Special Issue Paper

Statistics in Medicine

Received 3 November 2011, Accepted 30 May 2012 Published online 16 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5498

Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables

M. H. P. Hof*[†] and A. H. Zwinderman

In record linkage studies, unique identifiers are often not available, and therefore, the linkage procedure depends on combinations of partially identifying variables with low discriminating power. As a consequence, wrongly linked covariate and outcome pairs will be created and bias further analysis of the linked data. In this article, we investigated two estimators that correct for linkage error in regression analysis. We extended the estimators developed by Lahiri and Larsen and also suggested a weighted least squares approach to deal with linkage error. We considered both linear and logistic regression problems and evaluated the performance of both methods with simulations. Our results show that all wrong covariate and outcome pairs need to be removed from the analysis in order to calculate unbiased regression coefficients in both approaches. This removal requires strong assumptions on the structure of the data. In addition, the bias significantly increases when the assumptions do not hold and wrongly linked records influence the coefficient estimation. Our simulations showed that both methods had similar performance in linear regression problems. With logistic regression problems, the weighted least squares method showed less bias. Because the specific structure of the data in record linkage problems often leads to different assumptions, it is necessary that the analyst has prior knowledge on the nature of the data. These assumptions are more easily introduced in the weighted least squares approach than in the Lahiri and Larsen estimator. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: record linkage; matching error; regression analysis

1. Introduction

Computerized record linkage is an effective manner to aggregate information from different data sources when a unique identifier is not available. Partially identifying variables that are registered in both data sources, referred to as linking variables, are used to find records that belong to the same individual (match) or do not belong to the same individual (non-match). Because record linkage is based on existing data, it has the potential to answer research questions without the need for new data collection [1–4].

To determine whether a set of records belongs to the same individual, Fellegi and Sunter [5] have developed a theory for record linkage. Basically, comparison rules are defined for each of the linking variables, and for each combination of rows in both datasets, comparison vectors are made. The frequency of all comparison vectors along with the number of matches is used to determine the posterior probability of a match given a certain comparison vector.

A problem that arises with record linkage is the presence of wrongly matched records. Neter *et al.* [6] investigated the impact of these records and concluded that these records will substantially bias the estimation of the relationship between the response variable and covariates of interest. Therefore, Scheuren and Winkler (SW) have developed a method to correct for linkage error in linear models [7, 8]. The idea behind their method is that each observed pair of outcome and covariate (x, z) can be described in terms of the true values (x, y) and the bias (x, b), assuming that only the outcome y is biased by the linking process.

Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands

**Correspondence to: M. H. P. Hof, Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands.*

[†] *E-mail: m.h.hof@amc.uva.nl*

From the observed data, the outcome z with the highest posterior probability of a match is assigned to the true value y. In addition, the outcome z with the second highest posterior probability is assumed to be the wrong value b. Although simulations by SW showed that this method is able to correct regression estimates for linkage error bias, this method does not produce unbiased results. However, a modification of the SW estimator developed by Lahiri and Larsen (LL) does produce unbiased estimators [9]. Their proposed method uses the posterior probabilities as a weighting scheme to derive the expected value y from z for all rows of x.

To derive the unbiased estimator, LL used constraints on the data structure; the outcome variable and covariates are always located in separate datasets, both datasets have similar number of records, and all records from both datasets refer to the same population. However, these constraints are not met in most record linkage studies.

Chambers *et al.* [10] generalized the LL method. They proposed an estimation equation and derived a best linear unbiased estimator that performed somewhat better than the LL estimator. Although simulations showed that Chamber's estimator is useful for dealing with bias, the required constraints on the data structure are similar to LL.

In addition to the work of Chambers *et al.*, Kim and Chambers [11] tried to relax the constraints on the data with an extra weighting procedure in the analysis, which was based on the assumption that the dataset with the covariates x is a subsample of the population of interest and the dataset with outcomes z covers the entire population. However, their extension was based on the strong assumption that for all matches, the sign of the regression errors are available. This requires information on the process that governs the assignment of matches and non-matches, which is almost never available. Kim and Chambers also considered linkage situations with nonlinear relations between outcome and covariates.

In this article, we extend the LL method and also consider alternative estimators based on a weighted least squares (WLS) method to deal with bias introduced by linkage error. We investigate situations with multiple datasets and where the datasets describe different populations but contain a number of true matches. We investigate both linear and logistic regression problems. We evaluated performance of all methods with simulations.

2. Record linkage

Consider there are two datasets A and B containing respectively n and m records and both datasets contain records from the same individuals. Because there is no unique identifying variable per person, other less discriminative variables r must be used in the record linkage procedure such as date of birth, surname, place of residence, or sex. These variables are registered in both datasets and are referred to as linkage variables. Each linkage variable has its own discriminative power, determined by the number of variable values and distribution of variable values [12].

Because there is no prior knowledge on likely matches in both datasets, the strategy begins by comparing all records $A_i(i = 1, 2, \ldots, n)$ with all records $B_j(j = 1, 2, \ldots, m)$ leading to nm comparisons $g_{ij}(ij = 1, 2, \ldots, nm)$. This matrix contains measures of agreement for all k linking variables $r_l(l = 1, 2, \ldots, k)$ and can be described by the vectors $g_{ij} = (g_{ij1}, \ldots, g_{ijl}, \ldots, g_{ijk})$.

Comparisons g_{ijl} can be defined in different ways, but for the sake of simplicity, we use only dichotomous agreement/disagreement outcomes in this paper

$$g_{ijl} = \begin{cases} 1 & \text{if } r_{il} = r_{jl} \\ 0 & \text{if } r_{il} \neq r_{jl} \end{cases} \tag{1}$$

Notice that the number of unique patterns g_{ij} is 2^k, and for $k = 2$, the unique patterns are $(0, 1)$, $(0, 0)$, $(1, 0)$, and $(1, 1)$. All comparisons need to be divided into two groups: matches and non-matches. Therefore, we can specify $q_{ij} = p(\text{match}|g_{ij})$, that is, the probability of a match given g_{ij}. Basically, there are two types of strategies to determine q_{ij} [13]. In the deterministic approach, all linking variables or a subset of linking variables in both datasets have to agree to consider a record pair as a link. If the linking variables are highly discriminative, the q_{ij} values will be close to zero or one, depending on the comparison vector g_{ij} [14]. With the probabilistic approach, the probability of a comparison vector, g_{ij}, can be expressed in the mixture model [15]

$$p(g_{ij}) = \pi p(g_{ij}|\text{match}) + (1 - \pi) p(g_{ij}|\text{non-match}) \tag{2}$$

where π is the relative frequency of matches among the nm records pairs, for example, the probability that two random records are from the same person. The parameters from this model can be estimated using, for instance, an expectation–maximization algorithm [16].

Parameters need to be estimated for $p(g_{ij}|\text{match})$ and $p(g_{ij}|\text{non-match})$ for each unique comparison value g_{ij} in Equation (2). Because there are 2^k unique patterns for k linking variables, the number of parameters that need to be estimated may become impracticable. To decrease the number of parameters in the model, Fellegi and Sunter [5] suggested to assume independence between the comparison outcomes of each linking variable. Other assumptions and extensions of the mixture model that have been investigated are, among others, the introduction of approximate field estimators [17], the use of approximate string comparison [18], the introduction of clerical review in the estimation [15], and the addition of interactions among comparison fields [15, 19].

After estimating the parameters of the mixture model in (2), the posterior probability of a match given g_{ij} can be introduced in the regression of y on x by creating an $n \times m$ weighting matrix $Q = (q_{ij})$, where q_{ij} equals

$$q_{ij} = p(\text{match}|g_{ij}) = \frac{\pi p(g_{ij}|\text{match})}{\pi p(g_{ij}|\text{match}) + (1-\pi)p(g_{ij}|\text{non-match})} \tag{3}$$

3. Linear regression

3.1. *Lahiri–Larsen estimator*

After datasets A and B are linked together, the relationship between outcome variable y and covariates $X = (x_1, \ldots, x_p)$ may be estimated. Simple regression models are not applicable in this situation because the true pairs of (x, y) are not observed. Only all possible combinations of values with their corresponding posterior probability derived from the record linkage procedure are available.

Lahiri and Larsen showed that their assumptions generate unbiased estimators of a linear relationship. Covariates must be located in dataset A and the outcome in dataset B, and both datasets must contain the same number of records, therefore implying $n = m$. Kim and Chambers [11] relaxed the assumption of $n = m$ and showed that the LL estimator is still unbiased if the records from dataset A are a subset of the population described by dataset B, resulting in $n \leqslant m$. Notice that all records from dataset A must be located in dataset B. Now consider the following linear relation

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \quad i = 1, \ldots, n \tag{4}$$

where the βs are the unknown regression coefficients. In addition, we assume $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$, and $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for $j \neq i$, $j = 1, \ldots, n$.

Because it is unknown which record from dataset A belongs to which record in dataset B, the record pairs (X_i, z_i) are observed instead of the true (X_i, y_i). Therefore, the relation from LL is not based on $E(y)$ but on $E(z)$ and [7]

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i,\ j = 1, \ldots, n \end{cases} \tag{5}$$

Furthermore, LL required $\sum_{j=1}^{n} q_{ij} = 1, i = 1, \ldots, n$ and proposed the estimator of β

$$\widehat{\beta} = \left(X^T Q^T Q X\right)^{-1} X^T Q^T z \tag{6}$$

which is unbiased because

$$E(z_i) = E(E(z_i|y)) = \sum_{j=1}^{n} q_{ij} X_i \beta \tag{7}$$

Lahiri and Larsen proposed a parametric bootstrap procedure for variance estimation of $\widehat{\beta}$, which captures both the uncertainty of $\widehat{\beta}$ and also the uncertainty of Q caused by the estimation in the record linkage procedure.

Because the LL estimator gives unbiased regression coefficients, we propose a method that uses this estimator in more complex linkage situations. We no longer assume that all covariates X from (4) are located in the same dataset. The only assumption that is required is that the covariate datasets contain records that are a subset of the population described by the dataset containing y.

Suppose we have the situation in which all covariates x_k $(k = 1, \ldots, p)$ and the outcome are located in $p + 1$ separate datasets. This means that the direct use of (6) is not possible because the structure of the Q matrix becomes far too complex. Each separate covariate x_k has its own Q matrix with the outcome variable y. In addition, the covariance of x_k with x_s, for $s \neq i, s = 1, \ldots, p$ is also influenced by different Q matrices.

However, if the influence of all x_s is removed from x_k, it is still possible to estimate β_k with the LL method. By regressing all x_s on x_k using the LL method, residuals x_k^* are derived that are orthogonal to all other covariates. The vector x_k^* can be used in formula (6) to calculate the unbiased $\widehat{\beta_k}$. This method is iteratively performed for all p covariates x_k.

Concisely, this method works as follows. Consider the situation where an outcome and two covariates are divided over three datasets A, B, and C, where A contains the outcome y, B contains the first covariate vector x_1, and C contains the second covariate vector x_2. In addition, we define three Q matrices Q_{AB}, Q_{AC}, and Q_{BC} as the weighting matrices obtained from the record linkage procedures linking A to B, A to C, and B to C, respectively. We use the following algorithm to estimate the weights β_1 and β_2 of the regression model $\mathbb{E}[y|x_1, x_2] = \beta_1 x_1 + \beta_2 x_2$

$$\begin{gathered}\text{Estimate the regression parameter of } x_1 \text{ on } x_2\\ \beta_{(x_1 \sim x_2)} = \left(x_2^T Q_{BC}^T Q_{BC} x_2\right)^{-1} x_2^T Q_{BC}^T x_1\end{gathered}$$

$$\begin{gathered}\text{Calculate the residuals } x_1|x_2\\ x_{1\text{res}} = x_1 - Q_{BC} x_2 \beta_{(x_1 \sim x_2)}\end{gathered}$$

$$\begin{gathered}\text{Estimate the regression parameter of } X_{1\text{res}} \text{ and } X_2 \text{ on } y\\ \beta_{(y \sim x_{1\text{res}})} = \left(x_{1\text{res}}^T Q_{AB}^T Q_{AB} x_{1\text{res}}\right)^{-1} x_{1\text{res}}^T Q_{AB}^T y\\ \beta_{(y \sim x_2)} = \left(x_2^T Q_{AC}^T Q_{AC} x_2\right)^{-1} x_2^T Q_{AC}^T y\end{gathered}$$

$$\begin{gathered}\text{Derive the true regression coefficients}\\ \beta_1 = \beta_{(y \sim x_{1\text{res}})}\\ \beta_2 = \beta_{(y \sim x_2)} - \beta_{(y \sim x_{1\text{res}})} \beta_{(x_1 \sim x_2)}\end{gathered}$$

Another extension is the relaxation of the fact that $\sum_{j=1}^{n} q_{ij} = 1$, meaning that all records from A must have at least one match in B. Now consider a record linkage situation in which records in A do not necessarily have a true match in B. With perfect linkage variables, the sum of posterior probabilities in this row will be zero, and therefore, these rows cannot be transformed to sum up to one. Because we focus on situations with non-perfect linkage variables, the cumulative posterior probabilities of these rows will not be non-zero. However, if the discriminative power of the linkage variables is high, these sums will be relatively small. Therefore, we propose to only transform the rows in Q where it is likely that at least one match has been found. A likely match can be defined as combination of records from A and B with a probability higher than an (arbitrary) threshold λ and the Q matrix is transformed as

$$\sum_{j=1}^{n} q_{ij} = \begin{cases} 1 & \text{if} \max(q_{ij}) \geqslant \lambda \\ 0 & \text{if} \max(q_{ij}) < \lambda \end{cases} \tag{8}$$

Because non-matches are likely to have a low q_{ij} and matches a q_{ij} close to one, the λ value directly determines the number of true and false data pairs in the analysis. Choosing a λ value close to one will guarantee that few non-matches are included in the analysis. Conversely, choosing λ close to zero will result in many non-matches. Basically, the choice of λ is a trade-off between specificity and sensitivity of the categorization of matches and non-matches [20, 21].

3.2. *Weighted least squares estimator*

Again consider the linear relation defined in formula (4) and that the real pairs of covariates and outcomes are not observed. The two datasets A and B with n and m records are linked to each other, and the Q matrix is calculated. For now, we assume that the covariates X are located in **A**, and the outcome y in **B**.

To analyze the data, we propose a WLS approach, which requires some data restructuring. We consider all nm combinations of all n records in A and all m records in B. Define R as the operator matrix, which adds m multiples of X, and P as the operator matrix, which extends y n times as follows:

$$\begin{aligned}\mathbf{X} &= RX\\ \mathbf{y} &= Py\end{aligned}$$

$$R = \begin{pmatrix} \begin{array}{cc}
1 & \begin{array}{cc} & \begin{array}{ccccc} &&&& n \end{array} \\ \begin{array}{c} 1 \\ \\ \\ n \end{array} & \begin{pmatrix} 1&0&0&\dots&0\\ 1&0&0&\dots&0\\ 1&0&0&\dots&0\\ 1&0&0&\dots&0 \end{pmatrix} \end{array} \\
2 & \begin{array}{cc} & \begin{array}{ccccc} &&&& n \end{array} \\ \begin{array}{c} 1 \\ \\ \\ n \end{array} & \begin{pmatrix} 0&1&0&\dots&0\\ 0&1&0&\dots&0\\ 0&1&0&\dots&0\\ 0&1&0&\dots&0 \end{pmatrix} \end{array} \\
 & \vdots \\
m & \begin{array}{cc} & \begin{array}{ccccc} &&&& n \end{array} \\ \begin{array}{c} 1 \\ \\ \\ n \end{array} & \begin{pmatrix} 0&0&0&\dots&1\\ 0&0&0&\dots&1\\ 0&0&0&\dots&1\\ 0&0&0&\dots&1 \end{pmatrix} \end{array}
\end{array} \end{pmatrix}
\qquad
P = \begin{pmatrix} \begin{array}{cc}
1 & \begin{array}{cc} & \begin{array}{cccc} &&& m \end{array} \\ \begin{array}{c} 1 \\ \\ \\ m \end{array} & \begin{pmatrix} 1&\dots&\dots&0\\ \vdots&1&&\vdots\\ \vdots&&\ddots&\vdots\\ 0&\dots&\dots&1 \end{pmatrix} \end{array} \\
2 & \begin{array}{cc} & \begin{array}{cccc} &&& m \end{array} \\ \begin{array}{c} 1 \\ \\ \\ m \end{array} & \begin{pmatrix} 1&\dots&\dots&0\\ \vdots&1&&\vdots\\ \vdots&&\ddots&\vdots\\ 0&\dots&\dots&1 \end{pmatrix} \end{array} \\
 & \vdots \\
n & \begin{array}{cc} & \begin{array}{cccc} &&& m \end{array} \\ \begin{array}{c} 1 \\ \\ \\ m \end{array} & \begin{pmatrix} 1&\dots&\dots&0\\ \vdots&1&&\vdots\\ \vdots&&\ddots&\vdots\\ 0&\dots&\dots&1 \end{pmatrix} \end{array}
\end{array} \end{pmatrix}$$

where $\mathbf{X}$ and $\mathbf{y}$ contain all nm combinations (X_i, y_j) from the original X and y matrices. In addition, we need to restructure the Q matrix to an $nm \times nm$ diagonal weighting matrix W as follows:

$$W = \left\{ \begin{array}{cccccc}
q_1 \otimes I & \dots & \dots & \dots & \dots & 0 \\
\vdots & q_2 \otimes I & & & & \vdots \\
\vdots & & \ddots & & & \vdots \\
\vdots & & & q_i \otimes I & & \vdots \\
\vdots & & & & \ddots & \vdots \\
0 & \dots & \dots & \dots & \dots & q_n \otimes I
\end{array} \right. \tag{9}$$

where all rows from the Q matrix form the diagonal values for the weighting matrix W, where $\otimes$ is the elementwise multiplication function and I the $m \times m$ identity matrix. W can be introduced in the analysis and β can be estimated by minimizing the weighted sum of squares of the transformed data pairs

$$\begin{aligned}\widehat{\beta} &= \left(\mathbf{X}^T W \mathbf{X}\right)^{-1} \mathbf{X}^T W \mathbf{y}\\ &= \left(X^T R^T W R X\right)^{-1} X^T R^T W P y\end{aligned} \tag{10}$$

This is, however, a biased estimator under imperfect linkage, and the bias is

$$\text{bias}\left(\widehat{\beta}\right) = \left(\left(X^T R^T W R X\right)^{-1} \left(X^T R^T W P X\right) - I\right) \beta \tag{11}$$

where the bias depends on both β and the weighting matrix W and the estimate is unbiased if $R^T WR = R^T WP$ or $\beta = 0$. Notice that $R^T WR = R^T WP$ will only occur if all true pairs (X_i, y_i) have weight $w_{ii} = 1$ and all wrong pairs (X_i, y_j) have weight $w_{ij} = 0$ for $j \neq ij = 1, \ldots, m$. In good linkage situations, all w_{ij} will be small resulting in relatively low bias.

Although this estimator is biased in most linkage situations, its simplicity and the potential to improve its performance with some simple operations makes it an interesting alternative to the LL estimator. By manipulating the q_i vectors, the weights in the W matrix are directly affected. Similarly to the LL method, we assume that only *one* of the non-zero values in the vector q_i describes a true covariate and outcome pair. These properties allow us to specify the following operations for each q_i that contains *more* than one non-zero value:

1. Assign zero to all values of the vector, that is, $\sum_{j=1}^{m} q_{ij} = 0$.
   This operation removes all the records from A with more than one match in dataset B from the analysis. This operation ensures that all the records that accidentally have two or more matches are not included in the analysis. All wrong data pairs are removed from the analysis with this operation if the LL assumptions hold.
2. Weigh the vector with its cumulative probability of a match, that is, $\sum_{j=1}^{m} q_{ij} = 1$.
   Weighing the vector with its cumulative probability results in one weighted match for each record in dataset A. Because it is not known which one of the non-zero weights in q_i resembles the true match, all q_{ij} values are weighted to have a cumulative probability of one. Although this method does not remove wrongly linked data pairs from the analysis, it reduces the weight that is assigned to potential wrong data pairs.
3. Randomly select one non-zero value and assign zero values to the others.
   In this operation, one non-zero weight is randomly chosen as the true match, and its weight will be the original posterior probability q_{ij}. Zero is attributed to all other non-zero weights. Notice that this method can be performed an arbitrary number of times to reflect the uncertainty that is present in the q_i vector.

A disadvantage of using the WLS approach might seem that the required P, R, and Q matrices can become extremely large in certain linkage studies. For instance, the Rochester Epidemiology Project has linked 1,145,856 medical records to 486,564 individuals [4]. Analyzing these datasets with the WLS method will require matrices with unrealistic dimensions of nm (5.58×10^{11}) rows.

However, most of the comparisons between datasets A and B will have a weight q_{ij} close to zero and $q_{ij} < \lambda$. We remove these records from R and P and the associated rows and columns from W. The resulting R, P, and W matrices will be considerably smaller. In the case of the Rochester Epidemiology Project, the number of rows of P and R will be close to 1,145,856 instead of 5.58×10^{11}. This is of the same order as the LL method.

Similarly to the variance estimation of LL, the variance of $\widehat{\beta}$ can be derived from a bootstrap procedure. Furthermore, the WLS method can also be extended to fit situations with more than two datasets. To capture all the unique combinations of records from the available datasets, we need to multiply all datasets with permutation matrices. These matrices will have similar structures as the P and R matrices, and the greatest difference is that their inner matrices are multiplied in more dimensions instead of one. In addition, the W matrix can be calculated by using a generalized Fellegi–Sunter framework to calculate the posterior probabilities of a data pair combination for more than two datasets.

4. Logistic regression

Both the LL and WLS methods can be generalized to deal with logistic regression problems

$$\text{logit}(p(y_i = 1|x_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \qquad i = 1, \ldots, n \tag{12}$$

The regression estimates β can be calculated with an iterative reweighted least squares procedure [22], which maximizes the log-likelihood function with Newton–Raphson updates. For the LL estimator, one single update is

$$\begin{aligned} \text{LL}(\beta^{\text{new}}) &= \beta^{\text{old}} + \left(X^t Q^t \Omega Q X\right)^{-1} X^t Q^t \left(z - \widehat{p}\left(z|X, Q, \beta^{\text{old}}\right)\right) \\ \widehat{p}\left(z|X, Q, \beta^{\text{old}}\right) &= e^{QX\beta^{\text{old}}} \Big/ \left(1 + e^{QX\beta^{\text{old}}}\right) \\ \Omega &= \text{diag}\left(\widehat{p}\left(z|X, Q, \beta^{\text{old}}\right)\left(1 - \widehat{p}\left(Py|z, X, Q, \beta^{\text{old}}\right)\right)\right) \end{aligned} \tag{13}$$

and for the WLS method

$$\begin{aligned} \mathrm{WLS}(\beta^{\mathrm{new}}) &= \beta^{\mathrm{old}} + \left(X^t R^t \Omega R X\right)^{-1} X^t R^t \left(Wy - \widehat{p}(Py|RX, \beta^{\mathrm{old}})\right) \\ \widehat{p}\left(Py|RX, \beta^{\mathrm{old}}\right) &= e^{RX\beta^{\mathrm{old}}} \Big/ \left(1 + e^{RX\beta^{\mathrm{old}}}\right) \\ \Omega &= \mathrm{diag}\left(\widehat{p}\left(Py|RX, \beta^{\mathrm{old}}\right)\left(1 - \widehat{p}\left(Py|RX, \beta^{\mathrm{old}}\right)\right)\right) \end{aligned} \tag{14}$$

Both estimators are unbiased when the covariates are located in the same dataset. However, it is not possible to use formula (13) in the LL method to calculate $\widehat{\beta}$ if the covariates are located in more than one dataset. This is because the method proposed in Section 3.1 to calculate each $\widehat{\beta_k}$ separately does not give unbiased estimates, as the covariates are linked to the outcome variable with a nonlinear logit function. Therefore, the $\widehat{\beta_k}$ that is derived from $p(y = 1|x_k^*)$ will not be similar to β_k, because our proposed method implicitly assumes linear relations between outcome and all covariates.

5. Simulation

5.1. Scenarios

To measure the performance of the LL and WLS estimators, we performed simulations of different record linkage scenarios (Table I). We investigated situations with different locations of the covariates and the number of records in dataset A without a true match in dataset B. Scenario 1_a represents the situation that fits all the assumptions made by Lahiri and Larsen [9] and Kim and Chambers [11], required for unbiased estimators in the LL model. Covariates are located in A, the outcome in B, and all records from A have one true match in B. Therefore, the linkage procedure only introduced error in the outcome variable. In scenario 1_b, not all records from A are located in B. Because a number of matches were falsely identified in the record linkage procedure, wrong covariate and outcome pairs were introduced in the analysis.

In scenarios 2_a and 2_b, covariate x_1 was located in A and both the covariate x_2 and the outcome y in B. Similarly to scenario 1, in 2_a, all records from dataset A are located in dataset B, and some records from A were not located in B in scenario 2_b.

In the scenarios, either good or relatively bad linking variables were available, drawn from discrete uniform distributions. All linkage variables were free from error, and therefore, all true matches were present in the analysis. In the good situation, five linkage variables were present in both datasets with respectively 30, 8, 7, 4, and 2 unique values (13,440 unique patterns g). In the bad situation only four variables were available with 30, 8, 7, and 2 unique values (3360 unique patterns g). In the scenarios with 200 records in A and 800 in B, we would expect 12 wrong matches (6% of all matches) with the good linking variables and 48 wrong matches (19% of all matches) with the bad linkage variables. In addition, the maximum posterior probability of a match q_{ij} was approximately 0.9 with good linking variables and approximately 0.8 for the scenarios with bad linkage variables.

We maximized the likelihood to estimate the parameters from the Fellegi–Sunter model. In the simulation, both linear and logistic relations between one outcome measure and two covariates were simulated

$$y = f^{-1}(a + x_1\beta_1 + x_2\beta_2 + \epsilon) \tag{15}$$

Table I. Characteristics of all simulated scenarios.

	Dataset A			Dataset B	
Scenario	Regression variables	Number of records	Real matches	Regression variables	Number of records
1_a	x_1, x_2	200	200 ($A \subset B$)	y	800
1_b	x_1, x_2	300	200 ($A \subsetneq B$)	y	800
2_a	x_1	200	200 ($A \subset B$)	x_2, y	800
2_b	x_1	300	200 ($A \subsetneq B$)	x_2, y	800

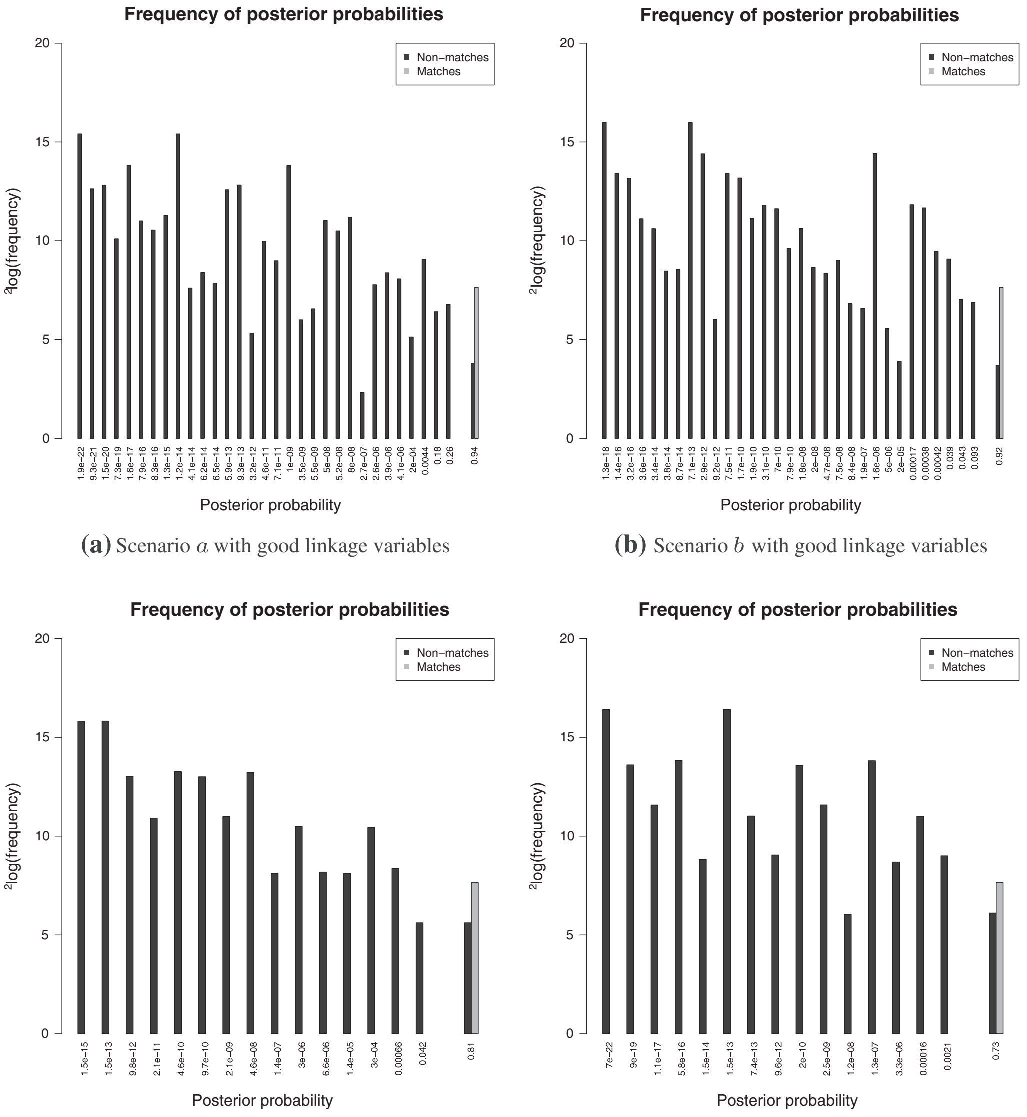
where f^{-1} is a link function, which is the identity function with continuous outcome y and the logit function with binary outcome y. We simulated an intercept $a = 0$, $\beta_1 = 1.5$, and $\beta_2 = -2$. The residuals ϵ were normally distributed with mean 0 and standard deviation 2. Two covariates x_1 and x_2 were drawn from a multivariate normal distribution with means 0, variances 1, and covariance 0.2.

5.2. *Regression models*

In linkage problems, Q often contains a large number of comparisons with a very low probability. To reduce bias, a simple pre-analysis procedure is to assign the value zero to all these highly unlikely data pairs. Because we simulated a fairly simple linkage problem, this phenomenon was clearly seen in all Q matrices (Figure 1). Notice that we assigned the value zero to all comparisons that did not have the highest probability of a match, resulting in the matrix Q'.

For all of the following analysis approaches, we calculated the bias, mean squared error, and coverage of the 95% confidence interval

LL: extended LL estimator.
WLS$_1$: use the Q' matrix to generate the weighting matrix W.



(a) Scenario a with good linkage variables

(b) Scenario b with good linkage variables

(c) Scenario a with bad linkage variables

(d) Scenario b with bad linkage variables

Figure 1. Frequencies of posterior probabilities in the Q matrix taken from one simulated run for scenario a and scenario b with either good or bad linkage variables.

WLS$_2$: all rows from Q' with one or more non-zero values are weighted so that $\sum_{i=1}^{n} q_{ij} = 1$, following the assumption that all records from A have one true match in B.

WLS$_{3a}$: assign the value zero to all entries in the rows from Q' with more than one non-zero value. This strategy is based on the assumption that there is a maximum of one true match for all rows A. If it is unclear which record from B is the true match for record from A, this record from A is discarded from the analysis.

WLS$_{3b}$: first perform the procedure suggested in WLS$_{3a}$. Afterwards, treat the removal of rows from Q as a missing data problem by estimating the covariate or outcome from B that should belong to the regression variables from A by multiple imputation.

WLS$_4$: randomly select one non-zero value in the rows from Q' with more than one non-zero value. To all other values, zero is assigned. This procedure is repeated an arbitrary number of times, which was 50 in our simulations.

For the variance estimation of the regression coefficients, we used the bootstrap procedure from LL. We performed all simulations in R [23] and repeated all scenarios 5000 times.

5.3. *Results*

We summarize the results of the simulations with continuous outcome and continuous covariates in Tables II and III. In all scenarios, the LL, WLS$_{3a}$, and WLS$_{3b}$ approaches gave the lowest bias with both good and bad linkage variables. The methods WLS$_2$, WLS$_4$, and especially WLS$_1$ gave more biased estimates.

In scenarios 1_a, all methods were comparable in performance with good linking variables. In the regression models, both βs were accurately estimated. The highest bias was present in the WLS$_1$ method and the lowest bias in the LL method (−0.083 compared with −0.001 in β_1 and 0.112 compared with

Table II. Simulation results for regression estimates in scenario 1 with continuous outcome and continuous covariates.

			β_1			β_2		
			$E(\widehat{\beta_1} - \beta_1)$	MSE	Coverage	$E(\widehat{\beta_2} - \beta_2)$	MSE	Coverage
Scenario 1_a	Good	LL	−0.001 (0.145)	0.021	0.951	0.003 (0.148)	0.021	0.946
		WLS$_1$	−0.083 (0.151)	0.021	0.951	0.112 (0.158)	0.030	0.933
		WLS$_2$	−0.045 (0.144)	0.020	0.943	0.061 (0.148)	0.023	0.926
		WLS$_{3a}$	−0.002 (0.147)	0.021	0.953	0.002 (0.148)	0.021	0.949
		WLS$_{3b}$	−0.003 (0.148)	0.022	0.949	0.004 (0.15)	0.022	0.945
		WLS$_4$	−0.045 (0.021)	0.020	0.951	0.061 (0.022)	0.023	0.933
	Bad	LL	−0.001 (0.158)	0.023	0.953	0.002 (0.159)	0.025	0.952
		WLS$_1$	−0.283 (0.163)	0.023	0.845	0.378 (0.169)	0.107	0.742
		WLS$_2$	−0.165 (0.150)	0.020	0.789	0.222 (0.154)	0.05	0.666
		WLS$_{3a}$	−0.001 (0.163)	0.025	0.948	0.004 (0.166)	0.027	0.945
		WLS$_{3b}$	−0.006 (0.170)	0.027	0.932	0.011 (0.173)	0.029	0.925
		WLS$_4$	−0.166 (0.023)	0.020	0.845	0.223 (0.024)	0.050	0.742
Scenario 1_b	Good	LL	−0.040 (0.154)	0.021	0.940	0.056 (0.154)	0.025	0.935
		WLS$_1$	−0.119 (0.157)	0.022	0.920	0.161 (0.159)	0.039	0.890
		WLS$_2$	−0.083 (0.152)	0.021	0.910	0.113 (0.152)	0.03	0.878
		WLS$_{3a}$	−0.041 (0.156)	0.022	0.936	0.058 (0.155)	0.026	0.932
		WLS$_{3b}$	−0.043 (0.157)	0.022	0.934	0.06 (0.157)	0.026	0.923
		WLS$_4$	−0.083 (0.023)	0.021	0.920	0.113 (0.023)	0.030	0.890
	Bad	LL	−0.150 (0.172)	0.025	0.866	0.201 (0.173)	0.052	0.808
		WLS$_1$	−0.395 (0.163)	0.025	0.570	0.526 (0.171)	0.183	0.340
		WLS$_2$	−0.295 (0.159)	0.023	0.493	0.393 (0.163)	0.112	0.274
		WLS$_{3a}$	−0.159 (0.180)	0.030	0.841	0.212 (0.181)	0.058	0.761
		WLS$_{3b}$	−0.165 (0.185)	0.031	0.824	0.219 (0.188)	0.061	0.732
		WLS$_4$	−0.296 (0.026)	0.024	0.570	0.394 (0.027)	0.113	0.340

Covariates x_1 and x_2 are located in dataset A and the outcome variable y in dataset B.

Table III. Simulation results for regression estimates in scenario 2 with continuous outcome and continuous covariates.

			β_1			β_2		
			$E(\widehat{\beta_1}-\beta_1)$	MSE	Coverage	$E(\widehat{\beta_2}-\beta_2)$	MSE	Coverage
Scenario 2_a	Good	LL	−0.003 (0.143)	0.010	0.853	0.004 (0.150)	0.021	0.951
		WLS_1	−0.046 (0.151)	0.021	0.951	0.120 (0.158)	0.025	0.932
		WLS_2	−0.025 (0.147)	0.02	0.942	0.065 (0.149)	0.022	0.924
		WLS_{3a}	0.001 (0.151)	0.022	0.945	0.004 (0.152)	0.023	0.949
		WLS_{3b}	0.001 (0.152)	0.022	0.943	0.004 (0.152)	0.023	0.947
		WLS_4	−0.024 (0.021)	0.02	0.951	0.065 (0.022)	0.022	0.932
	Bad	LL	−0.003 (0.151)	0.010	0.842	−0.001 (0.158)	0.023	0.961
		WLS_1	−0.138 (0.16)	0.022	0.936	0.396 (0.171)	0.045	0.695
		WLS_2	−0.084 (0.147)	0.02	0.901	0.231 (0.155)	0.029	0.627
		WLS_{3a}	0.001 (0.164)	0.026	0.953	−0.004 (0.163)	0.027	0.951
		WLS_{3b}	0.002 (0.169)	0.027	0.941	0.001 (0.169)	0.029	0.942
		WLS_4	−0.084 (0.022)	0.020	0.936	0.232 (0.024)	0.029	0.695
Scenario 2_b	Good	LL	−0.024 (0.143)	0.010	0.844	0.060 (0.155)	0.021	0.934
		WLS_1	−0.065 (0.153)	0.022	0.942	0.172 (0.162)	0.028	0.868
		WLS_2	−0.045 (0.148)	0.021	0.936	0.120 (0.154)	0.024	0.854
		WLS_{3a}	−0.023 (0.153)	0.022	0.946	0.061 (0.157)	0.024	0.923
		WLS_{3b}	−0.023 (0.154)	0.023	0.945	0.061 (0.158)	0.024	0.920
		WLS_4	−0.045 (0.022)	0.021	0.942	0.120 (0.024)	0.024	0.868
	Bad	LL	−0.080 (0.137)	0.010	0.790	0.214 (0.172)	0.025	0.787
		WLS_1	−0.185 (0.161)	0.023	0.862	0.548 (0.169)	0.060	0.271
		WLS_2	−0.143 (0.155)	0.022	0.811	0.412 (0.162)	0.044	0.210
		WLS_{3a}	−0.081 (0.176)	0.029	0.918	0.225 (0.183)	0.038	0.728
		WLS_{3b}	−0.083 (0.183)	0.030	0.907	0.231 (0.189)	0.040	0.717
		WLS_4	−0.143 (0.024)	0.022	0.862	0.412 (0.026)	0.044	0.271

Covariate x_1 is located in dataset A and both the covariate x_2 and the outcome variable y in dataset B.

0.003 in β_2). In addition, the MSE ranged from approximately 0.02 to 0.03 for all methods in both βs. The coverage of the 95% confidence interval of β_1 was accurately estimated in all methods. We also observed this for β_2, with exception of WLS_2 that slightly underestimated the coverage.

With bad linking variables, the differences between methods in scenario 1_a increased. Unbiased estimators were still achieved using the LL, WLS_{3a}, or WLS_{3b} method. However, the other methods were biased ranging from −0.165 (for WLS_2, WLS_4) to −0.283 (for WLS_1) for β_1. For β_2, the bias ranged from 0.223 (for WLS_2, WLS_4) to 0.378 (for WLS_1). In addition, this overestimation and underestimation decreased the coverage of the 95% confidence interval.

In scenarios 1_b, all methods gave biased results. The bias was approximately −0.04 for LL, WLS_{3a}, and WLS_{3b}, −0.11 for WLS_1, and both the WLS_2 and WLS_4 method underestimated β_1 by −0.08. We found similar differences between methods in the estimation of β_2, in which all methods overestimated the regression coefficient.

With bad linking variables in scenario 1_b, all estimations of the regression coefficients were highly biased. Similar to the other simulations, LL, WLS_{3a}, and WLS_{3b} gave the best estimations, followed by WLS_2, and WLS_4, and the worst estimations were obtained with WLS_1. In addition, the coverage of the 95% confidence interval was low and ranged from 0.49 to 0.87 for β_1 and from 0.27 to 0.80 for β_2.

The bias in scenarios 2 was comparable with scenarios 1 for all the correction methods. In scenario 2_a, the LL, WLS_{3a}, and WLS_{3b} methods gave unbiased results regardless of the discriminative power of the linking variables. The WLS_2 and WLS_4 methods were biased, and the WLS_1 method gave the highest biased estimates.

In scenario 2, the LL method structurally underestimated the coverage of the 95% confidence interval of β_1, because we ignored the uncertainty that arose with the stepwise procedure that removed the correlation between covariates.

With binary outcome data in scenario 1, we found similar differences for the WLS and the LL methods (therefore, the data is not shown). All methods were unbiased in scenario 1_a, and in scenario 1_b, the LL, WLS_{3a}, and WLS_{3b} approaches gave the lowest bias, and the other WLS methods performed worse.

With binary outcomes in scenario 2, the performance of all WLS approaches was similar to the scenarios with continuous outcome. The LL method was, however, biased in scenario 2 because of the nonlinear transformation problem recognized in Section 4. In scenario 2_a, where the LL estimator gave unbiased results for the linear model, the logistic version showed large bias.

6. Discussion

In this paper, we proposed a number of approaches to regression analysis of data derived from record linkage. We have extended the analysis method developed by Lahiri and Larsen [9] to relax its assumptions and investigated the use of WLS methods.

In record linkage problems, bias is introduced by wrong covariates and outcome pairs. LL made it possible to remove these pairs from the analysis with a number of strong assumptions on the structure of the data. However, in most record linkage situations, these assumptions are not met, and the LL estimator is then also biased.

The impact of wrong data pairs could be seen in the WLS approaches. In the WLS_1 approach, in which the least assumptions were made to decrease the impact of wrong data pairs, the bias was highest. By introducing all the LL assumptions in a WLS approach (WLS_{3a} and WLS_{3b}), we showed that WLS is also able to create unbiased results regardless of the discriminative strength of the linking variables. Furthermore, the performance of the WLS_{3a}, WLS_{3b}, and LL estimators were comparable if the LL assumptions were violated.

Our results showed that the regression coefficients can only be unbiased when the assumptions are valid, thus it is necessary that the analyst has prior knowledge on the nature of the data. Note that the structure of the data is not similar in all record linkage problems and different techniques are needed to exclude wrong covariate and outcome pairs from the analysis.

The WLS method has more flexibility than the LL method and could more easily be used in different situations. The W matrix can be modified to fit particular assumptions, whereas more complex procedures are necessary for the LL method [10, 11]. In addition, the WLS approach can be used in more general situations where covariates and outcomes are located in multiple datasets. The WLS method was able to give unbiased estimates for the relation between a binary outcome and covariates that were located in two datasets, whereas the LL method was biased.

Another approach to the data derived from record linkage is to use error in variables models, in which records that are linked to more than one record are considered to have some measurement error. The rows with more than one match are assumed to follow either a normal or uniform distribution, and for each row, the corresponding sufficient statistics may be calculated. The regression coefficients can be estimated with a Bayesian approach or likelihood maximization. This approach had similar characteristics as WLS_1, and therefore, we did not show its results.

Because record linkage problems are used in many different settings, more research is needed to measure the performance of the different estimators in other scenarios. Furthermore, more operations could be suggested for the W matrix in the WLS method to fit more specific linkage situations. Another problem that requires more attention is the bootstrap procedure to estimate the variance of the regression parameters. Because the calculation of the Q matrix is repeated in each iteration of the bootstrap procedure, it is computationally intensive and requires cluster or grid computing.

References

1. Newcombe HB. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford Medical Publications, Oxford University Press: Oxford, 1988.
2. Gill L, Goldacre M, Simmons H, Bettley G, Griffith M. Computerised linking of medical records: methodological guidelines. *Journal of Epidemiology and Community Health (1979-)* 1993; **47**(4):316–319.
3. Howe GR. Use of computerized record linkage in cohort studies. *Epidemiologic Reviews* 1998; **20**(1):112–121.
4. St Sauver JL, Grossardt BR, Yawn BP, Melton LJ, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. *American Journal of Epidemiology* May 2011; **173**(9):1059–68. DOI: 10.1093/aje/kwq482.
5. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association* Dec 1969; **64**(328):1183. DOI: 10.2307/2286061.

6. Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association* 1965; **60**(312):1005–1027.
7. Scheuren F, Winkler WE. Regression analysis of data files that are computer matched part I. *Survey Methodology* 1993; **19**(1):39–58.
8. Scheuren F, Winkler WE. Regression analysis of data files that are computer matched part II. *Survey Methodology* 1997; **23**(2):126–138.
9. Lahiri P, Larsen MD. Regression analysis with linked data. *Journal of the American Statistical Association* Mar 2005; **100**(469):222–230. DOI: 10.1198/016214504000001277.
10. Chambers R, Chipperfield J, Davis W, Kovacevic M. Inference based on estimating equations and probability-linked data. *Working Paper 18-09*, Centre for Statistical and Survey Methodology, University of Wollongong, 2009: 1–36.
11. Kim G, Chambers R. Regression analysis under incomplete linkage. *Working Paper 17-09*, Centre for Statistical and Survey Methodology, University of Wollongong, 2009: 1–30.
12. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 2002; **31**(6):1246–1252. DOI: 10.1093/ije/31.6.1246.
13. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology* May 2011; **64**(5):565–572. DOI: 10.1016/j.jclinepi.2010.05.008.
14. Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and vital statistics databases using deterministic methods. *BMC Health Services Research* Jan 2006; **6**:48. DOI: 10.1186/1472-6963-6-48.
15. Larsen MD, Rubin DB. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association* March 2001; **96**(453):32–41. DOI: 10.1198/016214501750332956.
16. Yancey WE. Improving EM algorithm estimates for record linkage parameters. *Proceedings of the Section on Survey Research Methods*, St. Pete Beach Florida, American Statistical Association, 2002; 3835–3840.
17. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. *Journal of Biomedical Informatics* Feb 2010; **43**(1):24–30. DOI: 10.1016/j.jbi.2009.08.004.
18. Porter EH, Winkler WE. Approximate string comparison and its effect on an advanced record linkage system. *Advanced Record Linkage System. U.S. Bureau of the Census, Research Report*, 1997; 190–199.
19. Winkler WE. Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, St. Charles Illinois, American Statistical Association, 1993; 274–279.
20. Sariyar M, Borg A, Pommerening K. Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics* Aug 2011; **44**(4):648–654. DOI: 10.1016/j.jbi.2011.02.008.
21. Belin TR, Rubin DB. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* June 1995; **90**(430):694. DOI: 10.2307/2291082.
22. Green PJ. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* 1984; **46**:149–192.
23. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2009. ISBN 3-900051-07-0.