Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Scotland's Census 2022

Resolving Multiple Responses: Identify Duplicates

August 2020

National Records of Scotland

Contents

1. Plain English Abstract

There are situations where people submit more than one response to the Census, for any number of reasons. Sometimes, this could be due to a miscommunication (for example, where two members of a household think they have not yet responded, so each submits a response), or a respondent attempts to respond to their Census online, creates a Census record, but is unable to complete it, subsequently submitting a paper response as well. When these types of situations occur, the population is 'overcounted'. This duplication occurs for both persons and households, and needs to be resolved to ensure that Scotland's population count is not overestimated. This is done in a process called Resolve Multiple Responses (RMR).

RMR first identifies duplicates within the dataset that potentially need resolution, or merging, into one. This paper outlines the methodology that will be used to link individuals to each other in Scotland's 2022 Census.

The linking methodology compares individual records within the same postcode. Records that appear to represent the same person would be possible duplicates, that is, the person has returned details multiple times. An administrative dataset is then used to quality assure such possible duplicate returns. This helps check whether there is more than one individual with these details within the postcode, which informs the decision on whether to resolve the entries from the statistical dataset into one. This process allows for more-accurate dealing of such duplicates, improving the quality of the census.

This paper covers the method of identifying duplicates to resolve. The process of resolution will be covered in a separate paper.

2. Abstract

People sometimes give multiple census responses in error, and these need to be resolved into a single record in order to avoid overestimating the population. The Resolve Multiple Responses (RMR) step in Census processing looks to resolve cases where this occurs within the same household or postcode. Where there are cases across different postcodes, these will be dealt with by the overcount correction methodology, later in processing.

This paper explores the proposed methodology outlining how these cases can be identified with the aid of administrative data. The census is linked to itself grouped on postcode and then using variables such as name and date of birth. It is then linked to an administrative data source for an extra layer of verification, so that those found are true duplicates to be resolved.

Note that it is preferable to err on the side of resolving too many, rather than too few, cases. This is because an undercount can be better dealt with during the estimation process than overcount. Undercount can be dealt with by Dual-System Estimation. Missing a true match (where two records relate to the same individual) would have a greater impact than resolving two records that were not the same individual.

This paper covers identifying records to resolve. The process of resolution will be covered in a separate paper.

Note: On 17 July 2020 Scottish Government announced the decision to move Scotland's Census to 2022 following the impact of the COVID-19 pandemic. The information included in this report reflects the methodology intended, at the time of publication, to be used in the 2022 Census. It is not expected that there will be any major differences between the methodology presented here and that used. However, some detail may change or be completed before or during census processing. Any major changes to the intended methodology will be described in an update here.

3. Introduction and Background

There are situations when people submit more than one response to the Census. This is sometimes for legitimate reasons; if a respondent is making a paper response but the household is larger than five people, for example, they are asked to submit another questionnaire (a continuation form). These situations do not result in overcount and pose no problems in terms of accuracy and data quality. However, there are also situations where someone responds to the Census more than once and it causes issues in data quality — where people, or even whole households, are duplicated in the dataset. This can occur for a number of reasons, for example:

1. Where someone in the household fills out the questionnaire and sends it off, but someone else in the household has already done this;

2. When someone changes their mind about what they want to include in their response, and submits a new one;

3. A respondent begins filling in the census return online but decides they would rather fill it in on paper. In such cases the information on the online return would be collected as an unsubmitted return;

4. A respondent begins filling in the census return online, but forgets their login details before completing it. They would then need to request a new Internet Access Code (IAC) and begin a new return[1]. Again, the information on the first return would be collected as an unsubmitted return;

5. Where a person gets confused about a paper response, and answer the individual questions for themselves for Persons 1–5 (each paper household questionnaire contains spaces for up to five people)

[1] For data security reasons, individuals cannot be given access to partially completed census returns over the phone.

This duplication creates what is called 'overcount', an inflation in the count of people or households. Overcount can lead to an overestimation of the population. Therefore, the process that looks at duplication of individuals at a location — Resolve Multiple Responses (RMR) — is an important part of data preparation. It would be preferable for estimation if RMR over-resolves, that is, resolve more responses than it under-resolves, because an undercount can be better dealt with in estimation. In statistical processing undercount can be dealt with by Dual-System Estimation[2], a widely used quality assurance step performed on censuses, though methodology might vary between countries.

The first step in RMR is to identify those responses which may be duplicates of this nature. The approach for Scotland's Census in 2022 looks to resolve cases where this occurs within the same household, and within the same postcode. RMR's purpose is to deal with duplication within households. However, it is possible that individuals from the same household could have their address recorded differently, if they edit their address on their return. In order to identify such cases RMR is done at postcode, rather than household level. Cases where a person appears at genuinely different locations cannot be dealt with using RMR, as it would not be known where the person should appear. Therefore, duplicates across postcodes will be dealt with using a separate process.

This paper explores the proposed methodology for identifying potential duplicates for RMR, outlining how cases can be identified with the aid of administrative data. This is done by linking the census to itself, blocking on postcode, on variables such as first name, last name, middle name and date of birth. Following this the linked census records could be linked to another administrative data source to guide the identification of 'genuine' duplicates. If two administrative data records were found that link to these census records then this would suggest that it was two distinct people and so the records should not be resolved. Examples where this might happen include cousins with the same name living in the same postcode area.

[2] For more information on Dual-System Estimation and how it is used in the census see https://www.scotlandscensus.gov.uk/documents/Scotland%E2%80%99s%20Census%202022%20-%20PMP001%20-%20Estimation%20and%20Adjustment%20Methodology%20(pdf).pdf.

This paper covers the methodology used in the linking/identifying aspect of Resolve Multiple Responses. Please note that the process of resolving the records will be published in a separate paper.

4. 2011 Method

In Scotland's Census 2011, there was no administrative data usage in statistical data processing. In RMR, households were instead looked at by combinations of different 'form types'[3], including looking for more than one submission based on specific form identification (each household or communal establishment was given a unique form ID). This was followed by searching for person-level duplication.

At the person level, linking (i.e. where people were considered matches) was based on the following criteria:

No condition on age
and
Date of birth matches on month and day or month and year (not missing)
and
First names and surnames match exactly (not missing)
and
Sex matches or sex is missing (can be missing on some or all records)

OR

At least one is over 30 (years old)
and
Date of birth matches exactly (not missing)
and

[3] Census Questionnaires were more commonly known as 'forms' in 2011. Form types describe the type of questionnaire one was issued — for example, most households were issued a Household Form, but there were forms for Continuation (of a household), Individuals and Communal Establishments as well.

Soundex[4] of first names and surnames match or names are missing (can be missing on some or all records)

and

Sex matches or sex is missing (can be missing on some or all records)

5. Proposed 2022 Method

The proposed method in this paper takes a new approach to that of 2011, and looks to link the census not only to itself in a more robust way, but also to use administrative data to help quality assure this process.

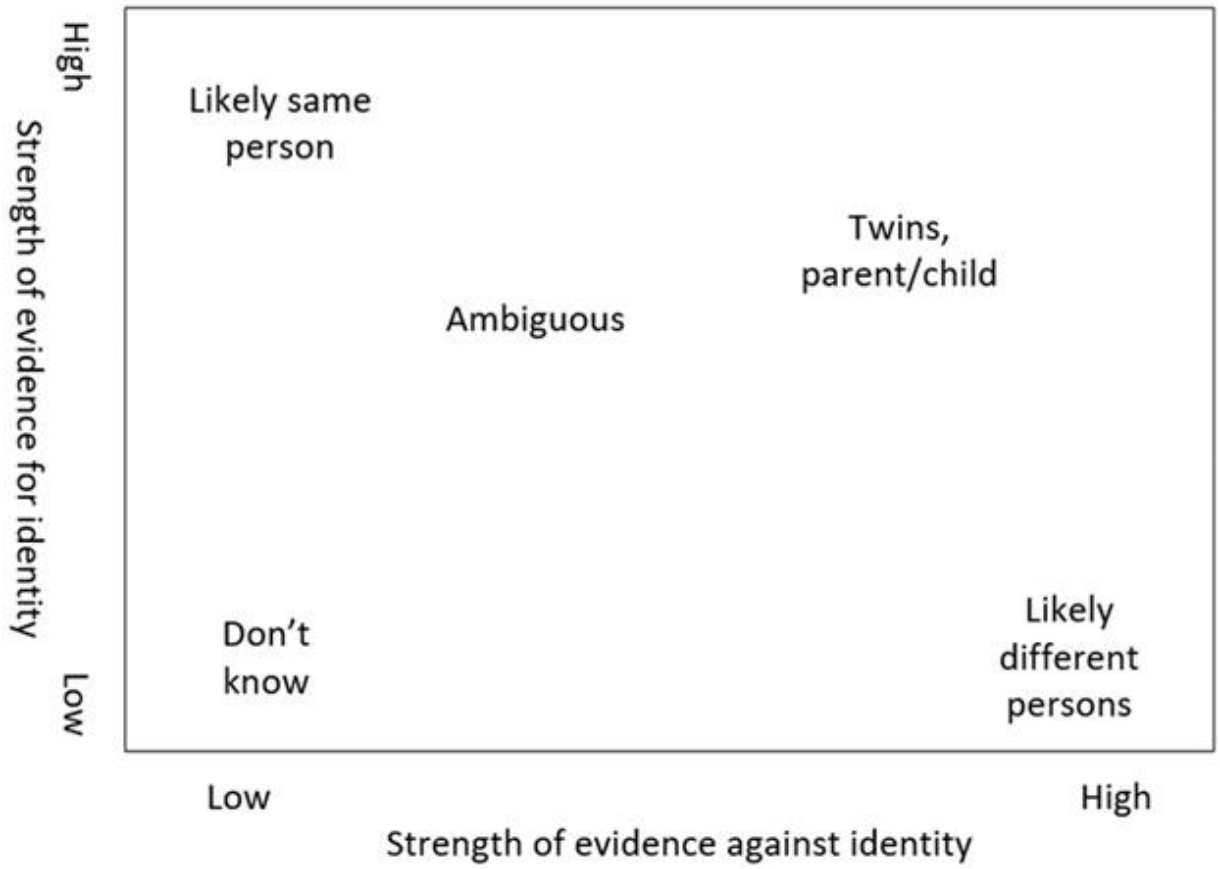
5.1 Census – Census Linking

This method links the census to itself using comparisons of the key variables: date of birth, name (first name, last name and middle name) and sex. Linking blocks[5] on postcode: the linking method only considers pairs of records in the same postcode. For each pair, a score is assigned for each linking variable, depending on how similar (accounting for differences such as typos, character level check, phonetic similarity) the records are on these linking fields. Then this is converted into evidence for and against the records being a match (that is, the two records representing the same individual). The scoring is done in the same way as for other census linking tasks, and was developed to reflect the judgements of a human reviewer. More information on this can be found in Annex 2 and Annex 3. Having separate scores for evidence for and evidence against a match allows cases where information is slightly different to be distinguished from cases where information is missing. Pairs where much of the information used for linking is missing would sit at the lower left of Figure 1, while those with conflicting evidence would appear at the upper right.

[4] See, for example, https://documentation.sas.com/?docsetId=lefunctionsref&docsetTarget=n1i9a3o4kciemhn1kpgutl20e4i0.htm&docsetVersion=9.4.

[5] When blocking, the records for linking are separated into blocks with the same value of some blocking variable(s). Links are only sought within (rather than between) blocks. There will then be no links where the linked records have different values for the blocking variable(s). See Steorts et al. (2014), for a discussion of blocking.

Figure 1 How links tend to appear in the parameter space of evidence for and evidence against a match.



The links are then categorised by the strength of their evidence for and against for the different variables. These categories include possible Parent/Child pairs or Twins. For example John Smith born 1960 and John Smith born 1991 living at the same location would be considered a parent and child pair. 78 categories (see Annex 1 for a breakdown of these) of links were developed. If the records are from the same questionnaire then there may be information on the relationship between the persons represented by the two records, for example that one is the parent of the other. If this information is present then it provides further evidence that the records represent distinct individuals, and so is considered when categorising the link.

A sample of the cases in each category was reviewed, and on the basis of this the categories (see Annex 1) were then classed into:

1. Automatically flagged for resolution
2. Passed for clerical review[6], or
3. Automatically discarded (records linked with these links will not be resolved)

Categories were placed in the automatic resolution set if the links were generally considered matches, in the automatic discarded set if the links were generally considered non-matches, and in the group for clerical review if there were a mixture or it was felt that some could be ambiguous.

Having so many categories allows decisions on whether or not to resolve a link, to be applied to all similar links. This then reduces the number of categories where the links need to be clerically reviewed. Even for categories where the links are passed to clerical review, if, during review, it is found that all the links are accepted, or all rejected, this can then be applied to all further links in that category.

Records are then assigned to a group so that all the records in a group link to each other directly or indirectly. Indirect links are when records do not link directly to each other, but do link via another record. (For example, if A links to B, and B to C, then A links indirectly to C.) Groups of linked census records would only be automatically resolved if every record in the group links to every other record in the group and the score and subsequent category of each link is strong enough not to require clerical review. Conversely, if any of the links in a group need reviewed then the whole group would need reviewed.

In cases where there is some ambiguity about the matches, for example, where a match is identified but there is potential information to indicate they are two different respondents, it may be better to resolve the responses. Coverage adjustment[7] would deal with the undercount that would result (rather than allowing these cases through to disrupt CCS matching).

[6] Clerical review is a process whereby an individual manually calls up the cases in question and looks at all relevant information to determine the outcome of a decision.
[7] For more information on Dual-System Estimation and how it is used in the census see https://www.scotlandscensus.gov.uk/documents/Scotland%E2%80%99s%20Census%202022%20-%20PMP001%20-%20Estimation%20and%20Adjustment%20Methodology%20(pdf).pdf.

As a quality assurance step a sample of the automatically accepted links will be clerically reviewed as well.

5.2 Census – Administrative Data Linking

The records in the groups to be resolved are then linked to the administrative data. If there are no administrative records, or only one administrative record linked to the group, then the group would be resolved as planned. However, if there were multiple administrative records, the group would be clerically reviewed alongside the administrative records.

This is an additional layer of quality assurance. If we have two census records in a group that is being considered for resolution then if there were also two matching records in the administrative dataset, this would suggest that there were in fact two distinct persons, and so the group of linked records should not be resolved to one person.

This method takes the records in the groups and links them to the administrative data source. This is done by blocking on postcode, and using the same methodology as the initial linking of census records to census records. The only difference is that there are no recorded relationships between the administrative data source records and the census records (as they are from independent datasets).

In addition there is a further search across the whole of the administrative data source, but only for cases where name and date of birth agree exactly. Such links are assigned to a special category '2A NHSCR different PC', all of which are sent for review. This would be useful if the person appears at a different location in the administrative dataset, and so would be missed when blocking on postcode.

Groups were flagged for automatic resolution when:

Every record in the group links strongly to every other record in the group, no more than one administrative data record links to the census records in the group, and one of the following three conditions is met:

1. The number of groups of records associated with a particular Individual Access Code (IAC) is equal to the number of usual residents indicated on the census form; or
2. The number of usual residents is missing on the questionnaire; or
3. The records in the group come from more than one IAC; and
    - All the records link very strongly[8] to every other record; or
    - All but one of the IACs in the group were unsubmitted returns.

6. Results using 2011 Census Data

The new RMR method was tested using the 2011 census data and an administrative data source. The administrative dataset used was the NHS Central Register (NHSCR, see Annex 4). This is a dataset of people who were born in Scotland, or have been registered with a GP in Scotland. This test assumed that records that had failed the Removing False Persons[9] (RFP) processing step (including the linkage to administrative data) would be removed from the dataset before being sent to RMR. Therefore, any records that would fail RFP are removed before the test.

Table 1 Groups of records identified linking on the test dataset that has the records which fail removed, broken down by the number of records in the group and whether they needed clerical review.

Number of grouped census records	Groups Passed for Clerical review	Groups accepted automatically without clerical review	Total
2	404	1,318	1,722
3	104	60	164
4	56	29	85
5	45	52	97
6+	35	0	35
Total	644	1,459	2,103

[8] Very strongly means links that have a category whose code begins with a zero (see Annex 1).
[9] See the methodology papers on Remove False Persons at https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0 for more information.

For 2011 census, we took one Processing Unit (PU, one of 10 roughly equal areas of Scotland) and ran it through the new methodology. Table 1 shows the number of groups of records that are passed for clerical review and how many are automatically passed for resolution. There were 2,103 cases found, of which 1,459 cases could be resolved without clerical review. This would equate to about 14,600 to review if applied to all 10 PUs which make up the 2011 census. The majority of the groups, 1,722, had two census records, indicating that the return had only one other census return where the same person had been listed. During this process a sample of cases in the groups accepted automatically were clerically reviewed indicatively to give confidence in, and quality assure, the method.

7. Results using 2019 Rehearsal Data

The methodology was also tested using the 2019 census rehearsal dataset.

One of the main findings from the rehearsal on RMR was that there were substantially more cases needing resolved than suggested by the 2011 data. This was partly due to the switch to online collection as the primary collection method. This resulted in more unsubmitted returns (possibly as respondents started a return but then forgot their password before submitting). In many instances records from unsubmitted returns were linked to records from a submitted return.

The administrative dataset used in rehearsal for linking was NHS Central Registrar (NHSCR) as of the 30 June 2019 (see Annex 4).

Census records are considered as a group if they link directly or indirectly via the categories of links for automatic acceptance or clerical review (see Annex 1).
The process was run on two different extracts of the rehearsal data. The first cut was a smaller dataset (44,420 records) and the second cut (51,080 records) included all household individuals from unsubmitted returns, even for those who did not complete individual forms. This occurred where information, including the date of birth, on the individual forms was missing, but the individual had been mentioned on the household form.

Using the first extract of the rehearsal data of 44,420 records, 517 groups were identified, of which 449 were for automatic acceptance when not using administrative data, 443 when administrative data was used. Using the second extract of 51,080 records, 1,886 groups were identified for resolution or clerical review (447 for automatic acceptance when not using administrative data, 445 when using administrative data). The reason for there being so many groups passed to review with the second extract was due to the large number of unsubmitted returns being included in the individual dataset. Such records would then link to records from a different return if the respondent submitted their details on a separate return. Links where one of the record has missing date of birth would not be categorised as any of the categories that can be automatically accepted.

Of the 1,439 groups for clerical review (before linking to administrative data) using the latter extract, 1,263 groups had all the links in the group the same category. In 763 of these groups (53 per cent of all groups to be reviewed) the link category was 3M, indicating that the name was exactly the same, but that sex and date of birth were missing on at least one of the linked records, and there was no relationship information between the records (for example because the records came from different returns). For categories such as these there will be limited benefit from clerical review (as all variables are either identical or missing).

To check the process, all 517 groups identified from the first rehearsal extract were manually clerically reviewed. This included all the groups that were passed for automatic acceptance in order to check the process. In the live run of the 2022 Census a sample of groups passed for automatic acceptance will be clerically reviewed for quality assurance purposes.

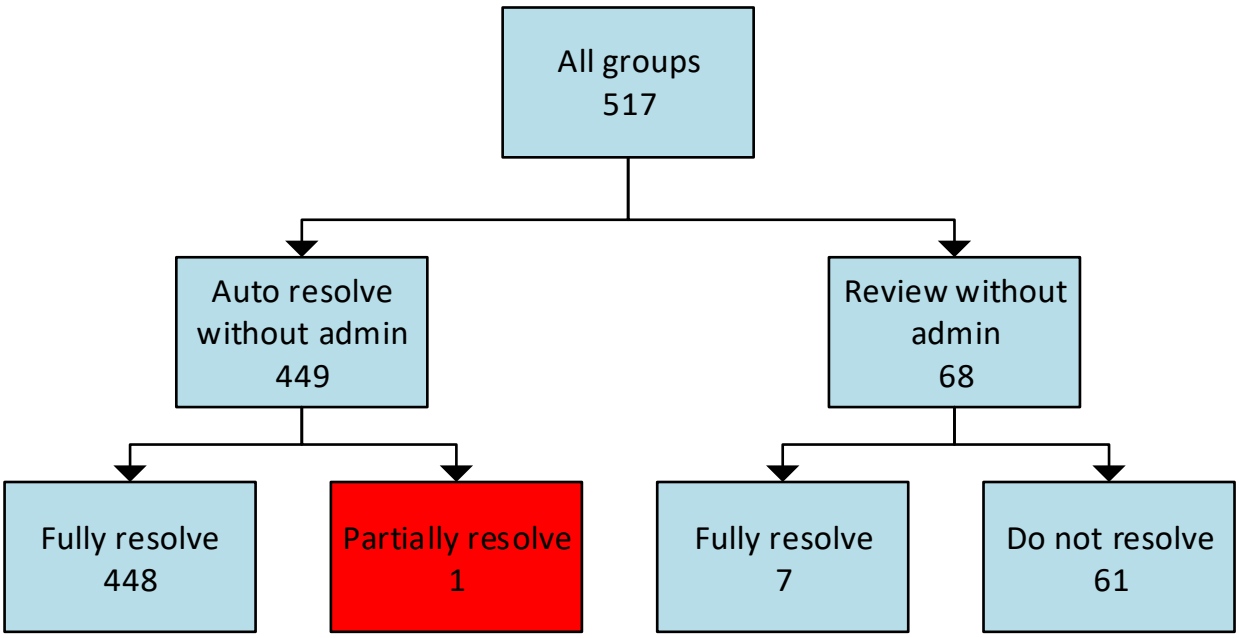


Figure 2 Number of groups by whether they were flagged for review or automatic acceptance (before using administrative data) and the results of the review for the first cut of the Census Rehearsal. The red box indicates where there is a conflict between the outcome of the automatic process and the clerical review.

The results without the administrative data is shown in Figure 2 (broken down by whether the process sent the group for review, and the results of the review). 449 groups were passed for automatic acceptance. When these were reviewed, the reviewer agreed that all but one of these groups should be fully resolved. For the 68 groups passed to review, seven were reviewed as needing to be fully resolved.

Table 2 Number of groups by whether they were flagged for review or automatic acceptance and the results of the review (with and without admin data) for the first cut of the Census Rehearsal. Shaded cells indicate where using admin data lead to a different outcome.

Results without admin data	Results of review without admin data	Results with admin data					Total
		No admin data	Auto	Review			
			Results of review with admin data				
			Fully Resolve	Do not resolve	Partially resolve	Fully resolve	
Auto	Fully resolve	5	437	1	0	5	448
	Partially resolve	0	1	0	0	0	1
Review	Do not resolve	17	0	37	1	6	61
	Fully resolve	0	0	0	0	7	7
Total		22	438	38	1	18	517

The clerical review of all these groups was repeated after the administrative data had been linked in, and this time the clerical reviewer had access to the information from the linked administrative data records. The results of this review for each of the final cells in Figure 2 are shown in Table 2. The results from both the process and the review were similar to those from using the rehearsal data on its own. This is encouraging as it suggests that the process using the rehearsal data on its own is fairly reliable. For one group the process had originally passed it for automatic resolution, but multiple administrative records were linked. It was then sent for clerical review, and the reviewer indicated it should not be resolved. This is the type of case for which linking the administrative data is intended to catch, avoiding census records being resolved when they should not be. This particular case was where there were two persons with the same name in the same postcode. The name, though fairly uncommon nationally, was prevalent locally. Methods are available from other tasks that could deal with such cases, and it is intended that these will be adapted to be applied to RMR.

There are also seven groups where the reviewer with only the rehearsal data indicated that the group should not be resolved, but were reviewed as needing resolved when the administrative data were included. This suggests that in some borderline cases a reviewer with just the census data might suspect there are multiple persons, and that multiple persons would then appear on the administrative data. However, once the reviewers know that there are not multiple linked records on the administrative data, then they might be more inclined to think that the census records do represent one person. This suggests that it would be useful for the reviewers to be aware of the results of the administrative data linking, even in cases where there were not multiple administrative data records linked. In these cases the reviewer would not necessarily need to see the administrative data records, they would just need to be informed that there were not multiple administrative data records linked to the records in the group. Similarly, the case that was passed for automatic resolution but the reviewer with just the census data indicated that it should only be partially resolved was indicated to be fully resolved when the NHSCR data were available to the reviewer.

8. Strengths and Limitations

This method has two main benefits compared with the 2011 method. The first is that the linking method is more thorough. While the 2011 only made two comparisons, the proposed method makes many more comparisons, allowing a greater range of problems to be detected and addressed.

Secondly, linking the administrative data improves quality assurance. If multiple administrative data records link to the group (and so there genuinely are multiple persons present) then this will avoid the group being resolved in error. Even when administrative data does not suggest there are multiple persons, this corroboration can increase the confidence reviewers have in their decision to resolve the group.

The main limitation to this method is the amount of clerical review required. Using the full cut of the rehearsal data with 51,080 records lead to 1,886 groups being identified (1,441 to be reviewed). With a population of 5,463,300[10], this would suggest around 200,000 groups being identified, with around 150,000 needing reviewed. This much review would be prohibitively time consuming.

However, having 78 categories of links means that all the links in a particular category will be very similar to each other. If a particular category had a great many links, and it was found that all the groups involving these links were getting the same outcome from review, then that outcome could be applied to all the other similar groups. It is therefore planned that the treatment of the link categories (that is, whether to automatically resolve, review or reject, as indicated in Annex 1) will be kept under review as clerical review progresses. It is expected that this would greatly reduce the amount of clerical review required, especially if many of the links had exact agreement on name, and missing date of birth on a unsubmitted return. In particular, just making a decision on how to deal with cases with exact agreement on name, but missing date of birth more than halves the number of groups to be reviewed.

[10] See the NRS 2019 Mid-year population estimates for Scotland.

An issue was identified when analysing the rehearsal data that affected linking in some rare cases. This related to names that are common in the local area. There are plans to address these issues before the method is used for the 2022 Census. It is likely that this will make use of methods developed to deal with similar issues in other census linking tasks.

The process also needs to be fitted into the end-to-end processing testing, which may raise further operational issues. Similarly how management of the clerical review process will happen is still being reviewed.

9. Conclusion

The methodology for RMR using administrative data to support the quality assurance process outlined in this paper is robust and an improvement on the 2011. This method improves the quality of the census records set, by reducing the potential for an over estimation of the population. The groups of records identified for resolution are quality assured against the administrative data which helps identify groups where the records should not be resolved as they represent distinct individuals. As the NHSCR was successfully used in the test on the 2011 data and the 2019 rehearsal, it is planned that the NHSCR will also be used in 2022.

10. References

National Records of Scotland (2020a), *PMP001: Estimation and Adjustment Methodology*, (online) available at: https://www.scotlandscensus.gov.uk/documents/Scotland_Census_2022_-_PMP005_-_Name_Reordering_Methodology_paper(2).pdf

National Records of Scotland (2020b), *Mid-2019 Population Estimates Scotland*, (online) available at: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/mid-2019

Philips, L., 2000, 'The double metaphone search algorithm', *C/C++ Users Journal*, vol. 18, no. 6, pp. 38–43

Steorts, R., Ventura, S., Sadinle, M. and Fienberg, S, 2014 'A Comparison of Blocking Methods for Record Linkage' in: Domingo-Ferrer J. (eds) *Privacy in Statistical Databases: Lecture Notes in Computer Science*, vol. 8744

Zhao, C. and Sahni, S. (2019) 'String correction using the Damerau-Levenshtein distance', *BMC Bioinformatics*, vol. 20, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6551241/

11. Annex 1: Descriptions of categories of links and which to accept

Cases from each category were manually reviewed by staff experienced in the RMR process to decide whether they should be for automatic resolution, clerical review or rejection. The results of these decisions are listed below.

11.1 Categories of links for automatic resolution

Groups where each record links to each other record in the group with a link in one of these categories will be automatically resolved.

0 A Exact: same gender, missing relationship

0 B Exact: missing middle name, same gender, missing relationship

0 C Exact: missing gender, missing relationship

0 D Exact: missing middle name, missing gender, missing relationship

1 4 Same: first and last exact, DoB diff 1 step, missing relationship

1 5 Same: first and last exact, DoB diff 1 step, relationship

1 6 Same: first and last exact, DoB diff 2 steps, missing relationship

1 7 Same: first and last exact, DoB diff 2 steps, relationship

1 9 Same: DoB exact, missing relationship

1 A Exact: same gender, relationship

1 B Exact: missing middle name, same gender, relationship

1 C Exact: missing gender, relationship

1 D Exact: missing middle name, missing gender, relationship

1 E Exact: diff gender, missing relationship

1 G Exact: missing middle name, diff gender, missing relationship

1 I Same: first and last exact, middle name similar, DoB exact, missing relationship

1 J Same: first initial, last exact, DoB exact, missing relationship

1 L Same: DoB exact, first initial, missing relationship

1 M Same: DoB exact, first exact, missing relationship

1 R Same: first initial, last exact, DoB diff 1 step, missing relationship

1 W Same: missing relationship

1 X Same: first and last exact, birthday different, year exact, missing relationship

1 Y Same: first, last exact, middle name more diff, DoB exact, missing relationship

11.2 Categories of links for clerical review

Groups that include links in these categories will be reviewed. If it is found that links in a particular category are predominantly accepted, or predominantly rejected, then the category can be reclassified either as to be automatically accepted, or automatically rejected.

1 Z Same: first and last exact, DoB 2 part missing, missing relationship
2 B Remaining: missing DoB, missing relationship
2 C Likely same: exact DoB, missing relationship
2 D Likely same: missing relationship
2 E Remaining: first exact, DoB exact, missing relationship
2 F Remaining: missing relationship
2 H Possible parent-child: first and last exact, age diff => 15 missing relationship
2 I Possible parent-child: missing relationship
2B A Same: first and last exact, DoB exact, relationship
2B B Same: DoB exact, first exact, relationship
2B C Same: DoB exact, relationship
2B H Likely same, exact DoB, relationship
2B I Likely same, relationship
2B J Remaining: relationship
3 G DOB similar, missing name, missing gender, missing relationship
3 H DOB similar, missing name, same gender, missing relationship
3 I DOB similar, missing name, diff gender, missing relationship
3 K DOB similar, missing name, same gender, relationship
3 M First, last exact, missing DOB, missing gender, missing relationship
3 N First, last exact, missing DOB, same gender, missing relationship
3 O First, last exact, missing DOB, diff gender, missing relationship
3 P First, last exact, missing DOB, missing gender, relationship
3 Q First, last exact, missing DOB, same gender, relationship
3 S Name same, missing DOB, missing gender, missing relationship
3 T Name same, missing DOB, same gender, missing relationship
3 V Name same, missing DOB, missing gender, relationship

3 W Name same, missing DOB, same gender, relationship

3 Y First name same, missing last name, missing DOB, missing gender, missing relationship

3 Z First name same, missing last name, missing DOB, same gender, missing relationship

4 A Don't know: missing name, missing DoB, missing gender, missing relationship

4 B Don't know: missing name, missing DoB, same gender, missing relationship

4 C Don't know: missing name, DoB exact, missing gender, missing relationship

4 D Don't know: missing name, DoB exact, same gender, missing relationship

4 E Don't know: missing name, DoB exact, missing gender, relationship

4 F Don't know: missing name, DoB exact, same gender, relationship

4 H Don't know: missing name, DoB partial agree, same gender, missing relationship

4 K Don't know: missing name, DoB similar, missing gender

4 L Don't know: missing name, DoB similar, same gender

4 M Don't know first partial agree

4 N Don't know last partial agree

5 A Don't know: missing name, missing DoB, missing gender, relationship

5 B Don't know: missing name, missing DoB, same gender, relationship

5 D Don't know: missing name, DoB exact, diff gender, missing relationship

5 Don't know: missing name, missing DoB, diff gender, missing relationship

5 E Don't know: missing name, DoB exact, diff gender, relationship

11.3 Categories of links for rejection

A sample of cases with the strongest of these links will be reviewed for quality assurance.

6 A Different: parent-child including relationship

6 B Different: parent-child

6 C Different: parent-child: missing DoB, same name

6 D Different: sibling – twin

6 E Different: sibling – other

6 F Different: relationship or gender diff

6 G Different: relationship and gender missing

6 H Different: relationship and gender diff

6 I Different: other, missing relationship

6 J Different: other, relationship

12. Annex 2: Scoring of Name Comparisons

This section discusses in detail how the for scores (which indicate the strength of evidence for two records representing the same person) and the against scores (which indicate the strength of evidence for two records representing the different persons) are calculated for the various components. There are a number of attempts to find evidence for a match. Each one will update the for and against scores only if that will strengthen the evidence for a match.

Missing Names

If name is missing on one or both records then the for and against scores are both 0. Otherwise if a name component is exactly the same between the two records then the for score is 50 (25 for middle name) and the against score is 0.

For first names there is also a check for the name being 'BABY' on both records. In this case the for and against scores are both set to 0 as the guidance (in 2011) indicated that unnamed infants should be recorded as 'BABY'.

Nicknames

Another check for first names is nicknames. Thus if we had 'Alexander' on one record and 'Sandy' on the other then it is quite plausible that these are the same person, even though the first name strings are quite different. To perform this check we make use of the nickname linking variable. That variable is set to a particular value for a range of names that have the same nickname. Thus if first was either 'Alexander' or 'Sandy' (or 'Alex', 'Xander', and others) then the nickname variable is set to 'Alexander'. (The name groupings were built up manually, assisted by exploring links between datasets where last name, date of birth and postcode agreed, but first name did not.) Thus if the first names differ between records but the nicknames agree then the against score is set to 0 and the for score is set to 20. Some of these are specific to a particular sex. Thus if the first name is 'Alex' then the nickname will be set to 'Alexander' if sex is male and 'Alexandra' if sex if female. There is also a second nickname variable that groups together more tenuous name groupings such as 'John' and 'Ian', which results in a for score of 10.

The nickname check also detects alternate spellings of the same name, such as 'Nicholas' and 'Nicolas'. This may be particularly important for Census Coverage Survey linking when data is reported verbally and spellings may not be confirmed. In total there are 189 groupings defined, and 45 more tenuous ones.

Character comparison for names

If none of these situations hold then the name components in the two records are compared at the character level using a method inspired by the Damerau–Levenshtein edit distance[11]. The characters in the name from one record are linked to those in the name from the other record. This is done by first comparing the characters at the same location in the strings. If these do not agree then this moves to adjacent letters, and then letters at a distance of two, and so on. Once this has completed there is a tidying up stage to ensure that adjacent letters are linked to letters at the same distance if possible.

Once the letters have been linked they are then analysed in order to identify the substitutions, transpositions, deletions, insertions and jumps would be required to transform one string into another. For each of these there is an associated score. These scores depend on the letters involved. For example if we need to insert a 'W' then that would attract a larger penalty than if we only need to insert a 'I' because a mark on a page may be mistaken for an 'I' in scanning, but is unlikely to be mistaken for a 'W'. Similarly for substitutions some changes are more plausible than others. Combinations like 'U' and 'V' can be easily confused, as can 'O' and 'D'. In total 50 such combinations are noted.

The scores from all the individual differences are then combined to give an overall score. That score is then converted to scores for and against the records being a match.

[11] See Zhao and Sahni (2019) and references therein.

Swapped first and last names

Sometimes people enter their names in an unexpected order. To account for this a comparison is made between the first name of one record and the last name on the other record and vice versa. If these both agree then the for scores for both first and last names are set to 40. If only one of these agrees then one of these scores is set to 40, while the other is set by doing the character comparison on the differing values. That is, if first_1 agrees with last_2 then the first for score will be 40, while the last for score will be set by doing a character comparison between first_2 and last_1.

Titles

If first name begins 'MR ' or 'MRS ' then that part is removed from the first name and stored in a variable called title. If the two records being compared both have 'MR' and 'MRS' respectively in their title variables, and their sex agrees with this information, then a penalty of 20 is combined with the for and against scores for first name.

Comparison to middle name

Some people go by what is officially their middle name. In order to successfully link these cases the first name for one record is compared with the middle name of the other. If this agrees then the for score for first name is set to 15 (unless it was already over 15). A similar check is also done between last name and middle name.

Compare name parts

Some people have double-barrelled first or last names. However they may go by only part of this. For example 'Sarah-Jane' may go by Sarah, or even Jane. To detect such cases we make use of other linking variables that pull out parts of names that are delimited by special characters. If these agree with the name from the other record then the for score is set to 25 (unless it was already over 25). This is done for first names and also for last names. In other comparisons special characters (including spaces) are removed before the comparison is made.

Comparing first letters of name or Double Metaphone code

The next check is to count the number of letters that agree at th
e start of the name from the two records. If so then the for score is set to be that given in Table 3. This covers a range from one letter agreeing to five (or more) letters agreeing. If only one letter agrees then this is treated differently, so that this method is used only if one record only has the initial (e.g. if one record had 'Peter' and the other had 'P', but not if the other was 'Paul'). These scores are only used if they result in a higher for score than would otherwise be. Another exception is when 3 or fewer letters agree and the names are distinct but common. For example if we had Mary and Margaret then the first three letters agree, but as the names are common then this is not used to score the similarity.

Table 3 The for scores assigned when the first part of the name agrees either on the name itself, or the Double Metaphone coding of it. If only one letter agrees then this method is only used if one of the records only has one letter. * When only 1 letter agrees on name then this is only used if one of the names only has one letter.

Number of characters agreeing	Score when characters agree in:	
	Name	Double Metaphone of name
5+	20	20
4	13	13
3	7	9
2	3	4
1*	10	-

Similarly the first characters of the Double Metaphone[12] are compared. The Double Metaphone is a phonetic code, so this allows for detection of cases where a name has been written differently, but sounds the same. This is another situation that may be particularly common for verbally reported data such as the Census Coverage Survey. As a character in the Double Metaphone code can relate to more than one letter in the original string, agreement on Double Metaphone can indicate stronger agreement than agreement with the same number of letters on the original string. Therefore these scores are slightly larger than the equivalents for the agreeing letters on the original name.

[12] The double metaphone was presented in Philips (2000).

There is an exception when comparing the last names on the original string or Double Metaphone. If the last name begins 'Mc' or 'Mac' then the count of the agreeing characters is reduced by 2 and 3 respectively. This is because names beginning this way are so common, while being very distinct. Therefore we would not want to say that MacDonald and MacPherson were as similar as Scalon and Scanlan.

Full name

Sometimes a space is missing between the first and middle name, meaning that the middle name gets concatenated onto the first name. Other times a space gets inserted between letters of the first name, meaning that part of the first name gets put as the middle name. Another issue is that the whole name can be entered in the first name field.

All these issues can be resolved by considering the full name, that is, the concatenation of first, middle and last names (with spaces and other special characters removed). This full name is one of the linking variables used. It is compared between the two records. If it is not exactly the same then a character comparison is done. This allows a for and against score to be calculated for the full name. If this score is better than the for scores for first and last name then the first and last for scores are amended using the full name for score.

13. Annex 3: Scoring of Sex and Date of Birth

Sex

If sex is missing on either record then the for and against scores are both zero. Otherwise if sex is the same then the for score is 5, while against score is 5 if the sex is different.

Date of Birth

If the day, month and year components either agree between the records, or are missing on one of the records, then we count the number of these components were at least one of the records is has missing information. The for score is then given by: $12(3-m)$, where m is the number of components that are missing on at least one of the records. The against score is 0 in such cases.

If the dates of birth are non-missing on both records, the years agree and the day and month agree with the month and day on the other record then the for score is 20 and the against score is 0. This is to account for cases where the date has been entered in American format on one of the records.

Table 4 Sets of digits that may be confused in scanning, and so are given a smaller difference penalty.

Set of Digits
2, 4, 5
8, 9
1, 7
3, 5, 8
2, 7
2, 3
5, 6
7, 9

If the two dates of birth are complete then the individual digits are compared. That is, the first digit of the day of birth from one record is compared with the first digit of the day of birth from the other record, then the second digit and so on. If the two digits are both in one of the sets given in Table 4 then we count this as a difference of 1. All other differences are counted as a difference of 2. (The particular sets of

digits are chosen to be those that are often confused in scanning, so are more likely to be the same than for other pairs of digits.) These differences are then totalled across the whole date of birth.

There is an exception for the century. If this differs between the records then it gets counted as a difference of 2, rather than comparing each digit. This is because people sometimes confuse the century in the year if they are used to writing, for example, 19-- instead of 20--.

Another exception is if a digit appears in a different position in the component. For example if day was 21 on one record and 02 on the other then it may just be that the '1' was missed on one side and a leading zero added. Such cases when one record has a leading zero would then get counted as a difference of 2, rather than 4.

The totalled differences (*d*) are then put into the following formula: $6(3 - d - 2m)$. If this is positive then it is used for the for score (with against score being 0), and if it is negative then the for score is 0 and the against score is the absolute value of the formula.

A final check is to count the number of components (day, month and year) that are different. If only one is different, then the against score is set to 0.

14. Annex 4: Information Governance

As with other linking to administrative datasets, this has been conducted in compliance with GDPR. The NHS Central Registrar was used as the administrative dataset for this quality assurance procedure, and the standard governance procedures were followed in this case. Only the Admin Data team will be working with this administrative data and it is only being used for quality-assurance processes.

More information on this can be found published on the website:

Data Protection Impact Assessment for use of NHSCR dataset
Quality Assurance report for use of NHSCR dataset for 2019

15. Annex 5: Glossary

Term	Definition
Link	Two records that have been connected
Match	Two records that represent the same individual
Non-match	Two records that represent different individuals
Clerical Review	A process where an individual statistician manually reviews particular cases in order to make decisions on how to proceed with the case (for example, remove it, merge it, move to next process). This generally happens with cases that are ambiguous in some respect. For links this may be when information across the records is similar but not identical.
Coverage Adjustment	A process involving linking to the Census Coverage Survey, where undercount in the census can be accounted for.
IAC	An Internet Access Code (IAC) is linked to an enumeration address or enumeration sub-address and is provided to the respondent. This can be either a household or a communal establishment enumeration address or enumeration sub-address. The IAC is used by the respondent when they log into the online instrument and associates the response with an enumeration address or enumeration sub-address. An enumeration address or enumeration sub-address can have more than one IAC.
Processing Unit (PU)	Used in the 2011 census. A processing unit (PU) is made up of one or more neighbouring council areas (CAs) and covers around 500,000 respondents. CAs were grouped in to PUs for practicalities around data processing.
Unsubmitted Return	An online unsubmitted return is an online return where the respondent has not completed the online collection process by submitting their questionnaire responses to National Records of Scotland (NRS).