Record linkage: a missing data problem

Harvey Goldstein[1,2] and Katie Harron[1]

[1] Institute of Child Health, University College London. [2] Graduate School of Education, University of Bristol

Introduction

The focus of this chapter is on record linkage as a tool for providing information on variables required for a statistical analysis. It is assumed that the data analyst has a primary data file, for example derived from a survey, clinical trial, registry or administrative database. For purposes of statistical modelling there are variables of interest (VOI) that are not recorded in the primary file, but reside in one or more secondary or linking data files to which the analyst has access. It is assumed that the information in the secondary files applies to the same individuals, or a well-defined subset, of those in the primary file. The aim is to transfer the values of the VOI from the individuals in the secondary file to the corresponding individuals in the primary file, and that this will be carried out by 'linking' each primary file individual record to the 'correct' VOI value in the secondary file.

This record linkage process is viewed as a means to an end; namely to carry out the appropriate statistical analysis. Thus, for example, the process of linkage can sustain ambiguity about whether a primary file record (PFR) actually links to the 'correct' secondary file record (SFR), for example by retaining the possibility of several links, so long as appropriate information about the associated variable values can be transmitted. This perspective departs from the traditional record linkage one where the emphasis is upon identifying the correct linking record. Where this cannot be ascertained unambiguously, a choice of a single record is nevertheless made and most of the theoretical apparatus supporting what is known as 'probabilistic record linkage' (PRL), as discussed elsewhere in this volume, is devoted to optimising such a choice. Naturally, in some cases, for example where there are administrative concerns, the emphasis may lie largely with the correct linking rather than statistical analysis, in which case techniques such as PRL will still be appropriate.

We develop our approach for the case where there is a single primary and a single secondary file, and then discuss extensions. Figure 1 shows a simple data structure that explains the essence of our approach.

In Figure 1, for purposes of analysis we would like to change all the 0s to Xs. If we restrict ourselves to the set B variables then, with suitable assumptions, we can carry out a statistical analysis where the 0s become converted to Xs via a suitable 'missing data' algorithm such as multiple imputation, the details of which we will elaborate later. For the set A variables this is not possible since every record has all values missing. If, however, we are able straightforwardly to link some, preferably most, of the individual records from the primary data file and the secondary data file unequivocally, using what are often referred to as deterministic linkage methods, then we will have a structure such as that in Figure 2 where records 2 and 3 have supplied the data for the set A variables.

We now see that this is simply another missing data structure and we can apply our existing missing data algorithms to obtain consistent estimates for our statistical models. In fact, in many cases as we will show, such an analysis will be adequate, without invoking any further linkage complexity. In the following sections we shall briefly refer to existing probabilistic record linkage (PRL) methods, show how these can be elaborated to improve performance using Prior Informed Imputation (PII), and discuss further extensions.

Figure 1. Primary data file with four records where the set B variables are recorded and the set A variables are located in a secondary data file. X represents a recorded variable value and 0 a missing value. Initially all the set A variables are missing and also some set B variables are missing as shown.

Record	Set A variables		Set B variables		
1	0	0	X	X	X
2	0	0	X	0	X
3	0	0	X	X	0
4	0	0	0	0	X

Figure 2. Primary data file with four records where the primary record file set B variables are recorded and the set A variable values for records 2 and 3 have been correctly transferred, unequivocally, via deterministic linkage with a secondary data file. X represents a variable with known value and 0 a missing value.

Record	Set A variables		Set B variables		
1	0	0	X	X	X
2	X	X	X	0	X
3	0	X	X	X	0
4	0	0	0	0	X

Probabilistic Record Linkage

The essence of PRL, (see Chapter 2) is as follows.

Identifying or matching variables (MV) on the individuals in both the primary data file (primary data file) and secondary data file (secondary data file) are used to 'link' each individual in the primary data file to the same individual in the secondary data file. In some cases, hopefully the majority, the link is unequivocal and we can carry across the relevant variable values from the secondary data file to the primary data file. In other cases there is some uncertainty about the link, due to missing or incorrect data in the MV. In these cases, a set of weights are estimated, based on the number of matches and discriminatory power of the matching variables. We note at this point that even where there is a 'perfect match' with agreement on all matching variables, such as age, name, date of birth etc., we may still encounter linkage problems. This might occur because the set of matching variables is insufficiently discriminatory so that several secondary data file records match a primary data file record or because there may be errors in both files that produce incorrect but coincidentally identical values. The second possibility is likely to be very rare and is typically ignored. In the first case with traditional PRL methods a choice will have to be made to select one of the chosen records to link, possibly at random and this implies that in some cases the wrong one will be chosen, so that the procedure itself is biased. As we show below, however, this can often be avoided or mitigated using PII.

Fundamental to PRL is the ascertainment of match weights. 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 refer to these as 'candidate' variables. For each of these primary data file records there will be a given pattern of MV agreement values (g). For example, for three binary matching variables we may observe a pattern g={1, 0, 1} indicating {match, no match, match}.[1] For each pattern we compute the probability of observing that pattern of MV values:

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

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

The traditional record linkage procedure then computes R=P(g|M)/P(g|NM) and a weight W=$\log_2$(R), so that for primary data file record i and a 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: estimating the joint probabilities is typically regarded as too complicated. 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 be allowed to vary between the blocks (e.g. age group (Newcombe 1995)).

The PRL methods propose a cut-off threshold for W, so that matches with W above this threshold are accepted as true matches. 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

[1] We may also encounter missing matching variable values. In this case the PDF record match status will always be equivocal and matching will take place using just the remainder. This assumes missingness is at random

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. We could then apply standard multiple imputation (MI) as described above, although this appears to be very rare in applications.

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 file. In such a case we could proceed by first linking the episodes in the secondary data file 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 PRL. 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. We shall return to this issue later, and propose a way of dealing with it. The second problem is that typically primary data file records that cannot be matched above a weight threshold are excluded from data analysis, reducing efficiency, although as suggested above this strictly is not necessary. The third problem occurs when the errors in one or more matching variables are associated with the values of the secondary data file variables to be transferred for analysis. Thus, Lariscy has shown that certain ethnic group surnames are more error-prone than others and this can lead to biased inferences for associated variables such as mortality rates (Lariscy, 2011).

Multiple Imputation (MI)

The use of MI for handling missing data in either response or predictor variables in statistical models is a common technique, and because it forms the basis for the PII method that we shall describe, we now outline how it operates. Further details can be found, for example in Carpenter and Kenward (2012).

Multiple imputation is used to replace missing values with a set of imputed values in the model of interest (MOI), for example a regression model. For each missing value, the posterior distribution of the value is computed, conditionally on the other variables in the MOI and any auxiliary variables, and a random draw taken from this distribution. Auxiliary variables are those that are associated with the responses in the imputation model but do not appear in the MOI. It is assumed that each missing value is missing at random, that is randomly missing given the remaining variables in the MOI and any auxiliary variables used. The standard application assumes that all the variables in the MOI have a joint normal distribution. In practice this involves setting up a model where the responses are variables with any missing data and predictors include other variables in the MOI and any auxiliary variables with fully known values. Generally, however, we cannot assume multivariate normality, for example where some variables are categorical. Goldstein et al (2009) propose a procedure that we refer to by the authors' initials GCKL, and which provides a means of dealing with this by transforming all variables to multivariate normality, carrying out the imputation and then transforming back to corresponding non-normal variable scales (Goldstein et al., 2009). A description of such a 'latent normal model' is given in Appendix A.

In practice an MCMC algorithm is used to sample a value from the conditional posterior distribution, for each value missing so that after every cycle of the algorithm we effectively have a complete data set consisting of a mixture of imputed and known data values. The algorithm is used to generate a number, *n*, of such complete data sets that are, as far as possible, independent by choosing MCMC cycles sufficiently far apart; in practice a distance apart of 250-500 will often be satisfactory. The value of *n* should be large enough to ensure the accuracy of the final estimates, which are obtained by fitting the MOI to each completed dataset and averaging over these according to the so called 'Rubin's rules'. A value of 5 is often used although between 10 and 20 completed datasets has been found by GCKL to be needed for multilevel data. Where the MOI is multilevel the multilevel structure should also be used in the imputation model. We shall refer to this procedure as standard MI. More recently Goldstein et al, extend this so that quite general models can be fitted within a single MCMC chain, rather than producing multiple datasets for analysis (Goldstein et al., 2014). In what follows we shall assume standard MI, although this more recent procedure will generally be faster as well as allowing more general models to be fitted.

Prior informed Imputation

In PRL the acceptance of a record as linked is determined by the weights *W* that depend on P(g|M) and P(g|NM). In PII we work instead with the probability of a match given the matching variable pattern, P(M|g), For primary data file record i and a candidate record *j* denote this by $\pi_{m,ij}(g)$. For primary data file record i, by default, we scale these probabilities to add to 1.0 and these scaled values are denoted by $\pi_{ij}(g)$. This assumes that the 'true' matching record belongs to this set of candidate records. We discuss the case when this may not be true below. In practice, to avoid very large files where many of the probabilities are very small a lower threshold can be chosen so that records with probabilities less than this are ignored.

For the set of variables, set A in Figure 1, to be transferred from the secondary data file, denote their distribution, conditional on the set B variables, by $f(Y^{A|B})$. This conditioning is derived from the joint distribution of the responses and any covariates in the MOI, as well as any auxiliary variables that may be used to satisfy the MAR assumption. Initial values will be derived from the unequivocal records that are transferred and the conditional distribution is updated at each cycle of the algorithm. This joint distribution is the latent normal distribution described above, that is with suitable transformations as necessary for non-normal variables. We now have a structure such as that in Figure 2, and we use an MCMC chain to impute the missing values. For each primary data file record i with no unequivocal match, at each iteration we carry out the following procedure.

We compute a modified prior probability which is the likelihood component, $f(Y^{A|B})$ multiplied by the prior, $\pi_{ij}(g)$, for associated (equivocal) secondary data file record *j*, namely $\pi_{ij} \propto f\left(y_{ij}^{A|B}\right) p_{ij}$. These are scaled to add to 1.0 and comprise the modified probability distribution (MPD) for each primary data file record, denoted by π_{ij} .

We first note that we should not simply sample records at random according to the MPD since this will result in some incorrect choices of the true record in a similar way to standard probabilistic linkage. Instead we propose that, as in standard probabilistic linkage, that a threshold is set for accepting a record as a true link. If no record exceeds the threshold then the data values are regarded as missing and standard MI is used. It is also possible that for any given record at some iterations we may find a record exceeding the threshold, whereas at other iterations standard MI is used. In general we would choose only high thresholds to minimise bias. For example, if a value of 0.95 is chosen, this implies that the ratio of the highest to next highest probability is at least 0.95/0.05$=20$. Given a high enough threshold, the proposed procedure will produce approximately unbiased estimates. It has the advantage of efficiency over PL in that that all records in the primary data file contribute to the estimation. Furthermore, conditioning on the values of further variables, including the matching variables as auxiliaries when computing the $f(Y^{A|B})$ can be expected to make the MAR assumption more reasonable. Incorporating the likelihood component in the MPD can be expected to lead more often to the correct record being the one to exceed a given threshold.

So far we have assumed that the true matching record is located within the secondary data file file. In some cases, however, this may not be the case. For example, if we wish to match a patient file to a death register in order to record survival status in the primary data file, some equivocal records might either indicate that the patient is still alive or that they are dead but not equivocally matched. Assume we know, or have a good estimate, of the mortality rate among our patients, say π_d . If a proportion of the primary data file $\pi_m < \pi_d$ are unequivocally matched then the probability that a randomly chosen remaining record in the LDF is not a death from the FOI sample is $\pi_r = 1-(\pi_d - \pi_m)$. We therefore multiply the $\pi_{ij}(g)$ by $1-\pi_r$ and add an extra pseudo-record with probability π_r with an associated code for a surviving patient. If a good estimate of the mortality rate is not available then a sensitivity analysis might be carried out for a range of plausible values.

We have assumed that record linkage is between two files. In practice, however, there may be several files to be linked. Without loss of generality we can assume that one of these is the main primary data file with several secondary data file files. One way to proceed is conditionally. Thus for each iteration of the algorithm we first carry out a PII for the primary data file and the first secondary data file, say $\text{secondary data file}_1$. Then, conditioning on the original set A variables and those carried over from $\text{secondary data file}_1$ we carry out a PII for the augmented primary data file and $\text{secondary data file}_2$ and so on. We assume that matching errors across linkages are independent. Alternatively, we can think of forming the joint prior distribution and hence a joint MPD over the complete set of secondary data file files, but this may become impractical when the number of secondary data file files is moderately large. In some cases we may have sets of matching variables that are common only to a subset of LDFs. Thus, we may wish to match patient records within the same hospital on different occasions, using a local hospital identifier, but which is not available for the main primary data file. In this case we would first carry out a PII choosing one of the hospital secondary data file files as the primary data file and then carry out a PII where the combined records are matched to the main file of interest. If there are matching variables common to all three files then the linkage of the linked hospital records to the primary data file will need to consider the matching probabilities associated with the three possible combinations of values across files for each matching variable.

Estimating matching probabilities

In general the prior probabilities $\pi_{ij}(g)$, are unknown. Given the overall probabilities of a true match occurring and a given observed pattern, these are proportional to the (inverse) probabilities

P(g|M) described above. Thus we could use these, as derived from a PRL analysis, suitably scaled. Goldstein et al. suggested that the weights w_{ij}, could be scaled to add to 1.0 (Goldstein et al., 2012). This uses information about the probability of a non-match and is equivalent to working with the scaled logit($\pi_{ij}(g)$) values, although in practice this may not make very much difference.

One of the difficulties with the use of PRL estimates is that they will be based upon the assumption of independence. An alternative approach, that avoids this, is to attempt to estimate the joint prior probabilities directly. This would involve empirical evidence about ways in which linking variable errors occur, for example based upon experimental situations. If a 'gold standard' data set is available where the correct matching status is known then the results of a trial involving coding and recording the records in the files would provide such data. We are currently pursuing research into this.

Example 1: Linking electronic healthcare data to estimate trends in blood-stream infection

Linkage of records between electronic health databases is becoming increasingly important for research purposes as individual-level electronic information can be combined relatively quickly and inexpensively (Jutte et al., 2011, Black, 2003). For example, linkage of national bloodstream infection (BSI) surveillance data with national audit data on admissions to paediatric intensive care units (PICU) could be used to provide enhanced monitoring of BSI rates in PICU. However, linkage errors occurring due to a lack of complete unique identifiers in these data sources could lead to biased outcomes measures. Therefore it is vital that the impact of linkage error on results is minimised. The following section describes linkage between data from the Paediatric Intensive Care Audit Network (PICANet) and microbiology data collected from hospital laboratories. Results from PRL and PII are compared.

Methods

Data

Admission data for children admitted to two PICUs between March 2003 and December 2010 were extracted from the PICANet database (Universities of Leeds and Leicester, 2012). Microbiology records for all positive bacterial isolates from blood were obtained from the two hospital laboratories covering the same time period. Deterministic linkage of PICANet and microbiology records was conducted based on unique identifiers (National Health Service (NHS) number, hospital number, name, date of birth and sex). The deterministic linkage was manually verified to ensure there were no false-matches or missed-matches and additional data from the hospital IT systems (e.g. examination of previous names) were used to clarify any uncertainties. This process provided a "gold-standard" dataset where the true match status of each record pair was known.

Although high quality data were available in the microbiology data obtained directly from the hospital laboratories in this study, identifiers available in the national infection surveillance data collected from hospital laboratories across England and Wales is of poorer quality. In the national system, coordinated by Public Health England, matching variables available are often restricted to date of birth, sex, and Soundex code (an anonymised phonetic code using for reporting surname to the surveillance system). To evaluate PRL and PII for linkage of such data, unique identifiers were removed from the microbiology data. Linkage using PRL and PII was then repeated using date of birth, sex and Soundex only.

VOIs

The purpose of the linkage was to transfer information on whether an admission was associated with a PICU-acquired BSI, defined as any positive blood culture occurring between two days after PICU admission and up to two days following PICU discharge. The VOI was therefore a binary variable representing PICU-acquired BSI. The crude rate of PICU-acquired BSI was calculated as the number of events per 1000 bed-days. Poisson regression models were fitted to the data to estimate the incidence-rate ratio for PICU-acquired BSI at PICU 1 compared with PICU2. Variables known to be associated with PICU-acquired BSI in these datasets were included in models (renal status, quarter-year at admission, age, admission type and admission source). Since the match status of each record pair was known in the gold-standard data, match probabilities and match weights were calculated directly.

Linkage methods

For PRL, thresholds for classifying record pairs as links or non-links are typically chosen by manual inspection of record pairs ordered by match weight and examination of the distribution of weights. This manual review is not feasible for national data, firstly due to the large numbers of records, and secondly due to the scarcity of identifying information on records. Alternatively, a single threshold that minimises the effect of error can be chosen, based on known error rates. As error rates are usually unknown, it might be possible to obtain estimates from a subset of "gold-standard" data where the true match status is known. In this example, a random 10% subset of data where the true match status was known was used to select a threshold that minimised the sum of linkage errors (false-matches + missed matches). An alternative would be to choose the threshold that minimised the net error (|false-matches – missed matches|).

For PII, records that had a match probability >0.9 were classified as unequivocal. For these records, a VOI value of 1 was accepted. For the remaining records, a MPD was derived from the likelihood in the unequivocally linked records and the prior distribution of match probabilities in the candidate linking records. The MPD threshold was set to 0.9, so that VOI values would be accepted if they exceeded this probability and standard MI would be performed if no value exceeded this threshold.

Results

In the gold-standard data, 1496 (7.1%) of the 20924 admission records extracted from PICANet linked to at least one microbiology record of PICU-acquired BSI. Based on results from data linked using date of birth, sex and Soundex only, the number of PICU-acquired BSI was estimated as 1316 (6.3%) using PRL and 1457 (7.0%) using PII.

The rate of PICU-acquired BSI was 12.88 (95% CI 12.23-13.53) per 1000 bed-days in the gold-standard data, 11.33 (95% CI 10.72-11.95) using PRL and 12.75 (95% CI 11.61-12.89) using PII.

The incidence rate ratio for PICU-acquired BSI at the first hospital compared with PICU-acquired BSI at the second hospital was 1.31 (95% CI 1.18-1.46) in the gold-standard data, 1.25 (95% CI 1.12-1.41) using PRL and 1.27 (95% CI 1.14-1.43) using PII.

Figure 3 compares the rate of PICU-acquired BSI estimated using PRL and PII.

Conclusions

Even in these relatively good-quality data, bias was introduced to results due to linkage of imperfect identifiers. For all outcome measures evaluated, PII provided less biased results than PRL.

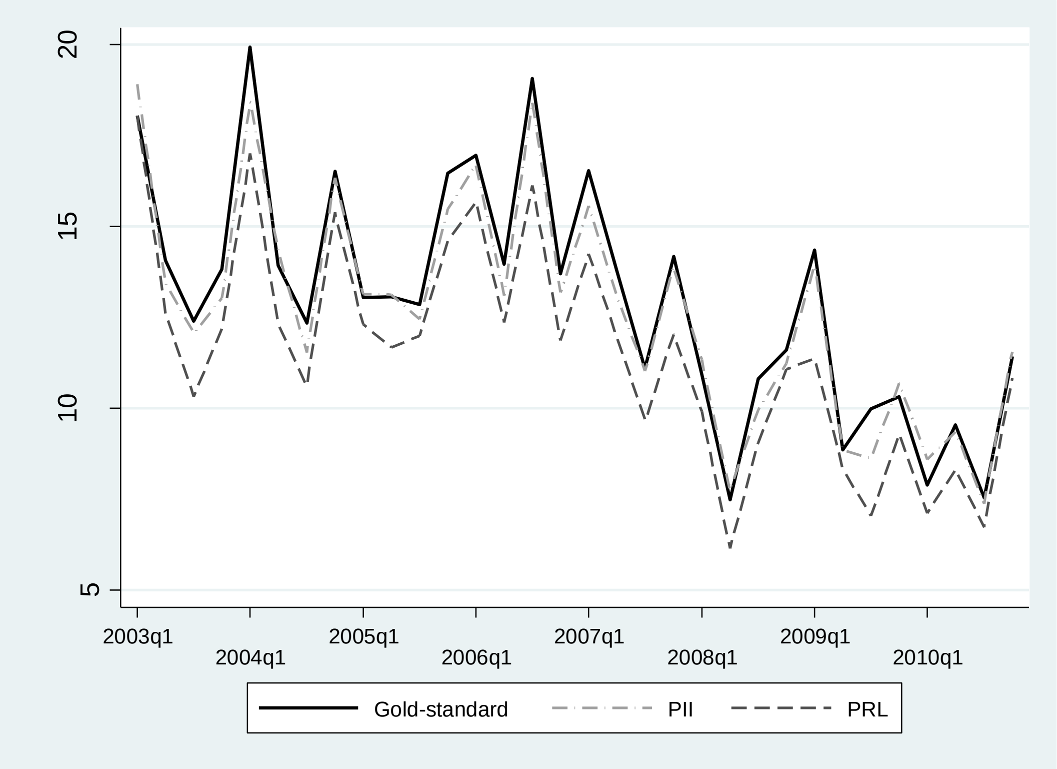

Figure 3: Comparison of prior-informed imputation (PII) and probabilistic record linkage (PRL) with gold-standard data: Rates of PICU-acquired BSI.

Example 2: Simulated data including non-random linkage error

Example 1 demonstrates that PII can reduce some of the bias resulting from linkage errors using the more traditional PRL approach to linkage. However, the impact of linkage error can vary according to the expected match rate (the number of primary file records expected to have a match to the secondary file). This is because for a given rate of linkage error (false-match rate or missed-match rate), the absolute number of false-matches or missed-matches is directly related to the number of true matches and non-matches.

The impact of linkage error also depends on the distribution of error. Linkage error can have a large impact on results when the probability of linkage is associated with particular variables or subgroups of data (see Chapter 4). For example, relative outcome measures can be affected by differences in data quality between subgroups (e.g. by hospital or by ethnic-group); outcome measures may be underestimated if the outcome is difficult to capture in linked data (e.g. under-representation of vulnerable groups due to missing identifiers (Lariscy, 2011, Ford et al., 2006)).

This example demonstrates the performance of PII in different linkage situations through the use of simulated data.

Methods

Simulated data

The primary file was generated by randomly sampling 10,000 values for Soundex, day of birth, month of birth, year of birth and sex from an extract of 112,890 PICANet records (for 2003-2010). Predictor variables were created by randomly sampling values for renal status, admission type (planned or unplanned), admission source (same or other hospital), length of stay, age in months, quarter-year at admission and Unit from the PICANet file. Predictor values were sampled jointly, to preserve the association between variables. Identifier values were then brought together with predictor values to create 10,000 complete simulated records in the primary file (Figure 4).

The linking file was created by selecting a number of records from the primary file to represent admissions that had experienced a PICU-acquired BSI. Associations between the predictor variables and the VOI (PICU-acquired BSI) were induced by selecting records according to the values of their predictor variables. Predictor variables were then removed from the linking file. Additional identifier and predictor values were sampled from PICANet and used to create "non-match" records which were added to the linking file so that in total, the linking file contained 10,000 records, some of which had a match in the primary file, and some of which did not (Figure 4).

To represent data quality found in administrative data, identifier errors and missing values were randomly introduced into each identifier in the linking file at a rate of 5%. The process was repeated to produce a set of 25 simulated linking files.

To explore the performance of PII with different match rates, the proportion of true-matches was set to 10%, 50% or 70% by selecting the corresponding number of records from the primary file to create each linking file.

To explore the performance of PII for handling non-randomly distributed error, the way in which error was introduced into simulated datasets was varied. Firstly, non-random error was introduced into the simulated datasets according to Unit: data from Unit 1 were set to be five times more likely to include error than data from Unit 2. Secondly, error was introduced according to the outcome: linking file records for children with a PICU-acquired BSI were set to be five times more likely to have error than records for children with no infection.

VOIs

The outcome of interest was the rate of PICU-acquired BSI and a secondary outcome was the absolute difference in adjusted rates between hospitals. The VOI was therefore the presence of PICU-acquired BSI, represented as a binary variable. Statistical analysis was performed as described in Example 1.

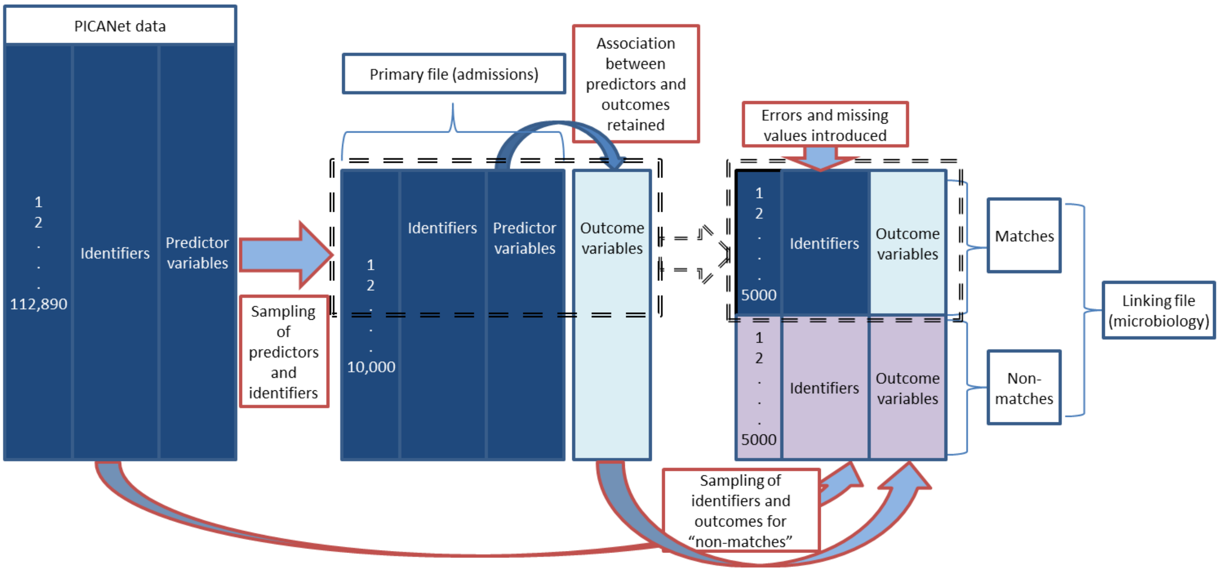
PICANet data
1
2
.
.
.
112,890
Identifiers
Predictor variables
Sampling of predictors and identifiers
Primary file (admissions)
1
2
.
.
.
10,000
Identifiers
Predictor variables
Association between predictors and outcomes retained
Outcome variables
Errors and missing values introduced
1
2
.
.
5000
Identifiers
Outcome variables
1
2
.
.
5000
Identifiers
Outcome variables
Matches
Non-matches
Linking file (microbiology)
Sampling of identifiers and outcomes for "non-matches"

Figure 4: Generation of simulated linking files (for 50% match rate example)

Results

Overall, most bias was introduced into results when linkage error was distributed non-randomly, i.e. associated with either the hospital or with having the outcome (PICU-acquired BSI). Estimates of the rate of PICU-acquired BSI were most biased when error was associated with the outcome of infection and PRL provided particularly biased results in this situation (Figures 5-7).

The amount of biased introduced also differed according to the underlying match rate. Using PII rather than PRL had the most dramatic benefit for high match rates.

Estimates of the difference in adjusted rates were not substantially affected by linkage error when the error was distributed randomly (Figure 8), as error was introduced to each hospital equally and the relative outcome was not affected. However estimates of the difference in rates were substantially biased by non-random error associated with the hospital, as errors in the data from one hospital led to apparent lower rates and therefore falsely inflated the difference between units (Figure 8).

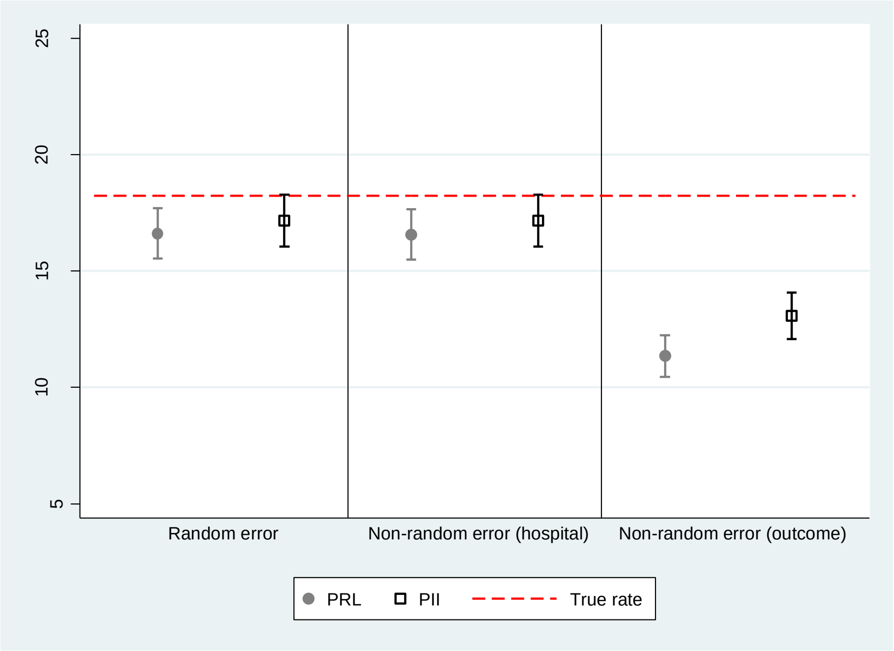

Figure 5: Comparison of prior-informed imputation (PII) and probabilistic record linkage (PRL) with simulated data and 10% match rate: Rates of PICU-acquired BSI

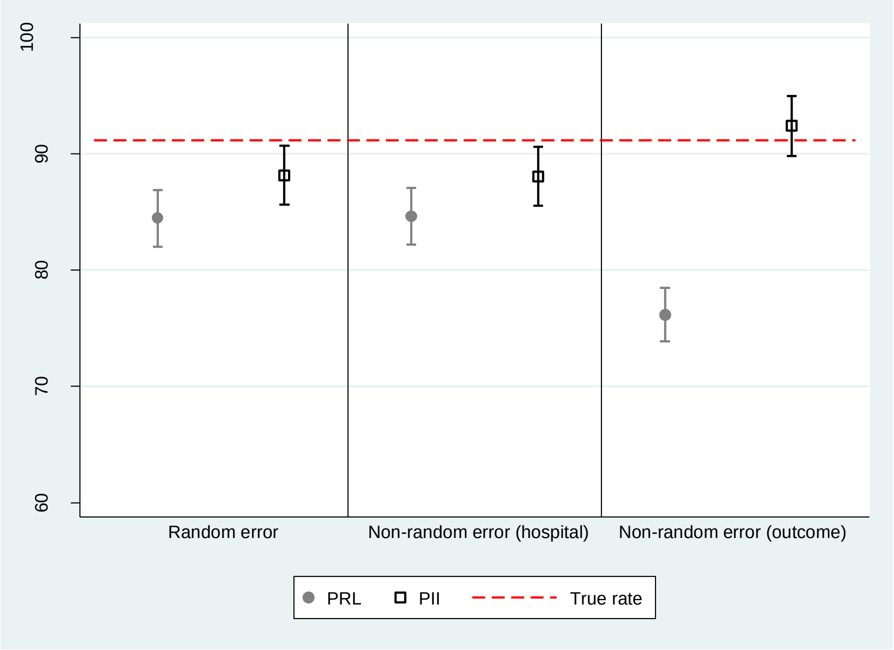

Figure 6: Comparison of prior-informed imputation (PII) and probabilistic record linkage (PRL) with simulated data and 50% match rate: Rates of PICU-acquired BSI

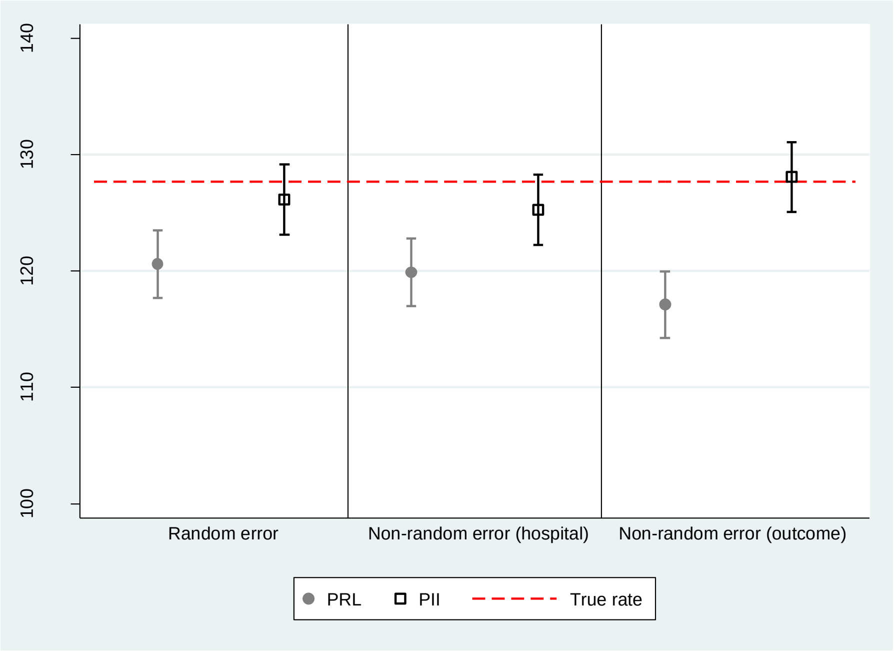

Figure 7: Comparison of prior-informed imputation (PII) and probabilistic record linkage (PRL) with simulated data and 70% match rate: Rates of PICU-acquired BSI

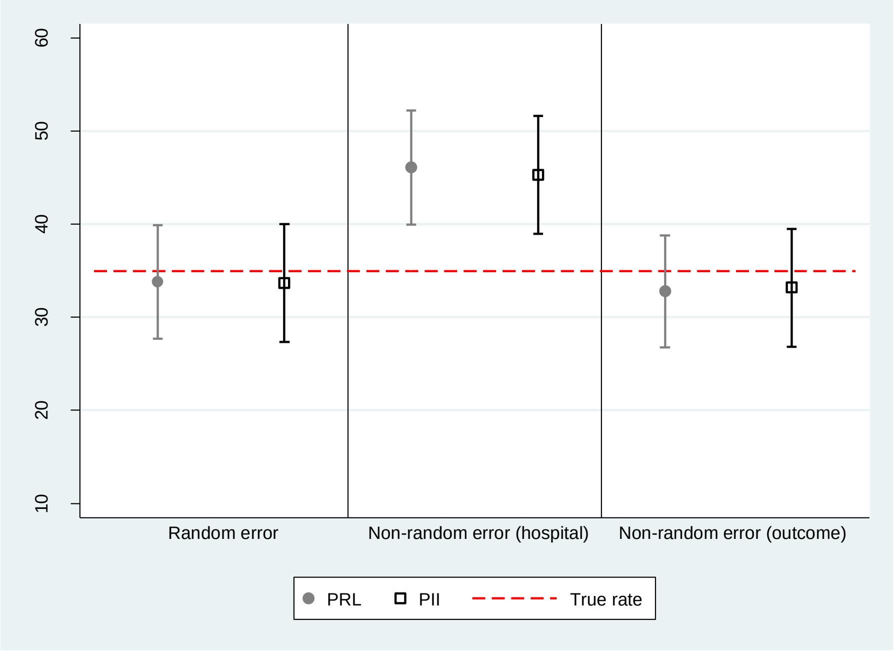

Figure 8: Comparison of prior-informed imputation (PII) and probabilistic record linkage (PRL) with simulated data and 70% match rate: Difference in adjusted rates of PICU-acquired BSI by hospital

Discussion

Existing methods that aim to adjust for linkage bias are generally limited to the context of regression analysis and rely on a number of assumptions(Chambers et al., 2009, Kim and Chambers, 2011, Scheuren and Winkler, 1997, Hof and Zwinderman, 2012). This chapter describes how viewing record linkage as a missing data problem could help to handle linkage error and uncertainty within analysis. The motivation for using imputation methods for data linkage is that the ultimate purpose of linkage is not to combine *records*, but to combine information from records belonging to the same individual. Measuring the quality of linkage then shifts from quantifying match rate and linkage error, to obtaining correct estimates for the outcomes of interest.

Linkage of simulated datasets with different characteristics (i.e. with non-random error or with different match rates) illustrated that linkage error affects different types of linkage and analysis in complex ways.

Non-random linkage error

Simulations highlighted that when data quality differs by unit (identifier errors more likely to occur in one PICU than another), relative outcome measures (e.g. differences in adjusted rates) were more likely to be affected than absolute measures (e.g. incidence rate). This finding has important implications for analysis of linked data in the presence of linkage error. For example, if linkage error is constant over time, estimated trends should be unaffected, even if absolute rates are over- or under-estimated. However if the aim is to compare groups, and linkage error is not constant between these groups, relative outcome measures are likely to be biased, either over-exaggerating differences or biasing to the null. This type of non-random identifier error could easily occur if data quality differs between institutions or in different groups of records (see Chapter 4).

Detailed understanding of the underlying quality of data to be linked could help to identify which populations are most vulnerable to bias due to linkage error. Furthermore, evaluation of linkage quality, e.g. through comparing linked and unlinked records, is of utmost importance so that linkage strategies can be tailored for different groups of records (Bohensky et al., 2010).

Strengths and limitations: handling linkage error

PII was not able to completely avoid bias due to linkage error. The causes of bias when using PII are the same as those occurring when using PRL. However, combining information from both the candidate records and the unequivocally linked records means that more often, the correct VOI value should be accepted. This was confirmed by the higher levels of bias occurring with PRL.

PII with a high MPD threshold performs well. This implies that where there are a sufficient number of unequivocal links (e.g. those identified through deterministic linkage), standard MI may be sufficient. This would provide a simple alternative to PRL and avoid the need for calculating match weights or probabilities.

PII (and standard MI) provide a more efficient means for analysis compared with PRL, as all records are retained for analysis. A further advantage of PII over PRL is related to standard errors. For probabilistic linkage, standard errors are calculated assuming that there is no error and are therefore falsely small (But see Chapter 5). For PII, the process of combining several imputed datasets means that uncertainty occurring during linkage is properly reflected, resulting in slightly larger standard errors compared with probabilistic linkage.

PII also provides advantages over existing methods for linkage bias adjustment. Previously described methods have been limited to simulation studies and have not been adopted in practice, possibly due to their complex nature and a lack of practical guidance for users. Conversely, PII has been evaluated both in simulated data and for the linkage of two large national administrative data sources and this linkage has been described in practical terms in several papers. Furthermore, PII can be implemented using REALCOM code implemented through the Stat-JR software developed by the University of Bristol (Charlton et al., 2012).

Implications for data linkers and data users

In order for research based on linked data to be transparent, data linkers (including trusted third parties) need to be willing to provide details of the linkage processes used to create linked datasets. Data users need to know what to ask for, in terms of information required to evaluate the quality of linkage. This means that linked data provided should not be restricted to only the highest-weighted link, but that information from other candidate links should also be made available, to facilitate sensitivity analyses and prior-informed imputation. Furthermore, characteristics of unlinked data should also be provided, to allow comparisons with linked data and identification of potential sources of bias.

As the size of datasets to be linked increases, manual review will become infeasible, even where sufficient identifying information is available. In these cases, alternative linkage methods will become even more important. The practicalities of storing multiple candidate links and associated match weights or match probabilities to facilitate sensitivity analyses need to be further explored. Graph databases (as opposed to traditional relational databases), could provide a technical solution to this problem, by storing records and links in the form of edges and nodes. This area will be explored in Chapter 7.

References

Black N: Secondary use of personal data for health and health services research: why identifiable data are essential. *J Health Serv Res Policy* 2003, 8(Supplement 1):36-40.

Bohensky M, Jolley D, Sundararajan V, Evans S, Pilcher D, Scott I, Brand C: Data linkage: A powerful research tool with potential problems. *BMC Health Serv Res* 2010, 10(1):346-352.

Carpenter J, Kenward M: Multiple imputation and its application: John Wiley & Sons; 2012.

Chambers R, Chipperfield J, Davis W, Kovacevic M: Inference based on estimating equations and probability-linked data. In. Edited by Centre for Statistical & Survey Methodology Working Paper Series. University of Wollongong.; 2009: 38.

Charlton CMJ, Michaelides DT, Cameron B, Szmaragd C, Parker RMA, Yang H, Zhang Z, Browne WJ: Stat-JR software. In.: Center for Multilevel Modelling, University of Bristol and Electronics and Computer Science, University of Southampton; 2012.

Clark D. Practical introduction to record linkage for injury research. Injury Prevention. 2004;10(3):186

Ford JB, Roberts CL, Taylor LK: Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. *Paediatr Perinat Ep* 2006, 20(4):329-337.

Goldstein H, Carpenter J, Kenward MG, Levin KA: Multilevel models with multivariate mixed response types. *Stat Model* 2009, **9**(3):173-197.

Goldstein H, Carpenter J, Browne, W.: Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and nonlinear terms. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 2014.

Goldstein H, Harron K, Wade A: The analysis of record-linked data using multiple imputation with data value priors. *Stat Med* 2012, 31(28):3481-3493.

Goldstein H, Kounali D: Multilevel multivariate modelling of childhood growth, numbers of growth measurements and adult characteristics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009, 172(3):599-613.

Hof MHP, Zwinderman AH: Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Stat Med* 2012, 31(30):4231-4242.

Jutte DP, Roos L, Brownell MD: Administrative record linkage as a tool for public health research. *Annu Rev Public Health* 2011, 32:91-108.

Kim, G. and Chambers, R. (2012). Regression analysis under probabilistic multi-linkage. Statistica Neerlandica, **66**, 1, 64-69.

Lahiri, P. and Larsen, M.D. (2005). Regression analysis with linked data. Journal of the American statistical Association, 100, 222-230.

Lariscy JT: Differential record linkage by hispanic ethnicity and age in linked mortality studies. *J Aging Health* 2011, **23**(8):1263-1284.

McGlincy, M. H. (2002). A Bayesian record linkage methodology for mulitiple imputation of missing links. ASA section on survey research methods, 2004.

Neter J, Maynes E, Ramanathan R. The effect of mismatching on the measurement of response error. Journal of the American Statistical Association. 1965;60(312):1005-27

Newcombe HB: Age-related bias in probabilistic death searches due to neglect of the "prior likelihoods". *Comput Biomed Res* 1995, **28**(2):87-99.

Rubin, D. B. (1987). Multiple imputation for non-response in surveys. Chichester, Wiley.

Scheuren F, Winkler W. Regression Analysis of Data Files that Are Computer Matched - Part I. Survey Methodology. 1993;19(1):39-58

Scheuren F, Winkler W: Regression analysis of data files that are computer matched - Part II. *Surv Methodol* 1997, 23(2):126-138.

Universities of Leeds and Leicester: Paediatric Intensive Care Audit Network National Report 2009 - 2011. In.; 2012.

Appendix A

The latent normal model

For multivariate data with mixtures of response variable types, GCKL show how such a response distribution can be represented in terms of an underlying 'latent' multivariate normal distribution (Goldstein et al., 2009). For ordered categorical variables and for continuously distributed variables, each such variable corresponds to one normal variable on the latent scale. For an unordered categorical variable where just one category is observed to occur, with *p* categories we have *p-1* normal variables on the latent scale. They also show how this can be extended to the multilevel case. An MCMC algorithm is used which samples values from the latent normal distribution.

This representation can be used to fit a wide variety of multivariate generalised linear models with mixed response types, and, after sampling the latent normal variables, reduces to fitting a multivariate normal linear model. The following summary steps are those used to sample values from this underlying latent normal distribution given the original variable values. At each cycle of the MCMC algorithm a new set of values is selected. Each such sampling step conditions on the other, current, latent normal values.

Normal response

If the original response is normal this value is retained.

Ordered categorical data

If we have *p* ordered categories we have an associated set of *p-1* cut points, or threshold, parameters on the latent normal scale such that if category *k* is observed a sample value is drawn from the standard normal distribution interval defined by the $(-\infty, 1), \text{if } k = 1, \ (p\text{-}1, \ \infty) \text{ if } k = p, \ \text{otherwise by} \ (k-1, k)$. The threshold parameters are estimated in a further step. In the binary case this corresponds to a standard probit model.

Unordered categorical data

If we have *p* unordered categories then we sample from a standard *p-1* multivariate normal with zero covariances, as follows. The final category is typically chosen as the reference category. A random draw is taken from the multivariate normal and if the category corresponding to the maximum value in this draw is also the one which is observed then the values in that draw are accepted. If all the values in the draw are negative and the last category is the one observed then the draw is accepted. Otherwise a new draw is made.

The procedure can be extended, with certain limitations, to discrete distributions such as the Poisson (Goldstein and Kounali, 2009) and to non-normal continuous distributions for which a normalising transformation exists, such as the Box-Cox family (Goldstein et al., 2009).

After all of these sampling steps have been completed we obtain a multivariate normal distribution. Where there are missing data values we can therefore use standard imputation procedures to impute the missing values, on the normal scales, and use the inverse set of transformations to those given above, in order to provide a randomly imputed value on the original scales.