NHS Central Register (NHSCR) Governance Board Meeting
24th October 2013

Linking the NHS Central Register Extract to the 2011 Census

1. Purpose

A previous paper considered by the Group (NEWG13B) reported initial work in linking the NHS Central Register (NHSCR) extract to Processing Units (PUs) 1, 2, 3 and 7 of the 2011 Census. This paper reports how the work had been taken forward since then. The two main changes are that the processing of individual PUs has been replaced by the processing of all PUs together, and the way of presenting the results has been changed.

2. Method

It will be recalled that in probabilistic record linkage, the probability that two records taken respectively from file A and file B refer to the same person is never considered to be either zero (it is impossible) or one (it is certain). All events are regarded in terms of the match probability (i.e. how likely it is that the two records refer to the same person). This reflects the fact that the data is always incomplete and/or error-prone and never supports absolute judgments. The measure normally used to assess the probability of a match is a log likelihood ratio but, without some knowledge of mathematics, this is not an easy figure to interpret. In the present method therefore it is replaced by the actual probability itself which is easier to interpret. It is supplemented by a second measure, which is the likely number of false links accepted during a given stage in the procedure. This figure is important because it refers to one of the two types of error which can be made during record linkage. One type is the false link which occurs when a pair of records which in truth refer to different people is accepted as being a match. The other type of error is the missed match where when a pair of records which in truth refer to the same person is not accepted, either because that record pair is not found or because it is found but, at a review stage, it is rejected as a match.

The two files to be matched consisted of 5,672,245 records in the NHSCR file and 5,000,834 records in the Census file. The former is not the totality of the records in the extract (which number around 9 million). For present purposes, those registered as having been alive and registered to a Scottish health board on 27 March 2011 were included. This may have led to the exclusion of a handful of people whose records were in fact in the Census file but the number would not affect the overall shape of the outcome. The Census file is clearly smaller than the population of Scotland and reflects the fact that not everyone who ought to complete a Census form actually does so.

This note outlines the various stages in the linking procedure. Different methods were used for the different stages so as to make the computation as efficient as possible. The following section describe each stage in turn and the outcome.

3. Stage 1: the deterministic stage

This stage does not use probabilistic linkage software. The reality of linkage is that some matches are easier to find than others and do not need computationally intensive methods. Even within this stage however there are gradations in the confidence with which a record pair can be said to be a match. The highest level of confidence is found when there is exact agreement between the records on all of the linkage fields being used. In the present case, the fields in stage 1 were first name, last name, date of birth, gender and postcode. For stage 1a, the definition of a link was that the pair had to show exact agreement on all five fields and that this five-way combination was unique in both files. This last requirement was introduced in order to "weed out" duplicate records as it can safely be assumed that two records in either file with the same combination of first name, last name, date of birth, gender and postcode refer to the same person and are duplicate records.

Stage 1a calculation identified a total of 2,922,199 record pairs which satisfied the criterion (58.4% of the records in the Census file). The probability that a record pair satisfying the criterion is a match is so close to one that it difficult to express the difference. If it is subtracted from one and multiplied by 2,922,199 (to give the expected number of false links in this stage), the result is less than one. We can say then that it is very likely that stage 1a identifies nearly three-fifths of the Census records, probably without making an error. So far so good: but the nature of linkage is that the easiest parts come at the start with the procedure becoming more difficult as the number of records remaining unlinked falls.

Stage 1b was also deterministic but employed a slightly looser criterion. This was that there should be exact agreement on first name, last name, date of birth and gender; that at least one of the postcodes should be missing so that no comparison is possible on this field; and that this four-way combination was unique in both files. For example, two male John Smiths born on the same day, where no other male John Smith born on this day was recorded in either file, would be accepted as a match. This criterion identified a further 777,204 record pairs (15.5% bringing the total to 74.0% of the Census records). However the price for accepting this looser criterion is an increased probability of a false link. This is now 0.00004 which means that there are four chances in a hundred thousand that a record pair satisfying criterion 1b is a false link. This is small but, multiplied by the number of such links found, suggests that about 30 of them may have been errors. It should be stressed at this point that the method is not exact – in fact record linkage is not an exact science whatever method is used – and where it has been necessary to make approximations, these have erred on the side of caution. Therefore this figure is an upper limit on the number of false links rather than an exact count but it gives an indication of how the likely error rate increases as the link criterion is relaxed.

Stage 1c takes the procedure a stage further. The criterion here differs from that of stage 1b only in that a pair is still accepted even if the post codes are both present and are not the same, provided that the first name, last name, date of birth and gender agree and are unique to both files. This now identifies a further 357,453 pairs (7.1% of the Census records, bringing the total to 81.1%). However the false link probability is now 0.00059 or 59 chances in a hundred thousand in which case there might be up to 213 false links introduced in this stage. While still a tiny proportion of the total, it is inexorably increasing as the criterion is loosened. It should be noted that this stage of the procedure treats postcodes as merely 'the same' or 'different' and that if they are different then they are no more similar than if they were selected randomly from the

population of postcodes. However examination of these 357,453 pairs indicates that this is not so. For each of these the number of letters they have in common at the beginning was counted. For example a value of three would mean that the postcodes started with the same three letters but the fourth letter was different. The value was zero (meaning that they started with different letters) for just 13%, a much smaller figure than would be expected if the postcodes were independent of each other. Values of 2, 3, 4, 5 and 6 each accounted for around 15% of the pairs which are much higher numbers than would be expected if the record pairs all referred to different people. It suggests that while the postcodes were not exactly the same, they were often similar. This is a general limitation of deterministic methods of linkage. While they are computationally very efficient, they are not able to take into account degrees of similarity between postcodes but only whether they are the same or not. The same problem applies to names and this is why the later stages of the procedure use the computationally intensive but much more sensitive probabilistic methods.

4. Stage 2: the probabilistic stage

At this point all the easy links have been found and it is necessary to introduce specialist probabilistic linkage software. First however it is necessary to reduce the size of the files to be matched to removing the NHS Central Register -census pairs for which decisions have already been made. Doing so leaves 1,633,735 records in the NHS Central Register (NHSCR) file and 944,732 in the Census file. This does not just make the job easier for this software – without it, it would not be feasible at all. The size of the computational load is proportional to the product of the two files to be linked. The original files would have given a load of 28 units (a unit being a potential 10^{12} pairs of records to be matched). After removing the records featuring on the pairs identified in stage 1, this figure falls from 28 to 1.54, a reduction of nearly 95%. This reduction brought the task within the capability of the Link Plus package produced by the US Centre for Disease Control and Prevention but nevertheless the run still took 36 hours to complete.

The result was a set of 943,152 record pairs. Note that this number is only 1,580 less than the number of records in the census file which indicates that almost every census record was linked to something in the NHSCR file, though some of these would be very weak links which would not be accepted in any final set of matched record pairs. Scanning these record pairs indicated that those with a log likelihood ratio (the mathematical idea referred to at the start of this paper) in excess of 18[1] were, apart from a very occasional exception, convincing links and the decision was made to accept them without further manual scrutiny. This identified a further 604,403 pairs (another 12.1% of the census records, bringing the total to 93.2%). Calculating the match probabilities for these indicated that around 108 of them were false positives, bringing the estimated total number of false positives so far to 351 out of about 4.66 million.

Footnote

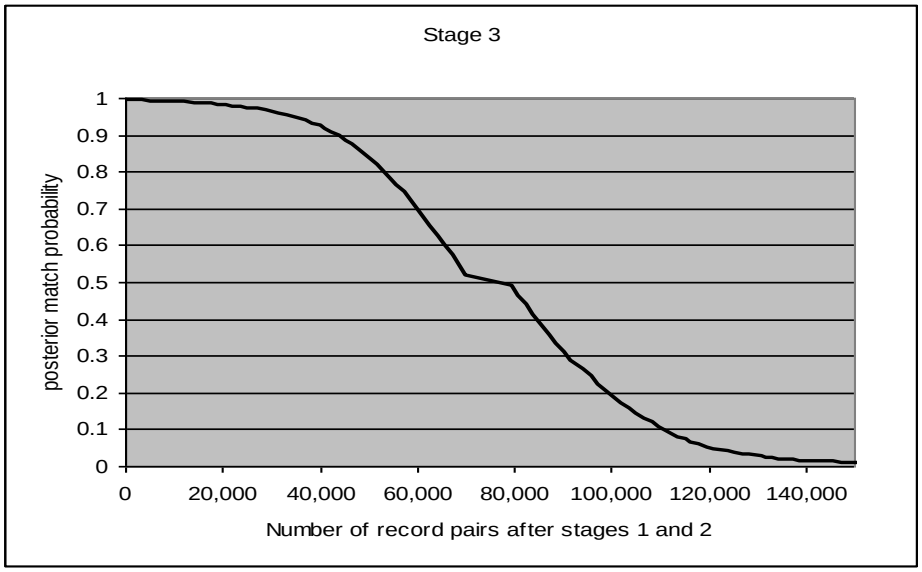
1) A log likelihood ratio of 18 corresponds to a match probability of 99.5% so this is the minimum value for record pairs accepted at this point.

The table below gives the number of record pairs accepted at each of the above stages, this number as a percentage of the number of Census records to be matched, and the estimated number of false links which might have been made.

Stage	number of links accepted	links as % of Census file	cumulative links accepted	cumulative links as % of Census file	estimated number of errors	cumulative estimated errors	Cumulative errors as % of all links
1a	2,922,199	58.43%	2,922,199	58.43%	0	0	0.000%
1b	777,204	15.54%	3,699,403	73.98%	30	30	0.001%
1c	357,453	7.15%	4,056,856	81.12%	213	243	0.006%
2	604,403	12.09%	4,661,259	93.21%	108	351	0.008%

5. Stage 3: the clerical review

The Link Plus package allows each NHSCR record to be linked to more than one Census record so such multiple links were removed by retaining only the pair with the higher match probability. This left 220,442 record pairs with a match probability of 99.5% or less (those with a probability above 99.5% having been "banked" at stage 2). These record pairs not "banked" after either the deterministic or probabilistic stages could then go to clerical review. The match probability can be calculated for each of these and used to sort the pairs into descending match order. The first 150,000 of these are plotted in the figure below.



The horizontal axis starts at zero but this excludes the 4.66 million pairs which have been 'banked' at earlier stages, so the actual values are from 4.66 million to 4.81 million. The probabilities plotted suggest that there are between 75 and 80 thousand true links amongst the 150 thousand pairs on the horizontal axis. These are approximately the last 1.5% of the links to be found, taking the total to around 94.7%. It is not desirable for this figure to rise to 100% as there were some people who returned a Census form but were not registered to a General Practitioner on Census Day. To link them to the extract would constitute a false positive. On the basis of the available evidence (which it must be admitted is not very rigorous), the ideal figure would be around 95% to 96%.

It is possible to estimate the person-hours which would be required to undertake this clerical review on the basis of a sample of 240 records which was independently reviewed by three reviewers. The result suggests that, assuming that a maximum

working regime would be four half-hour periods in any working day and that a working week is five days, the clerical review of all 220,000 record pairs remaining after stage 2 would require between two and three person-years.

6. Conclusion

6.1 Using efficient but simple deterministic linkage techniques, over 81% of the Census records can be linked to an NHSCR record with reasonable confidence.

6.2 Using more computationally intensive probabilistic linkage techniques, this percentage can be raised to over 93%.

6.3 To raise this still further closer to its probable maximum of 95-6% would require a clerical review of a scale which would involve a level of personnel expenditure which Demography Division could not at the moment sustain.

6.4 In general however the linkage exercise has been successful and has allowed the vast majority of people who returned a Census form and were on the NHSCR extract on Census Day to be identified.