Research Article

Statistics in Medicine

Received: 4 August 2016, Accepted: 2 March 2017 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7287

A scaling approach to record linkage

Harvey Goldstein,[a,c*†] Katie Harron[b] and Mario Cortina-Borja[c]

With increasing availability of large datasets derived from administrative and other sources, there is an increasing demand for the successful linking of these to provide rich sources of data for further analysis. Variation in the quality of identifiers used to carry out linkage means that existing approaches are often based upon 'probabilistic' models, which are based on a number of assumptions, and can make heavy computational demands. In this paper, we suggest a new approach to classifying record pairs in linkage, based upon weights (scores) derived using a scaling algorithm. The proposed method does not rely on training data, is computationally fast, requires only moderate amounts of storage and has intuitive appeal. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: scaling; record linkage; data linkage; correspondence analysis

Introduction

With the increasing availability of large datasets derived from administrative and other sources that contain records for the same underlying population of individuals, the successful matching of individuals via linking algorithms can provide rich sources of data for further analysis and has become a key issue in the management of publicly available data sources. Aside from ethical issues related to disclosure, the very large size of such datasets and variation in the quality of the identifiers used to carry out linkage raises important considerations. Existing 'probabilistic' linkage approaches often require 'training' data, i.e. a subset of data where the true match status is known, from which to estimate parameters for linkage algorithms. However, parameter estimation can be time consuming and demand large amounts of storage, and appropriate training data are often unavailable. The focus of our paper is therefore 'unsupervised' linkage. An overview of current approaches can be found in Harron, Goldstein and Dibben [1]. In this paper, we propose a new approach to deriving weights, or scores, for linkage that is computationally efficient, requires only moderate amounts of storage and has intuitive appeal.

Probabilistic record linkage

We begin with a review of existing approaches.

The standard algorithm for probabilistic record linkage derives from the original work of Fellegi and Sunter [2], and its history and development are outlined by Winkler [3]. Recent reviews of probabilistic record linkage can be found, for example, in Christen [4].

Record linkage involves a characterisation of the primary units of analysis as the pairs of the set $C = (A \times B)$ where A, the file of interest (FOI) is a sample of individuals from a well-defined population and B, the linking data file (LDF) is a second 'independent' sample that includes the same N individuals and possibly additional individuals. This results in a set of pairs of records, for which the same set of 'identifiers' (such as age, sex, identification number, name, etc.) is available. We note that there are other linkage scenarios where only a subset of individuals from the FOI is present in the LDF. While this will not affect our comparison of methods, it will affect the subsequent analysis of the matched data [1]. The term 'independent' is taken to mean that the identifiers are measured independently in each sample. Because of measurement errors, identifiers cannot always uniquely 'match' the individuals from the two samples. It is usually assumed that these errors are independent of the values of the identifiers or

[a] *University of Bristol, Bristol, U.K.*
[b] *London School of Hygiene and Tropical Medicine, London, U.K.*
[c] *University College London, London, U.K.*
**Correspondence to: Professor Harvey Goldstein, University of Bristol Graduate School of Education, Bristol BS8 1JA, U.K.*
[†] *E-mail: h.goldstein@bristol.ac.uk*

any other variables that may be measured on the individuals, although this assumption does not always hold [5]. Samples A and B, apart from the common identifiers, contain different variables that the data analyst wishes to bring together for modelling purposes.

The aim of probabilistic record linkage methods is to determine a set of weights, or scores, for the set C that allows a classification of the elements of C into 'matches', 'non-matches' or undecided matches, ranked according to the assigned weights. A threshold weight is traditionally chosen to classify these record pairs above which a pair is accepted as a match and below which a pair is accepted as a non-match which thus determines false-positive and false-negative error rates.

For each identifier or 'field', we define two conditional probabilities for each record pair i_A, i_B of C. For simplicity of exposition, we shall assume that A and B contain the same set of individuals, that there are p identifiers, indexed by j, and that we measure either agreement or disagreement rather than degrees of agreement. This will suffice to motivate our comments on the algorithm, although methods can be readily extended to other linkage scenarios. We have $m_j = P(\text{agreement on identifier } j | i_A = i_B)$, i.e. the probability of observing agreement given that it is the same individual.

$u_j = P(\text{agreement on identifier } j | i_A \neq i_B)$, i.e. the probability of observing agreement given that it is not the same individual.

The default assumption, which we discuss further, is now made that for any record pair these probabilities for each identifier are independent so that we can write the joint probability for the observed agreement/non-agreement values (y) of a pair, indexed by l as

$$P\left(y_{l1}, \ldots..., y_{lp}\right) = \left(\prod_j m_j\right)\pi_{i_A=i_B} + \left(\prod_j u_j\right)\pi_{i_A \neq i_B} \tag{1}$$

where $\pi_{i_A=i_B}$ and $\pi_{i_A \neq i_B}$ are, respectively, the probabilities that $i_A = i_B$ (a match) and $i_A \neq i_B$. (a non-match).

The standard analysis proceeds by writing down the 'likelihood' for the data as

$$\prod_l P\left(y_{l1}, \ldots..., y_{lp}\right) \tag{2}$$

and maximising it for the m_j, and u_j typically using an EM algorithm ('unsupervised' linkage) [3]. Alternatively, where 'training' data are available, parameter estimates can be derived from these using the known true match status ('supervised' linkage).

Equations (1) and (2) define a latent class model where the classifier is a function of $R = \left(\prod_j m_j\right) / \left(\prod_j u_j\right)$, usually $\log_2 R$, where these classifiers are summed over identifiers to give an overall weight to each pair. These weights are then used to classify a pair as a 'match' or 'non-match' according to whether a suitably chosen threshold value is exceeded. Typically, where all identifiers agree, a match is assigned; where all disagree, a non-match is assigned. A function of the weight can be treated, suitably scaled, as the probability of a match as in Goldstein, Harron and Wade [6].

Alternative approaches

In fact, (2) does not represent a true likelihood because the elements are not strictly independent; the observed identifier patterns for a set of record pairs associated with any given individual are related, because given the observed pattern for the one true match, the probability associated with all identical patterns will be that for a non-match. Tancredi and Liseo [7], among others, point this out.

These latter authors develop a fully Bayesian classification procedure for the estimation parameters and linkage decisions, which incorporates a model for the misclassification probabilities. They propose a simple version of the 'hit–miss' model for these misclassification probabilities, which assumes a particular form for the probabilities of observing any given value given an underlying true value, as well as making the assumption of independence among identifiers. The first of these assumptions requires that the misclassification mechanism is the same in both files, which in many examples is debateable, for example if data inputting processes differ across files. The distribution of the misclassified values is assumed to be the same as the 'true' distribution, which is also debateable, and so it is not clear whether such a model is likely to be reasonable in practice. Sadinle [8] extends this model to handle missing data and partial agreements. Hit–miss models for linkage have also been explored by other authors, but are not as frequently used in practice as the Fellegi–Sunter approach [9]. In principle, rather

than relying on a model, it would be possible to determine these misclassification probabilities empirically, but they are likely to be context specific and not easily generalisable.

These Bayesian procedures are attractive in that they provide a coherent statistical model that avoids the record pair independence assumption. In addition to concerns about the assumptions made, a practical drawback is that they are currently computationally intensive and may not be feasible for large datasets.

Machine learning approaches have also been used for linkage. Essentially, these use a training set to derive a classification into matched or unmatched groups. Where training data are unavailable, it may be possible to create a set using similar data where matching status has been determined, or possibly a subset of current data that has been subject to careful manual matching (see for example Ng and Jordan [10]).

In the next section we describe an alternative approach that is similarly motivated but conceptually and practically simpler.

A scaling procedure

The Felligi–Sunter 'likelihood' based and similar procedures do not, as we have pointed out, have the usual optimality properties associated with maximum likelihood estimation because the likelihoods are not true likelihoods. Nevertheless, they can be viewed as convenient algorithms for assigning weights that discriminate between the matching classes that units belong to. The Bayesian procedures likewise have the ultimate aim of labelling record pairs as matches or non-matches, with some cases where no decision is made. In the present paper, we propose an alternative procedure in order to derive weights, but based upon a scaling model first introduced by Healy and Goldstein [11] to assign weights or scores to observed stages of wrist bone maturity development in children who passed from completely immature to fully mature stages. It belongs to the class of procedures broadly known as correspondence analysis [12] that seek to assign weights or scores to discrete categories based upon the minimisation of a suitable loss function.

For the application to record linkage, we use the terminology given in [6]. Specifically, rather than setting up a formal statistical model, we define a loss function that is intuitively appealing and derive a procedure for its minimisation.

For each of p identifiers, j, we assume that we have several (ordered) states denoted by $k=1, \ldots k_j$ where 1 is the least agreement and k_j is the greatest level of agreement between the FOI and the LDF. In the simple binary case, there are two states (agree/not agree) for each identifier j. This gives a total of $K = \sum_j k_j$ categories over all identifiers. For example, binary agreement/disagreement on each of four identifiers would give a total of $K=8$ categories.

Where an identifier value is missing, if we assume that missing data occur completely at random, or at random conditionally on values of other identifiers or other record values, then we can draw a value at random from the appropriate posterior distribution, estimated separately for each file.

We seek to estimate a score x_{jk} for state k and identifier j, where the classifier for pair i now takes the form $z_i = \sum_j z_{ij}$, and $z_{ij}=x_{jk}$ if pair i has state k for identifier j. We fix the average classifier values for a definite match as $\sum_j x_{jk_j} = 1$ and for a definite non-match as $\sum_j x_{j1} = 0$. That is, the sum of the scores for each greatest level of agreement is 1, and the sum of the scores for each lowest level of agreement is 0. These scores are analogous to the 'weights' defined in existing methods. We shall elaborate our model to allow further, pre-defined, weights below. We define the following:

$$x_{(n\times 1)} = vec\left(x_{ij}\right),\ \ q_{(n\times 1)} = vec\left(q_j\right), q_j = \frac{1}{p}(1,\ \ 0, \ldots\ 0)_{(1\times kj)} r_{(n\times 1)} = vec\left(r_j\right),\ \ r_j = \frac{1}{p}(0, \ldots,\ \ 0,\ \ 1)_{(1\times kj)}$$

$\delta_{ijk}=1$ if observed agreement state is k for pair i identifier j, otherwise 0,
$N_{jk} = \sum_i \delta_{ijk}$, the number of pairs with agreement state k for identifier j
$N_{jklm} = \sum_i \delta_{ijk}\delta_{ilm}$, the number of pairs with agreement state k for identifier j and with agreement state m for identifier l

$A_{(K\times K)} = \frac{1}{p^2}\begin{pmatrix} (p-1)N_{11} & & \\ \vdots & \ddots & \\ -N_{11pk_p} & \dots & (p-1)N_{pk_p} \end{pmatrix}$, where diagonal elements of A are $\frac{(p-1)}{p^2}N_{jk}$ and off-diagonal elements are $-\frac{1}{p^2}N_{jklm}$.

We note, for computational purposes, that the matrix $A = S - Z$, where S is the $K \times K$ diagonal matrix with elements N_{jk}/p and Z is $K \times K$ symmetric with diagonal elements N_{jk}/p^2 and off-diagonal elements N_{jklm}/p^2.

$d_i = \frac{1}{p}\sum_j \left(z_{ij} - z_i\right)^2$, the average within-pair squared discrepancy between weights and

$D = \sum_i d_i = x^T A x$, is the total within-pair discrepancy.

We seek to minimise the within-pair discrepancy D subject to

$$q^T x = 0, r^T x = 1 \tag{3}$$

These 'end point' constraints are introduced to avoid the trivial solution where all the weights are zero and are appropriate in the context of assignment to one of two extreme classes (match, non-match) where the sum over all agreements is 1 and over all disagreements is 0.

This leads to the straightforward solution for x from the set of linear equations given by Goldstein [12]

$$2\mathrm{Ax} - \mathrm{q}\lambda - \mathrm{r}\mu = 0$$

which together with (3) leads to solving the non-homogeneous set of linear equations

$$A^* x^* = b,\ A^* = \begin{pmatrix} 2A & -q & -r \\ q^T & 0 & 0 \\ r^T & 0 & 0 \end{pmatrix}, x^* = \begin{bmatrix} x \\ \lambda \\ \mu \end{bmatrix},\ b = \frac{1}{p}\begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix} \tag{4}$$

We find $2D = \mu = -\lambda$. For convenience, by subtracting the 'non-agreement' score for each identifier from that identifier's scores, we can form a rescaled score vector so that the non-agreement score for each identifier is now 0 and the sum of the full agreement scores remains 1. In our example, we have multiplied all the scores by 100 for presentation purposes.

In fact, there is an infinity of possible constraint systems such as (2). The 'end point' constraint we have used recognises that complete agreement on all identifiers is associated with a maximum score equivalent to a matched record and complete disagreement on all identifiers is associated with a non-match.

In practice, with large datasets, the matrix A^* may not be well conditioned, in which case we can define

$$\mathrm{q_s} = \mathrm{sq},\ \mathrm{r_s} = \mathrm{sr},\quad \mathrm{b_s} = \mathrm{sb}$$

and substitute these values in (4). A suitable value of s could be the total number of units $N = \sum_{j,k} N_{jk}$.

We note that the storage requirements are modest, of the order K^2, the square of the total number of categories. For example, using binary agreement/disagreement on four identifiers would result in eight categories so that $K^2 = 64$. The matrix A is readily computed by cycling over the record combinations for each matrix cell. For each pair, we have p^2 simple comparisons, and the results of these are accumulated over the set of pairs. For the binary case, the comparison establishes whether (pairwise) agreement is present or not. Where we have more than two categories of agreement, for each comparison of identifiers, the agreement category has to be computed using a suitable algorithm, for example in terms of a 'distance' between the identifier values (e.g. using a string comparator). This has timing implications, but in general, it would seem that very large datasets can readily be handled. In the next section, we give an example of our procedure and compare its results with a traditional Fellegi–Sunter probabilistic linkage method.

An example of linking two files

The data are synthetic data generated from a dataset obtained from pediatric intensive care units in England and Wales with known matches, as described in [13]. The FOI has 7742 records, and the LDF has 10 000 records. Missing values are introduced completely at random into just one of the

identifiers in each of 19% of the records in the LDF. Errors in the identifiers are introduced randomly in such a way that the m probabilities for each identifier are 0.95. Because the missing values are introduced randomly and because the correlations among the identifiers are negligible (none is greater than 0.04 in absolute value), the imputed values to replace those missing are, for simplicity, sampled from the observed marginal distributions of each identifier.

The identifiers are day of the month, month, year (1991–2006) and gender. The final 4 years, 2003–2006, account for the majority of cases (13, 20, 27 and 31%, respectively).

Using the software LINKPLUS [14], we carried out an unsupervised probabilistic matching but incorporating the knowledge that the separate m probabilities for identifier agreement are all 0.95. The scaling algorithm was performed in MATLAB [15]. The MATLAB routine and also one written in R [16] are available from the corresponding author.

Table I shows, to the nearest integer, the estimated weights from the two algorithms. Day of the month, the most discriminatory identifier, has the highest weight for a match, followed by month, then year, then gender. Bearing in mind that we have generated only one synthetic dataset, we note that the weights have the same ordering and are approximately equivalent in terms of ranking record pairs. In fact, if we simply square and standardise the probabilistic weight estimates in row 2, we obtain weights very close to those given by our scaling method. When the weights are combined across identifiers, individual combinations may be ranked differently between methods and for each of the 16 possible agreement/non-agreement patterns, the overall pattern of weights with rankings is given in Table II.

We see a reasonable agreement between the rankings for each method. For most choices of threshold to classify record pairs as matches/non-matches, each method would produce the same result. Furthermore, the scales themselves are only invariant up to a monotonic transformation. For example, the weights in traditional probabilistic record linkage are derived from a logarithmic transformation of the ratio of the m_j and u_j, but other reasonable combinations of these parameters are possible. Because a monotonic but non-linear transformation of either scale would generally produce different rankings of the patterns, it is the ordering of the weights associated with each indicator that is the most appropriate basis for a comparison of methods. As we would expect, both procedures produce an ordering where the most discriminating variable, day, has the highest weight and sex the lowest.

Finally, because the data have actually been generated from known matches, we can compare the two procedures with respect to their closeness to the correct match status. In fact, only 1000 records in the FOI have matching records in the LDF so that only these records are used in the comparison. We estimated the probabilities by computing the proportion of times, for each identifier pattern, that the record pair was the correct match. Table III shows the estimated probabilities for each identifier pattern and an estimate of the simple correlation between these probabilities and the separate weights for each procedure, omitting the first (all identifiers disagree) and last (all identifiers agree) categories, because these are constrained. The traditional procedure gives a somewhat lower estimate than the scaling procedure for this example.

Extensions

We note that our agreement measure has been assumed to be categorical. In some cases, however, it may be effectively continuous, such as in measures of phonetic distance, or age. In such cases, the simplest approach is to categorise the scale into a small number of categories, and sensitivity analyses can be carried out to determine a satisfactory classification. For example, continuous values of the Jaro-Winkler [17] string comparator for measuring similarity between names could be categorised as <0.8, 0.8–0.9, 0.9+, etc. It would, in principle, be possible to consider a mixture of categorical and continuous variables where a particular functional form for the latter was assumed, for example linearity. We shall not, however, pursue this possibility here.

Table I. Comparison of agreement weights using scaling and traditional probabilistic matching.

	Day	Month	Year	Sex
Scaling estimates	53	22	19	7
Probabilistic estimates	32	27	26	15

Table II. Weights for each linkage pattern (day, month, year, sex). Ordered with respect to scaling model. Ranks in brackets.

Identifier linkage pattern	Scaling method	Traditional method
0 0 0 0	0 (1)	0 (1)
0 0 0 1	7 (2)	15 (2)
0 0 1 0	19 (3)	26 (3)
0 0 1 1	26 (5)	41 (6)
0 1 0 0	22 (4)	27 (4)
0 1 0 1	29 (6)	42 (7)
0 1 1 0	41 (7)	53 (9)
0 1 1 1	48 (8)	68 (12)
1 0 0 0	53 (9)	32 (5)
1 0 0 1	60 (10)	47 (8)
1 0 1 0	72 (11)	58 (10)
1 0 1 1	75 (12)	73 (13)
1 1 0 0	79 (13)	59 (11)
1 1 0 1	86 (14)	74 (14)
1 1 1 0	94 (15)	85 (15)
1 1 1 1	100 (16)	100 (16)

In traditional probabilistic record linkage, there is an assumption that the overall match probabilities are derived as a product over the identifiers of separate identifier probabilities. This implicit assumption of independence is mirrored by the scaling procedure use of a (possibly weighted) sum of weights assigned to each identifier, but there is no explicit assumption of statistical independence. Moreover, the procedure can be generalised, for example by combining identifiers so that all possible combinations are considered as a new set of categories.

For example, consider two identifiers, X, Y each with two categories. We may form the combined identifier which is the set of all possible category pairs for X and Y. We can then replace the separate identifiers by a new one XY with four categories. This can be done for a number of disjoint pairs and can be extended to sets of three or more categories if required. The analysis proceeds as before. In practice, there will be a limit to this procedure where category numbers become small or particular combinations may not exist, although it may be possible to combine cells with small counts. We can also

Table III. Probability that we obtain the correct match for each identifier linkage pattern, and correlation of these probabilities with the procedure weights.

Identifier linkage pattern	Probability the match is correct
0 0 0 0	*
0 0 0 1	*
0 0 1 0	*
0 0 1 1	*
0 1 0 0	*
0 1 0 1	*
0 1 1 0	*
0 1 1 1	0.072
1 0 0 0	*
1 0 0 1	0.013
1 0 1 0	*
1 0 1 1	0.071
1 1 0 0	*
1 1 0 1	0.071
1 1 1 0	0.048
1 1 1 1	0.695
Correlation for scaling procedure**	0.53
Correlation for traditional procedure**	0.26

*Indicates probability <0.01.
**The first and last (0000, 1111) categories are omitted.

carry out sensitivity analyses, trying different combinations of identifiers to examine changes to the estimates.

We have assumed that, a priori, our identifiers have equal status. In some cases, however, one or more identifiers may have low reliability, and in an extreme case be close to random noise. In such cases, we may wish to down-weight their role in determining the overall category weights. Suppose then that each identifier has a weight w_j, $\sum_j w_j = 1$. We now have

$$d_i = \sum_j w_j \left(z_{ij} - z_i\right)^2$$

which for the matrix A leads to the diagonal elements $w_j(1 - w_j)N_{jk}$ and off-diagonal elements $-w_j w_l N_{jklm}$ *and* $q_j = w_j(1, 0, \ldots, 0)_{(1 \times kj)}$, $r_j = w_j(0, \ldots, 0, 1)_{(1 \times kj)}$

Linking more than two files

We can extend the case of two files to several files as follows. We assume that one file is the primary FOI. We carry out separate linkages with each secondary LDF, possibly using a different set of identifiers in each case, and in each case derive a set of candidate weights for each record in the primary file. We further assume, without loss of generality, that a different set of variables from each is to be selected for transfer from each linking file to the primary file.

For each set of candidate weights, we may wish to set a threshold above which the record with the maximum weight is selected, and, as in the two file case, where the threshold is not exceeded, this will result in a set of missing data values. For these, records, alternative approaches such as prior-informed imputation can be used to carry over a set of variable values [6]. This procedure will not depend on the order of the files being linked but may be dependent on the choice of the primary file. In many cases, this choice may be a natural one, such as when a survey sample is being supplemented with data from administrative datasets. In other cases, such as the linking of several administrative datasets, the choice may not be obvious, and a sensitivity analysis, choosing different primary files, may be needed to explore sensitivity to the choice. The computational efficiency of our suggested procedure will often make this practically feasible.

Discussion

We propose a scaling approach to deriving match weights with which to classify record pairs in linkage. Our approach provides a measure of pattern agreement equivalent to the weight derived from the traditional probabilistic approach, and in our example, we obtain similar rankings of record pairs which will provide the same or similar linkage result, depending on choice of threshold. We also note that the estimated correlation between the probability of being a true match and the weights estimated from the scaling method is 0.53, compared to 0.26 for the traditional method, although we would not wish to generalise from this one example. We could also choose to stratify the agreement measure according to which identifiers agree. Thus, for example, we could choose to distinguish agreement on birth day and month within and between years, by introducing different weights for the two cases, or alternatively by forming a combined variable, say month and year.

Our approach does not rely on the availability of training data to estimate parameters, or on any distributional assumptions. Importantly, our proposed procedure is conceptually and computationally simpler than existing methods, makes fewer assumptions, yet captures implicitly the notion of identifier patterns as indicators of the propensity for a match and has the ability to handle large datasets in an efficient manner. A probabilistic interpretation can still be made if we consider the datasets as being 'sampled' from a notional population of similar datasets so that the resulting weights can be applied in other instances with similar data. Thus, for example, a non-parametric bootstrap procedure could be used to obtain interval estimates for the weights by resampling records with replacement.

We would note that using the EM algorithm for parameter estimation in the Fellegi–Sunter model will sometimes fail to converge or converge to a local maximum, even when a range of starting values are used. This is a drawback and may well be related to the fact that the 'likelihood' being maximised ignores the dependencies in the data. The algorithm becomes particularly unstable when the proportion of records with a true match is low.

Where training data are available with known match status, the latter is often treated as an outcome in a general linear model with the identifier agreement as predictors. In this case, however, we also encounter the problem that for any given record in the FOI, only one linking file record is a match, and the usual model assumption of independent outcomes is violated.

We have suggested that our procedure is computationally efficient, not involving a time-consuming iterative estimation procedure. Nevertheless, we have used a fairly small dataset in our example, and computing times will be lengthier with very large data files, and especially when agreement status is based upon degrees of agreement in more than two categories. Procedures for applying the algorithm to a random subset of the data could usefully be studied, and this is an area for further research. One possibility for very large samples is to approximate A by selecting a simple random sample of all possible pairs with sampling fraction s to compute the N_{jk}, N_{jklm}, and then rescaling these by s^{-1}. As in traditional record linkage, we can also introduce blocking on certain identifiers to reduce the number of computations. Both of these possibilities are topics for further research.

Finally, we note that once we move away from the traditional explicit model-based approach that relies upon optimal properties such as those associated with maximum likelihood estimates, we are confronted by a need to choose both the classifier function and the constraints to ensure identifiability. Different choices will lead to different solutions, and in particular to different rankings of pairs and thus different selections based upon thresholds. We have argued that our own choice of scaling procedure is based upon sensible criteria, but it would be useful to explore this further. We would also welcome sensitivity analyses using real life datasets, but this will often be limited by the amount of computational time involved. In general, linkage success depends on the choice of threshold for classifying record pairs. This is an ongoing area of research, as a choice of threshold needs to be optimal for a particular linkage scenario and substantive research question, where false-positives or false-negatives may have differing impacts. The extent of non-random linkage errors is also important. We are currently looking at the impact of different choices upon the inferences from the final substantive models fitted to the linked data.

Acknowledgments

We are most grateful to several referees for their comments on an early version of the paper. We would also like to thank Luigi Palla for helpful comments.

References

1. Harron K, Goldstein H, Dibben C (Eds). Methodological Developments in data Linkage. Chichester: Wiley, 2015.
2. Fellegi I, Sunter A. A theory for record linkage. *Journal of the American Statistical Association* 1969; **64**:1183–1210.
3. Winkler WE. Chapter 2: Probabilistic linkage. In Methodological Developments in Data Linkage, Harron K, Dibben C, Goldstein H (eds). Wiley: Chichester, 2015.
4. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection (Data-Centric Systems and Applications). Springer: New York, 2012.
5. Tromp M *et al.* Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. *Journal of the American Medical Informatics Association* 2008; **15**(5):654–660.
6. Goldstein H, Harron K, Wade A. The analysis of record linked data using multiple imputation with data value priors. *Statistics in Medicine* 2012:**31**(28):3481–3493.
7. Tancredi A, Liseo B. A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics* 2011; **5**:1553–1585.
8. Sadinle, M. (2016). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2016.1148612. (to appear).
9. Copas J, Hilton F. Record linkage: statistical models for matching computer records. J. *Royal Statistical Society, A* 1990; **153**(3):287–320.
10. Ng A, Jordan M. On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes. In Advances in Neural Information Processing Systems, 14, Dietterich TG, Becker S, Ghahramani Z (eds). MIT press: Cambridge, MA, 2002; P841–P848.
11. Healy MJR, Goldstein H. An approach to the scaling of categorised attributes. *Biometrika* 1976; **63**:219–229.
12. Goldstein H. The choice of constraints in correspondence. *Analysis Psychometrika* 1987; **52**(2):207–215.
13. Harron K, Goldstein H, Wade A, Muller-peabody B, Parslow R, Gilbert R. Linkage, evaluation and analysis of national electronic healthcare data: application to providing enhanced blood-stream infection surveillance in pediatric intensive care. *PloS One* 2013; **8**(12) e85278. doi:10.1371/journal.pone.0085278.
14. LINKPLUS software: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm.
15. Mathworks (2015). MATLAB. mathworks.com/products/matlab.
16. The R project for statistical computing. https://www.r-project.org/.
17. Winkler W.E. (2006) Overview of record linkage and current research directions. Research Report Series (Statistics #2006-2) Statistical Research Division, Washington DC: US Census Bureau.