT DIS >.70) | SEX | DOB | NAT CODE | ARRIVAL DATE IS NULL	85	0.5	0.5
5	NAME ST | SEX | DOB <2 | NAT CODE | ARRIVAL DATE IS NULL	1	0	0
6	NAME ST | SEX | DOB | NAT CODE	192	1.2	1.2
Total		3,456	21.6	21.6

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 2c: Details of linkage rates by matchkey for Exit Checks linkage
Pass 3 linkage rates by matchkeys

	Matchkey	Count	% Refugee (16,264)	% During time window (15,969)
1	NAME ST | SEX | DOB | NAT CODE | ARRIVAL DATE	0	0	0
2	NAME ST | SEX | DOB | NAT CODE	1	0	0
3	NAME ST | SEX | DOB | ARRIVAL DATE	0	0	0
4	COMMON NAME >2 | SEX | DOB | ARRIVAL DATE	0	0	0
5	NAME ST | SEX | DOB <2 | ARRIVAL DATE	0	0	0
6	NAME ST | DOB | ARRIVAL DATE	0	0	0
7	NAME (EDIT DIS >.70) | SEX | DOB | ARRIVAL DATE	0	0	0
8	NAME ST | SEX | DOB excl. one element | NAT CODE | ARRIVAL DATE	0	0	0
9	NAME ST | SEX | DOB excl. one element | ARRIVAL DATE	0	0	0
10	COMMON NAME >2 | SEX | DOB | ARRIVAL DATE +/- 7 days	25	0.2	0.2
11	NAME (EDIT DIS >.80) | SEX | DOB<2 | ARRIVAL DATE	1	0	0
12	NAME ST | SEX | DOB excl. one element | ARRIVAL DATE +/- 7 days	0	0	0
13	NAME (EDIT DIS >.70) | SEX | DOB | ARRIVAL DATE +/- 7 days	0	0	0
Total		27	0.2	0.2

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

Appendix 2d: Details of linkage rates by matchkey for PDS linkage
Pass 1 linkage rates by matchkeys

	Matchkey	Count	% Refugee (16,264)
1	NAME NT | SEX | DOB | POSTCODE	9	0.1
2	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE	0	0
3	NAME NT | DOB | POSTCODE	0	0
4	NAME NT | SEX | DOB | POSTCODE SECTOR	0	0
5	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE SECTOR	0	0
6	NAME NT | SEX | DOB | POSTCODE DISTRICT	0	0
7	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE DISTRICT	0	0
8	NAME NT | SEX | DOB | LOCAL AUTHORITY	0	0
9	FIRSTNAME | LASTNAME | SEX | DOB | LOCAL AUTHORITY	0	0
10	NAME NT | SEX | DOB | POSTCODE AREA	8,635	53.09
11	FIRSTNAME | LASTNAME | SEX | DOB | POSTCODE AREA	1,696	10.43
12	NAME NT | SEX | DOB | ARRIVAL DATE/REGISTRATION DATE	2,670	16.42
13	FIRSTNAME | LASTNAME | SEX | DOB	977	6.01
14	NAME (EDIT DIS >.70) | SEX | DOB | POSTCODE	5	0.03
15	NAME (EDIT DIS >.70) | SEX | DOB | LOCAL AUTHORITY	621	3.81
16	COMMON NAME >2 | SEX | DOB | POSTCODE	0	0
17	COMMON NAME >2 | SEX | DOB | LOCAL AUTHORITY	146	0.9
18	NAME NT | SEX | DOB <2 | POSTCODE	0	0
Total		14,759	90.75%

Source: Office for National Statistics - Refugee Integration Outcomes (RIO) data linkage pilot

11 . Related links

Developing standard tools for data linkage: February 2021
Methodology | Released February 2021
Discussing the seven functions that were created to automate part of the data linkage pipeline to improve efficiency and quality.

The UK government's approach to evaluating the Vulnerable Persons and Vulnerable Children's Resettlement Schemes
Research report | Released 13 December 2018
Outlines the UK government's approach to resettling refugees under the Vulnerable Persons and Vulnerable Children's Resettlement (VPR and VCR) schemes and the strategy for evaluating their delivery and effectiveness.

Home Office Indicators of Integration framework 2019
Framework | Released 3 June 2019
This indicators framework provides practical ways to design more effective strategies, monitor services and evaluate integration interventions.

Statistics for the public good: UK Statistics Authority five-year strategy (2020 to 2025)
Report | Released July 2020
Outlines the UK Statistics Authority's five-year strategy: Statistics for the Public Good. The previous strategy: Better Statistics, Better Decisions ran from 2015 to 2020.