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

Scotland's Census 2022

Census–Census Linking and Overcount Correction

October 2020

National Records of Scotland

Contents

1. Plain English Abstract

All households in Scotland are required to complete a census return for all usually resident persons. However, sometimes people get recorded at multiple locations or the wrong location. In order to avoid overestimating the population as a result of this, the census dataset is linked to itself. The linked records are then linked to administrative data. Using these links, it is possible to calculate how likely it is that a record represents a distinct genuine individual. From this, the total amount of duplication in the census can be estimated and corrected for.

2. Abstract

All households in Scotland are required to complete a census return for all usually resident persons. Respondents should be recorded at their place of usual residence, and so should appear on the census exactly once. However, some people may have more than one residence and so they may appear on the census at more than one location. A common example is that of children who live part of the time with each of their separated parents. If not accounted for, these effects will result in an overestimate in the population.

To account for such cases the census dataset is linked to itself. Links found between records in different locations are then considered (links between records at the same location are dealt with in the Resolve Multiple Returns (RMR) process). Unlike in the RMR process, the records in these links cannot be resolved. One reason is that it would be difficult to know at which location the individual should be recorded.

Therefore, in order to account for this overcoverage, probabilities will be calculated for each census record, indicating the likelihood that the record represents a genuine distinct individual. To find this, the records in the census links are linked to an administrative dataset. Using the number of links where both census records, one of the census records, or neither of the census records link to the administrative

dataset, probabilities that each link represents one or two individuals can be calculated. This in turn can be used to calculate the probability that each record represents a genuine distinct individual, which is then attached to the record. By considering the difference between the total number of census records, and the sum of the probabilities attached to them, the scale of overcoverage can be estimated. Testing on 2011 data suggests that 99.37 per cent of the records are considered to be genuine, non-overcounted records. This can then be accounted for in the census process of estimation, when population estimates are produced to account for undercoverage in the census dataset. This is an important consideration as bias should be kept as low as possible[1]. The effect is smaller than that dealt with by estimation or RMR, but around the same as the Remove False Persons task.

It was announced on 17 July 2020 that the date of Scotland's next census would change from 21 March 2021 to 20 March 2022, due to the impact of COVID-19 on vital preparations for the census.

[1] See www.scotlandscensus.gov.uk/documents/Statistical%20Quality%20Assurance%20Strategy.pdf.

3. Introduction and Background

Estimation is the process where the true population of Scotland is estimated from the number of records on the Census (see Figure 1). In producing population estimates from the Census, undercount is the primary issue where households do not complete a questionnaire. However, there are cases of overcount in Census returns, where there are extra records which should not have been included.

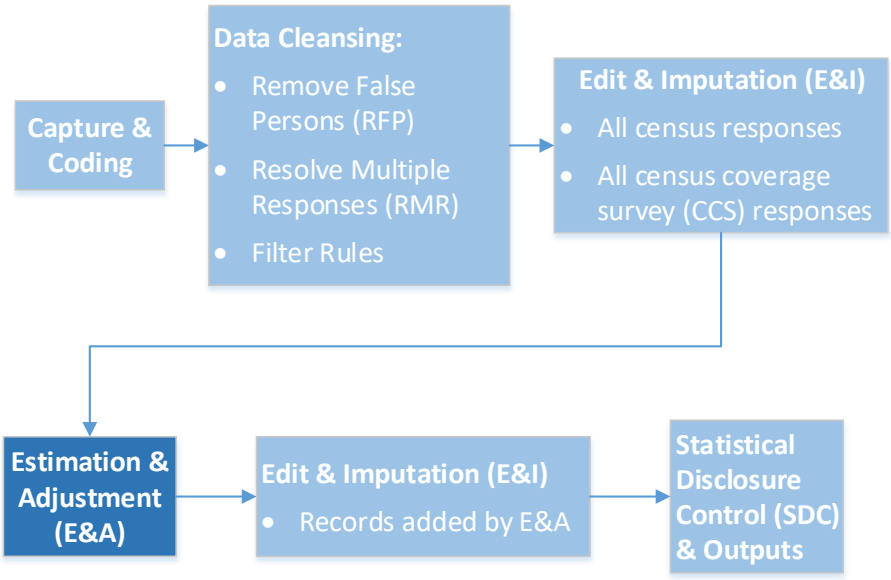

Figure 1 Where estimation fits into Statistical Data Processing.

There are four types of overcount:

- Type 1 — Duplication of individuals within the same location
  - These are duplicates where a person has either been included multiple times in the same household return, or in two or more separate returns for the same household.

- Type 2 — Individuals enumerated in more than one location
  - These are duplicates where a person has been included in more than one household return at different addresses, such as a child with separated parents included in the household of each parent.

- Type 3 — Individuals enumerated in the wrong location
  - These are cases where a person has been missed in the household where they should have been enumerated, but included in a household where they should not have been enumerated. This results in undercount in the area where they were missed, and overcount in the area where they were included.

- Type 4 — Erroneous returns
    - These can be returns which are fictitious or joke returns, as well as cases of babies that were born after Census day or individuals who died before Census day and as such should not have been included.
    - These are difficult to identify without additional field work or linking to vital events data.

If overcount is not identified and accounted for then this can lead to an overestimate in the population. Type 1 overcount will be identified and resolved in the Resolve Multiple Responses (RMR) process[2] as part of Data Cleansing. The Remove False Persons process[3] will deal with Type 4 overcount, to the extent that this is possible. This paper therefore covers methodology for dealing with types 2 and 3 overcount. Most of this paper discusses Type 2. Type 3 will be covered in Section 9.

To identify Type 2 overcount, this paper presents the method to link the census to itself, with the linked records then being linked to an administrative dataset. For each census–census link the number where one or both of the census records link to the administrative dataset is identified. Using this, the probability of each census record representing a distinct genuine individual is calculated. In addition, a contingency is presented for the case that administrative data is not available in 2022. This would use probabilities calculated for each category of link from the 2011 census and a previous administrative data.

To identify Type 3 overcount, this paper presents, in Section 9, how the Census–CCS links can be used to identify people who were enumerated in different locations between the Census and CCS.

This paper will then present how the numbers of records identified as overcount can be used to correct the population estimates.

[2] See the methodology papers on RMR at https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0 for more information.
[3] See the methodology papers on Remove False Persons at https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0 for more information.

4. 2011 Method

In 2011 only links where the name and date of birth agreed exactly were considered, and there was no checking against administrative data. (For the purposes of this document a link is a pair of records that have been connected.) This was applied to a sample of census records, ensuring the sample contained enough duplicates to give an acceptably low coefficient of variation. Running the process on the whole dataset would have been too computationally intensive.

The Census–CCS links were used to identify the level of Type 3 overcount (misplacement) in the Census, assuming that the location someone was captured in the CCS was the location where they should have been enumerated. As this can only be calculated within the CCS areas, it was scaled to the number of duplicates within the CCS areas.

Propensities for overcount (hereafter referred to as propensities) were calculated for Type 2 overcount, which could then be accounted for when estimating the true population. (The propensity is a measure of how much overcount is included in the census dataset. It is defined as the number of records in the dataset divided by the number of persons in the population these represent. For example a propensity of 1 indicates that there is no overcount (and so no modification is needed), while a propensity of 2 indicates that for each two census records there is one person in the population (and so the estimate needs to be divided by 2).) The propensities were calculated as in Equation 1.

$$\gamma = \frac{X}{X - E} \qquad 1$$

Where γ is the propensity, X is the overall count obtained in the census, and E is the erroneous count identified within the census (that is, the number of records that represent persons who have already been counted in the census). $X - E$ equals the expected 'true count' of the population.

The overcount propensities were stratified in to:

- 3–17 year olds
- 18–25 year old students
- 18–25 year old non-students
- 85+ year olds
- Everyone else

These groups roughly cover groups affected by several reasons for overcount, and so capture some of the variation in response patterns. 3–17 year olds may be counted at both parents' addresses if their parents live apart. Students may be recorded at both a term-time address and their parent's address. 18–25 year olds non-students may be more mobile than other groups and so be more likely to appear in multiple locations. People aged 85+ may appear at a home address and a communal establishment.

The inverse of the propensity, which is always less than 1, was applied to all records within that stratum to be used as the in-Census count for that record within Dual System Estimation (DSE). (DSE is a statistical process using the links between two independent datasets of the same population to estimate the total population. See the Estimation and Adjustment Methodology paper[4] for more information on DSE.)

[4] Available at https://www.scotlandscensus.gov.uk/documents/Scotland%E2%80%99s%20Census%202022%20-%20PMP001%20-%20Estimation%20and%20Adjustment%20Methodology%20(pdf).pdf.

5. Proposed 2022 Linking Method

In 2022 it is proposed to link the whole census against itself, rather than using a sample. However, the census is too large to search for duplicates without blocking[5]. With ~5,000,000 records an exhaustive search would require 2.5×10^{13} comparisons, which would take prohibitively long, even for the simplest comparison.

To avoid this, records are only compared if they agree exactly on name or date of birth. An initial comparison is made on date of birth (for pairs agreeing on name) or name (for pairs agreeing on date of birth). This can be used to filter out pairs that are clearly non-matches. The remaining pairs are then compared using a more-thorough method, which identifies links that could likely be matches.

Note that for the purposes of this document, a match is a link where the two linked records relate to the same individual. A non-match is where the two records relate to different individuals. Also, at various steps scores are calculated and thresholds applied. These have been developed in order to best replicate the judgements of a human reviewer.

The records in the links identified by the step above will be linked to an administrative dataset. The number of these links that have 0, 1 or 2 of the census records linking to the administrative dataset are then counted. Using these counts the probability that each census–census link represents a match ($P(M)$) can be calculated. From that the probability that each census record represents a distinct genuine individual ($P(g)$) can be calculated. Comparing the difference between the total number of census records, and the sum of these probabilities, gives an estimate of the total overcount in the census dataset. The probabilities can then be used in the DSE calculation, either directly to decrease the weight of particular census

[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.

records, or aggregated together to apply a downward overcount weight for all records within particular age groupings, reducing the estimates that are produced.

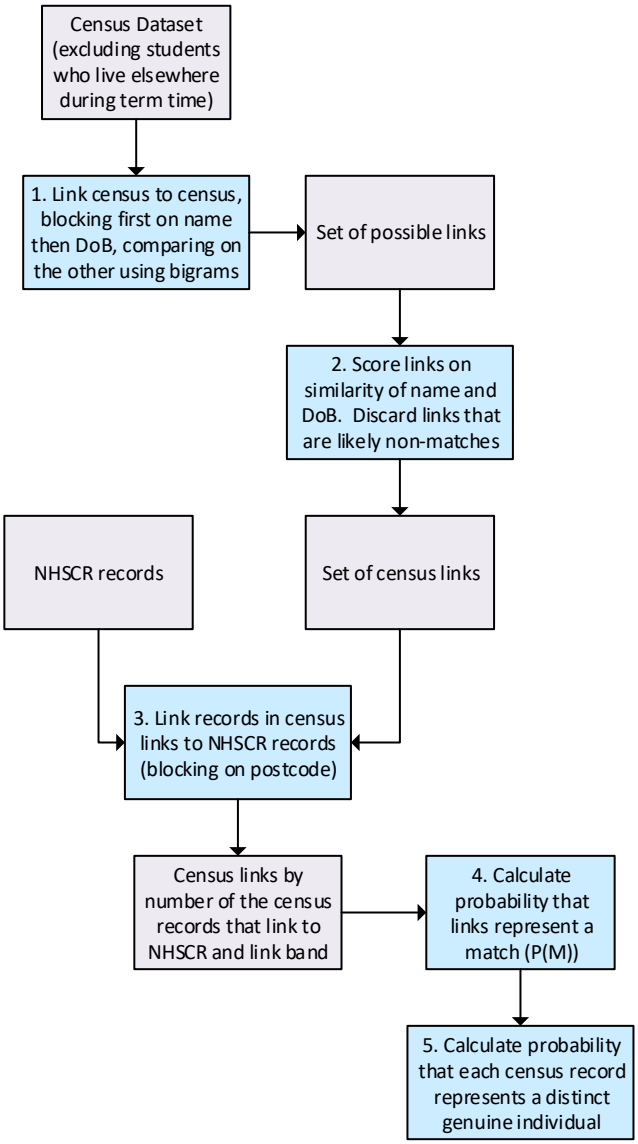

Figure 2 Flow of data (purple boxes) and steps (blue boxes) through the process.

The flow through these steps is summarized below, and in the flow chart in Figure 2. The numbering below, and in Figure 2, relates to the subsection number in Section 5 where each step is described in more detail.

1. Link census to census, blocking on name and date of birth
2. Score the census–census links and filter out weak links
3. Link the linked census records to an administrative dataset

4. For each link, calculate the probability that it represents a match ($P(M)$)
5. Assign to census records the probability that they represent a distinct genuine individual ($P(g)$)

5.1 Initial Census–Census Linking

The overcount needs to be estimated across all the census records that would be used for the population. This therefore excludes any skeleton records generated in adjustment to account for undercount. It also excludes any records where it has been indicated that the person is a student and lives at another address during term time. Students should be counted at their term-time address. Therefore any overcount from students being counted at a term-time address and a home address will be dealt with in this way, rather than by linking the records as is done for other duplicates.

There will be around 5,000,000 such records. Therefore blocking is used for the initial linking (only considering pairs of records where the two records have the same value for some blocking variable). Blocking should not be done too strictly on location, as the point of this exercise is to find the same person recorded in different locations. The other main linking components are name and date of birth. Linking blocked on name, and separately blocked on date of birth, could then be performed. This would greatly reduce the number of comparisons. The first step is done separately for each of the two blocking variables as described in Annex 5.

Even blocked on name or date of birth, doing a thorough comparison between the records may be prohibitively slow. Therefore to improve efficiency a much briefer comparison is made first, and the links that score well on that are then passed for the full comparison. For each link the two records are compared using a comparison variable. When blocking on date of birth, name is used for the comparison variable, and vice versa. The comparison is done using bigrams, as described in Annex 6. Using the results of the bigram linking, some links will be discarded, as likely representing non-matches.

The outputs from the two passes of the first step are then merged (and de-duplicated). At this stage any links between records in the same postcode are removed, as these should be dealt with by the RMR process, which would already have been carried out.

5.2 Detailed Comparison of Links

All of the links identified above are then scored. This method compares the first, middle and last names of the two records using the thorough comparisons described in Annex 1, and for each of these calculates a score indicating the strength of evidence for a match (the for score), and the strength of evidence against a match (the against score). It also calculates for and against scores for sex, as described in Annex 2, and for date of birth, also as described in Annex 2.

So that the links can be grouped later, each link is given a band reflecting each of:

- Strength of agreement on name
- Strength of date of birth agreement
- How rare the identifiable information is in the population

Table 1 Conditions for categorizing links into name bands.

Condition	Name Band
First for score = 50 AND last for score = 50 (exact agreement)	0
First for score ≥ 25 AND last for score > 25	1
First for score ≥ 20 AND last for score > 20	2
First for score ≥ 15 AND last for score > 15	3
First for score ≥ 10 AND last for score > 10	4
All other cases	5

This categorization into bands uses a simple mapping from the first name, last name (as described in Table 1) and date of birth scores (as described in Table 2).

Table 2 Conditions for categorizing links into date of birth bands.

Condition	Date of Birth Band
Date of birth for score = 36 (exact agreement)	0
Date of birth agrees if one has been recorded in American format	1
Date of birth for score ≥ 12	2
Date of birth for score ≥ 6	3
All other cases	4

Finally, the links are banded according to roughly how many people would be expected across Scotland to have the combination of name and date of birth. This can be used to distinguish common from unusual names. For example there would be less confidence in two records with 'John Smith' being a match than two records with 'Sarah-Jane Watt-Maxwell'. Therefore the number of times each name component appears in the dataset is used to calculated the expected number of people with the given characteristics. The following steps are used to calculate the expected number:

1. Expected = 5,400,000 (approximate population of Scotland)
2. If year of birth agrees and is not missing then expected = expected/80
3. If month of birth agrees and is not missing then expected = expected/12
4. If day of birth agrees and is not missing then expected = expected/30
5. Expected = expected times the maximum of 100 and the counts of the first name from each of the two records divided by 5,400,000
6. Expected = expected times the maximum of 100 and the counts of the last name from each of the two records divided by 5,400,000
7. Expected band = min(max(10+round($\log_2$ expected)),10)/2, where the round function rounds to the nearest integer

This results in a band ranging from 0 to 5, with 0 indicating the rarest values.

Any link is then discarded if:

- it looks like it may be a parent-child pair (where the difference in year of birth is 15 or more),
- first name, last name, or date of birth look different (i.e. the against scores for these variables are greater than zero), or
- it looks like a pair of twins (where the first names differ and are both common names (that is, it is expected that there are more than 100 people in Scotland with the name) that are not nicknames of each other, and there was evidence from the middle name or sex that the records represent different persons).

5.3 Linking to Administrative Dataset

The census records that are part of a link that is categorized as reasonably strong (that is, not one of the discarded links) are then linked to an administrative dataset. This is done blocked on postcode, and uses the thorough linking process.

The administrative dataset used to test against the 2011 Census was the National Health Service Central Register (NHSCR) 2011. This is a large administrative dataset of people who are registered with an NHS GP in Scotland or who were born in Scotland. It therefore has good coverage of Scotland's population. Persons who have moved away or have died are flagged and not included in the dataset to be linked. For 2022 it is planned that the NHSCR will also be used. A version will be extracted close to the census date.

The census links are not resolved down to a single record. Instead, an estimate is obtained for the number of records in the census dataset that do not represent distinct genuine individuals (label this n_e). Dual System Estimation (DSE) between the census and CCS estimates how many extra records need to be added to the census dataset in order to accurately represent the number of people in the population (call this n_a). To account for overcoverage and undercoverage, the number of records to be added to the census dataset should then be $n_a - n_e$. It is highly likely that the number of records to be added will exceed the amount of overcoverage in the census dataset, that is $n_a > n_e$. Therefore, exactly which records do not represent genuine individuals, is not required, only the total number. Similarly, it is not required to identify which links represent matches.

In order to estimate n_e, each census record can be assigned a probability that it represents a distinct genuine person. If n_c is the number of records in the census dataset (prior to skeleton records[6] being added in adjustment) and $P_i(g)$ represents the probability that record i represents a distinct genuine individual, then:

[6] Skeleton records are records added after estimation to bring the number of records in the census dataset up to the estimated population.

$$n_e = n_c - \sum_i P_i(g) \qquad 2$$

This process could be done separately for each DSE strata in order to more accurately determine the properties of the records added.

It is plausible that the strongest links are more likely to represent matches than weaker links. Rather than only considering the strongest links, it would be preferable to assign different probabilities to records that link strongly from those that link more weakly. This is done by grouping the links according to their bands for name, date of birth and expected number (discussed at Section 5.2 above). The probabilities for each group are then calculated separately.

Note that this also means that focus can be on more than just on the strongest links. If there was a large group of links, of which only a small proportion were believed to represented matches, then this information could still be used to reduce the estimate accordingly.

For a linked pair of census records, trying to link each of them to the administrative dataset leads to one of three possibilities:

0. Neither census record links to the administrative dataset
1. Exactly one census record links to the administrative dataset
2. Both census records link to the administrative dataset

Additionally, the link between the two census records either represents a match (that is, the two linked records represent the same individual) or a non-match (that is, the two linked records represent different individuals). It is also assumed that if a census record links to the administrative dataset in that location then it represents a genuine person in that location. So the two assumptions are:

1. When two census records link they represent either one genuine individual (a match) or two genuine individuals (a non-match)

2. When a census record links to an administrative dataset record then it represents a genuine individual

These assumptions imply, for example, that if both census records link to the administrative dataset then each census record represents a distinct genuine individual, and so the (census–census) link is a non-match.

Let g represent the case where the census record represents a distinct genuine person (and hence $\bar{g}$ represents when the record does not represent a genuine person). Let l represent the case where the census record links to the administrative dataset (and hence $\bar{l}$ represents when the record does not link). For example, $P(l|g)$ would represent the probability that a census record links to the administrative dataset, given that it represents a distinct genuine person.

Now, for a particular link between two census records, let M represent the case that this represents a match (that is, that the two linked records represent the same person). Thus, $\bar{M}$ represents the case where the link is a non-match, the records represent different persons (distinct genuine individuals). This is described in Figure 3.

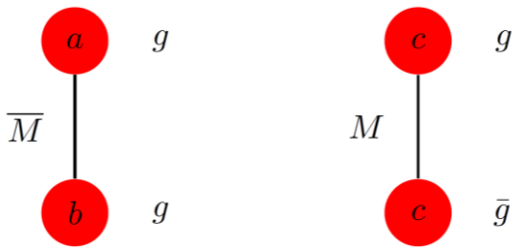

Figure 3 Diagrammatic representation of a match ($\boldsymbol{M}$) and a non-match ($\bar{\boldsymbol{M}}$). The red vertices represent census records, and the edges represent links. In a non-match the records represent different persons (*a* and *b*) and both represent distinct genuine persons ($\boldsymbol{g}$). In a match, only one of the census records represents a distinct genuine person ($\boldsymbol{g}$), as the other is not a distinct person ($\bar{\boldsymbol{g}}$), that is both records represent the same person, *c*.

Now, let:

0 represent the case where neither of the linked census records link to the administrative dataset

1 represent the case where exactly one of the linked census records link to the administrative dataset, and

2 represent the case where both of the linked census records link to the administrative dataset.

These three cases are described diagrammatically in Figure 4.

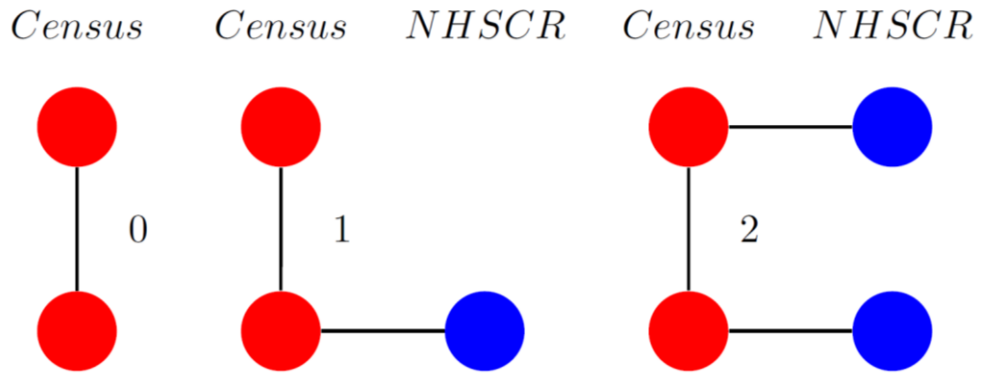

Figure 4 Diagrammatic representation of the **0**, **1** and **2** cases. Red vertices represent census records, while blue vertices represent administrative data (NHSCR) records. The edges represent links.

Note that a record representing a genuine person might not link because the linking process failed, or because there is not a matching record on the administrative dataset. Both of these possibilities are wrapped up in the probability $P(l|g)$. If most persons present in the population appear on the administrative dataset then the latter reason would not have much impact on $P(l|g)$. However, even if the administrative dataset was of lower completeness then the method would still hold, it would just be that $P(l|g)$ would have a lower value.

A more-serious problem would be if the coverage of the administrative dataset varied substantially for different portions of the population. For example, if the dataset included almost all adults but few children then it might be found that almost all non-matching pairs of census records would have either zero or two administrative data records linked to them (rather than one or two). This would affect how the

probabilities could be used to make inferences about how many of the pairs are matches. It is therefore important that the administrative dataset that is used has even coverage of all population subgroups. The NHSCR is therefore a good choice of administrative dataset, as it has high coverage for all parts of the population.

5.4 Calculate Probability that each Link is a Match

In what follows, use will be made of the following mathematical identities. Recall that as these are identities they hold for any random variable(s). A description of the mathematical notation used in this document is available in Annex 3.

$P(X\|Y) \equiv \frac{P(X \cap Y)}{P(Y)}$	3	definition of conditional probability[7]
$P(X \cap Y) \equiv P(X\|Y)P(Y)$	4	rearrange Equation **Error! Reference source not found.**
$P(X) \equiv P(X\|Y)P(Y) + P(X\|\bar{Y})P(\bar{Y})$	5	Total Probability Theorem[8]
$1 \equiv 1 \times P(Y) + 1 \times P(\bar{Y})$	6	suppose X in Equation **Error! Reference source not found.** were certain
$1 \equiv P(Y) + P(\bar{Y})$	7	simplify Equation **Error! Reference source not found.**
$P(Y \cap X) \equiv P(Y\|X)P(X)$	8	re-label Equation **Error! Reference source not found.**
$P(X \cap Y) \equiv P(Y \cap X)$	9	commutativity of intersection[9]

[7] http://mathworld.wolfram.com/ConditionalProbability.html
[8] http://mathworld.wolfram.com/TotalProbabilityTheorem.html
[9] https://mathworld.wolfram.com/Set.html

$P(X\|Y)P(Y) \equiv P(Y\|X)P(X)$	10	substitute in equations **Error! Reference source not found.** and **Error! Reference source not found.**
$P(X\|Y) \equiv \frac{P(Y\|X)P(X)}{P(Y)}$	11	rearrange to obtain Bayes' Theorem[10]

Consider the probability of a single census record linking to the administrative dataset. Let p be the probability that a census record representing a genuine person links to an administrative record (in the same postcode at least as strongly as some specified threshold).

Therefore:

$P(l\|g) \equiv p$	12	by definition of p
$P(\bar{l}\|g) + P(l\|g) = 1$	13	by Equation **Error! Reference source not found.**
$P(\bar{l}\|g) + p = 1$	14	substitute in Equation **Error! Reference source not found.**
$P(\bar{l}\|g) = 1 - p$	15	rearrange
$P(l\|\bar{g}) = 0$	16	by assumption 2
$P(\bar{l}\|\bar{g}) + P(l\|\bar{g}) = 1$	17	by Equation **Error! Reference source not found.**
$P(\bar{l}\|\bar{g}) = 1$	18	substitute in Equation **Error! Reference source not found.**

[10] http://mathworld.wolfram.com/BayesTheorem.html

Table 3 gives the probabilities of 0, 1 and 2 (of the census records linking to the administrative dataset), conditional on whether the link is a match or non-match. n_0, n_1 and n_2 are the number of links in the strata where 0, 1 or 2 of the census records (respectively) link to the administrative dataset, and n is the total number of links in the strata. Please see Annex 7 for the derivation of these probabilities.

Table 3 Probabilities of records in a census–census link linking to 0, 1 or 2 administrative dataset records, conditional on whether the census–census link is a match or non-match.

Status of census link	Number of census records that link to administrative dataset			
	0	1	2	Total
Match	$P(0\|M) = 1-p$	$P(1\|M) = p$	$P(2\|M) = 0$	1
Non-match	$P(0\|M) = (1-p)^2$	$P(1\|M) = 2p(1-p)$	$P(2\|M) = p^2$	1
Total	$\widehat{P(0)} = \frac{n_0}{n}$	$\widehat{P(1)} = \frac{n_1}{n}$	$\widehat{P(2)} = \frac{n_2}{n}$	$\frac{n_0+n_1+n_2}{n} = \frac{n}{n} = 1$

However, $P(M)$, the probability that the link is a match, is what is required. The number of links that have 0, 1 or 2 of the census records linked to the administrative dataset, n_0, n_1 and n_2, will be available. From the bottom row of Table 3 this gives us estimates for $P(0)$, $P(1)$ and $P(2)$. This means that by reversing the conditionals in the formulae in Table 3 using Bayes' Theorem, $P(M)$ can be estimated.

The equations above typically involve p, the probability that a genuine record links to the administrative dataset. However, this will be unknown[11]. However, using multiple equations, two equations can be found that involve p and $P(M)$. Substituting one into the other will eliminate p, giving an estimate of $P(M)$.

The first such relationship involves $P(M)$, p and $P(0)$. This is given below. Please see Annex 8 for the derivation of this.

[11] Note that p could be estimated by linking unlinked census records (which we could assume represent genuine persons) to the administrative dataset. However, this runs into problems as p may differ between unlinked records, and records that do link (and also records in links of differing bands).

$$P(0) = P(M)(p - p^2) + (1 - 2p + p^2) \qquad 19$$

With a relationship between $P(M)$, p and $P(0)$ (Equation 19) another relationship involving $P(M)$ and p is needed in order to eliminate p. Therefore, a relationship involving $P(M)$, p and $P(2)$ is sought. This is given below. Please see Annex 9 for the derivation of this.

$$p = \sqrt{\frac{P(2)}{1 - P(M)}} \qquad 20 \quad \text{as Equation 56}$$

Equations 19 and 20 can then be combined to estimate $P(M)$. The result is given by

$$\widehat{P(M)} = \sqrt{\frac{\hat{A}^2}{4} - \hat{A} + 2} - \frac{\hat{A}}{2} \qquad 21 \quad \text{as Equation 75}$$

where

$$\hat{A} = \frac{\left(\frac{n_0}{n} - \frac{n_2}{n} - 1\right)^2}{\frac{n_2}{n}} \qquad 22 \quad \text{as Equation 74}$$

For the derivation of this please see Annex 8.

Equation 21 represents the estimate of the proportion of all the census–census links in the stratum that are a match. The estimated total number of matches in the stratum is therefore $\widehat{n(M)} = n\,\widehat{P(M)}$ (that is, the estimated number of matches among n census–census links in a stratum is the observed number of census–census links in that stratum (n) multiplied by the estimated probability that each of those census–census links represents a match ($P(M)$)). $\widehat{n(M)}$ is also the estimate for the number of census records in the stratum that do not represent distinct genuine individuals, that is, the overcount.

5.5 Using $\widehat{P(M)}$ To Assign Probabilities to Census Records

There is now an estimate for the total overcount. However, probabilities need to be attached to individual census records because dealing with the overcount will be

done separately within each DSE strata, which will be different from the strata used here. Furthermore, each census–census link in the strata should not have the same probability of being a match. For one thing the census–census links where both census records link to the administrative dataset are non-matches (by assumption 2). Also matches are more likely than non-matches to have neither census record linking to the administrative dataset as one of them is not a (distinct) genuine person, and so (by assumption 2) cannot link to the administrative dataset. Correspondingly, census–census links with no links to the administrative dataset are more likely to be matches than census–census links with a link to the administrative dataset. Therefore the probability that a census–census link is a match can be calculated, given the number of the census records in it that link to the administrative dataset.

The relevant expressions are given in Table 4. For the derivation of these expressions please see Annex 11.

Table 4 Probability assigned to each census record depending on $\widehat{P(M)}$ and whether one of the census records links to the administrative dataset (NHSCR 2022).

$\widehat{P(M)}$	Neither census record links to the administrative dataset	One census record links to the administrative dataset		Both census records link to the administrative dataset
		Unlinked record	Linked record	
1	0.5	0	1	N/A
$0 < \widehat{P(M)} < 1$	$1 - \frac{\widehat{P(M)}\left(1 - \sqrt{\frac{n_2/n}{1-\widehat{P(M)}}}\right)}{\frac{2n_0}{n}}$	$1 - \frac{\widehat{P(M)}\sqrt{\frac{n_2/n}{1-\widehat{P(M)}}}}{\frac{n_1}{n}}$	1	1
0	1	1	1	1

Note that in cases where exactly one of the census records links to administrative data, in the limit of $\widehat{P(M)} = 1$ (when it is certain that the link represents a match), the linked census record would get a probability of 1, while the unlinked record would get

a probability of 0. This reduces to being effectively the same as resolving the two records together. (The only difference would be that the record would remain in the census dataset and there would be correspondingly fewer comparable made up records. Removing the record by resolving it would mean that it would not be included in the dataset and there would be a corresponding extra made up record. By giving the results as probabilities the data processing team have the flexibility to either resolve or to aggregate the probabilities to adjust estimation.)

In the limit of $\widehat{P(M)} = 0$ (when it is certain that the link represents a non-match) all the census records would represent distinct genuine persons, and so each should have a probability of 1. It can be seen that plugging $\widehat{P(M)} = 0$ into the formulae in Table 4, this does indeed return a probability of 1. (Although in practice $\widehat{P(M)}$ would not be estimated to be 0 if there were census–census links where one or both of the census records did not link to the administrative dataset. If all the linked census records linked to the administrative dataset then $n_2 = n$ and $n_0 = 0$. Plugging these into Equation **Error! Reference source not found.** gives $\hat{A} = 4$ and plugging that into Equation 21 gives $\widehat{P(M)} = 0$.)

Note that in order to have $\widehat{P(M)} = 1$ it must be the case that $n_2 = 0$. But if $n_2 = 0$ then $\hat{A}$ is ill defined so $\widehat{P(M)}$ cannot be calculated. If this happens in any stratum the options would be to either:

- Ignore the stratum (effectively giving all of those census records a probability of 1)
- Set $\widehat{P(M)} = 1$ (effectively giving all of those census records a probability of 0, 0.5 or 1 depending on whether the record links to the administrative dataset)

6. Results Using 2011 Data

6.1 Census–Census Linking

This process was run on the full 2011 census dataset. On the first pass the date of birth block returns 263,350 links, while the name block (first and last name) returns 826,550 links. After removing cases where the postcode is the same in each record (which would have been considered by the RMR process) this reduces to 256,965 and 823,667 respectively. Some of these are the same link, so the total number of links found at this stage is 1,059,848. These links are then all analysed in more detail. Links that that appear problematic are then discarded (see Table 5).

Table 5 Number of links discarded by reason.

Category	Number of links
Parent–child pair	190,790
Twins	77,451
Last name different	137,622
First name different	14
DoB different	210,236
Remaining	443,735
Total	1,059,848

The remaining 443,735 records are grouped by the name and date of birth bands, where a band 0 indicates exact agreement, and bands with larger numbers represent increasingly looser agreement (see Table 6).

Table 6 Number of links by band for name and date of birth. Shaded cells are those that the links in Table 7 are drawn from.

Date of birth	Name						
	0	1	2	3	4	5	Total
0	19,607	738	1283	2870	3653	7220	35,371
1	2012	85	323	911	1336	2499	7166
2	50,850	51	0	338	0	20	51,259
3	151,128	175	0	1037	0	78	152,418
4	195,864	214	0	1328	0	115	197,521
Total	419,461	1263	1606	6484	4989	9932	443,735

The combination of these two variables are used to decide which links to consider, and also to stratify the results (which is why they have been binned[12] in the first place). Further evidence and stratification is provided by the rarity of the identifiable

[12] That is, grouped. See https://mathworld.wolfram.com/Bin.html.

information, referred to as 'expected'. For example, there would be less confidence that two John Smith records were a match than two Sarah-Jane Watt-Maxwell records. Therefore the number of times each name component appears in the dataset is used to calculated the expected number of people with the given characteristics. This also makes use of the parts of the date of birth that agree.

Table 7 Number of links by expected, date of birth and name band. See text for explanation of the shading.

Expected	Date of birth	Name band					Total
		0	1	2	3	4	
0	0	5374	280	203	187	215	6259
	1	1	0	0	0	1	2
	2	49	1	0	0	0	50
	3	101	5	0	0	0	106
1	0	4057	139	173	235	391	4995
	1	1	0	0	1	2	4
	2	100	0	0	0	0	100
	3	217	5	0	4	0	226
2	0	4895	178	397	755	1180	7405
	1	1	0	1	1	10	13
	2	492	2	0	2	0	496
	3	1203	5	0	7	0	1215
3	0	3808	115	417	1112	1488	6940
	1	16	2	2	8	19	47
	2	3050	8	0	17	0	3075
	3	8701	17	0	43	0	8761
4	0	1473	26	93	581	379	2552
	1	110	4	27	72	132	345
	2	10,915	10	0	127	0	11,052
	3	32,110	34	0	383	0	32,527
5	0	0	0	0	0	0	0
	1	1883	79	293	829	1172	4256
	2	36,244	30	0	192	0	36,466
	3	108,796	109	0	600	0	109,505
Total	0	19,607	738	1283	2870	3653	28,151
	1	2012	85	323	911	1336	4667
	2	50,850	51	0	338	0	51,239
	3	151,128	175	0	1037	0	152,340
	Total	223,597	1049	1606	5156	4989	236,397

Once these bands are calculated for each link, any link that is particularly weak on any of these is discarded. Those with a name band of 5, or a date of birth band of 4 are discarded (leaving the shaded cases from Table 6). This then gives links by strata as given in Table 7.

At this point any link that is quite weak on more than one of the aspects is removed. In particular any link where the sum of the bands is greater than 4 is removed. This removes all the cases in the grey cells in Table 7, leaving 22,713 links.

6.2 Census–NHSCR Linking

Now the census records that are part of one of the 22,713 census–census links are linked to an administrative dataset. The dataset that was used was the NHSCR, as at census day 2011.

The census records are considered to have linked to the NHSCR if a link can be found that is at least as strong as the census–census link it is a part of. For example, suppose there were two census records that said OLIVIA WILSON, and OLIUIA WILSON. If the OLIUIA WILSON record was in the same location as the NHSCR record then insisting that the census–NHSCR link was an exact agreement would mean that it could be missed. However, if the census–census link had exact agreement, then that would suggest that there had not been an error. In such cases any differences to the NHSCR record might then be evidence that the NHSCR record represented someone else. The breakdown of the census–census links by the number of the census records that link to the NHSCR is given in Table 8.

Table 8 Number of census links by expected, date of birth and name band and the number of census records that link to the NHSCR with at least the strength with which each census–census link was formed. See text for explanation of the shading.

Exp	DoB	Neither census record links					1 census record links					Both census records link				
		Name band					Name band					Name band				
		0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
0	0	1911	106	68	61	59	3450	172	129	110	112	13	2	6	16	44
	1	1					0					0				
	2	15	1				28	0				6	0			
	3	29	2				48	3				24	0			
1	0	1376	48	41	50		2571	85	86	104		110	6	46	81	
	1	0					1					0				
	2	31					43					26				
	3	43					92					82				
2	0	1494	53	77			2855	96	159			546	29	161		
	1	0					0					1				
	2	73					187					232				
3	0	1066	33				2029	59				713	23			
	1	0					6					10				
4	0	389					763					321				

6.3 Calculation of Probabilities

The numbers in Table 8 give n_0, n_1 and n_2 for each of the strata. In Table 8 the numbers on the left indicate n_0, those in the middle n_1 and those on the right n_2. The different cells within each of these three parts indicate the different strata.

Table 9 Proportion of links in the strata estimated to be matches ($\widehat{P(M)}$) by expected, date of birth, name band. Cells shaded grey are based on fewer than 30 links.

Expected	Date of birth	Name band				
		0	1	2	3	4
0	0	0.99	0.98	0.93	0.78	0.38
	1					
	2	0.68				
	3					
1	0	0.94	0.89			
	1					
	2					
	3					
2	0	0.72	0.53			
	1	0.00				
	2					
3	0	0.46	0.33			
	1	0.38				
4	0	0.34				

Using the equations above, the proportion of the census–census links in each strata that are matches can then be estimated. These are given in Table 9. Reassuringly, the proportion estimated to be matches increases for stronger evidence (that is, where the band has a lower number).

Table 9 shows $\widehat{P(M)}$ for all the groups of links where it is defined. Discarding the cells were $\widehat{P(M)}$ is ill-defined, and those where there are fewer than 30 links (shaded cells), leaves 20,973 links (see cells shaded green in Table 7). This includes 6,679 links where neither census record links to the NHSCR (blue cells in Table 8), and 12,459 links where one of the census records links to the NHSCR (green cells in Table 8).

All the census records that do not link to any other census records would be assigned a probability of representing a distinct genuine person of 1, as there is no reason to believe that they do not. Similarly, census records that do link to another

census record, but where each links to the NHSCR would get a probability of 1 (by assumption 2). If exactly one of the linked census records links to the NHSCR then the record that does link would also get a probability of 1 (again by assumption 2). The method was developed in order to find the probabilities for the two other cases:

- The record that does not link to administrative data in census–census links where one of them does link to the administrative data
- Each record in census–census links where neither census record links to the administrative data (which is assumed to be the same for each record as there is no way to discriminate between them)

Table 10 For census records that do not link to the NHSCR, the probability assigned to them that they represent a genuine person, broken down by strata and whether one of the linked census records links to the NHSCR.

		Neither record links to NHSCR					One record links to NHSCR				
Expected	Date of birth	Name band					Name band				
		0	1	2	3	4	0	1	2	3	4
0	0	0.50	0.50	0.51	0.55	0.71	0.00	0.01	0.05	0.18	0.59
	1										
	2	0.58					0.26				
	3										
1	0	0.51	0.52				0.05	0.08			
	1										
	2										
	3										
2	0	0.56	0.63				0.23	0.42			
	1										
	2										
3	0	0.66	0.74				0.49	0.64			
	1										
4	0	0.73					0.62				

Table 10 shows the probabilities assigned to 2011 census records in these two cases. All other census records would have a probability of 1. These are the probabilities that would be used in the estimation process. 25,817 of the census records would have a probability less than 1. The sum of the probabilities across these records would be around 9,911. This suggests that across the whole census there are $25{,}817 - 9{,}911 = 15{,}906$ records that do not represent distinct genuine persons. Not accounting for this would lead to an overestimate in the population of around 0.3 per cent.

7. Records Linking Multiple Times

In Table 7 there are 22,713 links within the census 2011 dataset that are considered further. These links consist of 45,027 records. 99.2 per cent of these records link to exactly one other record. However there are some that link to two or three other records. This could cause problems when assigning probabilities to these records as the links may be of different strengths or the records they link to may differ in whether they link to the administrative dataset. Also, in some cases a record may link to multiple records that are in the same postcode. In live running this situation should not happen often as these would generally have been resolved into one record during RMR. This difference between the run here and the live running may introduce a bias if the probabilities calculated for this paper had to be applied during 2022 processing.

To address this, links were removed from the set of 22,713 links until a situation was reached where every linked record links to only one other record. This was done by counting the number of times each record appears in the set of links. The links were then swapped so that the record on the left had the higher (or same) occurrence. For each record ID on the left all but the strongest link (or whichever was first) was removed. This process was repeated until every record in the set of links appears exactly once. The number of records that appear in 1, 2 or 3 links after each of these passes is shown in Table 11. Once this process has completed 44,878 records remain in 22,439 links.

Table 11 Number of records that appear in 1, 2 or 3 links after passes to remove links from records the link multiple times.

Number of records each record links to	Number of Passes 0	1	2	3
1	44,647	44,870	44,878	44,878
2	361	131	6	0
3	19	0	0	0
Total	45,027	45,001	44,884	44,878

In live running this step could be carried out by hand. This will allow a clerical reviewer to say whether it was more plausible that all records are distinct people or if some are definitely the same person. For example, if three records linked to each other the reviewer might decide that one of the records was likely to be a different

person. They would then remove the links to that record, leaving just the link between the remaining records to go on to be assigned probabilities. Alternatively if they thought that two of the records were definitely the same person then these two records could be resolved by deleting one of the records. Again this would just leave one link between two records to continue to the next stage of processing. A final option would be to manually assign probabilities to the three records. Before being passed to review the records could be linked to the administrative dataset to aid the reviewers decision.

Table 12 Number of census 2011 links by expected, date of birth and name band and the number of census records that link to the NHSCR 2011 with at least the strength with which each census–census link was formed and with links removed so that only 1–1 links remain. See text for explanation of the shading.

Exp	DoB	Neither census record links					1 census record links					Both census records link				
		Name band					Name band					Name band				
		0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
0	0	1890	104	68	61	57	3433	172	127	109	111	13	2	6	16	43
	1	1					0					0				
	2	15	1				26	0				6	0			
	3	29	2				44	3				21	0			
1	0	1359	48	41	49		2560	85	85	102		109	6	45	79	
	1	0					1					0				
	2	31					41					26				
	3	38					90					80				
2	0	1470	53	76			2837	93	154			541	28	157		
	1	0					0					1				
	2	70					181					230				
3	0	1048	33				1996	58				701	23			
	1	0					6					10				
4	0	380					746					312				

Table 12, Table 13 and

Table 14 are the equivalents of Table 8, Table 9 and Table 10 respectively with the links removed as discussed above. For strata with a substantial number of links this the resulting probabilities are very similar to what was obtained previously.

Table 13 Proportion of links in the strata estimated to be matches ($\widehat{P(M)}$) by expected, date of birth, name band and with links removed so that only 1–1 links remain. Cells shaded grey are based on fewer than 30 links.

Expected	Date of birth	Name band				
		0	1	2	3	4
0	0	0.99	0.98	0.93	0.78	0.41
	1					
	2	0.64				
	3					
1	0	0.94	0.89			
	1					
	2					
	3					
2	0	0.72	0.52			
	1	0.00				
	2					
3	0	0.46	0.30			
	1	0.38				
4	0	0.35				

Table 14 For census 2011 records that do not link to the NHSCR 2011 the probability assigned to them that they represent a genuine person, broken down by strata and whether one of the linked census records links to the NHSCR 2011 with links removed so that only 1–1 links remain.

		Neither record links to NHSCR					One record links to NHSCR				
Expected	Date of birth	Name band					Name band				
		0	1	2	3	4	0	1	2	3	4
0	0	0.50	0.50	0.51	0.55	0.69	0.00	0.01	0.05	0.18	0.55
	1										
	2	0.59					0.32				
	3										
1	0	0.51	0.52				0.05	0.08			
	1										
	2										
	3										
2	0	0.56	0.64				0.22	0.44			
	1										
	2										
3	0	0.66	0.76				0.49	0.69			
	1										
4	0	0.72					0.62				

8. Results Using Rehearsal Data

In October 2019 a rehearsal of the census was conducted in parts of Glasgow, Dumfries and Galloway, and the Western Isles. The method was tested again using the 2019 rehearsal data. 46,891 records were used (after filtering out records with missing information and students who live at different addresses during term time).

214 census–census links were found, and Table 15 breaks these down by their categorization. The remaining 132 links were then linked to a 2019 version of the NHSCR.

Table 15 Number of links discarded by reason.

Category	Number of links
Parent–child pair	36
Twins	8
Last name different	1
First name different	0
DoB different	37
Remaining	132
Total	214

Table 16 shows the breakdown of the remaining 132 links by the name and date of birth band.

Table 16 Number of links by band for name and date of birth. Shaded cells are those that the links in Table 17 are drawn from.

Date of birth	Name band						
	0	1	2	3	4	5	Total
0	21	0	0	2	0	1	24
1	1	0	0	0	0	1	2
2	12	1	0	0	0	0	13
3	45	0	0	1	0	0	46
4	47	0	0	0	0	0	47
Total	126	1	0	3	0	2	132

Table 17 breaks them down further by the expected band. The green cell is the only one where $\widehat{P(M)}$ can be calculated. The grey cells indicate links that are discarded at this stage as they are too weak.

Table 17 Number of links by expected, date of birth and name band. See text for explanation of the shading.

Expected	Date of birth	Name band				Total
		0	1	2	3	
0	0	7	0	0	0	7
	2	0	1	0	0	1
1	0	5	0	0	0	5
2	0	9	0	0	1	10
3	0	0	0	0	1	1
	3	11	0	0	0	11
4	2	2	0	0	0	2
	3	12	0	0	0	12
5	1	1	0	0	0	1
	2	10	0	0	0	10
	3	22	0	0	1	23
Total	0	21	0	0	2	23
	1	1	0	0	0	1
	2	12	1	0	0	13
	3	45	0	0	1	46
	Total	79	1	0	3	83

Table 18 Number of census links by expected, date of birth and name band and the number of census records that link to the NHSCR with at least the strength with which each census–census link was formed. See text for explanation of the shading.

Expected	Date of birth	Neither census record links		1 census record links		Both census records link	
		Name band		Name band		Name band	
		0	1	0	1	0	1
0	0	0		7		0	
	2		1		0		0
1	0	0		5		0	
2	0	0		8		1	

Table 18 shows a further breakdown of the links by the number of the census records in the link that link to the NHSCR. The shaded cells here indicate those for which the specific probability of being a match can be calculated. Note that for other cells there are no corresponding links where both census records link to the NHSCR.

Table 19 Proportion of links in the strata estimated to be matches ($\widehat{P(M)}$) by expected, date of birth, name band.

Expected	Date of birth	Name band	
		0	1
0	0		
1	2		
1	0		
2	0	0.89	

Table 19 shows $\widehat{P(M)}$ by band. That many cells are blank highlights the problem that $\widehat{P(M)}$ cannot be calculated when $n_2 = 0$. This is likely to be less of an issue in live running (as was seen in the testing using the 2011 data). This is because there will be many more cases, and so it is less likely that $n_2 = 0$. However, Section 10.3 considers some options for handling cases where $n_2 = 0$. Table 20 shows the probabilities that are attached to each of the census records (that do not link to the NHSCR.

Table 20 For census 2019 records that do not link to the NHSCR 2019, the probability assigned to them that they represent a genuine person, broken down by strata and whether one of the linked census records links to the NHSCR.

		Neither record links to NHSCR		One record links to NHSCR	
Expected	Date of birth	Name band		Name band	
		0	1	0	1
0	0				
1	0				
2	0				
3	0			0.0	

9. Proposed Method for Correction for Overcount in Estimation

The overall proportion of overcounted records can be calculated using the probabilities from Census–Census linking, and the proportions of misplaced records identified in Census–CCS linking. Records would be omitted from this calculation where the term-time indicator question in either the Census or the CCS indicates that they lived at another address during term time. CCS records should also not be included as misplaced where the respondent has answered that they lived somewhere else on Census day.

As in 2011, the overcount propensity will be calculated separately for strata of different age groups and Hard to Count. For the Census-CCS links identifying Type 3 overcount, these will be counted in the stratum according to their Census variables.

For Type 2 overcount, the sum of the probabilities should give an approximation for the 'true count' of the population. This gives the denominator for the equation used in 2011, while the overall count in the Census is the numerator. These will be stratified by age group, though they do not need to be the same age groups as used in 2011. Equation 23 demonstrates this calculation for γ_2, the propensity for Type 2 overcount, using X, the total census count, and the probability of a record being a genuine person, P.

$$\gamma_2 = \frac{X}{\sum P} \qquad 23$$

For Type 3 overcount, the Census–CCS links made between different addresses can provide the erroneous count, E in Equation 1. Although it was assumed that the CCS records had the correct location in 2011, investigation of many of the records does not agree with this, and therefore we are assuming that the Census is right 50 per cent of the time (and the CCS is incorrect in these cases). Therefore we only consider half of the links as being misplaced Census records. The numerator for the propensity would be the number of Census records that link to a CCS record, and the denominator this total minus the erroneous count. This is demonstrated in

Equation 24, where X_{linked} is the total number of links between the Census and CCS, and E_{links} is the number of linked Census records that are included in the wrong location (Calculated as half of the links where the Census and CCS records are in different locations). As we cannot detect errors in unlinked records, this is only using the count of Census records that linked to a CCS record. This would assume that the rate of misplacement is the same for records that do not link to the CCS as those that do.

$$\gamma_3 = \frac{X_{\text{linked}}}{X_{\text{linked}} - E_{\text{links}}} \qquad 24$$

Any CCS records that were linked to a Census record in a different location would be considered out of scope, and not used as links in DSE.

$$\gamma_T = \gamma_2\gamma_3 \qquad 25$$

Multiplying these propensities together gives the overall overcount propensity (γ_T, see Equation 25), a measure of the number of records in the data for every one genuine record. When calculating the estimates with DSE, without overcount each census record would count as 1 as in Equation 26 (where n_{Census} is the number of records on the census, n_{CCS} is the number of records on the CCS, n_{both} is the number of records that represent people who appear on both, and $\widehat{N}$ is the estimate of the population in a particular stratum). Instead the inverse of this propensity is used, as in Equation 27 to give a modified population estimate $\widehat{N_m}$, based on the age group and hard to count index of each individual record, dampening the estimate accordingly. $\gamma_{T,i}$ is the total propensity for record i, as these may be different depending on which stratum the record is in.

$$\widehat{N} = \frac{(n_{\text{Census}} + 1)(n_{\text{CCS}} + 1)}{n_{\text{both}} + 1} - 1 \qquad 26$$

$$\widehat{N_m} = \frac{\left(1 + \sum_{i=1}^{n_{\text{Census}}} \frac{1}{\gamma_{T,i}}\right)(n_{\text{CCS}} + 1)}{n_{\text{both}} + 1} - 1 \qquad 27$$

9.1 Results using 2011 Data

Each of the propensities were calculated using the data from 2011 which had been linked using the new methods. Using the probabilities calculated from links to admin data, the value of $1/\gamma_2$ was 0.9968 without stratification. The value for $1/\gamma_3$ was 0.9968. Multiplying these together suggests that 99.37 per cent of the records are considered to be genuine, non-overcounted records.

This detects a slightly lower level of overcount compared to 2011, where the overall percentage was 99.37%. Calculating the Type 2 overcount proportion similarly to 2011, where each record has a 50% chance of representing a genuine person, but using the new links gives $1/\gamma_2$ of 0.9958, which when combined with Type 3 overcount gives 99.26 per cent.

Although the overall effect of the 2011 method and the proposed method are the same, the way they reached this adjustment differ. The proposed method would find a greater number of links, but the number it assigned to each record would be closer to the default of 1. In 2011, while any exact matches on name and date of birth would have been considered to be a duplicate, using probabilities allows for the possibility that the records are genuine different people, particularly where the link is between people with commonly used names.

10. Strengths and Limitations

As the level of identified overcount increased from 2001 to 2011, it is expected that it will not be an negligible issue in 2022 and may be more considerable with the move to online collection. Therefore, it is very likely that a correction would be required.

Assigning probabilities to individual records has several advantages over resolving linked records:

- It avoids having to choose which record to retain. While administrative data could be used to assist in determining the correct record to retain, this would risk removing a genuine person. There would be no guidance for which was correct in cases that did not link to administrative data, as well as out-of-date administrative data leading to the wrong record being removed.
- It allows for the possibility that the linked records are legitimately two different people with similar name and date of birth. This gives a more conservative estimate for the level of overcount, particularly where people who have a commonly occurring name have been linked.
- Being more conservative around dealing with links means that a greater set of links can be considered. Records in links with a reasonable chance of representing distinct individuals could not all be resolved, as many should be retained. By assigning probabilities, such cases can be considered, and assigned probabilities close to one.

Using individual-level probabilities also allows records to be grouped into arbitrary estimation strata, to use the results in DSE, while previous sample based methods could only be stratified.

There is an assumption that the level of misplacement captured within the Census–CCS linking is representative of the overall population. This could be explored by looking to repeat the calibration process used in 2011 between duplicates identified in the Census–CCS linking and the Census–Census linking.

10.1 Timings and Practicalities

The first pass for the date of birth block took half an hour to run, while the first pass for the name block took just three minutes. Scoring and categorizing the links found in the first passes took just three minutes. Linking the linked census records to the administrative dataset took around 20 minutes. Therefore, the whole process can run in under an hour.

As the probabilities can be calculated directly from the links to administrative dataset, there is no need for clerical review of individual links, so there is no further time requirement. The process will be run in the secure administrative data area in order to link to the NHSCR due to security protocols of working with third party data. The final table of probabilities will then be transferred from the administrative data area to the data processing area.

10.2 Detail to be Worked Out

The above analysis assumes that census records would link to at most one other census record. Thought needs to be given for how to handle cases where there are groups of more than two census records linked together.

Further thought should be given as to how to handle cases with different dates of birth. Most of the links that were ultimately used had exact agreement on date of birth. Therefore it may be simpler to only consider such links. This would simplify the code, although it would make little difference to the run time (or the results).

Another potential issue is that some within-postcode census links may not be cases that RMR concluded represented different persons. This may happen if a person appears in two distinct households in the same postcode. If the households also have different persons in them, then it would not be known which household the person should appear in, and so which record should be retained. These may then be left in the dataset, to be dealt with at this stage.

In order to include such links in the overcount process, a note would be needed of them from RMR, to avoid them being discarded with other within-postcode links. Additionally, care would need to be taken when linking to administrative data. For links across postcodes it can be assumed that if both census records link to administrative data then there are two distinct administrative data records. This is because the linking to administrative data blocks on postcode. If the census records are in the same postcode then it is likely that both census records would link to the same administrative data record. Therefore a count of the number of linked administrative data records would be needed.

10.3 Dealing With Strata Where $n_2 = 0$

The above analysis uses $\widehat{P(2)} = \frac{n_2}{n}$ in order to ultimately work out $\widehat{P(M)}$. However, n_2 may be zero, even in cases when $P(2) \neq 0$. In strata where this happens the proportion of matches cannot be estimated, as the calculation involves dividing by n_2.

There are ways round this. A different estimator could be used, such as those discussed in Brown et al. (2001). This could be $\widehat{P(2)} = \frac{2+n_2}{4+n}$.

Another method would be to find a prior for $P(2)$, perhaps by merging different adjacent strata in order to get enough cases so that $n_2 > 0$. Adjusting the prior using the information for the strata of interest would give a posterior distribution for $P(2)$. From this, an estimate for $\widehat{P(2)}$ could be found that would not be exactly 0.

It may be that the problems would not be too great as overestimates in some strata would cancel out with underestimates in others. Some different methods could be tested using Monte Carlo simulations to see which methods best recover the original population.

10.4 Contingency: Not Making Use of NHSCR 2022

There is a risk that NRS are not able to link the census to the NHSCR 2022 (or any other administrative dataset covering the Scottish population). In this situation there would be no further information about which links were matches and no way of determining which of the linked records represented the genuine person.

The best that could be done in such a situation would be to group the links according to their strength, as described above. For each group the estimated probability of a link being a match ($\widehat{P(M)}$), calculated using the 2011 census and NHSCR 2011 data, could be used to estimate the probability that each record represented a genuine person. These values are given in Table 9. These values have been saved and are held in place to be used as an estimate in case the administrative data is not available to support the 2022 census.

The disadvantages of this option when compared with the main method include:

1. It does not account for systematic differences between the 2011 census and the 2022 census. It is likely that different guidance, and the proportion of online returns mean that the likelihood of links being matches would be different between 2011 and 2022. Applying the 2011 probabilities to the 2022 data could therefore introduce a bias.
2. Distinction cannot be made between links that would link to 0, 1 or 2 administrative-data records. Therefore the same probability would need to be used for all links, which would be less discriminating.
3. Distinction cannot be made between the records in each link. If exactly one of the linked census records were to link to the administrative data then it would be known that the linked record represents a genuine person. Not having this information would result in each record being treated the same, and so again this would be less discriminating.

Without the values of $\widehat{P(M)}$ from 2011, there would likely be no way to estimate the probabilities of records being genuine. Clerical review could perhaps make a guess,

either on individual cases, or for a group of links. However this would be time intensive, and there would be no way to calibrate the results so would likely introduce a bias.

11. Conclusion

The method presented here corrects for type 2 and 3 overcoverage. Type 2 is when a person appears multiple times in the census dataset (not counting cases dealt with by RMR). Type 3 is cases where a person appears at different locations in the census and CCS. That leads to overcoverage because they will already be counted at the location of the census record, but estimation will consider them to have been missed at the location of the CCS record, and increase the estimate there.

The method should account for most of the overcoverage, without the need for clerical review. Linking to administrative data helps avoid correcting for pairs of records that are just people who happen to have similar names and dates of birth. If administrative data is not available then the contingency method would be used. This includes the same census–census linking exercise, but uses the results from the 2011 NHSCR linking to inform the probabilities for census records. Therefore the main method is recommended, with the contingency method to be used in the event that administrative data is not available.

12. References

Brown, L., Cai, T. and DasGupta, A., (2001) 'Interval Estimation for a Binomial Proportion', *Statistical Science*, vol. 16, no. 2, pp. 101–133

Porter, E. and Winkler, W. (1997), *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, [online] available at: https://www.census.gov/library/working-papers/1997/adrm/rr97-02.html

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/

Annex 1: 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[13]. 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.

[13] 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 21. 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 21 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[14] 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

[14] 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.

Annex 2: 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 22 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 22 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.

Annex 3: Notation

The following explains the mathematical notation used in this document.

Notation	Explanation	Example	Explanation of example
$P(x)$	Probability that event x occurs or state x is true.	$P(0)$	Probability that neither of the linked census records links to an administrative record
$\|$	Conditional on, or given that.	$P(0\|M)$	Probability that 0 is true conditional on / given that M is true. That is: probability that neither of the linked census records links to an administrative record conditional on / given that the two linked census records represent a match (the same individual)
$\bar{x}$	Negation of x, or not x. This is equivalent to $\neg x$. If x is true then $\bar{x}$ is false and vice versa.	$\bar{M}$	The case that the two linked census records are not a match.
$\hat{x}$	Estimation of x (using available data).	$\widehat{P(M)}$	Estimate of the probability that the two linked census records represent a match (the same individual)

Annex 4: Glossary

Term	Definition
Match	Two records that relate to the same individual
Non-match	Two records that do not relate to the same individual
Link	Two records that have been connected
RMR	Resolve Multiple Returns. This is a census–census linking process that is blocked on postcode.
NHSCR	National Health Service Central Register. This is an administrative dataset of people who are registered with an NHS GP.
DSE	Dual-System Estimation. A statistical process using the links between two independent datasets of the same population to estimate the total population. See the Estimation and Adjustment Methodology paper for more information on DSE.
Propensity	A measure of how much overcount is included in the census dataset. It is defined as the number of records in the dataset divided by the number of persons in the population these represent. For example a propensity of 1 indicates that there is no overcount (and so no modification is needed), while a propensity of 2 indicates that for each two census records there is one persons in the population (and so we need to divide the estimate by 2). The propensity will be defined for particular strata, although the stratification will differ from that used for DSE.

Annex 5: Blocking

This annex describes the blocking method for the initial linking.

After loading in the data the blocking variables are set up. There are two blocking strategies used, one based on name, the other on date of birth. Using the name and date of birth exactly as presented would mean that matches where the name and date of birth were both slightly different would never be found. When the location, name and date of birth were all different the differences in at least one of these would need to be very minor for us to still believe that they represented the same person. This might include components in different order. Thus having John Smith and Smith John this would not provide evidence of a non-match, while John Smith and Jon Smith might start to suggest there were different people (especially if other information were also different). Another minor difference would be to miss out middle names. So John Smith and John Robert Smith could well be the same person. The name blocking variable is therefore taken to be the concatenation of first and last names, but these are first sorted alphabetically before being concatenated. So John Smith and Smith John would both appear as JohnSmith in the blocking variable. Similarly Robert Jones and Jones Robert would both appear as JonesRobert in the blocking variable (as Jones precedes Robert alphabetically).

A common difference in dates of birth is that sometimes they are entered in American format (MM/DD/YYYY). Again records that differ in this regard can be placed into the same block by first sorting the day and month component before concatenating the three components into a date of birth. This would mean that 01/02/1980 would be placed in the same block as 02/01/1980. The day and month are only swapped if both numbers are 12 or smaller.

By making these two changes to the blocking variables means that records can still link even if there are minor differences in date of birth, name and location. If there were more substantial differences in both name and date of birth then it is unlikely that that records in different locations would be considered to represent the same

person. Therefore this method should be able to efficiently consider all the cases of interest.

When running the blocking, all the records with the same value for the blocking variable are loaded into RAM (a SAS temporary array). To set up the arrays the number of records in the largest block is needed. Each combination of records within the block is then looped through.

Annex 6: Bigram Comparison

This annex describes the comparison used in the first stage of linking. When blocking on date of birth the comparison is made using name, and vice versa.

To compare the values of the comparison variable the values are converted to bigrams. The bigrams of a string are all the pairs of letters that appear together in order. For example 'JOHNSMITH' would be converted to eight bigrams: 'JO', 'OH', 'HN', 'NS', 'SM', 'MI', 'IT' and 'TH' and 'JONSMITH' would be converted to seven bigrams: 'JO', 'ON', 'NS', 'SM', 'MI', 'IT' and 'TH'. These bigrams are sorted alphabetically and stored in a temporary array when the data is loaded. The two sets of bigrams are then scanned through, keeping count of the number that are the same. In the above example there are six bigrams common to the two sets: 'JO', 'NS', 'SM', 'MI', 'IT' and 'TH'. A distance measure[15] is then calculated using this count and the number of bigrams in each set: $d = 1 - \left(\frac{2c}{b_1+b_2}\right)$, where *c* is the number of bigrams in common b_1 and b_2 are the number of bigrams in set 1 and 2 respectively. The example above would then have $d = 1 - \left(\frac{2\times6}{8+7}\right) = 1 - \frac{12}{15} = 0.2$. If the strings were identical then all the bigrams would be the same and $d = 0$. Cases where the distance measure is less than some threshold are saved to a dataset to be considered further. When doing comparisons on full name the threshold used is 0.4, while for dates of birth a threshold of 0.3 is used.

[15] See Porter and Winkler (1997).

Annex 7: Derivation of Expressions in Table 3

This annex derives the expressions used in Table 3. These are the probabilities that for census–census links that are matches, or for those that are non-matches, 0, 1 or 2 of the census records link to the administrative dataset.

First, the probabilities when the link is a match ($P(0|M)$, $P(1|M)$ and $P(2|M)$) are derived. When the census–census link is a match then, by assumption 1, one of the records represents a distinct genuine person (g) and the other does not ($\bar{g}$). When neither links to the administrative dataset then both the genuine and non-genuine record do not link to the administrative dataset. Therefore:

$$P(0|M) = P(\bar{l}|\bar{g})P(\bar{l}|g) \quad (28)$$ by assumption 1

substitute in equations **Error! Reference source not found.** and **Error! Reference source not found.**

$$P(0|M) = 1 \times (1-p) \quad (29)$$

$$P(0|M) = 1 - p \quad (30)$$

When exactly one of the census records links to the administrative dataset then either the genuine record links and the non-genuine record does not link, or the genuine record does not link and the non-genuine record links to the administrative dataset. Therefore:

$$P(1|M) = P(l|\bar{g})P(\bar{l}|g) + P(\bar{l}|\bar{g})P(l|g) \quad (31)$$ exactly 1 of the records links

substitute in equations **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.**

$$P(1|M) = 0 \times (1-p) + 1 \times p \quad (32)$$

$$P(1|M) = p \quad (33)$$

When both of the census records link to the administrative dataset then the genuine record links to the administrative dataset and the non-genuine record links to the administrative dataset. Therefore:

$$P(2|M) = P(l|\bar{g})P(l|g) \quad (34)$$ both records must link

$$P(2|M) = 0 \times p \quad (35)$$ substitute in equations **Error! Reference source not found.** and **Error! Reference source not found.**

$$P(2|M) = 0 \quad (36)$$

That $P(2|M) = 0$ is a reflection of our assumption that records linking to the administrative dataset are genuine. But if the link is a match then one of the records does not represent a distinct genuine person. Therefore, it is not possible for both records of a match to link the administrative dataset.

Now the probabilities for when the link is a non-match ($P(0|\bar{M})$, $P(1|\bar{M})$ and $P(2|\bar{M})$) are derived. If a link represents a non-match then both of the records are genuine. Therefore:

$$P(0|\bar{M}) = P(\bar{l}|g)P(\bar{l}|g) \quad (37)$$ neither record links

$$P(0|\bar{M}) = (1-p)(1-p) \quad (38)$$ substitute in equation **Error! Reference source not found.**

$$P(0|\bar{M}) = (1-p)^2 \quad (39)$$

$$P(1|\bar{M}) = P(l|g)P(\bar{l}|g) + P(\bar{l}|g)P(l|g) \quad (40)$$ exactly 1 of the records links

$$P(1|\bar{M}) = p \times (1-p) + (1-p) \times p \quad (41)$$ substitute in equations **Error! Reference source not found.** and **Error! Reference source not found.**

$$P(1|\bar{M}) = 2p(1-p) \quad (42)$$ collect terms

$P(2\|\bar{M}) = P(l\|g)P(l\|g)$	43	both records must link
$P(2\|\bar{M}) = p \times p$	44	substitute in equation **Error! Reference source not found.**
$P(2\|\bar{M}) = p^2$	45	

This concludes the derivation of the probabilities shown in Table 3.

Annex 8: A Relationship Involving $P(M)$, p and $P(0)$

$P(0) = P(0\|M)P(M) + P(0\|\bar{M})P(\bar{M})$	46	from Equation **Error! Reference source not found.**
$P(0) = P(0\|M)P(M) + P(0\|\bar{M})(1 - P(M))$	47	apply Equation **Error! Reference source not found.**
$P(0) = (1-p)P(M) + (1-p)^2(1 - P(M))$	48	substitute in equations **Error! Reference source not found.** and **Error! Reference source not found.**
$P(0) = (1-p)P(M) + (1 - 2p + p^2) - P(M)(1 - 2p + p^2)$	49	expand
$P(0) = (1 - 2p + p^2) + P(M)(1 - p - 1 + 2p - p^2)$	50	collect terms
$P(0) = P(M)(p - p^2) + (1 - 2p + p^2)$	51	cancel

Annex 9: A Relationship Involving $P(M)$, p and $P(2)$

$P(M\|2) + P(\bar{M}\|2) = 1$	52	from Equation **Error! Reference source not found.**
$\frac{P(2\|M)P(M)}{P(2)} + \frac{P(2\|\bar{M})P(\bar{M})}{P(2)} = 1$	53	apply Equation **Error! Reference source not found.**
$\frac{0 \times P(M)}{P(2)} + \frac{p^2 P(\bar{M})}{P(2)} = 1$	54	substitute in equations **Error! Reference source not found.** and **Error! Reference source not found.**
$\frac{p^2(1 - P(M))}{P(2)} = 1$	55	apply Equation **Error! Reference source not found.**

$$p = \sqrt{\frac{P(2)}{1 - P(M)}}$$ 56 rearrange

Annex 10: Derivation of $P(M)$, the Probability that a Link is a Match

Using the two equations involving $P(M)$ and p (equations **Error! Reference source not found.** and **Error! Reference source not found.**) one can be substituted into the other to eliminate p.

$P(0) = P(M)\left(\sqrt{\frac{P(2)}{1-P(M)}} - \frac{P(2)}{1-P(M)}\right) + 1 - 2\sqrt{\frac{P(2)}{1-P(M)}} + \frac{P(2)}{1-P(M)}$	57	substitute Equation **Error! Reference source not found.** into Equation **Error! Reference source not found.**
$P(0) - 1 = \sqrt{\frac{P(2)}{1-P(M)}}(P(M) - 2) + \frac{P(2)}{1-P(M)}(1 - P(M))$	58	collect terms
$P(0) - 1 = \sqrt{\frac{P(2)}{1-P(M)}}(P(M) - 2) + P(2)$	59	cancel
$P(0) - P(2) - 1 = \sqrt{\frac{P(2)}{1-P(M)}}(P(M) - 2)$	60	rearrange
$(P(0) - P(2) - 1)^2 = \frac{P(2)}{1-P(M)}(P(M) - 2)^2$	61	square
$\frac{(P(0) - P(2) - 1)^2}{P(2)} = \frac{P(M)^2 - 4P(M) + 4}{1-P(M)}$	62	rearrange and expand
$A \equiv \frac{(P(0) - P(2) - 1)^2}{P(2)}$	63	define A

$A = \frac{P(M)^2 - 4P(M) + 4}{1 - P(M)}$	64	substitute Equation 63 into Equation 62
$A - AP(M) = P(M)^2 - 4P(M) + 4$	65	rearrange
$A - 4 = P(M)^2 + P(M)(A - 4)$	66	collect terms
$A - 4 = \left(P(M) + \frac{A-4}{2}\right)^2 - \left(\frac{A-4}{2}\right)^2$	67	complete the square
$\left(P(M) + \frac{A-4}{2}\right)^2 = A - 4 + \left(\frac{A-4}{2}\right)^2$	68	rearrange
$P(M) + \frac{A-4}{2} = \sqrt{A - 4 + \left(\frac{A-4}{2}\right)^2}$	69	root
$P(M) = \sqrt{A - 4 + \frac{A^2}{4} - 2A + 4} - \frac{A-4}{2}$	70	rearrange and expand
$P(M) = \sqrt{\frac{A^2}{4} - A + 2} - \frac{A}{2}$	71	cancel

$P(0)$ and $P(2)$ can now be estimated by using the observed number of census links where neither census record links to an administrative record and the observed number where both records link to an administrative record (see Table 3). These estimates can then be used to estimate $P(M)$.

$$A = \frac{(P(0) - P(2) - 1)^2}{P(2)} \qquad 72 \quad \text{Equation 63}$$

$$\hat{A} = \frac{\left(\widehat{P(0)} - \widehat{P(2)} - 1\right)^2}{\widehat{P(2)}} \qquad 73 \quad \text{estimate}$$

$$\hat{A} = \frac{\left(\frac{n_0}{n} - \frac{n_2}{n} - 1\right)^2}{\frac{n_2}{n}} \qquad 74 \quad \text{from Table 3}$$

$$\widehat{P(M)} = \sqrt{\frac{\hat{A}^2}{4} - \hat{A} + 2} - \frac{\hat{A}}{2} \qquad 75 \quad \text{estimate from Equation 71}$$

Equation 75 represents the estimate of the proportion of all the census–census links in the stratum that represent a match. The estimated total number of matches in the stratum is therefore $\widehat{n(M)} = n\,\widehat{P(M)}$ (that is, the estimated number of matches among n census–census links in a stratum is the observed number of census–census links in that stratum (n) multiplied by the estimated probability that each of those census–census links represents a match ($P(M)$)). $\widehat{n(M)}$ is also the estimate for the number of census records in the stratum that do not represent distinct genuine individuals, that is, the overcount.

Annex 11: Using $\widehat{P(M)}$ To Assign Probabilities to Census Records

This annex derives the expressions in Table 4, the probabilities attached to each census record.

For the case where neither census record links to the administrative dataset:

$$P(M|0) = \frac{P(0|M)P(M)}{P(0)} \quad (76)$$

from Equation **Error! Reference source not found.**

$$P(M|0) = \frac{(1-p)P(M)}{P(0)} \quad (77)$$

substitute in Equation **Error! Reference source not found.**

$$P(M|0) = \frac{P(M)\left(1-\sqrt{\frac{P(2)}{1-P(M)}}\right)}{P(0)} \quad (78)$$

substitute in Equation **Error! Reference source not found.**

$$\widehat{P(M|0)} = \frac{\widehat{P(M)}\left(1-\sqrt{\frac{n_2/n}{1-\widehat{P(M)}}}\right)}{\frac{n_0}{n}} \quad (79)$$

estimate

For the case where one of the census records links to the administrative dataset:

$$P(M|1) = \frac{P(1|M)P(M)}{P(1)} \quad (80)$$

from Equation **Error! Reference source not found.**

$$P(M|1) = \frac{pP(M)}{P(1)} \quad (81)$$

substitute in Equation **Error! Reference source not found.**

$$P(M|1) = \frac{P(M)\sqrt{\frac{P(2)}{1-P(M)}}}{P(1)} \quad (82)$$

substitute in Equation **Error! Reference source not found.**

$$\widehat{P(M|1)} = \frac{\widehat{P(M)}\sqrt{\frac{n_2/n}{1-\widehat{P(M)}}}}{\frac{n_1}{n}} \qquad 83 \quad \text{estimate}$$

For the case where both of the census records link to the administrative dataset:

$$P(M|2) = \frac{P(2|M)P(M)}{P(2)} \qquad 84 \quad \text{from Equation } \textbf{Error! Reference source not found.}$$

$$P(M|2) = \frac{0 \times P(M)}{P(2)} = 0 \qquad 85 \quad \text{substitute in Equation } \textbf{Error! Reference source not found.}$$

Once the probability of a pair of census records being a match is available, this needs to be converted to the probability that each record represents a (distinct) genuine person (call this $P(g)$). When both census records link to the census, assumption 2 suggests that each census record is genuine. Therefore:

$$P(g|2) = 1 \qquad 86 \quad \text{by assumption 2}$$

When one of the census records links to the administrative dataset then the record that links to the administrative dataset needs to be treated differently from the one that does not link. Therefore call the probability that the census record that links to the administrative dataset represents a distinct genuine person $P_l(g|1)$. Similarly, call the probability that the census record that does not link to the administrative dataset represents a distinct genuine person $P_{\bar{l}}(g|1)$. In particular the record that links will be considered genuine (by assumption 2). The other census record will be genuine just in case the link is a non-match. Therefore:

$$P_l(g|1) = 1 \qquad 87 \quad \text{by assumption 2}$$

$$P_{\bar{l}}(g|1) = \widehat{P(\bar{M}|1)} \qquad 88$$

$$P_{\bar{l}}(g|1) = 1 - P(\widehat{M}|1) \quad (89)$$

apply Equation **Error! Reference source not found.**

$$P_{\bar{l}}(g|1) = 1 - \frac{\widehat{P(M)}\sqrt{\frac{n_2/n}{1-\widehat{P(M)}}}}{\frac{n_1}{n}} \quad (90)$$

substitute in Equation **Error! Reference source not found.**

When neither census record links to the administrative dataset then, like the case where both records link, the two records are treated identically. When the link is a match then exactly one of the two census records represents a genuine person. Therefore the total probability of the two census records representing a genuine person is 1. Thus the probability of each record representing a genuine person will be 0.5 (by symmetry). This probability will increase linearly as $P(\widehat{M}|0)$ decreases, until it reaches 1 when $P(\widehat{M}|0) = 0$. Therefore

$$P(g|0) = 1 - \frac{P(\widehat{M}|0)}{2} \quad (91)$$

$$P(g|0) = 1 - \frac{\widehat{P(M)}\left(1 - \sqrt{\frac{n_2/n}{1-\widehat{P(M)}}}\right)}{\frac{2n_0}{n}} \quad (92)$$

substitute in Equation **Error! Reference source not found.**

These results lead to the expressions in Table 4.

Annex 12: 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