Research Article

Statistics in Medicine

Received 21 June 2011 Accepted 10 June 2012 Published online 17 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5508

# The analysis of record-linked data using multiple imputation with data value priors

Harvey Goldstein,[a,b*†] Katie Harron[a] and Angie Wade[a]

**Probabilistic record linkage techniques assign match weights to one or more potential matches for those individual records that cannot be assigned 'unequivocal matches' across data files. Existing methods select the single record having the maximum weight provided that this weight is higher than an assigned threshold. We argue that this procedure, which ignores all information from matches with lower weights and for some individuals assigns no match, is inefficient and may also lead to biases in subsequent analysis of the linked data. We propose that a multiple imputation framework be utilised for data that belong to records that cannot be matched unequivocally. In this way, the information from all potential matches is transferred through to the analysis stage. This procedure allows for the propagation of matching uncertainty through a full modelling process that preserves the data structure. For purposes of statistical modelling, results from a simulation example suggest that a full probabilistic record linkage is unnecessary and that standard multiple imputation will provide unbiased and efficient parameter estimates. Copyright © 2012 John Wiley & Sons, Ltd.**

**Keywords:** linking errors; missing data; multiple imputation; prior informed imputation; record linkage

## 1. Introduction

The linking of records from disparate sources is becoming increasingly important with the advent of large, typically administrative databases and the facility to link individual records across these. Record linkage can be an efficient and cost-effective way of combining datasets to increase the amount of information available to a researcher with the goal often being to carry over additional variables of interest (VOI) to a primary file—for example, transferring demographic data to a clinical dataset. The process is more complex when either there is no unique identifier (e.g. National Health Service or social security number), or errors exist in the identifiers that are used. Probabilistic linkage procedures, which assign weights to candidate linking records on the basis of the degree of matching of identifying variables, are often used [1].

Missing data, lack of available identifying information and lack of discriminatory identifiers can lead to two types of error in probabilistic linkage. The first source of error—true matches that fail to link—occurs when a record cannot be matched to another with a high enough degree of certainty. In this situation, information about possible values of the VOI is lost. The second source of error—false matches that link erroneously—occurs when a probabilistic match is chosen but the link is wrong. This implies uncertainty about the correctness of a link and the values of the variables associated with that link. This 'linkage error', akin to measurement error, should be incorporated into subsequent analysis of record-linked data, but this is complicated and it is not routinely carried out [2, 3]. That even relatively small linkage errors can result in the introduction of substantial bias has long been recognised [4], yet little research into methods that adjust for these errors in analysis has been published.

[a] *Medical Research Council Centre of Epidemiology for Child health, University College London Institute of Child health, London, WC1N 1EH, U.K.*

[b] *Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, BS8 1JA, Bristol, U.K.*

*Correspondence to: *Harvey Goldstein, Medical Research Council Centre of Epidemiology for Child health, University College London Institute of Child health, London, WC1N 1EH, U.K.*

† *E-mail: h.goldstein@bristol.ac.uk*

The separation of linkage and analysis processes is commonplace, and even encouraged, to help preserve confidentiality. However, this results in the uncertainty in linkage being ignored in subsequent analysis, where less-than-certain matches are carried over to analysis as though the match was in fact perfect. Furthermore, important information can be lost when only one match is chosen. For example, consider a particular record in a primary file [the file of interest (FOI)] having 10 candidate matches in a secondary file [the linking data file (LDF)], one of which is male and nine of which are female. If the male record has the highest probabilistic weight, even by a very small margin, it will be considered the correct match, and the information contained in the discarded nine records will be lost.

For these reasons, we view traditional probabilistic methods as potentially losing efficiency. They may also introduce bias into subsequent data modelling by ignoring the uncertainty associated with linked variable values so that these values will incorporate 'measurement errors' leading, for example, to underestimation of regression coefficients. In the present paper, we propose an extension of the existing methodology that helps to avoid the problems associated with linkage error in traditional probabilistic record linkage. We propose a new procedure based upon an extension of multiple imputation (MI) techniques as well as the use of standard MI itself [5].

Other authors have recognised these problems with record linkage and formulated modelling procedures to deal with them. Thus, for example, Lahiri and Larsen [6] consider the case of a linear regression model where the response is effectively only in the FOI and predictors are all transferred from the LDF. Given matching weights and corresponding matching probabilities, they show how to obtain unbiased estimates for the parameters and their standard errors. Kim and Chambers [7] extend this to the case where more than two files are to be linked. These approaches, however, are restricted to linear regression models, assume that the response is located in the FOI and rely upon the assumption that the linkage probabilities are independent of the values of the variables being linked. The procedures described in the succeeding text are intended to relax these assumptions and provide a general procedure that can produce more efficient and less biased estimates.

In the next section, we describe the traditional probabilistic linkage method, followed by our extension of this and a description of the MI we propose to use for the procedure. This is followed by an example using simulated data based on corruption of a real data set. These simulations thus create files with known error rates to be linked and analysed using the traditional technique and the proposed new technique, yielding estimates that can then be compared with the true known quantities.

## 2. Traditional probabilistic linkage methods

Probabilistic record linkage procedures typically involve at least two files—a primary FOI that contains most of the data we wish to analyse, usually derived from routinely collected data or surveys, and secondary files that contain additional data on some or all of the individuals in the FOI that we wish to incorporate into our statistical analyses. We shall assume that we have only one secondary file and refer to it as the LDF, although the designation of one file as the FOI and one as the LDF is a formal convenience. We use identifying or matching variables (MV) on the individuals in both files to 'link' each individual in the FOI to the same individual in the LDF. In some cases, hopefully the majority, the link is unequivocal, and we can carry across the relevant variable values from the LDF to the FOI. In other cases, there is some uncertainty about the link because of missing or incorrect data. In these cases, we estimate a set of weights on the basis of the number of matches and discriminatory power of the MV.

Each individual $i$ in the FOI that cannot be linked perfectly to a record in the LDF is assigned a weight $w_{ij}$ corresponding to each candidate linking record $j$ in the LDF. We choose a common cut-off for these weights to satisfy criteria related to sensitivity and specificity, and we choose the maximum weight above this threshold, with corresponding records being regarded as linked [7]. 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 LDF file. In such a case, we could proceed by first linking the episodes in the LDF file (de-duplication) so that each individual is represented by a single (longitudinal) record and then linking these records to those in the FOI. 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. For further details on probabilistic linkage, see [8].

## 3. An extension to probabilistic linkage allowing for linkage error

We propose that a probability distribution for the VOI is derived from the matching weights used in probabilistic linkage. Thus, for each FOI record there is one or more LDF records, each with an associated probability of being a correct match. The unequivocal matches are those with a single LDF record and associated probability of 1.0. We suggest that this probability distribution is used within an extended MI framework for missing data.

## 4. Multiple imputation

Multiple imputation [5] 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. We assume 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, even though this is not the case when, for example, categorical variables are involved. We use a procedure [9] 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. A description of such a 'latent normal model' is given in Appendix A. In practice, this involves setting up a multivariate normal response 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.

In practice, a Markov chain Monte Carlo (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 we obtain by fitting the MOI to each completed dataset and averaging over these according to the so-called 'Rubin's rules' [5]. A value of 5 is often used, and GCKL has found between 10 and 20 completed datasets to be adequate for multilevel data [9]. 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.

## 5. Imputation for record linkage

In record linkage, we can consider the variables that are brought across, imported from the LDF to the FOI during linkage, as having 'partially known' or 'probabilistically determined' values. That is, where we have a known correct record match, the values imported are correct. When the link is made 'probabilistically' with some uncertainty attached, the imported values can be thought of as 'missing' but with an associated probability distribution that is in general a function of the set of linking probabilities for the records. GCKL discuss how such a probability distribution can be combined with the known data values in the FOI and refer to the procedure as 'prior informed imputation' (PII), where the probability distribution assumes the role of a prior distribution for the unknown missing values. We propose to make use of such a distribution, combined with likelihood estimates, to impute the missing values.

We shall assume that for each record in the FOI there exists at least one record in the LDF that has a non-zero weight. If this is not the case, then the missing data are imputed using standard MI. We shall also assume that, for each record in the FOI, the LDF does contain the 'true' record match. We discuss in the succeeding text the case where there is a non-zero probability ($q$) that the true record match is not present in the LDF.

We assume that the linkage process has produced a set of FOI records where the matching is known to be correct—typically the majority. PII therefore applies only to the remainder—the equivocal matched records. Where a record is selected, as an equivocal or unequivocal match, but a variable value is missing from the LDF, then we carry out standard MI for the missing value.

Table I shows an example where there is a single FOI (set B) record with two variables from the LDF (set A) to be transferred from one of four LDF candidate (equivocal) records. The estimated linkage

**Table I.** Two set A variables to be transferred to a set B record from four candidate LDF records with associated linkage probabilities.

| Record | Set A variables | | Pr(correct) |
|---|---|---|---|
| 1 | 0 | 0 | 0.02 |
| 2 | X | X | 0.03 |
| 3 | X | 0 | 0.80 |
| 4 | X | X | 0.15 |

Known variable values marked X. Missing values marked 0.
LDF, linking data file.

probabilities, derived from the weights of a probabilistic linking (see the succeeding text) are also given. Standard probabilistic record linkage would first decide on a lower threshold for acceptance of a record as a link. Thus, in the present case, if this was 0.75 then record 3 would be transferred. The missing value could then be imputed in the standard way. We see, however, that the (estimated) probability, over repeated applications, that the record with the highest probability is the correct record, is only 80%. Thus, in 20% of cases we would expect a wrong record to be transferred.

Alternatively, we could apply a standard MI procedure to impute the values for all the four LDF records, ignoring the probabilities. For this to be satisfactory, we require the following assumptions to hold.

1. The variables to be used in the MOI are present among the set B variables.
2. Conditionally on the set B values, and possibly auxiliary variables, the probability of a matching error is unrelated to the values of the set A variables.

The second, independence, assumption is required because if this is not true then the conditional distribution of the set A variables in the equivocal records will not be the same as the conditional distribution of the set A variables in the unequivocal records. Because, generally, the imputation will be based largely upon the relationship between the set A variables and the set B variables in the unequivocal records, failure of this assumption will lead to biases. This assumption is essentially the ‘missing at random’ (MAR) assumption that applies more generally to any missing values.

If these assumptions can be satisfied, then the full data structure is preserved, and a standard MI analysis will yield consistent estimates. Where the proportion of equivocal records is small, this will often be satisfactory. This procedure, however, ignores information about the matching probabilities attached to the LDF records in equivocal cases, and taking account of these can be expected to yield greater statistical efficiency. In the next section, we summarise the standard procedures for probabilistic linkage and show how these allow us to obtain matching probabilities. We then propose a PII procedure that takes account of the matching probabilities and also can be extended to more complex data linkage problems.

## 6. Estimating linkage weights and matching probabilities

From the probabilistic record linkage, for each record $i$ in the FOI, we have a set of weights $w_{ij}(j = 1, \ldots, m)$ over the $m$ candidate records indexed by $j$ in the LDF. These are defined in the succeeding text. Unequivocal records in the FOI will have a single matched record in the LDF. We assume initially that the $w_{ij}$ are independent of the VOI. If not, then we will have a dependency created between the MV and the joint distribution of the VOI. We shall explore this in our simulation example.

Consider each FOI record with a given set of MV agreement values $(g)$. For example, for three binary MV, we may observe a pattern $g = \{1, 0, 1\}$ indicating {match, no match, match}. For each pattern, we compute the probability of observing that pattern of MV values:

A) Given that the MV values should match $P(g|M)$, that is, it is a true link.
B) Given that the MV values should not match $P(g|NM)$, that is, it is a false link.

The traditional record linkage procedure then computes $R = P(g|M)/P(g|NM)$ and a weight $W = \log_2(R)$ [7], so that for FOI record $i$ and the candidate LDF record $j$, we have the weight $w_{ij}$. These typically are averaged over the candidate LDF records to give a weight $w_i$, essentially an ‘independence’ assumption.

Initial estimates of $P(g|M)$, $P(g|NM)$ come from known record matches or other datasets, and these can be updated as more matches and non-matches are fully determined. 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 [10]).

The traditional method proposes a cut-off threshold for $W$ to be chosen, 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.

For our purposes, we require a set of probabilities

$$p_{ij} = f\left(w_{ij}\right), \quad \sum_{1}^{n_i} f\left(w_{ij}\right) = 1 \tag{1}$$

where $n_i$ is the number of candidate records for FOI record $i$. For convenience, we may choose $f$ as the identity function so that (1) becomes

$$p_{ij} = w_{ij} / \sum_{j=1}^{n_i} w_{ij} \tag{2}$$

but other choices are possible, and this is an area for further research.

## 7. Prior informed imputation

Following a matching procedure, we obtain a set of probabilities as given by (1) that we shall assume are scaled to sum to 1.0, as in (2). We discuss in the succeeding text cases where this becomes modified. Thus, each LDF record $j$ attached to FOI record $i$ has a set of variables $\{y_{ij}\}$ and a probability $p_{ij}$, and the set of these probabilities $p_i$ comprises the prior distribution for FOI record $i$. We shall also assume, without loss of generality, that all the variables follow a joint multivariate normal distribution. If this is not the case for some of the observed variables, then a joint multivariate normal (MVN) distribution can be obtained using the procedures described by GCKL [9]. In the case, for example, of categorical variables, imputed values will be back-transformed to their original scales. 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. In practice, it will often be convenient to ignore those records that have no match on any matching variable. For these records, we will regard the variables to be transferred as missing and carry out a standard MI.

For the set of variables (set A in the preceding text) to be transferred from the LDF, denote their distribution, conditional on the set B variables, by $f(Y^{A|B})$. The conditioning is on the responses and any covariates in the imputation model and includes variables from the LDF that are treated as auxiliary predictor variables in the imputation model. This conditional distribution is also multivariate normal. For each FOI record $i$, we compute a modified prior probability, which is the likelihood component $f(Y^{A|B})$ multiplied by the prior $p_{ij}$ for associated (equivocal) LDF record $j$, namely, $\pi_{ij} \infty f\left(y_{ij}^{A|B}\right) p_{ij}$. The (normalised) set $\pi_i$ comprises the modified probability distribution (MPD) for each FOI record. Our proposed procedure is as follows.

We first note that we should not simply sample records at random according to the MPD as this will result in incorrect choices of the true record in a similar way to the standard probabilistic linkage. Instead we propose that, as in standard probabilistic linkage, a lower threshold is set for accepting a record as a true link, and if any records exceed this, we choose that with the largest probability. If no record exceeds the threshold, then we regard the data values as missing and use standard MI. The choice of threshold is not as crucial as in standard probabilistic linkage because even when we choose it to be very high, the unlinked record data is still utilised via standard MI. The largest gain can be expected to arise when the probability of a link is associated with the values of the variables to be transferred. When the MAR assumption discussed earlier holds, then given a high enough threshold, the proposed procedure will produce unbiased estimates, but standard probabilistic linkage will not. Furthermore, conditioning on the values of the MV as auxiliaries in the FOI can be expected to make the MAR assumption more reasonable. Incorporating the likelihood component in the MPD can also be expected to lead more often to the correct record exceeding a given threshold. We shall explore these issues further in our simulation example.

So far, we have assumed that the true matching record is located within the LDF 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 on the FOI, any equivocal records might indicate either that the patient is still alive or that they are dead but not equivocally matched. Assume that we know or have a good estimate of the mortality rate among our patients, say $\pi_d$. If a proportion of the FOI 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 $p_i$ by $1 - \pi_r$ and add an extra pseudo-record with probability $\pi_r$ with an associated code for a surviving patient. If a good estimate of the mortality rate is not available, then we might carry out a sensitivity analysis 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 FOI with several LDFs. One way to proceed is conditionally. Thus, for each iteration of the algorithm, we first carry out a PII for the FOI and LDF$_1$, then conditioning on the original FOI variables and those carried over from LDF$_1$, we carry out a PII for the FOI and LDF$_2$ 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 LDFs, but this may become impractical when the number of LDFs is moderately large. In some cases, we may have sets of MV 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 FOI. In this case, we would first carry out a PII choosing one of the hospital LDFs as the FOI and then carry out a PII where the combined records are matched to the main FOI. If there are MV common to all three files, then the linkage of the linked hospital records to the FOI will need to consider the matching probabilities associated with the three possible combinations of values across files for each matching variable.

## 8. A simulation example

We have simulated a data set to investigate the properties of our proposed method. To create the simulated LDF, we used 1080 paediatric intensive care admission records for children under 16 years of age, sampled from Paediatric Intensive Care Unit (PICANET). We chose the MV to be sex, month of birth, year of birth (1995–2009) and Soundex of surname. We included equal numbers of records for each month and year of birth to simplify probability calculations. Our interest is in the association between a predictor and 'time to infection' using a regression model. By removing the predictor from each (simulated) FOI and leaving it in the LDF together with corrupted MV, we can study the effects on the model parameters given different forms of linkage.

We simulate time to infection $(y_i)$ from

$$y_i = 0.5 + 0.5x_i + e_i, \quad x \sim N(0,1), \quad e \sim N(0,1) \tag{3}$$

We corrupt the LDF by introducing independent errors in the MV as follows:

(1) A random 4% of each birth month changed, with equal probabilities for any other birth month.
(2) A random 4% of each birth year changed, with equal probabilities for any other birth year.
(3) Four per cent of each gender's values changed.
(4) Ten per cent Soundex codes changed with changed values not corresponding to a correct value for any other individual.

The values for these errors are based on the analysis of error rates found in a manual linkage study. This introduction of known error levels allows us directly to estimate the required probabilities without having to implement the full probabilistic linkage process. For example, we can directly derive $P(\text{true match}|\text{FOI record is female and LDF male}) = 0.0485$ (in the case where there are equal numbers of boys and girls) because we know that 4% of gender values are incorrect. In the present case, the proportion of boys is 0.55, and the corresponding probability is 0.033. Likewise, the probability of a true match where the FOI record is a girl and the LDF record is a girl is 0.952. See appendix B for details. For each set of candidate records for a FOI record that is not unequivocally matched, because the errors are introduced independently, we can compute an overall matching probability. The set of these over the candidate records, after combining with the likelihood in the case of PII, is then scaled to sum to 1.0.

Results are based upon 100 simulated datasets and are compared with the full complete analysis with the known parameter values in (3). When exploring the properties of the standard probabilistic record linkage procedure, the set of probabilities determined by (2) is obtained for all equivocal records in the LDF. Unequivocal matches are first removed from the LDF, and then when an equivocal record from the LDF is chosen to be matched, it is also removed from the set of equivocal records. If any one of these exceeds a chosen threshold, then the associated VOI value is chosen and carried to the FOI. We have chosen three thresholds, 0.50, 0.30 and 0.10, corresponding to proportions of records in the final analysis of 90.7, 91.4 and 93.3, respectively. For a threshold of 0.5, only one LDF record can exceed the threshold, only three can exceed the threshold of 0.3 and nine the threshold of 0.1.

Secondly, we obtained results by treating all equivocal matches as missing data with a standard MI using five completed datasets. Finally, we performed the PII analysis as described in the preceding text, also using five completed datasets. All the MCMC analyses use a burn in of 250, and thereafter a completed dataset is chosen every 100 iterations to ensure approximate independence. In all analyses, the standard errors are estimated from the variation over simulated datasets. Software has been written to carry out PII, being an extension of the REALCOM software (http://www.bristol.ac.uk/cmm/software/realcom/imputation.html).

Table II compares results given by the three methods.

For the standard probabilistic matching, the downward bias in the estimate of the regression coefficient increases from just 1% when 90.7% of records are selected to about 10% when the threshold is lowered so that 93.3% of records exceed it. For PII, we have used probability thresholds of 0.45, 0.20 and 0.13 to obtain approximately the same number of selected records as in standard probabilistic matching. We see a small bias that does not increase as the threshold decreases. It appears that the combination of the likelihood and a prior based upon probabilistic linkage weights more consistently provides the correct choice of LDF record. For the procedure where standard MI is used for all equivocal matches, we obtain essentially unbiased and efficient parameter estimates.

## 9. Informative matching

In the previous simulations, we have assumed that the probability of matching on MV values is independent of the VOI. In many cases, however, this will not be so. For example, different hospitals may have different distributions of the VOI, and the quality of recorded identifying information may also differ between centres, inducing a lack of independence between the probability of matching and the VOI. It has been shown that results of analysis based on linked data can be misleading when the probability of matching is related to the VOI [11].

Therefore, for the model given by (3), whenever any field of the LDF is corrupted, we select a random value of $X$ from $N(0, 1)$ to replace the existing value. This will result for the equivocal records in a zero correlation between $X$ and $Y$. For the standard imputation procedure, this will not induce any change. The results, corresponding to those in Table II, are given in Table III.

It is clear from Table III that standard probabilistic record linkage performs considerably worse than in the uninformative case. PII, as expected, also performs worse although better than standard probabilistic record linkage, and standard MI performs best giving unbiased and efficient estimates.

For both informative and uninformative matching in our simulation examples, we compute the standard errors in the usual way, in the case of imputation using ‘Rubin’s rules’ [5], and the standard errors are similar to those obtained from the standard deviation between simulations. The standard error estimates in Tables II and III are the between-simulation estimates.

## 10. Discussion and conclusions

Linkage error can have important impacts on subsequent analysis of linked data. Our results confirm that standard probabilistic linkage methods lead to biased estimates because of error associated with choosing only the match with the highest probabilistic weight. This bias increases as the threshold for acceptance increases. Where there is an association between the variables used in the MOI and the linkage error probabilities, the bias is increased further. Lariscy [11] gives an example where such errors have led to incorrect inferences about comparative mortality rates. Another important issue with existing probabilistic record linkage methods is that for very large files, which are becoming increasingly common, the use of manual checking of equivocal records becomes prohibitive. Incorporating the (conditional) likelihood associated with candidate records into the choice of LDF matching record (PII) improves the bias and, in our simulation example, provides acceptable estimates where there is no association between the linkage probabilities and the LDF variables to be transferred. Where there is an association, a noticeable amount of bias is introduced, although less that in the standard probabilistic record linkage case. The use of standard MI, however, outperforms PII and, at least for the purposes of statistical modelling, generally will be the procedure of choice.

**Table II.** Standard and prior informed multiple imputation and probabilistic record linkage with different threshold choices.

| | [Mean % total records selected] (s.e.)<br>Burnin = 250, iterations = 400 | | | | Threshold probability [mean %<br>total records selected] (s.e.). | | |
|---|---|---|---|---|---|---|---|
| | Standard MI | Imputation with PII PT | | | Standard probabilistic record linkage | | |
| Parameter<br>(true value) | Standard MI<br>[80.4] | PII PT = 0.13<br>[93.1] | PII PT = 0.20<br>[91.0] | PII PT = 0.45<br>[90.5] | 0.1 [93.3 ] | 0.3 [91.4 ] | 0.5 [90.7 ] |
| $\beta_0$ (0.5) | 0.504 (0.0033) | 0.501 (0.0034) | 0.500 (0.0032) | 0.501 (0.0030) | 0.499 (0.0034) | 0.499 (0.0033) | 0.500 (0.0035) |
| $\beta_1$ (0.5) | 0.503 (0.0034) | 0.490 (0.0032) | 0.490 (0.0035) | 0.490 (0.0033) | 0.451 (0.0032) | 0.485 (0.0033) | 0.495 (0.0034) |

Uninformative matching. Standard errors in brackets. Number of simulations = 100.
MI, multiple imputation; PII, prior informed imputation; PT, probability thresholds.

**Table III.** Standard and prior informed multiple imputation and probabilistic record linkage with different threshold choices and with informative matching.

| | [Mean % total records selected] (s.e.)<br>Burnin = 250, iterations = 400 | | | | Threshold probability [mean %<br>total records selected] (s.e.). | | |
|---|---|---|---|---|---|---|---|
| | Standard MI | Imputation with PII PT | | | Standard probabilistic record linkage | | |
| Parameter | Standard MI | PT = 0.13 | PT = 0.20 | PT = 0.45 | 0.1 [93.4 ] | 0.30 [91.3 ] | 0.5 [90.6 ] |
| (true value) | [80.3] | [93.0] | [91.1] | [90.6] | | | |
| $\beta_0$ (0.5) | 0.501 (0.0028) | 0.505 (0.0038) | 0.49 (0.0032) | 0.495 (0.0032) | 0.500 (0.0037) | 0.501 (0.0032) | 0.492 (0.0031) |
| $\beta_1$ (0.5) | 0.497 (0.0033) | 0.455 (0.0036) | 0.452 (0.0034) | 0.453 (0.0034) | 0.398 (0.0035) | 0.427 (0.0031) | 0.444 (0.0035) |

Standard errors in brackets. Number of simulations = 100.
MI, multiple imputation; PII, prior informed imputation; PT, probability thresholds.

If records are to be linked, for example for the purpose of maintaining administrative data files, then our results suggest that PII is superior to standard probabilistic linkage techniques while having reasonable performance in statistical analysis, especially where there is little or no association between the matching probabilities and the LDF variables to be transferred. In our simulation example, we have transferred just one variable. We would expect the performance of PII to be improved, perhaps considerably, as more variables are transferred from the LDF, because the use of the (conditional) likelihood would be expected to select the correct linking record more often as we increase the number of such variables that have significant partial associations with the remaining variables. More generally, if such LDF variables can be identified, an efficient strategy would be to transfer these for use with PII, even where they are not featured in the MOI. Further work on this is currently being undertaken.

Several further issues remain for investigation. In some models, we may wish to use one or more MV in the substantive model, and in this case we would suggest incorporating such variables in the imputation model. More generally, conditioning on MV may also improve the performance of PII where MV are associated with LDF variable values. Further work is desirable where the analysis of interest is more complicated, for example a generalised linear model. There is also the issue of linking together three or more files, with possibly different sets of MV, that needs further investigation.

In conclusion, we suggest that for purposes of statistical modelling, probabilistic linkage techniques may be unnecessary and that it suffices to identify just those records with an unequivocal match. For other purposes, our findings suggest that a standard probabilistic linkage procedure is enhanced by taking account of the likelihood associated with the records to be transferred, using PII, and that this may be suitable also for general statistical modelling. In particular, we suggest that researchers should be more aware of and acknowledge the presence of error introduced through record linkage and should take such error into account in subsequent use of linked data.

## Appendix A

### *The latent normal model*

For multivariate data with mixtures of response variable types, GCKL [9] show how such a response distribution can be represented in terms of an underlying ‘latent’ multivariate normal distribution. 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 that 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 $(-\infty, 1)$ if $k=1$, $(p-1, \infty)$ if $k=p$, 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 to discrete distributions such as the Poisson [12] and to non-normal continuous distributions for which a normalising transformation exists, such as the Box–Cox family [9].

After all of these sampling steps have been completed, we thus have a multivariate normal distribution to deal with. 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 in the preceding text in order to provide a randomly imputed value on the original scales.

## Appendix B

Suppose that $A$ is the event that is the observed pair of values for a given matching variable chosen from the FOI and the LDF. We shall assume that the LDF MV are those that contain errors. There are 1080 records in each file. Where we refer to a 'match', we are concerned with a match on a matching variable value rather than a true 'record match'.

(1) **Soundex**. For the Soundex matching variable, we have a 10% error rate, and A is the event of no match. If $q_i$ is the proportion of records in the FOI having the distinct value $i$, for our dataset we have

$$\Pr(\text{match}) = \Pr(\text{Soundex values are really the same}) = \frac{\sum_{\text{distinct values of } i}(q_i N)^2}{N^2}$$

$$= \sum_{\text{distinct values of } i} (\mathrm{q_i})^2 = 0.0028$$

$$\Pr(\text{record non-match}) = 0.9972$$

$$\Pr(A|\text{record match}) = 0.1$$

$$\Pr(A|\text{record non-match}) = 1$$

So that

$$\Pr(A) = 0.9975$$

By Bayes theorem, we have

$$pr(\text{match}|A) = pr(A|\text{match})pr(\text{match})/pr(A)$$

So that $pr(\text{match}|A) = \frac{0.1*0.0028}{0.9975} = 0.00281$ and

$$\Pr(\text{Match}|\bar{A}) = 1.0$$

(2) **Gender**. For sex, let A be the event that the FOI is a boy and the LDF is a girl, both with error rates $=$ 0.04. Let the proportion of true boys be $m_1$, which is 0.55 in our dataset. Then we have

$$\Pr(\text{boy observed in LDF}) = m_1 + 0.04\,(1 - m_1) - 0.04m_1 = 0.92m_1 + 0.04$$

$$\Pr(\text{girl observed in LDF}) = 1 - 0.92m_1 - 0.04$$

$$\Pr(\text{match|A}) = \Pr(\text{girl in LDF is true boy}) = 0.04m_1/(1 - 0.92m_1 - 0.04) = 0.0485$$

Likewise, we compute

$$\Pr(\text{match}|\bar{A}) = \Pr(\text{girl in LDF is true girl}) = 0.952$$

where $\bar{A}$ is the event of FOI girl and LDF girl.

Similarly, if A is the event that the FOI is a girl and the LDF is a boy. We have

$$\Pr(\text{match|A}) = \Pr(\text{boy in LDF is true girl}) = \frac{0.04\,(1 - m_1)}{0.92m_1 + 0.04} = 0.033$$

and

$$\Pr(\text{match}|\bar{A}) = \Pr(\text{boy in LDF is true boy}) = 0.967$$

(3) **Month of Birth**. For month of birth, let the true proportion in each month (January to December) be denoted by $h_1, \ldots, h_{12}$ with error rates 0.04. let A be the event that the FOI has month $j$ and the LDF has month *not* $j$. We assume that when an error occurs it is equally likely to be recorded as any of the remaining months. We have

$$\Pr(\text{Observed month } j \text{ in LDF}) = h_j - 0.004h_j + \sum_{k \neq j} 0.04h_k/11$$

$$\Pr(\text{match}|A) = (\text{Month not } j \text{ is really } j) = 0.04h_j/\left\{1 - \left(h_j - 0.04h_j + \sum_{k \neq j} \frac{0.04h_k}{11}\right)\right\}$$

To simplify the computations, we choose the FOI with equal numbers (90) in each month so that

$$\Pr(\text{match}|A) = 0.00364$$

and

$$\Pr(\text{match}|\bar{A}) = 0.96$$

(4) **Year of Birth**. For year of birth, let the true proportion in each of the 15 years (1995–2009) be denoted by $h_1, \ldots, h_{15}$ with error rates 0.04. Let A be the event that the FOI has year $j$ and the LDF has year *not* $j$. We assume that when an error occurs it is equally likely to be recorded as any of the remaining years. We have, similarly as for month of birth,

$$\Pr(\text{Observed year } j \text{ in LDF}) = h_j - 0.04h_j + \sum_{k \neq j} 0.04h_k/14$$

$$\Pr(\text{match}|A) = (\text{Year not } j \text{ is really} j) = 0.04h_j/\left\{1 - \left(h_j - 0.04h_j + \sum_{k \neq j} \frac{0.04h_k}{14}\right)\right\}$$

To simplify the computations, we choose the FOI with equal numbers (72) in each year so that

$$\Pr(\text{Match}|A) = 0.002857$$

and

$$\Pr(\text{match}|\bar{A}) = 0.96$$

For each non-match and match on the matching variables, we multiply the appropriate probabilities together. For each FOI record without an unequivocal record match, we scale these probabilities to sum to 1 and form a prior vector from the associated VOI values and scaled probabilities.

We note that in general there are non-zero probabilities that an observed match using all matching variables is actually a non-match. However, the probability that this occurs is generally very small and in our simulation is actually zero because the errors for Soundex do not correspond to possible values.

## Acknowledgements

We are very grateful to Ruth Gilbert, James Brown, Berit Muller-Pebody and anonymous referees for comments on drafts of this paper. We also wish to thank John Hartley for contributing microbiology data; Roger Parslow and the PICANet team for contributing PICU data; and Paul Lock, Joanna Bell and Thomas Fleming for facilitating the data retrieval. The work was supported by a grant from the Medical Research Council G0400546 and also by a grant from the Economic and Social Research Council, RES-062-23-2265, as part of the National Centre for Research Methods programme. Katie Harron was funded by the National Institute for Health Research Health, Technology Assessment (NIHR HTA) programme (project number 08/13/47).

## References

1. Clark D. Practical introduction to record linkage for injury research. *Injury Prevention* 2004; **10**(3):186.
2. Scheuren F, Winkler W. Regression analysis of data files that are computer matched—Part I. *Survey Methodology* 1993; **19**(1):39–58.
3. 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 Services Research* 2010; **10**(1):346.
4. 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):1005–1027.
5. Rubin DB. *Multiple Imputation for Non-response in Surveys*. Wiley: Chichester, 1987.
6. Lahiri P, Larsen MD. Regression analysis with linked data. *Journal of the American Statistical Association* 2005; **100**:222–230.
7. Kim G, Chambers R. Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica* 2012; **66**(1):64–79.
8. McGlincy MH. A Bayesian record linkage methodology for mulitiple imputation of missing links. *ASA Section on Survey Research Methods, 2004*. http//www.amstat.org/section/srms/Proceedings/y2004/file/Jsm2004-000683.pdf (Accessed June 23, 2012).
9. Goldstein H, Carpenter J, Kenward M, Levin K. Multilevel models with multivariate mixed response types. *Statistical Modelling* 2009; **9**(3):173–197.
10. Newcombe HB. Age-related bias in probabilistic death searches due to neglect of the “Prior Likelihoods”. *Computers and Biomedical Research* 1995; **28**(2):87–99.
11. Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *Journal of Aging and Health* 2011; **23**(8):1263–1284.
12. Goldstein H, Kounali D. Multivariate multilevel modelling of childhood growth, numbers of growth measurements and adult characteristics. *Journal of the Royal Statistical Society, A* 2009; **172**(3):599–613.