Measurement of Type I and Type II Record Linkage Error

Dean Resnick[1], Jana Asher[2]
[1]National Opinion Research Center, 4350 East-West Highway, Bethesda, MD 20814
[2]Department of Mathematics and Statistics, Slippery Rock University, 1 Morrow Way, Slippery Rock, PA 16057

Abstract
Measurement of Type I and II error associated with record linkage is an ongoing research topic. Most authors have either relied on a "gold standard" identifier for measuring error, post-linkage data validation through finding implausible or impossible linked data (e.g., two or more death records), sensitivity analysis, or comparison of linked and unlinked data to explore biases. This paper presents an alternative method for measurement of both Type I and Type II error through a direct estimation of the probability of a match, dependent on *m* and *u* probability values determined during the matching process. Simulated data are used to show the accuracy of the method in estimating both types of error.

Key Words: record linkage; administrative records; type I error; type II error; empirical Bayes

1. Introduction

Record linkage—that is, the analysis of joined data sets in order to infer which pairs represent the same entity—is widely used in multiple fields, including census coverage measurement, longitudinal medical studies, economics research, genealogy, and survey enhancement.[1] However, a universally accepted standard for measuring the quality of a specific record linkage method does not exist. The most comprehensive outline for such a standard was published in Bohensky et al. (2011); they propose a set of 14 guidelines for appraising the quality of record linkage which are grouped into four domains: 1) the quality of the datasets to be linked; 2) data preparation and choice of linkage variables; 3) linkage process and technology used; and 4) ethical considerations. Measurements of linkage quality are included in the third domain, and false positive and false negative rates are specifically suggested; however, methods for measurement of these error rates are not outlined. A framework given in Gilbert et al. (2017) for information that should be provided with linked data sets also lists estimates of linkage error rates and methodology for those estimates as essential information; use of positive predictive value (1-false discovery rate) or specificity (1-Type I error rate),[2]

[1] For specific references, please see Asher (2017).

[2] Because the literature on record linkage spans a wide variety of fields, different measures of linkage error are found in different papers. The most common measure related to false positive linkage in health-related research is the positive predictive value, which is equal to 1 – the false discovery rate. This measure includes, in its denominator, the number of links rather than the number of true matches; it measures the total number of true positives out of the total number of links. In the health context, the positive predictive value is the best measure (a patient is much more interested in the probability they have a disease given that they test positive for it than the probability they test positive for a disease given that they have the disease). In other contexts, such as census coverage measurement, the interest is in the probability that the records link given they are not a true match, or the Type I error rate (1–specificity); in this case the denominator is the total number of true nonmatches. Because the authors of this paper work primarily in a

sensitivity (1-Type II error rate), and the F-measure[3] are suggested, again without specific methodological guidelines.

An exhaustive review of the tremendous literature around record linkage is close to impossible—literally tens of thousands of papers have been published since Dunn introduced the topic in 1946 and Fellegi and Sunter formalized probabilistic linkage in 1969.[4] Limiting the review to papers that focus on linkage error measurement still yields a wide array of papers outlining multiple potential methods, from use of a gold standard to simulation methods to complex modeling procedures.

The remainder of this introduction will introduce the reader to the most common methods used for the estimation of Type I/false discovery and Type II linkage error rates given in the literature. Section 2 will first describe a mathematical model for the direct estimates of the overall probability of a match, and then describe the creation of a synthetic data set that is subsequently used compare this method to use of a partially available gold standard. Section 3 will outline the results of the test, comparing the two methods. Finally, Section 4 will discuss future research directions.

1.1 Gold Standard

A gold standard, in this context, is a unique identifier, typically available for only a subset of the data records to be matched, that can be used to verify the results of a probabilistic matching algorithm. Because the unique identifier is available for only a subset of the records, it can not be used as part of the probabilistic record linkage process. After the matching process occurs, the data are matched again deterministically using the gold standard variable, and the results of the two matches are compared. Records that are matched by the probabilistic record linkage process but not the deterministic process are false positives; those matched by the deterministic process but not the probabilistic process are false negatives. From counts of these matches, Type I and Type II error rates can be determined.

The literature contains multiple examples of gold standard-based linkage quality measurement; however, most of the papers found during our literature review come from a cohort of researchers in England. Hagger-Johnson et al. (2015) uses the Paediatric Intensive Care Audit Network (PICANet) Patient Identification Number as a gold standard to test a pseudonymization algorithm for maintaining confidentiality during record linkage. They then study in detail which variables in the dataset are associated with the greatest levels of false matches.

Harron et al. (2014), Harron et al. (2017) and Aldridge et al. (2015) all use National Health Service (NHS) numbers—unique identifiers assigned to individuals within the United Kingdom—as a gold standard to assess the quality of a record linkage process. Harron et

---

demographic context, we focus on the measurement of Type I and Type II error; however, the methods given in this literature review are relevant regardless of the specific error measurements preferred.

[3] The F-measure, first proposed for use with record linkage by Christen and Goiser (2007), balances the sensitivity (1–Type II error) and the positive predictive value by creating a weighted harmonic mean of the two.

[4] This paper is intended for readers that are very familiar with the Fellegi-Sunter model and the basic steps of probabilistic record linkage; see Herzog et al (2007) for an introduction to these topics.

al. (2014) first uses NHS number to create a gold standard linkage between data from the Birmingham Children's hospital and the pediatric intensive care unit (PICU) of the Great Ormond Street hospital. They then use those data sources to simulate data collected under different conditions (e.g., national infection surveillance data, which would not contain unique identifiers or other important variables). The simulated data is linked, and the gold standard data used to assess the quality of the linkage. They are able to show that estimates created from the matched data are most biased by non-random error.

Harron et al. (2017) focuses on an automated linkage between mothers and their babies across administrative data records from NHS hospitals in England. The gold standard data were available through obstetrics records containing both the mother and baby NHS numbers in a single record.

Aldridge et al. (2015) tests an enhanced matching system by matching reports of tuberculosis in the Public Health England database to positive tuberculosis tests from reference laboratories in the United Kingdom. In this case, the gold standard NHS number data are present in only some of the records; therefore, statistical properties of the records containing NHS are compared to those of records missing NHS. Differences were noted in the following populations: those 65 and older, men, those missing ethnicity data, those missing whether they were born in the UK or not, and those missing social risk factor data—that is, whether or not they used drugs or alcohol, were homeless or had spent time in prison. As such, the false discovery and Type II error rates reported – which were extremely low – were most likely biased.

Moore et al. (2014) of Australia, use measures of sensitivity (1–Type II error) and specificity (1–Type I error) to assess record linkage between population-based studies of inmates in NSW and the Australian National Death Index. Their gold standard is the results of a sub-study of New South Wales (NSW) prison inmates which provides vital status at the end of the study period. They then use the calculated sensitivity and specificity to create an adjustment factor for the total count of inmate deaths.

Randall et al. (2018), also of Australia, probe the issue of differential linkage quality across different demographic groups in more detail. Each of four separate administrative health datasets is de-duplicated using an automated probabilistic record linkage algorithm, and the results are compared to a previously created "truth-set" of the same data that has undergone probabilistic record linkage, intensive manual review, and multiple quality assurance processes. The authors find significant differences in linkage quality for individuals born after 1980 but note that differences in linkage quality varied over gender and socioeconomic characteristics in only one or two of the four datasets. In three of the datasets, linkage error was higher in remote geographic areas. Their overall conclusion is that linkage error rates across demographic groups is highly list-dependent in this context.

While most research on linkage error has occurred in high-income countries (i.e., North America, Australia, and western Europe), Rentsch et al. (2018) focus on data from Tanzania, an area where linkage errors are likely to be more frequent and have a greater effect on subsequent analyses due to poorer data quality. They create a gold standard database using a point-of-contact interactive record linkage (PIRL); within this method linkages are created between records during researcher's interactions with the person whose records are being linked. Using the PIRL-created data, they then attempt an automated record linkage on the same records to determine error rates. The automated match, at a minimum match score threshold, yields a false discovery error rate of 39%.

They also found significant differences between estimates created with the correctly matched data and the false-match data, suggesting that significant biases are created through the automated record linkage.

While gold standard studies have offered a window into the types and extent of biases created by linkage error, there are some significant drawbacks to this method. The most significant is that the gold standard data are almost always only available for a subset of the population, and missing gold standard data is an indication of lower data quality overall for a record. Studies that have compared characteristics of the records for which gold standard data are available to the records with no gold standard data have noted significant differences in both the socio-economic profile of the two groups and also the resulting estimates for the two groups. The gold standard method is therefore believed to be underestimating the level of error caused by the record linkage process. An additional issue is the assumption of accuracy within the gold standard data; although the gold standard data are of very high quality, they also contain errors which effect the accuracy of the Type I/false discovery and Type II error rate calculations.

1.1.1 Direct comparison of different record linkage techniques/Sensitivity analysis
Like gold standard studies, direct comparisons across different record linkage techniques compare links made through one method to the links made through alternative methods; Type I/false discovery and Type II error rates are calculated by noting the links that are not consistent across the methods or by using auxiliary gold standard data. For example, Monga and Patrick (2001) compare record linkage based on different combinations of variables (first name, middle name, last name, and date of birth), designating one combination as the reference method (last name, first name, date of birth) and the others as comparison methods. They then focus on the methods that provide supersets of links when compared to the reference method, and manually review the additional links produced by the comparison methods to determine that the reference method missed 4% of the links caught through the comparison methods.

Harron et al. (2017) completes a sensitivity analysis by altering either the linkage algorithm or match weight thresholds and studying the impact on the results of the record linkage process. They find increasing the match weight threshold in probabilistic linkage reduces the likelihood of false positives (Type I error) and increases the likelihood of false negatives (Type II error). They then compare linkage results for multiple variables across four linkage algorithms—gold standard linkages, probabilistic linkage, high-threshold probabilistic linkage, and deterministic linkage only—measuring Type II error rate and positive predictive value. Not surprisingly, they find positive predictive value is maximized for deterministic linkage only, while Type II error rate is minimized for original probabilistic linkage. By comparing the statistics for multiple variables across the linkage techniques, they show that there is little bias in this example caused by changes in linkage algorithms.

Hagger-Johnson et al. (2017-2) uses the combination of a gold standard and a comparison of two linkage techniques (deterministic linkage versus deterministic linkage followed by a probabilistic linkage step) to explore linkage error. Specifically, a reference standard dataset is created from Hospital Episode Statistics (HES) data collected from the NHS hospitals in England and then compared to the results from the two linkage techniques applied to a subset of variables from the same data source. Overall, the addition of the probabilistic step was shown to increase sensitivity (1 – Type II error) without significant changes in specificity (1 – Type I error).

1.2 Data dependent analysis

Some of the methods for exploring Type I and Type II linkage error given in the literature are specific to the particular type of data being linked. Harron et al. (2017-2) proposes exploring linkage quality by identifying implausible scenarios within the matched data; for example, a hospital admission linked to a death record where the date of the hospital admission follows the date of death indicates a false match, or the linking of multiple death records to a single hospital visit record. Using this method is highly dependent on the type of data being linked, and it is not exhaustive—a hospital admission could still be incorrectly linked to a subsequent death. Calculations of Type I and Type II error rates via this method would therefore be underestimates. However, in combination with other methods, implausible data can still help determine if a record linkage technique contains serious flaws.

A specific example of this technique is given in Hagger-Johnson et al. (2015-2). Record linkage is used to find multiple admittances to the hospital in the UK HES data for infants from 2011-2012 and for adolescents from 2005-2011. Each episode of care is assigned a unique HESID. Six scenarios are chosen to indicate a potential false match, including multiple births with the same HESID, re-admission after death, simultaneous admission at different hospitals, and adolescent admissions coded as births. They found .1% possible false matches (false discovery rate), and determined that there was a differential pattern to false matches, with missing gestational age, preterm birth, or Asian ethnicity being most strongly associated with false match status among babies, and male gender, younger age, mixed ethnicity, or missing ethnicity being most strongly associated with false match status among adolescents. The authors also note significant variation among hospitals' false match rates.

A different example of data-dependent analysis of Type I error is given by Blakely and Salmond (2002). They note that in the case where there can be only one true match per individual between two datasets and one dataset's records greatly outnumber the other dataset's records—for example, in the linkage between census records and mortality records in the New Zealand Census-Mortality Study (NZCMS)—then analyzing the duplicate matches yields information about false match rates. The system they use balances the information provided by the match weights of the duplicates and the count of duplicates to create a decision rule for tallying false positives. They then determine the positive predictive value for different match weight cutoff values; the estimated false discovery rates vary from < 1% for the top 9 cutoffs, down to 98% for the lowest cutoff value. Because Type II error is not calculated, no method for optimizing the cutoff value is proposed.

1.3 Simulation Studies

Simulation studies have been used in multiple contexts to test error rates for a specific record linkage process. In simple terms, synthetic data are developed with known match patterns; the matching process is implemented, and then the results of the probabilistic match process are compared to the known match pattern to estimate linkage error rates.
Hagger-Johnson et al. (2016) and Trentin et al. (2018) both create synthetic data to test record linkage processes. Hagger-Johnson et al. (2016) test the algorithm used to create HES in England and find that the number of missed matches is reduced by using a deterministic match followed by a probabilistic match. Trentin et al. (2018) focus on the need to create synthetic data that reflects the naming characteristics of Brazilian families; they determine that standard data generation techniques do not have the flexibility required in the Brazilian context to create quality synthetic data for testing record linkage processes.

1.4 Model dependent analysis

The techniques discussed so far rely on empirical methods to determine linkage error rates. There is a body of papers, however, that focus on modeling methods for calculating linkage error; these papers start with the Fellegi-Sunter model (Fellegi and Sunter 1969) and use the match weights generated by that procedure [5] to create estimates of the match probabilities assuming conditional independence of the data fields; linking error rates can be derived from the estimated match probabilities. From this information, the match cut-offs can be optimized to balance Type I and Type II error rates.

Up through the 1990s, most efforts to model match probabilities and error rates followed one of two approaches: direct or indirect. Typically, in the direct approach, an indicator variable of match status (match or nonmatch) is the dependent variable in a logistic regression with match weights as the predictor variable. In the indirect approach, discriminant analysis is used to model the match weights as iid under the following model:

$$\begin{aligned} f(W_i, Z_i|\phi, \lambda) &= h(W_i|Z_i, \phi)g(Z_i|\lambda) \\ &= h(W_i|Z_i, \phi)[\lambda^{Z_i}(1-\lambda)^{(1-Z_i)}] \end{aligned}$$

where W_i is the match weight and Z_i is the indicator variable for match status. The first component of the model is a conditional normal distribution of W_i with different means but a common variance. The second component is the marginal probability of Z_i; λ and ϕ are assumed independent a priori (Belin and Rubin 1995). The posterior distribution of Z_i is used to estimate match probabilities. However, both of these approaches are still dependent on manual review to determine appropriate match cut-off values.

Belin and Rubin (1995) were the first to combine the previous modeling efforts into an unsupervised process (i.e., requiring no clerical data).[6] They use a single mixture model for the purpose of estimating error rates. The underlying assumption is that the observations (match weights for each pair) will either fall into the distribution representing the true matches (f_T) or the one representing the false matches (f_F); f_T and f_F each represent a different transformed-normal distribution; that is, each distribution of match weights is normal after different Box-Cox transformations are applied to them. The final likelihood for the mixture model takes the form:

$$\begin{aligned} &L(\lambda, \mu_T, \sigma_T^2, \varpi_T, \gamma_T, \mu_F, \sigma_F^2, \varpi_F, \gamma_F | \, W_1, \dots, W_n; Z_1, \dots, Z_n) \\ &\qquad = \prod_{i=1}^{n} [\lambda f_T(W_i | \mu_T, \sigma_T^2, \varpi_T, \gamma_T)]^{Z_i} [(1-\lambda) f_F(W_i | \mu_F, \sigma_F^2, \varpi_F, \gamma_F)]^{(1-Z_i)} \end{aligned}$$

where $\mu_T, \sigma_T^2, \mu_F, \sigma_F^2$ are the means and variances of f_T and f_F, respectively, ϖ_T, γ_T, ϖ_F, γ_F are the parameters for the Box-Cox transformations, and the Z_i s are missing/unknown. To implement a modeling procedure, Belin and Rubin use the results of a previous linkage which has been thoroughly reviewed as a "gold standard" to create estimates for the Box-Cox parameters and the ratio of the variances. This allows them to utilize an expectation-maximization (E-M) protocol to estimate the remaining parameters. False match rates (Type I error) for specific match cut-offs can then be estimated. However,

[5] The original Fellegi-Sunter model did not include explicit measurement of linkage error rates.
[6] In the computer science literature, this is an example of unsupervised machine learning.

this method is reliant on the quality and type of data; in cases where f_T and f_F overlap too much the usefulness of the method is compromised.

Larsen and Rubin (2001) expand this methodology to accommodate a wider variety of record linkage situations, but demonstrate the method using census data. Their method relies on a three-component mixture model; in their example, one component represents pairs that agree on all or almost all fields (the true match cluster), one component represents pairs that mostly agree on household characteristics but mostly disagree on personal characteristics (one non-match cluster), and the last component represents pairs that mostly disagree on all fields (the second non-match cluster). After initial model parameter values are selected based on expert opinion, they iterate between E-M fitting of the mixture model and clerical review to allow optimal parameter estimation.[7] They point out that mis-specified models in the initial step are corrected by later clerical review steps, so the parameter estimation is based solely on the current linkage process (and not previous data).

Winglee et al (2005) use a simulation method (SimRate) to generate data based on multinomial distributions similar to the ones we use in our match rate method in Section 2 of this paper; however, their approach is based on separate estimated cumulative probability functions for the matched and nonmatched pair weights. They determine Type I and Type II error levels at thresholds along the cumulative functions and choose the final threshold to minimize both types of error. They caution that their method works for data with dependencies between fields only if SimRate generates similar data incorporating any dependencies. Finally they compare their method to the one proposed by Belin and Ruben (1995) and found different Type I error estimates across the methods.

Winkler (2007) and Winkler (2014) reframe the modeling of linkage error in terms of Naïve Bayes classification/machine learning and also use census data to demonstrate the modeling procedure. Winkler's model partitions the possible data agreement patterns[8] into three classes similar to the three-component mixture model in Larsen and Ruben: matches within household (true matches) non-matches within household (nonmatches), and non-matches outside household (nonmatches). However, Winkler's model incorporates individual probabilities of agreement for each field and uses the Naïve Bayes framework to incorporate prior distributions for the parameters to be estimated. The model is constructed so that either semi-supervised or unsupervised learning is possible; however, note that some form of training data is required. To improve the classification results, Winkler uses an augmented fitting technique – an E-M algorithm that has been constrained to a closed convex region on the parameter space (EMH). In other words, it applies constraints on the values the parameters can take; for example, ensuring the sum of the probabilities of the four possible outcomes (True positive, false positive, true negative, false negative) sum to one.

Fiegenbaum (2016) uses a supervised machine-learning approach to develop correct matching across historical census records. His algorithm requires a relatively small manually coded training set. He starts by training a probit model, using bootstrap samples from his training data, to assign each pair a probability of being a true match. As a second

[7] In the computer science literature, this is an example of semi-supervised machine learning.

[8] The data agreement pattern is the agreement results across the fields being used during the record linkage. For example, in a linkage using four fields (e.g., first name, middle name, last name, and date of birth), agreement pattern 1111 would mean every field matches and agreement pattern 0000 would mean every field does not match.

stage, he then selects out pairs that have the highest probability of being a match and have no high-probability competing pairs; that is, pairs that contain the same record. To select his cutoff thresholds, he cross-compares the true positive rate and the positive prediction rates for the algorithm, maximizing a weighted sum of the two measures.

Finally, Tuoto (2016) returns to the initial Fellegi-Sunter model but adds a step in which estimation of linkage error occurs through a logistic regression with auxiliary fields (i.e., fields not used in the original linkage) as the independent variables. Using a training sample, she calibrates the variables in the logistic regression model, and then applies the calibrated model to the record linkage results, generating match probabilities for all pairs. From these, linkage error rates can be calculated.

2. Methodology

2.1 Mathematical Justification for Estimates of Match Probabilities

The Fellegi-Sunter process is a probabilistic record linkage. It starts with the assumption of two lists—for our purposed, call them list A and list B. Every record in list A must be checked against every record on list B to determine if there is a match. If there are N_A records on list A and N_B records on list B, then there are $N_A \times N_B$ pairs of records. However, only a small subset of the possible pairs represents the same underlying entity.

Let a match be a pair that represents the same underlying entity. For each pair, the process will compare several identifiers (e.g., first name, last name, date of birth, address, etc.), and determine whether they agree or not. Agreement of a particular identifier does not guarantee that the pair is a match – for example, agreement on gender category provides very little information as to match status. Keeping this in mind, let

$$u_i = \Pr(\text{identifier i agreement} \mid \text{non-match}) > 0$$
$$m_i = \Pr(\text{identifier i agreement} \mid \text{match}) \approx 1$$

The *u*- probabilities are greater than zero because of the possibility of spurious agreement (e.g., two people with the same first name). The *m*- probabilities are not exactly one because we allow error in the identifiers.

By Fellegi-Sunter, the agreement weight for identifier *i* is $\frac{\ln(\frac{m_i}{u_i})}{\ln(2)}$, and the non-agreement weight for identifier *i* is $\frac{\ln(\frac{(1-m_i)}{(1-u_i)})}{\ln(2)}$. A pair weight is then the sum of the agreement and non-agreement weights for its identifiers. The pairs are then listed in order by their pair weights; a threshold value is found above which all pairs are believed to be matches, and a second threshold value is found below which all pairs are believed to be non-matches. Between these two thresholds, clerical review or another process is used to determine whether the pairs are believed to be matches.

For clarity, we label a pair as a "link" when it is determined to represent the same entity by the Fellegi-Sunter process. Think of a "link" as being the estimated status, and a "match" as being the true status. Type I error is then the pairs that are a link but not a match, and Type II error is the pairs that are not a link but are a match.

More detail can be found in Fellegi and Sunter (1969).

2.1.1 Calculating the Match Probabilities for Specific Identifier Agreement Vectors
Our method is an extension of the Fellegi-Sunter process; it can either be completed after the *m*- and *u*- probabilities are calculated as an auxiliary procedure, or be incorporated into the record linkage (i.e., E-M algorithm for fitting the *m*- and *u*- probabilities).

Let $i \in \{1, \cdots, n\}$ be the index for field i.
Let $j \in \{1, \cdots, 2^n\}$ be the index for agreement vector j.
Let a_{ij} = 1 if there is agreement on field i for agreement vector j;
0 if there is non-agreement on field i for agreement vector j
Let $A_j = \{a_{1j}, a_{2j}, \cdots, a_{nj}\}$ be agreement vector j

Then $\forall A_j$,

$$P(Aj|Match) = \prod_{i=1}^{n} {m_i}^{a_{ij}} (1 - m_i)^{(1-a_{ij})}$$

and

$$P(Aj|Nonmatch) = \prod_{i=1}^{n} {u_i}^{a_{ij}} (1 - u_i)^{(1-a_{ij})}$$

and recall from above,

$$\text{Pair Weight}(A_j) = \log_2\left(\frac{\prod_{i=1}^{n} {m_i}^{a_{ij}} (1 - m_i)^{(1-a_{ij})}}{\prod_{i=1}^{n} {u_i}^{a_{ij}} (1 - u_i)^{(1-a_{ij})}}\right) \quad (1)$$

Now let

$N_j =$ the count of pairs with agreement pattern A_j

$N_{Pairs} = \sum_j N_j =$ the known count of pairs

$X_j =$ the unknown count of agreement vector *j* matches among all N_j pairs

$X_{Match} = \sum_j X_j$
$=$ the unknown count of matches among all pairs

Note that:

$$\{X_1, \cdots, X_{2^n}\} \sim \text{Multinomial}\left(X_{Match}, \prod_{i=1}^{n} {m_i}^{a_{i1}} (1 - m_i)^{(1-a_{i1})}, \cdots, \prod_{i=1}^{n} {m_i}^{a_{i(2^n)}} (1 - m_i)^{\left(1-a_{i(2^n)}\right)}\right)$$

And:

$$E(X_j) = X_{Match} \prod_{i=1}^{n} {m_i}^{a_{ij}} (1 - m_i)^{(1-a_{ij})}$$

There is a hierarchy in that X_{Match} also follows a probability distribution as follows:

$$X_{Match} \sim \text{Binomial}(N_{Pairs}, p)$$

where p is an unknown hyperparameter. In an empirical Bayes or full Bayes context, we can then estimate p as

$$\mathrm{P(Match)} = p = \frac{\mathrm{X_{Match}}}{\mathrm{N_{Pairs}}} \tag{2}$$

To extend this methodology to a full Bayesian framework, the joint distribution of the $(\mathrm{N_j}-\mathrm{X_j})$s can be included and a prior distribution for p can be added as a third layer in the hierarchical model.

By Bayes Theorem, $\forall \mathrm{A_j}$,

$$\begin{aligned}\mathrm{P}\big(\mathrm{Match}|\mathrm{A_j}\big) &= \frac{\mathrm{P}\big(\mathrm{A_j}|\mathrm{Match}\big)\cdot \mathrm{P(Match)}}{\mathrm{P}\big(\mathrm{A_j}\big)} \\ &= \frac{\prod_{i=1}^{n} {m_i}^{a_{ij}}\,(1-m_i)^{(1-a_{ij})}\cdot \frac{\mathrm{X_{Match}}}{\mathrm{N_{Pairs}}}}{\frac{\mathrm{N_j}}{\mathrm{N_{Pairs}}}} \\ &= \frac{\prod_{i=1}^{n} {m_i}^{a_{ij}}\,(1-m_i)^{(1-a_{ij})}\cdot \mathrm{X_{Match}}}{\mathrm{N_j}}\end{aligned}$$

Similarly,

$$\mathrm{P}\big(\mathrm{Nonmatch}|\mathrm{A_j}\big) = \frac{\prod_{i=1}^{n} {u_i}^{a_{ij}}\,(1-u_i)^{(1-a_{ij})}\cdot (\mathrm{N_{Pairs}}-\mathrm{X_{Match)}}}{\mathrm{N_j}}$$

Then,

$$\begin{aligned}\mathrm{Odds}\big(\mathrm{Match}|\mathrm{A_j}\big) &= \frac{\mathrm{P}\big(\mathrm{Match}|\mathrm{A_j}\big)}{\mathrm{P}\big(\mathrm{Nonmatch}|\mathrm{A_j}\big)} \\ &= \frac{\prod_{i=1}^{n} {m_i}^{a_{ij}}\,(1-m_i)^{(1-a_{ij})}}{\prod_{i=1}^{n} {u_i}^{a_{ij}}\,(1-u_i)^{(1-a_{ij})}}\cdot\frac{\mathrm{X_{Match}}}{\mathrm{N_{Pairs}}-\mathrm{X_{Match}}}\end{aligned}$$

and by (1)

$$\begin{aligned}\log_2\big(\mathrm{Odds}\big(\mathrm{Match}|\mathrm{A_j}\big)\big) &= \log_2\Big(\frac{\prod_{i=1}^{n} {m_i}^{a_{ij}}\,(1-m_i)^{(1-a_{ij})}}{\prod_{i=1}^{n} {u_i}^{a_{ij}}\,(1-u_i)^{(1-a_{ij})}}\cdot\frac{\mathrm{X_{Match}}}{\mathrm{N_{Pairs}}-\mathrm{X_{Match}}}\Big) \\ &= \mathrm{Pair\ Weight}\big(\mathrm{A_j}\big) + \log_2\Big(\frac{\mathrm{X_{Match}}}{\mathrm{N_{Pairs}}-\mathrm{X_{Match}}}\Big)\end{aligned}$$

and

$$\mathrm{Odds}\big(\mathrm{Match}|\mathrm{A_j}\big) = 2^{\mathrm{Pair\ Weight}\left(\mathrm{A_j}\right)+\log_2\left(\frac{\mathrm{X_{Match}}}{\mathrm{N_{Pairs}}-\mathrm{X_{Match}}}\right)}$$

Using the equation for the conditional odds, we can calculate that:

$$\mathrm{P}\big(\mathrm{Match}|\mathrm{A_j}\big) = \frac{\mathrm{Odds}\big(\mathrm{Match}|\mathrm{A_j}\big)}{1+\mathrm{Odds}\big(\mathrm{Match}|\mathrm{A_j}\big)}$$

and then

$$\mathrm{P(Match)} = \sum_{j} \mathrm{P}\big(\mathrm{Match}|\mathrm{A_j}\big) \tag{3}$$

Therefore, we can equate (2) and (3), find

$$X_{Match} = N_{Pairs} \sum_{j} \frac{2^{\text{Pair Weight}(A_j)+\log_2(\frac{X_{Match}}{N_{Pairs}-X_{Match}})}}{1+2^{\text{Pair Weight}(A_j)+\log_2(\frac{X_{Match}}{N_{Pairs}-X_{Match}})}}$$

and we can solve empirically for X_{Match} and find $E(X_j)$ for all j. We assume that either the *m*- and *u*- probabilities were previously estimated during the matching procedure, or will be estimated in conjunction with *p*.

Using X_{Match} and the $E(X_j)$s, we can determine the probabilities of matching for each agreement vector A_j, and then use that information to determine the proportion of links with that agreement vector that are true matches. This allows a Type I error calculation for each agreement vector that represents a link, and a Type II error calculation for each agreement vector that does not represent a link. Overall Type I and Type II error for the record linkage process can then be determined.

2.2 Simulation Methodology

2.2.1 Creating the synthetic data

The first step in making the simulation was to set the number of comparison fields, which we varied systematically (as will be described in Table 1, below) from 4 to 10 by 2. Then for each of the comparison fields, we randomly assigned the *m*- and *u*-probabilities as random draws from the following distributions:

$$m_i = \mathrm{P}(a_i = \text{Agree}|\text{Match}) \sim \text{Uniform}(.85, 1)$$
$$u_i = \mathrm{P}(a_i = \text{Agree}|\text{Non-match}) \sim \text{Uniform}(0, .1)$$

These ranges were chosen, based on our experience, to be similar to the parameter estimations in actual record linkage analyses.

Next we set the number of total pairs, and the percentage of pairs that were matches. For total number of pairs we varied the count from 200,000 to 20 million; for proportion of matches we varied from 0.0025 to 0.05. Although in practice it is more likely to find a higher proportion of matches, we wished to test all areas of the parameter space to understand when the model works well and when it is likely to fail.

For each of the matched pairs to be generated, we simulated the agreement statuses for each of the fields by Monte-Carlo draw from the distribution Bernoulli(m_i), where *i* is the comparison field index. Likewise for the non-matched pairs, we simulated the agreement statuses for each of the fields from the distribution Bernoulli(u_i). Note that the distributions for each of the comparison fields are independent.

Table 1: Parameters varied to simulate used of Gold Standard and Exact Match Probability Estimation.

Parameter	# Values Tested	Values tested
# of Comparison Fields	4	4, 6, 8, 10
Total Number of Pairs	5	200,000; 500,000; 2 million; 5 million; 20 million
Proportion of Matches	4	0.0025, 0.005, 0.01, 0.025
ID Present Proportion	5	0.1, 0.3, 0.5, 0.7, 0.9
Basal Error Rate	5	0.0025, 0.005, 0.01, 0.02, 0.05

Next, to simulate situations where the comparison field data is unavailable, we assigned missing status to the comparison fields at a rate of *i* x .02, where *i* is the index for the comparison fields as above. Thus, based on a random draw from Bernoulli(.02) we set the a_{1j} status (whether originally being Agree or Disagree) to missing. For the second comparison variable the draw was from Bernoulli(p=.04), and so forth.

Several additional parameters that relate to gold standard error estimates were also generated at this point. Since these were simulated pairs, we knew which pairs had been simulated as matches and which had been simulated as non-matches. The gold standard reflected this classification with the variable ID_Agree, (where, ID_Agree = 1 for matched pairs, and ID_Agree = 0 for non-matches), with several adjustments meant to simulate the limitations in the level of information the gold standard provides.

First, we specified that only a fraction of the pair records had two unique ID fields (one from each of the files being merged) that were available for comparison. For each simulation, this fraction was set at a level that varied systematically (from 0.1 to 0.9 by 0.2) over the simulation runs, as given in Table 1. If a draw from a Bernoulli(ID Present Proportion) = 0 then we set the ID_Agree field to missing, regardless of whether it was associated with a matching or non-matching pair.

Additionally, we knew from practical linkage experience that in some cases the unique IDs disagree even for a matched pair (i.e., records for the same person generally should having agreeing SSNs, but in the case of transcription errors, they do not). We termed the rate of such errors the basal error rate, and we simulated it by switching from ID_Agree = 1 to ID_Agree = 0 when a draw from a Bernoulli(Basal-Error-Rate) = 1.

Across the five parameters defined in Table 1, there are 4 x 5 x 4 x 5 x 5 = 2000 possible combinations. For each simulation run, each unique combination of parameter values was repeated 20 times. These 20 repeated runs were not identical; the *m*- and *u*- probabilities for each comparison field were set by random draw during each run as described above. For each run of the simulation, the *m*- and *u*- probabilities differed, the random draws made from them to generate the comparison fields differed, and the random draws to set the ID_Agree variable (i.e., to apply missingness and transcript error) differed. Therefore, even when all of the five systematically varied parameters had the same values, in each of the 20 runs for that combination of parameter values, results differed.

2.2.2 Running the simulation

To simulate the estimation of *m*- and *u*- probabilities we applied the E-M algorithm, which also generates an estimate of the total number of matched pairs (X_{Match}) among all the pairs. By assuming independence of the comparison variables, it is a straightforward to compute a match proportion, P(Match) for each of the realized comparison vectors (including those that have missing values). Then, for each simulation, we assigned pairs as links based on whether this probability was greater than a linking match-probability threshold, which was varied systematically, from 0.60 to 0.95 by 0.05.

In other words, for a given simulation, we would first determine links meeting the P(Match)>0.60 threshold. Next we would determine links meeting the 0.65 threshold, which would include all pairs with an agreement pattern estimated at 0.65 or greater, and likewise repeat this process up to the 0.95 threshold. While this link acceptance threshold can be thought of as an additional simulation parameter, unlike for the other parameters,

the 8 levels of the link acceptance threshold, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95 are all run on the exact same simulated pairs.

2.2.3 Error computation

For each simulation run and each link acceptance cut-off, we can compute the true Type I and Type II error rates based on the known match status. That is, for Type I error, we are looking at what proportion of the links (i.e., pairs with match probability greater than the link acceptance threshold) are not matches. For Type II error, we are looking at what proportion of the matches are not linked.

Again, for each simulation run and link acceptance cutoff, we can estimate the Type I and Type II error rates by two methods: gold standard or match probability.

For the gold standard method we compute the Type I error by computing

$$1 - \sum_{\substack{\text{Links} \\ \text{above} \\ \text{cutoff}}} \frac{\text{ID\_Agree}}{\text{Count of Links above Cutoff}} - \left(1 - \sum_{\substack{\text{Links in} \\ \text{maximum} \\ \text{agreement} \\ \text{category}}} \frac{\text{ID\_Agree}}{\text{Count of Links in Maximum Agreement Category}} \right) \tag{4}$$

In (4), the second term is the proportion of links with ID_Agree=1 out of all the links above the cutoff threshold. The third term is an estimate of the Basal Error rate; taken as the complement of the proportion of links with ID_Agree=1 in the category for the agreement vector Aj with the highest pair weight.

An example might clarify. Assume we have six comparison variables and when all of these variables have status=AGREE then

$$\sum_{\substack{\text{Links in} \\ \text{maximum} \\ \text{agreement} \\ \text{category}}} \frac{\text{ID\_Agree}}{\text{Count of Links in Maximum Agreement Category}} = 0.995$$

Then the basal error rate would be estimated as 1 – 0.995 = .005. Assume we then calculate the

$$\sum_{\substack{\text{Links} \\ \text{above} \\ \text{cutoff}}} \frac{\text{ID\_Agree}}{\text{Count of Links above Cutoff}} = 0.982$$

for a cutoff threshold of 0.90. The estimated Type I error rate is then

$$1 - 0.982 - 0.005 = 0.013.$$

For the gold standard, we compute the Type II error by first finding the difference between the estimated total of matches among all the pairs (which is computed as $\sum$ ID_Agree for all pairs) and the estimated total of matches among the links (which is computed as the $\sum$ ID_Agree for links above the cutoff). We then divide this difference by the estimated total matches among all the pairs.

For the match probability method, we compute the Type I error rate as the estimated number of matches not linked divided by the estimated number of matches overall. Both values are calculated by $\sum$ P(Match) for the pairs within the category. For the match probability method, we compute the Type II error rate by first finding the difference between the estimated count of matches and the estimated count of matches accepted as links, and then dividing by the estimated number of matches.

The basis for evaluating the results of the simulation is comparing the estimates of Type I and Type II error to the true error rates for the two methods under analysis, gold standard and match probability. At this stage of the analysis, we believe that a graphical comparison of estimated and true error rates is most useful in exploring the limits and strengths of the two proposed methods.

3. Simulation Results

3.1 Type I error

For the estimation of Type I error, we produced the plots given in Figure 1. Please note the Method 1 is the Gold Standard Method and Method 2 is the Match Probability Method.

The overall plots across the entire parameter space show only very limited correlation between the estimates and actual values of Type I error. While both plots appear to have an upward trend, their level of dispersion from the optimal distribution directly on the 45° line would make their use in actual record linkage analysis difficult. However, we find that by reducing the parameter space to include only those simulations that have 8 or 10 comparison variables and 2,000,000 pairs under analysis, with 2.5% of them being matches, that by-and-large the fit of estimated Type I error from the match probability analysis is quite good, falling in a tight band around the y = x line.

However, for the gold standard method, the fit is still questionable. These results show that for common record linkage scenarios (i.e., with many comparison variables and matched pairs), the match probability method has potential to be used an estimator of Type I error.

3.2 Type II error

Turning to Type II error, we see it is that for the full parameter space, it is the gold standard that is producing fairly precise estimates, but that the match probability method is quite diffuse.

However, again we follow-up to this analysis by the same reduction in parameter space that was applied to the Type I error, and here we see it is the gold standard error estimates that have a relatively imprecise relationship with the true error rates, but that for the match probability method, with the exception of certain strays, the fit is quite nice.

Overall, we see that the match probability method shows potential as a source of error estimates if care is taken to apply it within the correct parameter space.

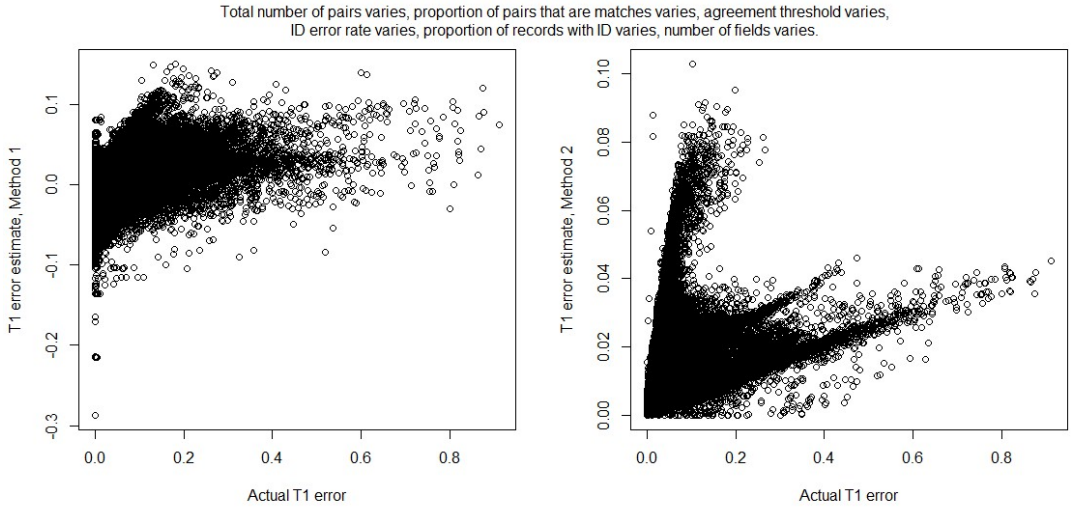

Figure 1: *Comparison of actual Type I error rates to estimated Type I error rates for two methods of estimation.*

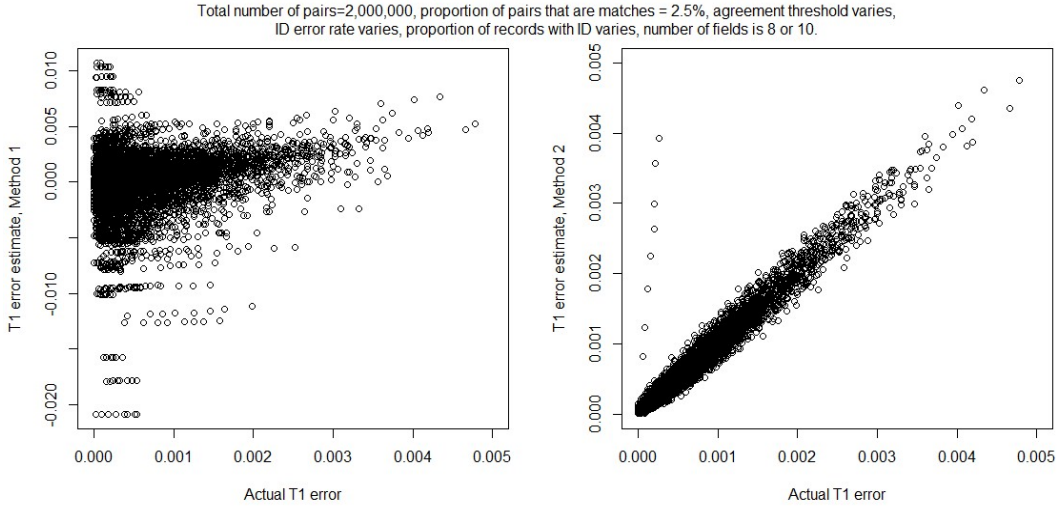

Figure 2: Comparison of actual Type I error rates to estimated Type I error rates for two methods of estimation; constrained simulation parameter space.

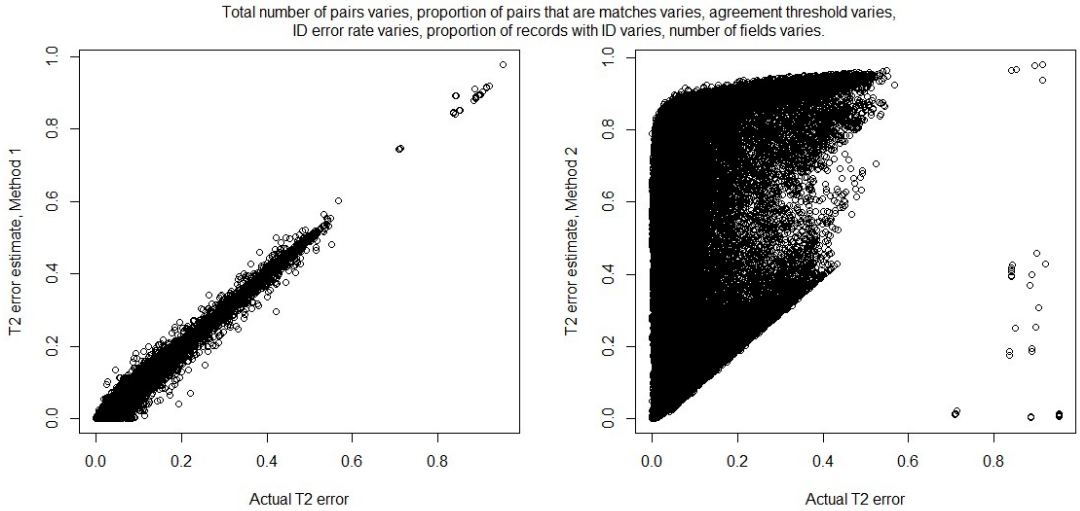

Figure 3: Comparison of actual Type II error rates to estimated Type II error rates for two methods of estimation.

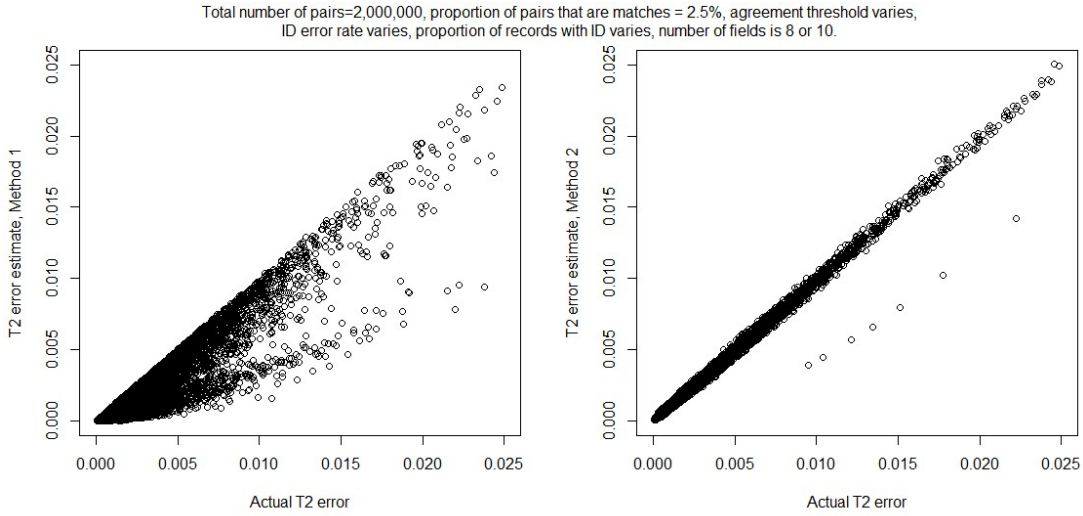

Figure 4: Comparison of actual Type II error rates to estimated Type II error rates for two methods of estimation; constrained simulation parameter space.

4. Discussion

It is rather puzzling why the gold standard performed so badly at estimating the Type I error. One possibility is that it is imprecise because it requires a good estimate of the basal error rate (because it is estimated by subtracting the basal error rate from the raw error rate) and generally this was not available. Also, it is probably impacted by the size of the basal error rate being nearly as big as or bigger than the overall error rates. But this may be more a feature of the way the simulation passes were constructed that something that is necessarily the case in real-world linkages.

For the match probability method, the imprecision of Type I estimates appears tied to instances when there is relatively little data available to make estimates of P(Match), such as when there are only a few comparison variables or the total number of matches is small. Fortunately, in real-world record linkage problems we have encountered, these conditions are rare.

4.1 Bayesian Extension

Extending the work done thus far to a fully Bayesian model is not difficult. If we take

$$\{\mathrm{X}_1, \cdots, \mathrm{X}_{2^n}\} \sim \mathrm{Multinomial}(\mathrm{X}_{\mathrm{Matc}}, \prod_{i=1}^{n} {m_i}^{a_{i1}} (1-m_i)^{(1-a_{i1})}, \cdots, \prod_{i=1}^{n} {m_i}^{a_{i(2^n)}} (1-m_i)^{\left(1-a_{i(2^n)}\right)})$$

And for $\mathrm{Y}_j = \mathrm{N}_j - \mathrm{X}_j$, take

$$\{\mathrm{Y}_1, \cdots, \mathrm{Y}_{2^n}\} \sim \mathrm{Multinomial}(\mathrm{N}_{\mathrm{Pairs}} - \mathrm{X}_{\mathrm{Match}}, \prod_{i=1}^{n} {u_i}^{a_{i1}} (1-u_i)^{(1-a_{i1})}, \cdots, \prod_{i=1}^{n} {u_i}^{a_{i(2^n)}} (1-u_i)^{\left(1-a_{i(2^n)}\right)})$$

and

$$\mathrm{X}_{\mathrm{Match}} \sim \mathrm{Binomial}(\mathrm{N}_{\mathrm{Pairs}}, p)$$

as the first two levels in the model, then we can take $\mathrm{N}_{\mathrm{Pairs}}$ to be provided by the data and set prior distributions for the *m*- and *u*- probabilities as well as *p*. In this paper we have used Uniform distributions to generate potential *m*- and *u*- probabilities; in the Bayesian context it would make more sense to take the priors for all the hyperparameters as Beta distributions, based on previous knowledge of the behavior of these parameters for particular fields. The prior for *p* could be taken as noninformative. Exploration of this model remains for future work.

Acknowledgements

The authors thank our colleagues at the National Opinion Research Center, the National Center for Health Statistics, and the Department of Mathematics and Statistics at Slippery Rock University for their support.

References

Aldridge, R. W., Shaji, K., Hayward, A. C., & Abubakar, I. (2015). Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies. *PloS one*, *10*(8), e0136179.

Asher, J. (2017, August). *A Cross-Disciplinary Review of Record Linkage Methodologies.* Presentation at the Joint Statistical Meetings, Baltimore, MD.

Belin, T. R., & Rubin, D. B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, *90*(430), 694-707.

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Ibrahim, J., & Brand, C. (2011). Development and validation of reporting guidelines for studies involving data linkage. *Australian and New Zealand journal of public health*, *35*(5), 486-489.

Christen, P., & Goiser, K. (2007). Quality and complexity measures for data linkage and deduplication. In *Quality measures in data mining* (pp. 127-151). Springer, Berlin, Heidelberg.

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nations Health*, *36*(12), 1412-1416.

Feigenbaum, J. J. (2016). *Automated census record linking: A machine learning approach*.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L. C., Smith, P., ... & Goldstein, H. (2017). GUILD: guidance for information about linking data sets. *Journal of Public Health*, *40*(1), 191-198.

Hagger-Johnson, G., Harron, K., Fleming, T., Gilbert, R., Goldstein, H., Landy, R., & Parslow, R. C. (2015). Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ open*, *5*(8), e008118.

Hagger-Johnson, G., Harron, K., Gonzalez-Izquierdo, A., Cortina‐Borja, M., Dattani, N., Muller‐Pebody, B., ... & Goldstein, H. (2015-2). Identifying possible false matches in anonymized hospital administrative data without patient identifiers. *Health services research*, *50*(4), 1162-1178.

Hagger-Johnson, G., Harron, K., Aldridge, R., Fu, B., Setakis, E., Goldstein, H., & Gilbert, R. (2017). Combining deterministic and probabilistic matching to reduce data linkage errors in hospital administrative data. *International Journal of Population Data Science*, *1*(1).

Hagger-Johnson, G., Harron, K., Goldstein, H., Aldridge, R., & Gilbert, R. (2017-2). Probabilistic linking to enhance deterministic algorithms and reduce linkage errors in hospital administrative data. *Journal of innovation in health informatics*, *24*(2), 891.

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., & Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC medical research methodology*, *14*(1), 36.

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. *International journal of epidemiology*, *46*(5), 1699-1710.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017-2). Challenges in administrative data linkage for research. *Big data & society*, *4*(2), 2053951717745678.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

Larsen, M. D., & Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, *96*(453), 32-41.

Monga, H. K., & Patrick, T. B. (2001). Error estimation in linking heterogeneous data sources. Health Informatics Journal, 7(3-4), 135-137.Randall, S., Brown, A., Boyd, J., Schnell, R., Borgs, C., & Ferrante, A. (2018). Sociodemographic differences in linkage error: an examination of four large-scale datasets. *BMC health services research*, *18*(1), 678.

Moore, C. L., Amin, J., Gidding, H. F., & Law, M. G. (2014). A new method for assessing how sensitivity and specificity of linkage studies affects estimation. *PloS one*, *9*(7), e103690.

Rentsch, C. T., Harron, K., Urassa, M., Todd, J., Reniers, G., & Zaba, B. (2018). Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural Tanzania. *BMC medical research methodology*, *18*(1), 165.

Trentin, V., Bastos, V., Costa, M., Camargo, K., Sobrino, R., Guillen, L. C., & Coeli, C. (2018). Synthetic data generator for testing record linkage routines in Brazil. *International Journal of Population Data Science*, *3*(4).

Tuoto, T. (2016). New proposal for linkage error estimation. *Statistical Journal of the IAOS*, *32*(3), 413-420.

Winglee, M., Valliant, R., & Scheuren, F. (2005). A case study in record linkage. *Survey Methodology*, *31*(1), 3-11.

Winkler, W. E. (2007). Automatically estimating record linkage false match rates. *Statistics*, *5*.

Winkler, W. E. (2014). Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(5), 313-325.