Scoping Paper SP5: Some issues in linking large data sets (Harvey Goldstein)

There is a relatively large literature in recent years about how to carry out and interpret linked datasets. One of the features of linkage practice, however, is how much of this literature is effectively ignored. The methodological part of this literature has been concerned with developing algorithms that overcome some of the deficiencies of traditional methods. Briefly, these deficiencies arise from errors in the identifiers to link files, such as name, date of birth, residence, identification codes such as social security or NHS number. The key issue is that a failure to match records, or matching the wrong records is not random but depends on individual characteristics. For example ethnic minorities often have higher non-match rates because of a greater propensity for errors to occur in spelling of names etc. It has been shown that this can sometimes lead to considerable biases in subsequent analyses. Nevertheless, in the UK, for example, major linkage organisations such as the Health and Social Care Information Centre (HSCIC) responsible for linking health service datasets, are still wedded to traditional methods.

One of the issues here is that it is typically difficult or impossible to obtain information about the linkage procedures used, whether the linkage has been carried out by public bodies such as HSCIC or outsourced to commercial organisations, in the latter case partly for reasons of commercial competitive secrecy. Thus the quality and hence utility of linked data will be unknown with obvious implications for inference. Thus, one issue that could be taken up is that of transparency so that the users of linked data can have full access to the methods used and an assessment of quality.

The availability of very large linked datasets provides great opportunities for important social and medical insights. Yet this availability carries dangers if the limitations are not understood, and especially when standard statistical procedures are either ignored or misunderstood. Just because data are big does not imply that the statistical uncertainties associated with parameter estimates can be ignored. We already see this problem when estimates for school or hospital performance are quoted in the media.

Finally, let me say something about modern methods of linkage.

In terms of methodology, probabilistic record linkage methods have been developed over recent decades to address the general problem of linkage failure that occurs due to the presence of errors in the identification variables used to carry out the linkage. They have been stimulated by the realisation that such errors can lead to biases and by the need to handle very large data sets of varying quality.

To illustrate how probabilistic linkage operates consider the basic case where data are assumed to consist of two files, a primary file A and a secondary file B, whose individual records are to be matched using linking variables (or fields) such as name, address, identification number etc. A special case of data linkage is deduplication, where records belonging to the same individual within the same file are linked over time. An example of such deduplication is the HESID algorithm, which attempts to assign the same HESID to hospitalisation records belonging to the same patient over time, within Hospital Episode Statistics. Subsequently the resulting ‘longitudinal’ file may be linked to other data such as GP records.

In the basic case suppose A is a research database and B an administrative database. In the simplest case it is assumed that each record in A truly matches no more than one record in B. If B contains multiple records for single individuals it is assumed that these will have been merged together, as in HES, Typically, the matching process first of all selects those records where there is perfect agreement on all linking variables – so called deterministic matching. Variants, such as used in HES allow alternative combinations of matching variables, but are still deterministic. Following these matches this leaves records in each file where matching is uncertain and it is these records that are the focus of probabilistic linkage algorithms. Thus, for each primary data file record that is not unequivocally linked (on all matching variables) there will be in general several associated secondary data file records, that is, those that agree on at least one of the matching variables. We may refer to these as 'candidate' variables, where there is an implicit assumption that the reason for the lack of a perfect match arises from an error associated with the linking variable values. For each of such primary data file records there will be a given pattern of matching variable agreement values, say g. For example, for three binary matching variables we may observe a pattern, g = {1, 0, 1} indicating {match, no match, match}. For each pattern we wish to compute the probability of observing that pattern of values:

A) Given that it is the correct link: P(g|M)

B) Given that it is not the correct link: P(g|NM)

Probabilistic record linkage procedures compute R=P(g|M)/P(g|NM) and a weight W=$\log_2$(R), so that for primary data file record i and a given 'candidate' record j we obtain the weight w_{ij}. Initial estimates of P(g|M), P(g|NM) come from known record matches or other datasets and these are updated as more matches and non-matches are allocated in an iterative procedure. In practice these weights are determined separately for each matching variable and averaged, essentially assuming that the probabilities associated with the matching variables are independent. The size of the contribution to the weight from each matching variable depends on the discriminatory power of the variable, so that agreement on NHS number makes a larger contribution than agreement on sex. If the dataset is large it may be more efficient to divide the individuals into mutually exclusive blocks (e.g. age groups) and only consider matches within corresponding blocks. P(g|M) and P(g|NM) may also be allowed to vary between the blocks.

Existing probabilistic record linkage methods propose a cut-off threshold for W, so that any match with a weight above this threshold is accepted as correct. This threshold is typically chosen to minimise the percentage of 'false positives'. Where several exceed the threshold, the one with the highest weight is chosen. If no candidate record reaches the threshold then no link is made. Thus, at the end of the process the linked file will have some records with missing variable values where links have not been made. This procedure is generally part of the software algorithm where the linker just has to specify the relevant threshold weight and does not necessarily involve manual review of 'equivocal' cases.

Variations on this procedure occur when the linking is one-to-many or many-to-many. For example, we may wish to link a birth record to several admission episodes for an individual within a single hospital secondary data file. In such a case we could proceed by first linking the episodes in the secondary data file (de-duplication) so that each individual is represented by a single (longitudinal) record and then linking these records to those in the primary data file. We may also have a many-to-many case where, for example, multiple, unmatched educational events such as test scores for individuals are to be linked to a set of unmatched health records. Again, we might proceed by 'de-duplication' of data within the educational and within the health files and then linking across.

There are certain problems with this basic procedure which is the focus of current research. The first is the assumption of independence for the probabilities associated with the individual matching variables. For example, observing an individual in any given ethnic group category may be associated with certain surname structures and hence the joint probability will not simply be the product of the separate probabilities. A second typical problem is that primary data file records that cannot be matched above a weight threshold are excluded from data analysis, reducing efficiency and introducing bias if this is associated with the characteristics of the variables to be analysed. A third problem occurs when the errors in one or more matching variables are associated with the characteristics of the variables to be analysed. This non-random linkage error can lead to biases in the estimates from subsequent analyses.

Both the second and third problems apply to deterministic matching where biases will occur whenever errors and failures to match are non-random. Probabilistic procedures in general will reduce such biases (because linkage variable errors are non-random in general) and increase efficiency, and this has been the main motivation for their development. In addition there is the recognition that with deterministic linkage the quality of the linkage is effectively determined by the file having the lowest accuracy.

A forthcoming volume looks at these developments with contributions from leading researchers (Harron et al., 2015). It includes a discussion of the research based on Bayesian models being conducted by the Record Linkage Methodology Group at UCL. This work directly addresses the bias problem by quantifying and recording the linkage uncertainties, without the need for manual review, and carrying these through to the data analysis stage where methods for dealing with them can be employed. A full description can be found in Goldstein et al (2012).

Harvey Goldstein

September 2015

Reference

Harron, K., Goldstein, H. and Dibben, C. (2015). *Statistical models for data linkage: methodological developments.* Chichester, Wiley.

Goldstein, H., Harron, K., and Wade, A. (2012). The analysis of record linked data using multiple imputation with data value priors. Statistics in medicine, 31. DOI: 10.1002/sim.5508