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

Scotland's Census 2022

Remove False Persons – Methodology

August 2020

National Records of Scotland

Contents

1. Plain English Abstract

There are a number of processing issues which occur when receiving Scotland's Census responses. For example, on paper questionnaires scanners can sometimes pick up dust on a blank question and record it as a valid mark, thereby creating a response. However, these marks do not relate to an actual answer from a genuine individual. Generally, these issues affect a small portion of responses received, as most of them arise from errors in scanning of paper questionnaires — the majority of responses in Scotland's Census 2022 will be online. Nonetheless, such issues can falsely increase the number of people counted in the Census (often referred to as overcount). For that reason, there is a part of statistical data processing called Remove False Persons (RFP), which looks at the possibility of a non-genuine person being made into a person record.

In 2011, the RFP process included one check, a '2 of 6' rule, where a record must contain at least two of six key variable groups for it to be considered valid. Records which do not pass this check are not taken through further processing on the basis that it is unlikely that the record relates to a genuine person.

This process worked fairly well, but for Scotland's Census in 2022, the Remove False Persons methodology will refine what was previously used in 2011 by adding two additional steps, which will a) review certain records that do not originally pass this check for further scrutiny, and b) review information left by a respondent in the name field(s) for clues as to the nature of the record. This in turn will reduce the burden on later processes to account for such records (for example, reducing the amount of overcount adjustment that Estimation and Adjustment makes down the line).

2. Abstract

A raw census dataset sometimes contains blank or mostly-blank records, which may not belong to a genuine person. Such records are usually created at the data capture stage for a number of reasons, such as scanners recording paper-dust as tick or text, or respondents crossing through individual forms on the paper questionnaires (meaning that if the score runs through a tick or text box, it gets picked up as a character and a record is created). In such cases, no person is related to these records.

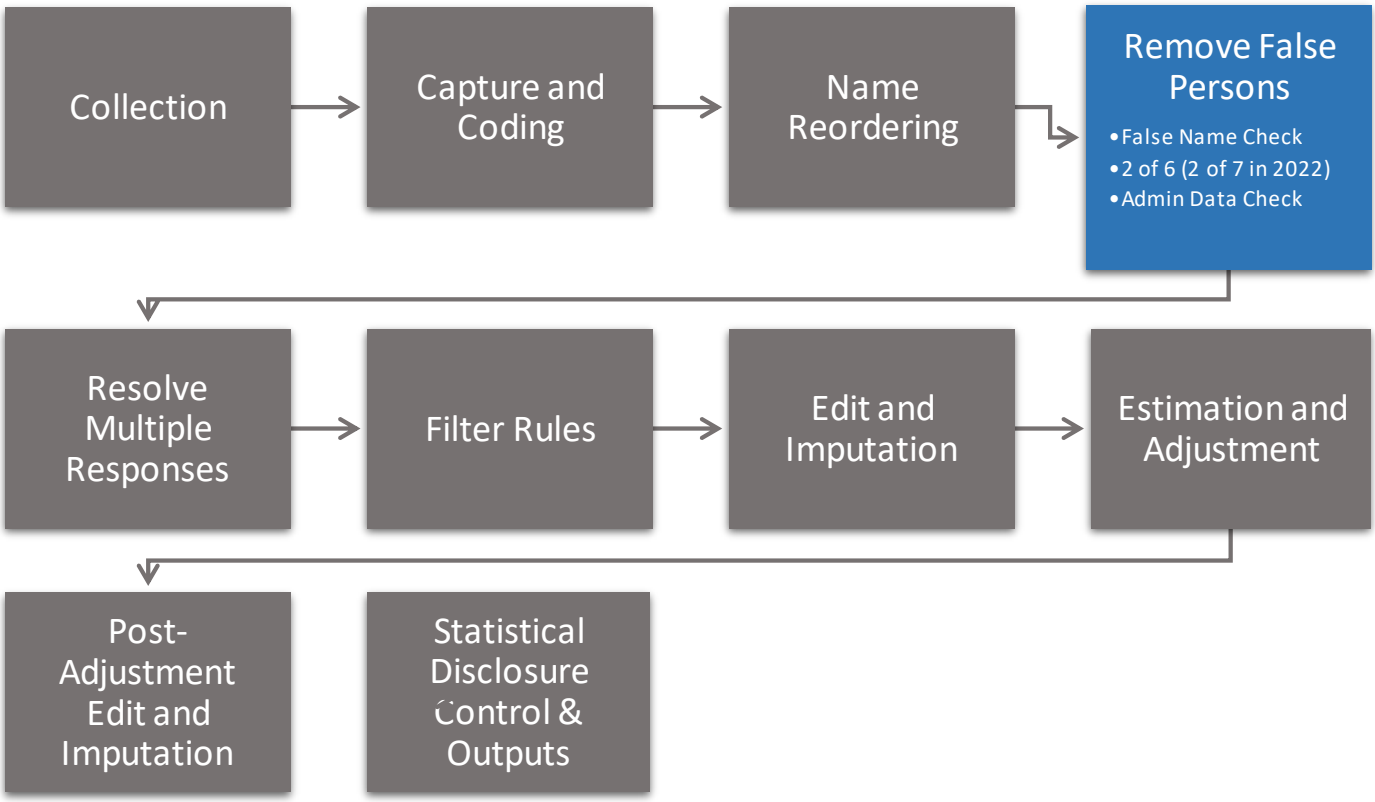
However, keeping them in a census dataset creates overcount, and burdens further statistical processes which are required to adjust for it. Scotland's Census therefore has a data cleansing step called 'Remove False Persons' to account for this. In 2011, this step checked a record for at least two of six key variable groupings, where one **must also** be a name or date of birth.

For 2022, this methodology is to be expanded to include the addition of two new checks or filters — one for potentially false name strings (such as 'anonymous' or 'no one'), as well as an administrative data check which looks at cases where minimal information is given, but there is potentially an indication of a genuine person (for example, a name but nothing else). This should refine the process further to catch the most obvious 'false' records in the dataset, which in turn lessens the burden on statistical processes further down the line.

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. Background and Introduction

Diagram 1: Simplified Overview of the Census Data Journey[1]



3.1 Issues and Historical Development

Remove False Persons, or RFP, is a data cleansing process mostly run by the Statistical Methods and Data Processing team. It essentially deals with spuriously created records in the census dataset, in a preliminary effort to reduce overcount.

The primary rule, often referred to as the '2 of 6' (named for its requirement to contain valid responses for at least two of six specific variable groups as an indication of a genuine individual), is a part of a process called Remove False Persons. A variation of this rule was first implemented in 2001. Evaluation of the 1991 Census discovered records which were created erroneously — records which had virtually no information on them. The information that was there was

[1] Exact sequencing of data flows are still to be finalised.

inconsistent and sparse, usually in the form of an odd character (for text answers) or a single tick, and the 2 of 6 (or a variation thereof) was created and implemented for the 2001 Census.

Such spurious records are created for a myriad of reasons, mostly due to some type of scan or capture error. Some of the most common instances include:

- Scanners registering dust and debris from guillotined questionnaires as marks on a page, and recording them as tick or text;

- Questions, or sometimes whole pages being crossed out by a respondent — if the cross runs through a tick or text box, this can be picked up as a legitimate answer to a question;

- The writing of 'N/A' on questions which respondents were not required to answer.

On paper questionnaires, there are also instances found where respondents skipped pages, usually those at the end of one person form and the start of the next person form. The respondent would continue to fill out the questionnaire, not noticing the discrepancy — which meant that one person's information was spread across two records, again resulting in overcount.

Prior to 2011, there was no online questionnaire mode (and this was limitedly used in 2011), so historically the development of RFP was aimed at addressing issues which arise from completion of the Census via paper. For Scotland's Census 2022, the main completion mode has shifted from being majority paper to majority online, and as such census questionnaires submitted through the Online Collection Instrument (OCI) automatically passes the primary RFP rule as a feature built into its design.[2]

[2] Validation, i.e. error messages, will prevent the progression or submission of a questionnaire without name and date of birth (the essential component in this rule), and will also pipe the first entered name to all other appropriate name fields, thus passing 2 of 6 - or, in 2022, 2 of 7 - automatically.

Scotland's Census 2022 will also capture responses from the OCI at the end of the collection phase which have been left idle (i.e. not submitted), to ensure that forgotten submissions are not discounted.[3] Generally, these are respondents who either cannot get back into the questionnaire to complete it (and in these cases may submit another response), or respondents who find the online questionnaire too frustrating to use and abandon them, submitting a paper response instead. Although respondents will be prompted to complete their questionnaires online or create a password so they can come back later, if unsubmitted responses are found at the end of the collection phase, by nature these records will be incomplete to varying degrees. Thus, records collected by the OCI will also be subject to RFP steps in addition to the paper questionnaires received during the Census in 2022. Remaining cases which pass the RFP stage but have other records associated with them will be dealt with by the more rigorous Resolve Multiple Responses (RMR) processing step.

The variation of the 2 of 6 Rule used in the 2001 Remove False Persons process was found to have worked relatively well, and was implemented again for the 2011 Census (as the 2 of 6). It accomplished the main goal — to identify those records which were clearly false and largely blank, (i.e. there was not enough information to say that it came from a plausible individual), and prevented them from passing into further processing.

3.2 Remove False Persons for Scotland's Census in 2022

For the Census in 2022, there are checks in addition to the main rule (2 of 6) proposed, to enhance the method further by removing *or* retaining records with greater precision.

The introduction of a check for strings or phrases in the name field can assist in determining why a record was created. Because the name question comes first on a questionnaire and is also a text field, a respondent sometimes includes a note as to

[3] Note that secondary, or duplicate submissions which may stem from this will be dealt with in a separate statistical cleansing process called Resolve Multiple Responses, and is not covered here.

why the rest of the form could not be completed. Searching for commonly used notes or phrases (e.g. 'no one', 'n/a', 'none') would allow processing to further filter records which otherwise pass this criteria for the 2 of 6. As an illustration, a record may pass RFP if a respondent writes, 'Not Applicable' in the name section, and crosses through subsequent pages/questions, registering tick marks. However, since the record does not pertain to a person, this should not be included in the count.

There may also be individuals who decide to complete the Census, but do so in an anonymous way – for example, by obscuring their name or date of birth by writing, 'Anonymous' or '00/00/0000'. In such situations, the information contained on the record is valid, but processing cannot plausibly relate it to a genuine individual. In these cases it may be best to prevent these records from passing the check, and allow adjustment[4] to account for them in subsequent statistical processing. This is of particular concern with increased awareness around data privacy issues in the general public, and as such more responses in 2022 may reflect those who choose to return their census questionnaire in this manner.

There are some, though comparatively fewer situations where the Remove False Persons filters are too aggressive. Such records do not pass *either* the 2 of 6 or the name check, but may still contain just enough information to potentially be genuine people. For these, further quality assurance will be provided by comparing them against administrative data sources. This would confirm that the record is indeed related to an individual, and thus should be retained within the census dataset.

It should be noted that in the 2022 Census, there are three places where respondents are asked to list their name, and the order in which their names appear is important for household alignment (so which responses pertain to which individual can be sorted). Out of alignment, it is possible that certain sections (which, for

[4] More information on the adjustment process can be found in the Estimation and Adjustment Methodology paper
https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202022%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf

example, may be blank) pertain to an incorrect individual. To address this, the Remove False Persons step comes after a 'Name Reordering' processing step[5]. Cases that would have failed RFP due to names being misaligned throughout the questionnaire will be realigned in name reordering, and should therefore go on to pass the RFP stage.

This paper details the method used for the Remove False Persons process in 2011 and examines the changes proposed above. Analysis using the 2011 Census dataset (and lately the 2019 Rehearsal dataset) has been done in order to quantify the scale of the primary issue – ensuring that blank or mostly blank records which do not pertain to a genuine individual do not falsely inflate the census count. The further checks proposed are about ensuring that records from legitimate individuals are retained, offering a granularity of accuracy that still allows for the reduction of burden on later statistical processing. This work has additionally been used to extrapolate run timings and potential resource or time savings which could be applied to the 2022 Census.

3.3 2011 Method

The Remove False Persons process in 2011 consisted of one deterministic (or rules based) filter called the '2 of 6' – so named for the requirement of a person record to have validly filled at least two of six key questions (variable groups) on the census questionnaire (see Table 1 below). A record passing this check meant it went on through further processing, while not passing the check meant the record was removed from the main census dataset, as it was deemed unlikely that the record was generated by a genuine individual.

[5] More information can be found at:
https://www.scotlandscensus.gov.uk/documents/Scotland_Census_2022_-_PMP005_-_Name_Reordering_Methodology_paper(2).pdf

__Table 1__: The Variable Groupings in 2011's 2 of 6 Rule

	Variable Group and Description	Validity
1	Household Name First or Last name in the *household member listing* of the Census questionnaire	Valid name fields in 2011 included *any* character in the name fields (i.e. initials are acceptable)
2	Person Name First or Last name in the *person section* of the Census questionnaire	As above
3	Date of Birth fields	Valid date of birth in 2011 included a valid month **OR** a valid year
4	Variables which describe relationship	Any indication of a valid relationship (tick response)
5	Sex	Any tick response
6	Marital Status	Any tick response

One of the passing 2 of 6 variable groupings **must also** have been name on either section (on the household table or on the person form) or date of birth. For example, if two tick questions such as sex and marital status were filled but no others, the record would not pass.

In 2011, only the relationships of the record in question *to* others in the household were considered at RFP, as those relationships *from* the record were derived in later processing (after RFP). For consistency, the analysis in the sections below follows the same approach. However, in the 2019 rehearsal and in Census 2022, relationships both *to and from* other members in the household will be provided right from the point of coding. This means that for 2022, *any* relationship provided will count as having data present at RFP (and therefore valid).

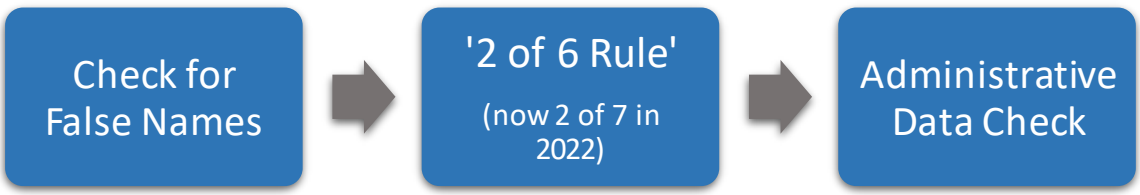
In certain situations[6], a person's living circumstances naturally excludes criteria from applying to the record. For instance, if an individual was living alone, they would not be required to fill in the relationship matrix, and thus the relationship criteria would naturally not apply. However, adjustments to the number of variables necessary to pass the filter were not made (e.g., only requiring 1 of the 6 variable groups), nor were variables substituted for different ones. The requirement of needing at least 2 of the aforementioned variable groups remained, and one of these must be a name or date of birth. In this example, it would essentially mean that a 'single person household' is subject to a rule equal to 2 of 5.

In the 2011 Census, the use of administrative data was an untested concept and therefore unavailable. However, over the last decade, there has been an increased use of administrative data to support the processing of high quality statistical outputs. It was felt that the opportunity to use such quality assurance techniques needed to be explored for Scotland's Census 2022.

4. Proposed 2022 Method

The proposed sequence of steps with the Remove False Persons process is:

***Table 2**: The Order of Steps in the Remove False Persons Process for 2022*[7]



[6] Please see Appendix for other scenarios where this applied.

[7] When we refer to 2 of 6 here, we mean the variation of the rule that the record is naturally subjected to as described in Section 3.3. In 2022, the base 2 of 6 rule will become 2 of 7 as there is an additional (name) variable group on the 2022 questionnaire. Please refer to the rest of Section 4, specifically 4.2 for more details.

Each step will be addressed in the proposed running order, below.

4.1 Check for False Names

In the 2011 Census, it was found that there were sometimes strings that were not actual names written into those fields. The content of these 'false names' were often dependent upon what the respondent was attempting to accomplish, but were generally for two purposes:

1) To assist in 'completing' the questionnaire, i.e. those responses which, in themselves, contain the reasons for writing responses - such as 'No one here' or 'I live alone', and

2) Purposely disguised names, in an effort to complete the questionnaire anonymously, such as 'anonymous', 'anon', and sometimes 'N/A'

In 2022, this may impact the Census dataset in two ways:

1) Cause a record to pass the 2 of 6 Rule where it would not otherwise have been wanted to accept the record (which occurred in 2011 as well), and

2) Hinder the ability to match records in subsequent processes, e.g. Census to Census Coverage Survey

To investigate such issues, analysis conducted on the 2011 Census dataset at the first processing step out of Coding located records with chosen strings in the person section name fields. These strings are common ways for a respondent of a questionnaire to express information, but do not attribute themselves to a genuine individual:

- ANON
- ANONYMOUS
- N/A
- NO ONE
- NONE

The check also accounted for several variations, either from different spellings or poor data captures. For example, respondents may use NA as a form of N/A or, NOONE may appear as a version of NO ONE. However, it is important to note that both NA and NOONE are proper names. As such, although the check itself is an automated process, it will be necessary to send records which do not pass this check to a reviewer for clerical review[8] (also called manual review or inspection), in order to distinguish between those which are properly false names (or miscaptured) to those which are genuine.

4.1.1 Results from the Test on 2011 Census Data

The test on 2011 data found 541 records which were potentially false in the first name field (Table 3), and 112 records which were potentially false in the last name field (Table 4; in both cases, using the name field from the individual section of the questionnaire). The dataset was then taken from the automated check and passed to a reviewer, who manually inspected the records in the dataset, and placed them into the following categories:

[8] Clerical review is a process whereby an individual manually calls up the record in question and looks at all relevant information to determine the outcome of a decision. For example, in this check, a clerical review would involve looking at the dataset of names to make the three determinations as outlined above, mark the record according to process and move on. Clerical review can involve looking at the dataset OR the scanned image of the record in question (note that online questionnaires will not have corresponding images). Also sometimes referred to as manual review.

Table 3: *Number of Potentially False Names in First Name field*

Name String	Frequency	% of *pername1* Total
ANON	*[9]	*
ANONYMOUS	*	*
N/A	219	40.48
NO ONE	*	*
NONE	128	23.66
False Name Variations[10]	179	33.09
TOTAL	**541**	**100.00**

Table 4: *Number of Potentially False Names in Last Name field*

Name String	Frequency	% of *pername2* Total
ANON	0	0.00
ANONYMOUS	0	0.00
N/A	42	37.5
NO ONE	12	10.71
NONE	0	0.00
False Name Variations[11]	58	51.79
TOTAL	**112**	**100.00**

In a live-running situation, it would be up to the reviewer to determine if these names are truly false, or a variation of a false name that is actually genuine (e.g, 'NA'). If false, the record would be flagged and changed to *missing or invalid* before being put through the 2 of 6 rule.

[9] Small frequencies recorded - under 10

[10] Includes several potentially false name variations from poor capture – however, these names are also potentially genuine and so are not listed here.

[11] As footnote (9) above

Another part of the 2011 test was taking a processing unit[12] to inspect for trends and variations other than the more obvious ones (as in the test) in name fields. These were not analysed in detail for the test above as they were often one-offs, but gives an idea of similar patterns and responses to look for.

Development of the 2022 filter will take variations of such strings into account. This sample list includes the following strings (strings appear as captured, with no changes to typo or capture mistakes):

- 1 PERSON LIVE ALO
- AS W3 PAGB 4
- AS IN 1 OF H13
- AS PAGE 6
- ASH3 PAGE 4
- NI/A
- NO OTHER PERSON L
- NO PERSON 3
- NO PERSON 4
- NO PERSON 5
- NO PERSON TWO
- OCCUPIER
- SAME
- SAME AS FIRST
- SAME AS ONE
- SAME AS PAGE 4
- SAMEAS H3/PAGE4
- SEE #
- SEE H3
- SEE INDIVIDUALS
- SEE PERSON 2
- SEE PERSON 4

Many false names appear only in one name field (either first name or last name, but respondent behaviour shows this is usually first name, likely as this is the first text field that appears on the questionnaire). This example list was created by searching for those name groups missing last name and scanning for strings in the first name field that are unlikely to be names.

[12] In 2011, a processing unit was a subset of Census or Census Coverage Survey data that was processed together, through all stages of data processing.

4.2 Results from the Test on 2019 Rehearsal Data

Performing a similar test to the 2011 Census data on the 2019 Rehearsal data did not yield useful results that could be used for the purposes of analysing respondent behaviour in respect to the name check. Although the dataset contained roughly 44,000 records, there were only a handful of records where *any* of the name fields contained the more commonly used 'false' strings. This was somewhat expected as a natural result of a voluntary survey, as the Rehearsal was.

Testing and development of the process continues, however; all data sources to hand (2011, rehearsal and synthetic) can be used to test this process.

4.2.1 Clerical Review for the False Name Checks

In all cases, all records flagged will be reviewed, in order to:

a) detect patterns in respondent behaviour (e.g. are there many more 'anonymous' records than expected?),

b) to ensure it is not a real name (for example, while N/A is often written in place of 'not applicable', it may also be mis-captured as NA or NIA – both genuine names), and

c) determine which *type* of false name it is. This is particularly important in the case of N/A variations, where respondents can feasibly write this both for a case of 'there is no person 3' and as a method of obscuring identity.

A test run of the false name dataset created for this analysis showed that it would be possible to review approximately 20 records in image format per hour. However, as names are the only concern for this particular filter, a review of the dataset containing flagged records for review should be sufficient to make accurate judgement calls about the legitimacy of many of the names. This method of manual review allowed all 541 records (from the earlier indicated test) to be checked in

approximately 60 minutes. These records can then be flagged appropriately (see below sections for detail) and passed to the next step in the Remove False Persons process, the 2 of 6 Rule.

4.2.2 Check for False Names - Resolution (No Genuine Person)

Those names which have 'NO ONE LIVES HERE' (for example) are generally done by householders who wished to communicate that there was not another person in the household. They intend to be helpful or ensure the respondent fully complies with 'completing' the Census questionnaire. In 2011, 2 of 6 was fully automated, and so having phrases such as these may have caused the record to pass this check where otherwise it should not have been retained; for example, cases where 'no one' was written for name and paper capture believed the sex variable was ticked. This overcount would subsequently have to be addressed in Estimation and Adjustment methodology.

Another concern is that where the names such as 'NO ONE' appears, these records can't be highlighted in the subsequent matching processes required for estimation purposes (for example, Census - Census Coverage Survey linking).

An indication of no genuine person, where a householder intended to communicate that there is no one present, should be treated differently than those who wish to obscure their information. In these cases, householders effectively state that there is not an individual to tie to the record, and as such the name field should be changed to 'missing' before the record is put through the 2 of 6 rule, flagging the record to indicate why this was done. The record can then proceed through 2 of 6, where the record should subsequently be removed.

4.2.3 Check for False Names - Resolution (for Privacy Reasons)

The difference for this type of name check compared to those cases above is where the householder is willing to state that there are people on the premises, but want to obscure their information so that the record is not personally identifiable. Usually,

this means that they have entered false names such as 'anonymous', 'householder' or 'occupier', or (less commonly) have written their birthdate, for example, as 01/01/0101. However, these records may be also willing to complete the subsequent questions as they contain less personally identifying information, and this behaviour was seen in some 2011 returns.

As an indication of a genuine person, records should not be outright discarded, as the quality of such information, though low at times, is still preferable to full skeleton records[13]. This type of situation is of particular concern with the increase of awareness around data privacy issues in the general public, and it is anticipated that some people in 2022 will choose to return their Census in this manner.

Because these records may *potentially* be of use for later processing, they will be flagged and allowed to pass to through the 2 of 6 Rule to give them a chance at being retained. However, these records also cause concerns for statistical matching purposes. As such, it is only once they have passed the 2 of 6 Rule that the names would then be set to missing. From there, the flag would indicate that the record should be reviewed by the administrative data check.

4.3 2 of 6 Becomes 2 of 7 in 2022

The 2011 2 of 6 method has become the core process — not only for Scotland's Census, but also for ONS and NISRA[14] — for resolving issues surrounding spurious records, and will continue to be used it in 2022. The process is programmed and run in SAS, a commonly used statistical programming language that is accessible by all Census statisticians and in the wider Scottish Government. It is a powerful system that allows for the quick processing of procedures such as statistical data cleansing.

[13] Skeleton records are records created in the statistical Adjustment process, to make up the households and people that were missed in the Census (calculated at Estimation). These records are created with a minimal subset of variables – hence the name, 'skeleton'. They are later 'fleshed out' at a secondary Edit and Imputation stage.

[14] Office for National Statistics, who look after the Census for England and Wales, and Northern Ireland Statistical Research Agency, which administers the Census for Northern Ireland.

There are some changes that are necessary, however, to make this rule work with the structure of the Scotland's 2022 Census, aside from changing data structures and formats:

1. Names (both first and last) will now be captured in three sections rather than two:
    a. Household Member's Table on the household section (as in 2011)
    b. Relationship Matrix Table (new for 2022)
    c. Person/Individual section of the questionnaire (as in 2011)

Since name is one of the primary criteria that a record must have to pass the rule, the 2 of 6 will expand to take this new section into account; thus, the full criteria a record is checked against becomes **2 of 7** (please see Table 5, below).

Table 5: *The Variable Groupings for the 2022 2 of 7 Rule*

	Variable Group and Description	**Validity**
1	Household Name First or Last name in the *household member listing* of the Census questionnaire	Valid name fields in 2022 include *any* character in the name fields (i.e. initials are acceptable)
2	**New** Relationship Matrix Name First or Last name found on the *relationship section* of the Census questionnaire	As above
3	Person Name First or Last name in the *person section* of the Census questionnaire	As above
4	Date of Birth fields	Valid date of birth in 2022 include a valid month **OR** a valid year
5	Variables which describe relationship	Any indication of a valid relationship (tick)
6	Sex	Any tick response
7	Marital Status	Any tick response

2. Age routing for the Marital Status question

In 2011, the Census asked the Marital Status question of all people, including those under the age of 16. For the 2022 Census, this will not be the case. This means that if a respondent is filling in the Census online, and is under the age of 16, they will be routed past the Marital Status question without the opportunity to complete it. When completing a paper questionnaire, a respondent is also directed to skip the question, although the possibility remains that they may still answer. This becomes another scenario where, due to the specific circumstances of the individual, the criteria for this rule is altered slightly, as Marital Status will not apply to under 16s. However, the essence of the rule remains – 2 (of now a remaining 6) variables are required to be filled, and one must be either a name or date of birth to pass, similar to the single person household example used earlier.[15]

Once the 2 of 7 rule is applied, those records which pass the filter (those which have at least a name or date of birth, and any other of the variables) move along for further processing – the next stage of which would be the administrative data check (run by the Census Admin Data team) in Remove False Persons. The records which do not pass are removed from the primary census dataset.

In some cases, a household may be completely 'emptied' of person records - for example, if only sex is contained on each of the person records, they will not pass the 2 of 7 Rule. In such situations, the person records are discarded while the household records are carried through further processing. The Edit and Imputation process ensures that missing questions on the household record are imputed, and adjustment subsequently uses the now-empty households as placeholders in the Adjustment process, filling them with person records if and where appropriate.[16]

[15] Please see Appendix for a list and description of all scenarios.

[16] More information on Estimation and Adjustment methodology can be found in this paper. https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf

4.4 Administrative Data to Quality Assure the RFP process

There are some cases where a record may fail the 2 of 7 check, but includes some key information (name or date of birth) that would enable the record to link to an administrative data source. By coupling this with the postcode information from the associated questionnaire, a corresponding record found in the administrative dataset would provide an indication that the Census record represents a genuine person, even though it would normally fail the 2 of 7 check. As such, the aim for Census 2022 is to implement such a check on these types of records, so if a corresponding administrative data record is found, the census record will be retained instead of being removed as it would be normally.

If neither name or date of birth is provided however, any link to an administrative data record would not be strong enough to be confident that the return represents a person who has been found in the administrative dataset. Therefore, this check will *only* be performed on those records which a) have not passed the 2 of 7 Rule, and b) have either a name or date of birth recorded.

The administrative linking process is also written in SAS, and makes use of linking methods developed for other census processing tasks. As described below the linking method will consider records in the same postcode. When name is available the pairs are scored, and the scores used to categorise the links. These in turn are used to decide what to do with each link.

4.4.1 Blocking

When linking is carried out, each Census record that has not passed the '2 of 7' rule is compared with each administrative data record that is in the same postcode as the census record. Without both name and date of birth, it is not possible to confirm that records from different postcodes represented the same person, so only exact agreement on postcode is considered.

If date of birth is available, then it is also used to block[17] the linking. That is, links are made between the census records and administrative data records only when they agree *exactly* on postcode and date of birth. This is because, without name, it is not possible to confirm that records with different dates of birth represent the same person (even if their dates of birth are similar and they are in the same postcode). When name is available the blocking is only done on postcode, so that slight differences in how the name is recorded can be accounted for.

4.4.2 Scoring

Census records where the name is available are compared with each administrative data record in the same postcode. For each such pair of records, the similarity on their names is measured and scored for each of first, middle and last names. 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 (see Section 8.2 in the Annex for the details of this). For each name component there is a score for the evidence of the pair being a match and a score for the evidence against the pair being a match. The scores from the different components are totalled separately for the evidence for and evidence against, to give a total score for the evidence for a match and the evidence for a non-match.

Cases where there is just date of birth do not need to be scored as only identical dates of birth are considered.

4.4.3 Categorisation

Once the various scores have been calculated, each link is categorized into groups. These groups are arranged according to the strength of the evidence for and against a match. Again, the categorisation was developed to reflect the judgements of a

[17] 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.

human reviewer, and are shared with other census linking tasks (although adjusted to account for date of birth being missing, as other tasks make use of date of birth.).

Each category is given a distance score ranging from 0 (exact agreement) to 9 (most likely a non-match). Any link with a distance score of 5 or more are deleted at this stage. The remaining links are grouped into those that require clerical review and those that do not. Links with a distance score of 2–4 are placed in the group for clerical review. Links with a distance score of 0 or 1 will not need reviewed. The categories with a distance score of 4 or lower are:

- 0 Exact
- 1 Same (A)
- 2 Same (B)
- 2 Goes by middle name
- 4 Likely same (A)
- 4B Name same, missing DoB

For further detail on categorisation, see Section 8.5 in the Annex.

Note that there is a possibility that there are two similar census records (one failing RFP and the other passing it), and that the administrative data record links to both of them. If the census records related to the same person, then ideally the record failing RFP should not be retained, even though it linked to an administrative data record. However, this should not be a problem. Following RFP is the Resolve Multiple Returns (RMR) task, which identifies census returns at the same location that appear to relate to the same person (ie, a duplicate response), and resolves them into one record. Therefore, there is no need to check for such cases when deciding to keep a record that fails RFP.

4.4.4 Results from the Test on the 2011 Census Dataset

The original 2011 RFP process was re-run on the 2011 Census data, and then passed to the Admin Data Team. The test was run on a single processing unit (PU) consisting of circa 500,000 records. There were 712 records which did not pass the 2 of 6 check, but had either a valid date of birth (49 records) or name (663 records, with either a first or last name in any of the name sections). These were then linked to an administrative dataset to try to determine if there were any genuine people. The administrative dataset used was the NHS Central Register (NHSCR[18]), corresponding to Census Day 2011 (March 27, 2011).

Results from Linking on Date of Birth

The first part of this method is the Date of Birth (DOB) linking, on records that fail RFP but have DOB recorded. As noted in the methods section on blocking (Section 4.4.1), all records on the census dataset were compared against each record in the administrative dataset in the same postcode (referred to as a 'postcode block') with the same DOB.

Of the 49 cases in the test which failed RFP but had a valid entry for DOB, 11 records linked exactly on DOB and postcode to an administrative record. These 11 records would then have been retained in the census dataset, instead of being removed. If the links were multiplied by a factor of 10 (ie, the test was run on approximately 500,000 person records, or approximately 1/10th the population of Scotland), this could be extrapolated up to approximately 110 links *if* applied to the full 2011 Census - an extra 110 records that potentially would have been kept in the Census dataset, instead of being removed.

In order to assess the overall impact, consideration was required to determine if these records would persist through the rest of processing. In all 11 cases, it turned out there was another census record in the same postcode with the same date of

[18] Only variables that are required to do this matching purpose are requested through the data sharing agreement and only for the purpose of quality assurance. See Section 8.3.

birth. Furthermore, there is no evidence from the NHSCR that the census records are distinct. Therefore, these 11 records (110 scaled up) would indeed have been resolved into another record at the Resolve Multiple Responses process (which looks at duplicate responses), and so the part of the Admin RFP check that liked to DOB would have had no impact on the dataset.

Results from Linking on Name

Of those cases which had a valid name, 389 linked to at least one administrative data record. In 20 of these cases the link was not strong enough to be automatically accepted and so needed to be reviewed. Scaling this up to the full census would result in about 3,890 records being retained, and about 200 cases for clerical review. At around one case reviewed a minute, this could be completed within half a working day.

As with the linking on date of birth (detailed above), these cases were checked against the other census records, so find out how many would still be retained following RMR. 43 of the census records did not have any corresponding census records. This suggests that across the whole census, around 430 records of all that originally did not pass RFP (0.009 per cent of all census returns) would ultimately have been retained in the census dataset.

4.4.5 Results from the Test on Rehearsal Data

The rehearsal data set was 51,080 records, thus a much smaller (around 100 times) sample than what is expected in the live run of the 2022 Census.

Table 3 Rehearsal records by the number of the RFP variables that are missing.

Number of 2 of 7 variables missing	Number of records
0	26,410
1	10,738
2	1,145
3	5,062
4	2,472
5	653
6	1,826
7	2,774
Total	**51,080**

Table 3 shows the breakdown of the 51,080 record by how many of the 6 RFP variables are missing. The 1,826 records where all but 1 are missing are those of interest. These will fail the RFP test, but may have information that can be used for linking. Table 4 breaks these 1,826 records down by which variable is not missing. It can be seen that there are only 21 records with only date of birth, but there are 1,770 records with name (either on the household form, individual form or on the relationship matrix).

Table 4 Rehearsal records with all but one of the 7 RFP variables missing, broken down by which variable was not missing.

Variable not missing	Number of records
Relationship	17
Marital status	14
Sex	4
Name (household form)	438
Name (individual form)	782
Name (relationship matrix)	550
Date of birth	21
Total	**1,826**

Results from Linking on Date of Birth

Of the 21 records with only date of birth, only 1 had a complete date of birth. That record linked to the NHSCR (with the same date of birth in the same postcode).

Scaling this up to the approximate full census would suggest of order 100 such cases.

To assess the overall impact of this the census record is compared with other census records to find out if it would likely get resolved at RMR anyway. It was found that there was a matching census record with the same date of birth in the same postcode. Therefore, in this case the administrative data check on the RFP records would not likely have an overall impact on the census records.

Results from Linking on Name

Of the 1,770 records that have name but fail RFP, 5 link to NHSCR, none of which would be passed to clerical review. The cases passed were checked and all were found to be acceptable. This suggests that around 500 records would be kept. With a full census dataset there may be some links to review, but there are unlikely to be a great many, and so it is likely the clerical review could be completed in one person day.

Note that all of these cases were from paper returns. With online returns the name field automatically completes the three name variables, and so any return with name would automatically pass 2 of 7.

Checking back against the other census records, it is found that none of the linked NHSCR records links to another census record. This suggests that all of the RFP records that would be retained as a result of the administrative data check would be retained beyond RMR to be included in the census dataset.

Once the administrative data check, including manual review, has been completed, a flag is added to the dataset to indicate those records that the Admin Data team feel indicate a genuine record. When this is passed back to the Statistical Methods and Data Processing team, the records are set to pass Remove False Persons with the main census dataset, which then proceeds to the next step in processing (RMR).

5. Strengths and Limitations

As previously mentioned, on the whole the 2 of 6 Rule used in 2011 worked relatively well to prevent records not attributable to a plausible individual from passing to the end of census processing, thereby minimising the risk to overcount. Running this procedure is quick and can be done iteratively, as it does not have many dependencies. Since the nature of records which are not attributable to an individual tend to be blank or mostly blank, this also prevents additional burden for Edit and Imputation processes down the line.

Another strength is that using the 2 of 7 Rule in 2022 (or variations thereof) is accepted practice in other UK Census office's processing steps, which brings the base Remove False Persons process into alignment with rest of UK data cleansing procedures. It is of note, however, that the Office for National Statistics (ONS) and Northern Ireland Statistics and Research Agency (NISRA) will not be using the administrative data check in their methodology, due to the resource required for clerical review. However, as the gain from the administrative data check is fairly minimal, comparability in processing techniques can be still be made in terms of addressing what each agency does in handling records with minimal information, and so resources are collaborated upon in the development of the process.

There are, however, some minor issues with the procedure that could use refinement. As an automated check, the 2 of 7 Rule cannot prevent respondents from entering information other than a name in these fields. Since name and date of birth are key variables used in later processing — allowing for matching to occur with the Census Coverage Survey in the Estimation and Adjustment process, for example — a check for commonly used strings which conveys information other than a name would fine tune the process, as it assists in accurately removing records which do not pertain to a genuine individual, and would otherwise inflate the Census population count. In 2011, this overcount was mitigated in the Estimation and Adjustment procedure, whereas in 2022 using these checks will allow us to reduce the burden on this process.

There are also times where the 2 of 7 Rule may remove records too aggressively. Using administrative data to review ambiguous Census records provides a gain in accuracy with minimal effort or resource outlay (for the clerical review). It allows records which are from genuine individuals to be counted in the Census, which would otherwise be removed.

It is important to note, however, that while the trade-off between accuracy and resource while running the clerical reviews has been tested and thought to be minimal, two clerical procedures to the Remove False Persons process (one on the Data Processing side for false name strings, and one on the Admin Data team for the administrative data check), where in 2011 there weren't any. Further, we are unable to fully test clerical review procedures in a live running environment until 2022.

To minimise the impact, we will continue to test using 2011 data, synthetic data and rehearsal data, and we will be holding our own end-to-end processing test prior to the live census.

6. Conclusion

The move from a primarily paper to a majority-online Census collection has been designed so that records which come through the Online Collection Instrument should be from a genuine respondent. Although these will make up the vast majority of records, the Remove False Persons process is still required for both paper questionnaires and returns which are not fully submitted through the OCI. The 2 of 7 and False Name Checks are important statistical data cleansing steps that are necessary to prevent a large amount of overcount from burdening later processing, and ensures that resources are not wasted on records which do not pertain to a genuine individual.

While this method worked largely well in 2011, there are some cases where a small number of removed records may have been genuine people. In the 2011 Census, the equivalent to just over 1 per cent of records were removed; it is possible that *some* of these returns may have included people who provided limited information.

By removing these records there is a risk that both the error and the bias in the final population estimate could be increased. At this stage (RFP), the key problem to resolve is to determine which of the records failing the main check (2 of 7) are genuine returns and should be included. The use of administrative data, through the linkage of available datasets on the key variable groups name and date of birth, could provide an effective resolution to the issue. If a failing record is found in the independent administrative dataset, the indication this provides allows us to include the record in the census and thus improves the quality of the census data. As the NHSCR was successfully used in the testing, it is planned that the NHSCR will also be used as the administrative dataset in 2022. Although the gain in accuracy for this particular process is thought to be small, the linking methodology outlined in this paper is part of a greater linking exercise using Administrative Data that Scotland's census intends to perform in 2022[19], and thus is a beneficial by-product of which requires little additional resource for the improvement.

[19] Please see upcoming methodology papers on Name Re-Ordering, Resolve Multiple Responses, Date of Birth Check, and Census-CCS Matching. Links to be provided once published on this webpage: https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0.

7. 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), *PMP005: Name Reordering Methodology*, (online) available at: https://www.scotlandscensus.gov.uk/documents/Scotland%E2%80%99s%20Census%202022%20-%20PMP001%20-%20Estimation%20and%20Adjustment%20Methodology%20(pdf).pdf

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. (ed.) *Privacy in Statistical Databases: Lecture Notes in Computer Science*, vol. 8744, pp. 253–268

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/

8. Annex

8.1 Scenarios for an Adjusted 2 of 6 in 2011 / 2 of 7 in 2022

Single Person Households

Households which contained only one individual were exempt from the relationship criteria (as there is no other person to have a relationship to within the household) and so became **2 of 5**. In 2022, this will be 2 of 6.

Communal Establishments Persons

Communal Establishment person questionnaires did not have a household section, so the household member's table was not applicable, nor did it contain the relationship matrix. With these criteria removed, a communal establishment person became subject to a **2 of 4**. In 2022, this will be 2 of 5.

Unsubmitted Online Returns (Where another household return does exist)

The criteria for an unsubmitted online returns where another household questionnaire existed was changed to a **2 of 4** rule; these also had the household name variables, and relationship condition removed. In 2022, this category will not exist; they will be subject to the full **2 of 7**.

Online Returns (Submitted *and* unsubmitted)

All online returns (except for the type of unsubmitted returns noted above) **automatically passed** 2 of 6, as when household name was entered, the person names were *automatically* populated to online person questionnaires for convenience and consistency. In 2022, this will be the same, as fully completed and submitted online returns will automatically pass the **2 of 7,** since name and date of birth are required for submission. Unsubmitted returns will be created at the point a name is entered (and so will have all name fields populated), thus also automatically passing 2 of 7.

8.2 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[20]. 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.

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

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.

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 the start of the name from the two records. If so then the for score is set to be that given in Table 5. 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 5 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[21] 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

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

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.

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.

8.3 *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 our website:

Data Protection Impact Assessment for use of NHSCR dataset

Quality Assurance report for use of NHSCR dataset for 2019

8.4 *Glossary*

Term	**Definition**
Clerical Review	A process where an individual statistician manually recalls and reviews the record in question in order to make decisions on how to proceed with the record (ie, remove it, merge it, move to next process, etc). This generally happens with records which are ambiguous in some respect - for example, there is text written in the name field, but may not actually reflect a person and rather information instead.
Link	Two records that have been connected
Match	Two records that represent the same individual
Non-match	Two records that represent different individuals
Strong Link	Matches within the postcode with a full name match or date of birth match.
Processing Unit (PU)	In 2011, a processing unit was a subset of Census or Census Coverage Survey data that will be processed together through all stages of data processing. In 2022, some statistical processes will not use a 'processing unit' but will process data as it comes in (iteratively).

8.5 *Categorization of Links*

Once the for and against scores have been calculated for each component for each link, the links are placed into one of the categories shown in Table 6. Note that when these are used for linking census records with missing date of birth, not all categories will be used.

Table 6 List of categories used to class the links along with a brief description of the condition used to place them and the strength associated with the category. The categories are presented in order of the priority in which they are assigned. That is, links are only assigned to a given category if they do not meet the conditions for any preceding categories.

Strength	Name	Description of Condition
0	Exact	All components agree exactly and non-missing
7	Different – parent-child	Age difference ≥15, first and last for >0
6	Different – twin	Last for >15, DoB for >0 no evidence of match from first name
1	Same	Fairly strong evidence for match from first, last and DoB, no evidence against from gender or middle name
2	Same 2	As Same, but slightly weaker evidence
2	Goes by middle name	DoB, last and gender agree exactly and non-missing, first from one record agrees exactly with middle from other
4	Likely same (A)	Total for >70, total against =0, total for – last for >20
4	Female last diff	Female, fairly strong evidence for match from first and DoB, and last against >0
5	Non-female last diff	As Female last diff but without condition on being female
5	DoB same, missing name	DoB for >10, age difference <14, name missing on one record
4	Name same, missing DoB	First for ≥20 and last for ≥20 and total for >50, DoB missing on one record
5	Likely same (B)	Total for >45, total against =0, total for > last for + 15
6	Likely same (C)	Total for >20, total against =0, total for > last for + 10
7	Don't know	First, middle, last, and DoB all missing on one or both records, gender the same or missing on one or both records
7	Don't know diff gender	As don't know but without condition on gender
7	Don't know first partial agree	Middle, last and DoB all missing on one or both records, first names exactly the same to the length of the shorter string (e.g. Tom and Tomas)
7	Don't know last partial agree	As Don't know first partial agree but with condition on last

7	Likely different	Total for >50, total against <20
7	Probably different	Weak evidence against from first, last or DoB, total for > total against
8	Different – sub	Weak evidence against from up to two of first, last and DoB
9	Different other	Evidence against from first, last and DoB
7	Remaining	Any records not assigned to any of the above categories