International Statistical Review

International Statistical Review (2023), **91**, 3, 368–394 doi: 10.1111/insr.12535

Improving Probabilistic Record Linkage Using Statistical Prediction Models

Angelo Moretti[1] and **Natalie Shlomo**[2]

[1]*Department of Methodology and Statistics, Utrecht University, Sjoerd Groenmangebouw, Padualaan 14 3584 CH Utrecht, The Netherlands*
[2]*Social Statistics Department, University of Manchester, Oxford Road, Manchester, M13 9PL, The United Kingdom*
Correspondence: Angelo Moretti, Department of Methodology and Statistics, Utrecht University, Sjoerd Groenmangebouw, Padualaan 14, 3584 CH Utrecht, The Netherlands. Email: a.moretti@uu.nl

Summary

Record linkage brings together information from records in two or more data sources that are believed to belong to the same statistical unit based on a common set of matching variables. Matching variables, however, can appear with errors and variations and the challenge is to link statistical units that are subject to error. We provide an overview of record linkage techniques and specifically investigate the classic Fellegi and Sunter probabilistic record linkage framework to assess whether the decision rule for classifying pairs into sets of matches and non-matches can be improved by incorporating a statistical prediction model. We also study whether the enhanced linkage rule can provide better results in terms of preserving associations between variables in the linked data file that are not used in the matching procedure. A simulation study and an application based on real data are used to evaluate the methods.

Key words: Linkage errors; matching variables; propensity scores; predictions.

1 Introduction

The aim of record linkage is to merge statistical units across different data sources based on a set of common matching variables. In the deterministic approach, a set of rules is used to classify pairs of records as matches and non-matches. This requires *exact* agreement based on specified matching variables; for example, the National Insurance number, the UK equivalent of a tax file number and surname agree exactly to be declared a match (Harron *et al*. 2016). However, matching variables may appear with errors, variations and missing values; hence, probabilistic record linkage is used. In probabilistic record linkage, records are matched based on a *probabilistic rule* that the records belong to the same unit (see Newcombe *et al*. 1959; Fellegi & Sunter 1969).

At the end of the record linkage process, a linked dataset is produced, and further statistical analysis can be carried out. Record linkage is an important tool that can be used to study relationships between information contained in administrative data and other data sources. For instance, record linkage is used in many applications in official statistics, including population size estimation at National Statistical Institutes (NSIs) through a linkage of a post-enumeration survey to the census. Many NSIs are considering replacing traditional censuses with linked population registers, administrative data and surveys.

In probabilistic record linkage, the statistical quality of the linked dataset depends on the matching variables used in the record linkage process. If matching variables have low errors and are able to distinguish between units, the record linkage process will be of high quality. However, users may be restricted on the use of matching variables due to privacy and confidentiality. For example, first and last names and addresses would not be available on a dataset for users. In addition, matching variables can be affected by missing data and variations. Thus, the matching variables that are ultimately used may have low power in determining the class of matches and non-matches under the procedure of probabilistic record linkage.

In this paper, we focus on the role of statistical prediction models to enhance the probabilistic record linkage procedure. This is motivated by the fact that the linked file may be used for complex statistical analyses where relationships between variables should be preserved, particularly for those variables that may not be included in the matching process. Indeed, this idea was initially studied by Scheuren & Winkler (1993) and Scheuren & Winkler (1997). They highlighted that 'it is important to conceptualise the linkage and analysis steps as part of a single statistical system'. They showed that a more integrated record linkage approach with linear regression modelling may improve the linkage and, hence, we also focus on the quality of the estimates calculated from the linked data file.

In Section 2, we provide an overview of deterministic and probabilistic record linkage approaches, particularly with details on the Fellegi & Sunter (1969) framework, which will form the basis of investigating the enhanced strategies in Section 3. For the purpose of covering the literature, we also present in Section 2 other approaches of data linkage. The focus of this paper is to enhance the Fellegi & Sunter (1969) framework, and we propose to include predictions from statistical models in the set of matching variables in Section 3. In Section 4, we present the results of a simulation study that we designed to investigate the behaviour of the proposed enhanced record linkage strategies based on the Fellegi & Sunter (1969) framework. In Section 5, we show an application based on real data from the 1991 Israel Income Survey. We conclude in Section 6 with conclusions and future directions.

2 Overview of Record Linkage

In this section, we provide a description of deterministic and probabilistic record linkage and also mention other developments in these areas, particularly when training data may be available.

2.1 Deterministic Record Linkage

The term record linkage was introduced by Dunn (1946) in the context of linking medical records at an individual level. In particular, the Dominion Bureau of Statistics in Canada developed a system where information related to names was added into punch cards, and then lists were printed for verification and subsequent review by Canadian agencies. At that time, this approach was more cost-effective than matching the paper files manually.

In deterministic record linkage, also known as exact or rules-based record linkage, the records present in two files must agree *exactly* on every character of every matching variable so that the linker can conclude that they correspond to the same unit (Roos & Wajda 1991; Shlomo 2019). A set of rules based on exact agreement/disagreement results between corresponding fields in potential record pairs are applied (Grannis *et al.* 2002).

This approach is generally used when a high-quality identifier is available, which can discriminate the units, such as an ID number from tax or national insurance. As an example, the community health index is used in the Scottish Record Linkage System (Harron *et al*. 2016). If an ID number is not available, matching variables, such as age and place of residence, can be concatenated in order to create a unique ID number. For example, Clark *et al*. (2019) use a deterministic record linkage approach to link ambulance and emergency department data in the UK based on the ambulance incident number and the vehicle shift number. However, matching variables may be prone to error (Grannis *et al*. 2002; Shlomo 2019). Therefore, it is important to highlight that deterministic record linkage techniques are typically prone to missed-matches because any recording errors (e.g. spelling error) or missing values can prevent a set of identifiers from agreeing. However, false-match rates are usually low (Grannis *et al*. 2002).

There have been approaches to allow for errors and variations in matching variables, also known as fuzzy matching. One way to quantify differences for textual attributes such as names or addresses is via phonetic codes or string distance functions (Cohen *et al*. 2003). The Jaro–Winkler distance performs well for the comparison of strings that have less than nine characters (Winkler 1990, 2006).

There are other modifications of deterministic linkage available in the literature, for example, stepwise deterministic record linkage where a succession of rules are used and the $n - 1$ deterministic approach, where a link is made if all but one of a set of n identifiers agree and linkage is based on partial identifiers (Maso *et al*. 2001; Abrahams & Davy 2002; Mears *et al*. 2010). Abrahams & Davy (2002) link maternity records to Office for National Statistics (ONS) birth records, whereas Mears *et al*. (2010) apply record linkage in the context of stroke patient care.

Deterministic record linkage approaches are characterised by their simplicity and computational scalability. They can be used as baseline comparison methods or as the first part of a blocking stage in probabilistic record linkage where blocks are created to reduce the search space of possible matches. Only the record pairs within the block are examined for possible matches. Another recent development proposed by Chipperfield *et al*. (2018) is the calculation of quality measures of precision and recall for a deterministic record linkage.

2.2 Probabilistic Record Linkage

Newcombe *et al*. (1959) was the first work that proposed a probabilistic approach to record linkage and the first automatic computer-based approach. In their application, they linked 34 138 birth records from 1955 British Columbia to 114 471 records of marriage from 1945. The linked file was then used in further studies (Newcombe & Rhynas 1962; Newcombe 1965; Newcombe & Tavendale 1965). The authors introduced a weight that is assigned to each pair of records. In particular, a frequency analysis of the data is carried out to calculate a weight for each variable based on the ratio of the number of agreements on the value of a variable for matched pairs compared with the number of agreements on the value of a variable for unmatched pairs (or random agreements). The weights for each variable are then aggregated to obtain an overall weight for the record pair. A large weight means that there is more chance of a correct match, whereas a small weight means that there is little chance of a correct match. The magnitude of the contribution to the weight depends on the discriminatory power of the identifiers (Zhu *et al*. 2009). The matching variables, however, are subject to errors so there can be an erroneous agreement on the value of the variable, which is also reflected in the weight (Shlomo 2019).

Empirical literature comparing the deterministic and probabilistic approaches have shown consistent improvements of probabilistic techniques over deterministic methods (Tromp *et al.* 2011; Dusetzina *et al.* 2014; Sadosky *et al.* 2015; Chen *et al.* 2018; Avoundjian *et al.* 2020).

In the next section, we provide more details of the Fellegi & Sunter (1969) probabilistic record linkage framework, which is the focus of our research.

2.3 Fellegi and Sunter Framework of Probabilistic Record Linkage

The traditional probabilistic record linkage is based on the Fellegi and Sunter (F&S) approach that formalised Newcombe's ideas mathematically and placed it into a decision theory framework (Fellegi & Sunter 1969).

Consider two datasets A and B to be linked with dimensions n_A and n_B, respectively, and let the records in each dataset be denoted by $a \in A,\ b \in B$. The record pairs are denoted by the product space $A \times B = \{(a,\ b);\ a \in A,\ b \in B\}$. The pairs are to be classified into three classes: true matches (*M*), true non-matches (*NM*) and potential matches, which will need to undergo clerical review (Fellegi & Sunter 1969; Herzog *et al.* 2007).

We denote by $X(a)$ and $X(b)$ the set of K matching variables for entity a in A and b in B, respectively. We assume that there is no duplication on $X(a)$ and $X(b)$ in the datasets. Based on these, the goal is to find a set of matches $M = \{(X(a),\ X(b))|a = b\}$ and a set of non-matches $NM = \{(X(a),\ X(b))|a \neq b\}$. [Correction added on 26 December 2022, after first online publication: the extra vertical line '|' has been removed in the preceding equations.]

Under this set up, we define a comparison vector γ for each record pair and define the comparison space $C{:}X(a) \times X(b) \rightarrow \Gamma$, composed of comparison vectors $\gamma \in \Gamma$ representing an agreement pattern. In the simplest agreement pattern case, we define 1 for agree and 0 for disagree on a value of the match variable *k, j = 1, …, K*. For example, if we consider three matching variables with a binary agreement pattern, we can have the following vector for pair *j*: $\gamma^j = \left(\gamma_1^j, \gamma_2^j, \gamma_3^j\right) = (1,\ 0,\ 1)$, which means agreement on variable 1, disagreement on variable 2 and agreement on variable 3.

The F&S framework defines the *m*-probability as the conditional probability that a record pair *j* has an agreement pattern γ^j given that it is a match (*M*), denoted as $m = P(\gamma^j|M)$, and the *u*-probability as the conditional probability that a record pair *j* has an agreement pattern γ^j given that it is not a match (NM), denoted as $u = P(\gamma^j|NM)$. Finally, let $P(M)$ be the marginal probability of a correct match.

The probability of interest is the match probability given an agreement pattern γ^j: $P(M|\gamma^j)$. According to Bayes' theorem, this is the posterior probability calculated as follows:

$$P(M|\gamma^j) = \frac{P(\gamma^j|M)P(M)}{P(\gamma^j)} = \frac{P(\gamma^j|M)P(M)}{P(\gamma^j|M)P(M) + P(\gamma^j|NM)(1 - P(M))}$$
$$= \frac{1}{1 + \frac{P(\gamma^j|NM)(1 - P(M))}{P(\gamma^j|M)P(M)}}. \tag{1}$$

Fellegi & Sunter (1969) introduce the following decision rule: The agreement (likelihood) ratio $R(\gamma^j) = P(\gamma^j|M)/P(\gamma^j|NM)$ is defined as the test statistic (overall score) for record pair *j*, because maximising the likelihood ratio is the same as maximising the posterior probability of $P(M|\gamma^j)$. Therefore, one can simply order the likelihood ratios $R(\gamma^j)$ and choose an upper cutoff W^+ and a lower cutoff W^- for determining the correct matches and correct non-matches according to ex ante error bounds on false matches and false non-matches. The linkage rule $F{:}\Gamma \rightarrow \{M,\ C,\ NM\}$ maps a record pair *j* comparison value to a set of three classes—matches

(*M*), non-matches (*NM*) and a set of undecided cases for manual clerical review (*C*)—defined as follows:

$$F: \begin{cases} \gamma^j \epsilon M \;\; if \; R(\gamma^j) \geq W^+ \\ \gamma^j \epsilon NM \;\; if \;\; R(\gamma^j) \leq W^- . \\ \gamma^j \epsilon C \;\;\; otherwise \end{cases} \tag{2}$$

The F&S framework assumes conditional independence across matching variables. This means that the errors associated with one matching variable are independent of the errors associated with another matching variable. Under conditional independence, the *m*- and *u*-probabilities can be decomposed according to the *K* matching variables as follows:

$$P(\gamma^j|M) = P(\gamma_1^j|M) \times P(\gamma_2^j|M) \times \ldots \times P(\gamma_K^j|M) \text{ and}$$

$$P(\gamma^j|NM) = P(\gamma_1^j|NM) \times P(\gamma_2^j|NM) \times \ldots \times P(\gamma_K^j|NM).$$

The likelihood ratio for record pair *j* then becomes

$$R(\gamma^j) = \frac{P(\gamma^j|M)}{P(\gamma^j|NM)} = \frac{P(\gamma_1^j|M) \times P(\gamma_2^j|M) \times \ldots \times P(\gamma_K^j|M)}{P(\gamma_1^j|NM) \times P(\gamma_2^j|NM) \times \ldots \times P(\gamma_K^j|NM)}.$$

Taking the log transformation, the overall score based on the likelihood ratio for record pair *j* is the sum:

$$\log[R(\gamma^j)] = log\left(\frac{P(\gamma_1^j|M)}{P(\gamma_1^j|NM)}\right) + log\left(\frac{P(\gamma_2^j|M)}{P(\gamma_2^j|NM)}\right) + \ldots + log\left(\frac{P(\gamma_K^j|M)}{P(\gamma_K^j|NM)}\right). \tag{3}$$

We acknowledge that in real data applications, there might be dependencies across attributes; hence, the conditional independence assumption may be violated. However, a large variety of record linkage applications showed good quality of the linked products even under this assumption (see, e.g. Herzog *et al*. 2007).

As noted by Christen (2012), one of the challenges of the probabilistic record linkage is the estimation of the *m*- and *u*-probabilities. In practice, these may be known from manual assessment of the quality of the databases that are to be matched or from a manual evaluation of a previous linkage of the same datasets (Herzog *et al*. 2007). One approach to estimating these probabilities as well as the prior probability $P(M)$ is the expectation–maximisation (EM) algorithm (Dempster *et al*. 1977; Winkler 1988; Grannis *et al*. 2003; Winkler 2006).

2.3.1 Expectation–maximisation algorithm

We consider the following decomposition of the probability of agreement for record pair *j*:

$$P(\gamma^j) = P(\gamma^j|M)P(M) + P(\gamma^j|NM)(1 - P(M)). \tag{4}$$

The left hand of 4 is the proportion of the agreement patterns across all possible pairs.

Assuming a simple agree/disagree $\{1,0\}$ pattern for each matching variable, the *m*-probability for a matching variable *k* in record pair *j* is distributed as a Bernoulli random variable:

$$P(\gamma_k^j|M) = m_k^{\gamma_k^j}(1 - m_k)^{1 - \gamma_k^j}. \tag{5}$$

Under the conditional independence assumption, we can write

$$P\left(\gamma^j|M\right) = \prod_k m_k^{\gamma_k^j}(1 - m_k)^{1 - \gamma_k^j}.$$

Similarly, the u-probability for a record pair j is

$$P\left(\gamma^j|NM\right) = \prod_k u_k^{\gamma_k^j}(1 - u_k)^{1 - \gamma_1^j}.$$

The unknown parameters are m_k, u_k for each matching variable k and $P(M)$. The EM algorithm is as follows:

- E-step: The indicator value is estimated for the true match status, denoted by $g_m^j = 1$ if pair j represents the same entity (set M) or 0 otherwise; $g_u^j = 1$ if pair j does not represent the same entity (set NM) or 0 otherwise.

  Applying Bayes' theorem, we obtain the following estimates:

$$\widehat{g}_m^j = \frac{\widehat{p}\prod_k \widehat{m}_k^{\gamma_k^j}(1 - \widehat{m}_k)^{1 - \gamma_k^j}}{\widehat{p}\prod_k \widehat{m}_k^{\gamma_k^j}(1 - \widehat{m}_k)^{1 - \gamma_k^j} + (1 - \widehat{p})\prod_k u_k^{\gamma_k^j}(1 - u_k)^{1 - \gamma_k^j}},$$

$$\widehat{g}_u^j = \frac{(1 - \widehat{p})\prod_k \widehat{u}_k^{\gamma_k^j}(1 - \widehat{u}_k)^{1 - \gamma_k^j}}{\widehat{p}\prod_k \widehat{m}_k^{\gamma_k^j}(1 - \widehat{m}_k)^{1 - \gamma_k^j} + (1 - \widehat{p})\prod_k u_k^{\gamma_k^j}(1 - u_k)^{1 - \gamma_k^j}},$$

where $\widehat{p}$ denotes initial values for the probability of a match $P(M)$.

- M-step: The values of three probabilities are updated as follows:

$$\widehat{m}_k = \frac{\sum_j g_m^j \gamma_k^j}{\sum_j g_m^j}, \quad \widehat{u}_k = \frac{\sum_j g_u^j \gamma_k^j}{\sum_j g_u^j}, \quad \widehat{p} = \frac{\sum_j g_m^j}{R},$$

where R is the number of record pairs.

These new estimates can be replaced in the E-step and iterated until the difference between the probabilities at iteration $t - 1$ and iteration t is below a small threshold (until convergence). To initiate the starting values of the EM algorithm, one can use an evaluation from a previous linkage of similar datasets.

2.4 Other Developments in Probabilistic Record Linkage

More recent work improves the classical record linkage approach described in Fellegi & Sunter (1969). Larsen & Rubin (2001) discuss eliminating the conditional independence assumption. Smith & Shlomo (2014) expand earlier work of Winkler (1990) to exploit similarities between values by replacing the binary agreement patterns with a multinomial agreement pattern according to bins defined by string comparators and the record linkage parameters are estimated under the multinomial EM algorithm.

Training data may be available, and in this case, they can be used to estimate the probabilities for the F&S probabilistic record linkage, accounting for relationships between the matching fields, and provide estimates of error rates and cutoff thresholds (Winkler 2006). If a training dataset is available, one can set a rule-based learning system covering all training records in the training data (Sarawagi 2008). However, finding the optimal set of such rules in the given training dataset is intractable; thus, rule-based learning algorithms are based on heuristic approaches in practice (Christen 2012). These are discussed by Sarawagi (2008) in detail. A disadvantage of the approaches that are based on rules is that in case of unseen variations in the training data that are not covered by any rule, a new rule has to be created. This can be time consuming, because it requires adjustments. Furthermore, comprehensive training data files are needed and can be expensive to generate (Prasad *et al*. 2009; Christen 2012).

Having training data also allows for classification techniques in the context of supervised-learning record linkage. In the learning stage, a tree is constructed recursively, where the first tree is empty. In each step, an attribute resulting in the split of the training dataset is selected. This is done in such a way that matches are moved into one branch of the tree and non-matches into the other one. An early application of this approach for record linkage can be seen in Cochinwala *et al*. (2001), where two data files of customer records were linked. Here, a training dataset was used to train a Classification and Regression Tree classifier (Breiman *et al*. 1984). Christen (2008a, 2008b) developed an automatic classification approach for record linkage based on a support vector machine that maps the training data into a vector space in such a way that the records from the two classes (matches or non-matches) are separated.

In the Bayesian framework, Fortini *et al*. (2001) proposed a Bayesian approach to record linkage between two datasets. Their approach can be interpreted as a Bayesian version of the F&S framework. A prior on the number of matching pairs is considered, together with a prior on the matching configuration matrix, indicating the linkage structure between the two datasets to link. They also propose a Dirichlet prior on the m- and u-probability distributions. These parameters are then estimated via Markov chain Monte Carlo (MCMC). Sadinle (2014) proposes a Bayesian F&S technique for deduplication, relying on comparison vector data. Here, a prior on the matching configuration matrix is considered, imposing transitive closure; that is, the records are partitioned into groups where they are believed to refer to the same entity.

An important challenge of Bayesian F&S is their computational burden (Binette & Steorts 2020). In recent work, McVeigh *et al*. (2019) considered this problem by proposing a blocking approach based on simpler probabilistic record linkage techniques.

Steorts *et al*. (2014) and Steorts *et al*. (2016) develop a fully hierarchical-Bayesian approach to entity resolution in presence of categorical latent attributes assuming a data distortion model. They formulate an efficient hybrid Metropolis-within-Gibbs MCMC algorithm for estimating these models [Split and MErge REcord linkage and Deduplication (known as SMERED)]. This allows for full quantification of uncertainty around the number of latent individuals and the clustering structure of records. Additionally, Steorts (2015) proposes approaches for both categorical and noisy string data, proposing a string pseudo-likelihood and an empirically motivated prior [i.e. the empirical Bayes (EB) method].

As mentioned by Winkler (2006), a representative training dataset is rarely available in practice. Indeed, there are difficulties in generating, obtaining and selecting a training data representative of the actual data that are to be linked. Without the availability of a training data, the F&S approach has been widely adopted together with the EM algorithm and the literature has shown that it performs reasonably well (Xu *et al*. 2021).

There has also been some attention devoted to secondary data analysis of linked data, considering that in the vast majority of cases, data analysts are not involved in the linking process. Thus, they do not have information on the probabilities underpinning the record linkage; rather, they only have access to the final linked file. Chambers (2009) and Kim & Chambers (2012a, 2012b) assume a model for linkage errors requiring non-sensitive information on the performance of the record linkage and propose bias correction methods. Chambers & Kim (2016) discuss the measurement error issues that arise from record linkage focusing on linear regression models. Chipperfield & Chambers (2015) and Scholtus *et al*. (2022) develop a method to take into account linkage errors in analysing linked categorical data.

3 Enhancing Probabilistic Record Linkage Based on Statistical Models

There has been some work in the literature where statistical models were proposed to improve the quality of a linked file following probabilistic record linkage (see Armstrong & Mayda 1992; Winkler 1992, 1993; and Thibaudeau 1993). In these articles, log-linear models with interaction terms are proposed in order to take into account dependencies between matching variables. More recently, Xu *et al*. (2019) apply latent class models with a conditional dependence structure informed by the true match status of manually reviewed record pairs as training data. In one case, where the variables have poor discriminating power, the conditional dependence models return improved matching accuracy compared to the traditional F&S model. Daggy *et al*. (2014) also discussed and evaluated the use of conditional dependency statistical models in record linkage applications.

In this section, we introduce two approaches for enhancing probabilistic record linkage where we add matching variables into the F&S probabilistic record linkage procedure described in Section 2.3. The aim is to improve the decision theory for declaring a correct match and to improve the multivariate relationships in the linked data including those variables that may not have been used in the matching procedure.

3.1 The Use of Propensity Scores in Record Linkage

Propensity score matching is a widely used matching technique to estimate causal treatment effects and it finds applications in many fields (Dehejia & Wahba 1999; Perkins *et al*. 2000; Hitt & Frei 2002; Caliendo & Kopeinig 2005).

According to Rosenbaum & Rubin (1983), the propensity scores are used to draw causal conclusions from observational studies where a ‘treatment’ is not randomly assigned. It is defined as the conditional probability of a data subject receiving a treatment instead of the control given observed covariates common to both groups. Another way to look at propensity scores is that they create a balance between treatment and control groups and can be viewed as a data-dimension reduction approach accounting for the values of the covariates.

Propensity scores have been used to match records that do not belong to the same entity in statistical matching. For example, Kum & Masterson (2010) used propensity scores to statistically match two datasets to estimate the distribution of income and wealth in the USA. Here, we assess whether adding propensity scores to the set of matching variables can improve the decision rule in probabilistic record linkage. The reasons why this might be the case are two-fold: Firstly, as a data reduction approach based on the correlation structure of variables in the dataset, we conjecture that it could add power to the decision to declare a correct match on a record pair by including a second-order statistic. Secondly, we can add other variables to the propensity score model, particularly interactions and continuous variables, that may not be traditionally used as matching variables and improve the quality of the linked dataset.

Here, we add a stratification of the propensity scores and include it as an additional matching variable.

The procedure is as follows.

We first stack file A and file B and define a new variable denoted by I as follows: $I_a = 1, \forall a \in A$ and $I_b = 0, \forall b \in B$.

The new data file is called S with dimension $n_A + n_B = n_S$ and indexes $i = 1, \ldots, n_S$.

Next, the following logistic regression model is estimated on S:

$$P(I = 1|\boldsymbol{X}, \boldsymbol{\beta}) = \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{X}\right)}{1 + \exp\left(\boldsymbol{\beta}^T \boldsymbol{X}\right)}, \tag{6}$$

where $\boldsymbol{X}$ contains variables common to both file A and file B and may include some matching variables and other relevant variables or interactions: $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}(a) \\ \boldsymbol{X}(b) \end{bmatrix}$. [Correction added on 26 December 2022, after first online publication: the extra vertical line '|' in equation (6) has been removed in this version.]. From model 6, we obtain the estimated predicted probabilities, denoted by $\widehat{p}_i$ in file S as follows, where $\boldsymbol{X}_i$ relates to $\boldsymbol{X}$ variables for $i = 1, \ldots, n_S$:

$$\widehat{p}_i = \frac{\exp\left(\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i\right)}{1 + \exp\left(\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i\right)}. \tag{7}$$

Once predicted probabilities are estimated, files A and B are separated again, and the propensity scores are attached to each file and discretised into strata to be used as an additional matching variable.

3.2 The Use of Linear Model Predictions in Record Linkage

Again, motivated by the statistical matching literature (Rässler 2002; Moriarity & Scheuren 2003; D'Orazio *et al.* 2006), we introduce the use of a linear prediction model to add predictions as additional matching variables into the F&S probabilistic record linkage, particularly if one dataset has a potential match variable Y whilst the other dataset does not. Our motives are similar to adding the propensity score prediction defined in Section 3.1: (1) add power to the decision to declare a correct match on a record pair with more matching variables; (2) potential to preserve relationships between variables in the merged dataset, including variables that may not have been included as an original matching variable.

We assume that a continuous variable Y is available in file A only, Y_i for $i = 1, \ldots, n_A$. For example, when the aim is to link an Income Survey (file A) to administrative or health records (file B), the administrative records may not have an explicit income variable. In this case, we can link income from the Income Survey to a predicted income in the administrative file according to a regression model established on file A on a set of common covariates from the administrative records. In this way, the record linkage can be seen as compensating for a missing data problem (Goldstein & Harron 2016). As mentioned, the statistical matching literature have shown that the use of an explicit parametric model, that is, a linear regression model, helps in preserving the correlation structure of the variables in the linked dataset. This is important, because users will not be able to estimate relationships of variables that are not observed jointly. This strategy is usually adopted in statistical matching but to the best of our knowledge not in record linkage problems.

In this approach, first, the following linear regression model is estimated for a known Y on file A:

$$Y_a = \boldsymbol{X}_a \boldsymbol{\beta} + e_a, \quad e_a \sim N\left(0, \sigma^2\right) \tag{8}$$

with the usual assumptions of the linear regression model. Once the estimate of $\boldsymbol{\beta}$ is obtained, that is, $\widehat{\boldsymbol{\beta}}$, this is used to predict Y in file B using the same set of covariates in file B for $i = 1, \ldots, n_B$. For the record linkage, Y in file A and its predictions $\widehat{Y}$ in file B are discretised into strata to be used as an additional matching variable.

In this approach, we focus on the scenario where we aim to carry out downstream regression models in the linked dataset where the outcome Y is on one file, covariates X appear in both files and there may be additional covariates W in one or both files. By including an additional matching variable of Y and prediction $\widehat{Y}$ obtained from the common X covariates, we aim to assess whether we can improve the quality of the final regression model on the linked dataset. We also assess the case where prediction models and models of analysis may not be aligned. Another extension to this strategy is to define the additional matching variable on strata defined by *both* sets of predictions (where predictions are also obtained in file A using the same $\widehat{\boldsymbol{\beta}}$) similar to the notion of predictive mean matching (Rubin 1986).

4 Simulation Study

This simulation study is designed to compare three record linkage strategies presented in Sections 2 and 3 using original matching variables, matching variables with propensity score strata and matching variables with a $Y\,(\widehat{Y})$ strata. We first evaluate the strategies in terms of how well we determine correct matches. This is done by comparing decision matrices that define how well we estimate the parameters of the record linkage, for example, *m*- and *u*-probabilities, and the precision and recall measures as defined in Table 2. To compare across the strategies, we normalise the final match weights for each pair and use the same threshold for determining correct matches for each strategy. Second, we assess the quality of the linked dataset with respect to preserving correlations and regression modelling in the final linked dataset, assuming that users will use the linked file for further statistical analyses.

4.1 Generating the Population

We generate a population of size $N = 50\,000$ with the following variables for $i = 1, \ldots, N$:

- $x_{1i} \sim Pareto(2\,000, 16\,205)$,
- $x_{2i} \sim Pareto(3\,000, 40\,000)$,
- $x_{3i} \sim Pareto(10\,000, 50\,000)$.

We also introduce two extra variables, related to x_{1i}, x_{2i} and x_{3i}, via a linear model to evaluate the role of the statistical prediction model in the probabilistic record linkage. Two cases, small R^2 and medium R^2, are considered as follows:

<u>Small</u> R^2:

- $y_{1i} = 0.5 - 0.003x_{1i} + 0.5x_{2i} + 0.06x_{3i} + e_{1i}, \quad e_{1i} \sim N(0,\ 2\,500),\ R^2 = 0.30.$
- $y_{2i} = 0.09 - 0.010x_{1i} + 0.4x_{2i} + 0.18x_{3i} + e_{2i}, \quad e_{2i} \sim N(0,\ 5\,700), \quad R^2 = 0.13.$

<u>Medium</u> R^2:

- $y_{1i} = 0.2 + 0.20x_{1i} + 0.1x_{2i} + 0.2x_{3i} + e_{1i}, \quad e_{1i} \sim N(0,\ 2\,500), \quad R^2 = 0.45.$
- $y_{2i} = 0.3 - 0.9x_{1i} + 0.35x_{2i} + 0.55x_{3i} + e_{2i}, \quad e_{2i} \sim N(0,\ 5\,700), \quad R^2 = 0.51.$

We also generate two additional variables. One of these is generated from a discrete Uniform distribution (*dUnif*), that is, $x_{4i} \sim dUnif(20, 40)$, and is used as a blocking variable. The other variable is generated as binary variable as follows: $x_{0i} \sim Bernoulli(0.2)$ to enable an interaction term. We assume that there is no error in either of these variables.

The variables used as matching variables in the probabilistic record linkage are rounded to produce integer values. We note that there are no duplicates on the set of matching variables and we use a one-to-one matching approach. Table 1 shows the number of categories of the variables that were generated in the population.

4.2 Simulation Steps

The simulation consists of the following steps:

1 *Sampling*: From the population, select a 1:40 simple random sample without replacement denoted by s of size $n = 1\,250$, for $s = 1, \ldots, S$, $S = 500$.
2 *Perturbation*: Under two settings: 10% and 20% of the records of the variables x_{1i}, x_{2i} and x_{3i} are perturbed using a lag operator, so that the value of the variable for unit i takes the value of the variable for unit $i - 1$ (in case of unit $i = 1$, this is unchanged). The reason why we use the lag operator to perturb file B is because we want to ensure real values in the distributions. The perturbation is carried out according to all the possible agreement pattern profiles as follows: perturb x_{1is}, perturb x_{2is}, perturb x_{3is}, perturb x_{1is} and x_{2is}, perturb x_{1is} and x_{3is}, perturb x_{2is} and x_{3is}, perturb x_{1is} and x_{2is} and x_{3is}. The perturbed variables are denoted by $x_{1is}^{pert}, x_{2is}^{pert}$ and x_{3is}^{pert}.
3 *Files creation*: Two data files are created, A and B: $A = (x_{1is}, x_{2is}, x_{3is}, y_{1is}, x_{4is}, x_{0is})$ and $B = (x_{1is}, x_{2is}, x_{3is}, y_{2is}, x_{4is}, x_{0is})$, where $x_{1is} = x_{1is}^{pert}$, $x_{2is} = x_{2is}^{pert}$, $x_{3is} = x_{3is}^{pert}$ for $i \in B$. x_{4is} and x_{0is} are not perturbed. x_{4is} is used as a blocking variable. Without loss of generality, we assume that $n_A = n_B = n$.

Table 1. *Numbers of categories of the variables generated in the population.*

Variable	*Number of categories*
X_0	2
X_1	7 559
X_2	10 118
X_3	21 986
X_4 (blocking variable)	21
$Y_1^{R^2=0.30}$	12 767
$Y_2^{R^2=0.13}$	21 284
$Y_1^{R^2=0.41}$	13 615
$Y_2^{R^2=0.50}$	24 640

Table 2. *Decision matrix example for sample s.*

		True status	
		Non-matches (NM)	*Matches (M)*
Decision	Not linked pairs (NL)	$\mathrm{NL\_NM_s}$	$\mathrm{NL\_M_s}$
	Linked pairs (L)	$\mathrm{L\_NM_s}$	$\mathrm{LM_s}$
Recall (sensitivity)		$\mathrm{LM_s}/(\mathrm{LM_s} + \mathrm{NL\_M_s})$	
Precision		$\mathrm{LM_s}/(\ \mathrm{LM_s} + \mathrm{L\_NM_s})$	

4 *Record linkage:* Strategies A, B and C are evaluated in this study and details are as follows.

Strategy A: Probabilistic record linkage with X_1, X_2, X_3 as matching variables.
Strategy B: This procedure involves several steps:

i Sample B is stacked beneath sample A.
ii Create an indicator variable $I_{is} = 1$ if $i \in A$, $I_{is} = 0$ if $i \in B$.
iii We consider two scenarios and estimate the following models: Model_B0: $logit(p_{is}) = \beta_0 + \beta_1 x_{1is} + \beta_2\, x_{2is} + \beta_3\, x_{3is}$, and Model_B1: $logit(p_{is}) = \beta_0 + \beta_1 x_{1is} + \beta_2\, x_{2is} + \beta_3\, x_{3is} + \beta_4 (x_{0is} \times x_{1is})$ for $i = 1, \ldots, n_A + n_B$, with $x_{0is} \times x_{1is}$ being an interaction term. Strategy B with the interaction term from Model_B1 is called B1 in the results section.
iv Obtain the propensity score as follows:
$\widehat{p}_{is} = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{1is} + \widehat{\beta}_2\, x_{2is} + \widehat{\beta}_3\, x_{3is}\right) / \left[1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{1is} + \widehat{\beta}_2\, x_{2is} + \widehat{\beta}_3\, x_{3is}\right)\right]$ (without interaction term, and in case of an interaction term, we add $\widehat{\beta}_4 (x_{0is} \times x_{1is})$). Note that the propensity scores are calculated on the perturbed X variables in file B.
$\widehat{p}_{is}$ is then stratified in $q = 10$ quantiles and this new variable is denoted by $\widehat{p}^*_{is}$.

v Attach $\widehat{p}^*_{is}$ with $i = 1, \ldots, n_A$ to A and $\widehat{p}^*_{is}$ with $i = 1, \ldots, n_B$ to B.
vi Proceed with probabilistic record linkage with the following matching variables X_1, X_2, X_3 and $\widehat{p}^*_{is}$ under the two scenarios: Strategy B and Strategy B1.

Strategy C: The following steps are involved:

i We again consider two models and estimate the linear models on file A: Model_C0: $y_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2\, x_{2i} + \beta_3\, x_{3i} + e_1$ and Model_C1: $y_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2\, x_{2i} + \beta_3\, x_{3i} + \beta_4 (x_{0is} \times x_{1is}) + e_1$ for $i = 1, \ldots, n_A$,$e_i \sim N\left(0, \sigma^2_{e_1}\right)$. Strategy C with the interaction term is called C1 in the results section.Obtain the prediction of y_{1is} on file B for $i = 1, \ldots, n_B$:$\widehat{y}_{1is} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1is} + \widehat{\beta}_2\, x_{2is} + \widehat{\beta}_3\, x_{3is}$ (without interaction term, and in case of the interaction term, we add $\widehat{\beta}_4 (x_{0is} \times x_{1is})$ to the model).Note that the predictions are calculated on the perturbed X variables in file B.
ii Proceed with probabilistic record linkage with the following matching variables $X_1, X_2, X_3,\ Y_1$ ($\widehat{Y}_1$ infile B) where Y_1 $\left(\widehat{Y}_1\right)$ is transformed into 10 quantiles when used as a matching variable, under the two scenarios: Strategy C and Strategy C1.
iii This strategy is repeated for the case of medium R^2 and small R^2.

5 *Creating the linked file*: The linked file, denoted by LF, is a matched dataset obtained via one-to-one forced matches with a sample size of $n_{LF} = 1\ 250$. It contains $X_1, X_2, X_3,\ Y_1, Y_2$ and X_0. Note that Y_2 is not used as a matching variable in the record linkage and is present in file B only.

The steps 1–4 are repeated for s = 1, …, S, with S = 500.

In order to show fair comparisons among the strategies across the 500 samples, we use the same decision threshold after normalising the final match weight for the pairs on a scale [0,1]. The normalisation is based on the following, where j denotes the pair and k is the matching variable: For $W_k = \log(x_k) = \log(m_k/u_k)$ if corresponding variable k matches across sources, or $W_k = \log(x_k) = \log(1 - m_k/1 - u_k)$ if corresponding variable i does not match across sources, then for pair j, the overall final match weight is $W_j = \sum_k W_k$. Let $W_0 = E/(A \times B - E)$ where E represents the number of expected matches. The normalised final match weight is $pr_j = e^{W_0 + W_j}/(1 + e^{W_0 + W_j})$.

Here, we set a standard threshold of 0.0001 in order to ensure consistency across all strategies and enable the comparison of strategies based on this common threshold. This is opposed to the common practice of identifying optimal thresholds according to test data. In addition, the one-to-one forced record linkage is performed using an optimisation by van der Laan (2018). In particular, it seeks to optimise the total weight of the selected records under the constraint that each record can be selected only once.

In order to evaluate the impact of the different strategies on regression model estimates in the linked file (LF), we estimate the following linear regression models: Model_LF0 $y_{2is} = \beta_1 x_{1is} + \beta_2 x_{2is} + \beta_3 x_{3is} + e_i$ and a full model: Model_LF1 $y_{2is} = \beta_1 x_{0is} + \beta_2 x_{1is} + \beta_3 x_{2is} + \beta_4 x_{3is} + \beta_5 y_{1is} + \beta_6 (x_{0is} \times x_{1is}) + e_i$, where $(x_{0is} \times x_{1is})$ denotes the interaction term and we also include the variable y_{1is} for $i = 1, \ldots, n_{LF}$. Therefore, after linkage, we can use unperturbed X variables and also include Y_1 in the model of interest, thus showing the advantage of carrying out the record linkage for conducting further statistical analyses. In the linked file, X_1, X_2, X_3, Y_1 are taken from A, whereas Y_2 is taken from B. Furthermore, we also derive the analysis model Model_LF0 on file B only and not the linked file (LF) (denoted Model_orig): $y_{2is} = \beta_1 x_{1is} + \beta_2 x_{2is} + \beta_3 x_{3is} + e_i$ to demonstrate the advantage of having the linked file for statistical analyses.

Note that the same models are estimated on the population where the obtained regression coefficients are assumed to be the true values for comparison to calculate the bias.

All the computations and analyses in this article are produced using *R* software. Particularly, the record linkage is performed using the program developed by van der Laan (2018). The package is available from the CRAN (van der Laan 2018) and the written program is also available on his GitHub page at the link https://github.com/djvanderlaan/reclin.

4.3 Performance Evaluation

The Absolute Relative Bias (ARB) is used as a performance measure in our simulation study and calculated as follows:

Absolute Relative Bias (ARB)

$$ARB(\theta) = \left| \frac{\sum_{s=1}^{S} \left(\widehat{\theta}_s - \theta\right)}{\theta} \right|, \tag{9}$$

where $\widehat{\theta}_s$ denote an estimator for the true parameter in the population θ in sample s, for example, the estimator of the correlation coefficient or regression model parameters.

Also, in the results section, we present the average values, across the simulations of the estimates, that is,

$$\widehat{\theta} = \sum_{s=1}^{S} \widehat{\theta}_s / S. \tag{10}$$

In order to evaluate the performance of the strategies, the decision matrix for each repetition s is produced and then averaged values across all $S = 500$ samples. Table 2 shows the example decision matrix for sample s. We evaluate the record linkage using two quality measures: Recall (sensitivity) and the Precision as defined in Table 2. All indicators are averaged across the samples s to provide summaries.

In order to evaluate whether the association between variables in the linked file LF_s is preserved, compared with the population, the correlations between Y_1 and Y_2 and correlations between Y_2 with the auxiliary variables X_1, X_2, X_3 are calculated in the linked file for each sample and the average across the samples are described in Section 4.4. We also compare the coefficients of the regression model Model_LF0 using Y_2 as the dependent variable (not involved in the linkage process) with explanatory variables X_1, X_2, X_3 averaged across the samples as well as the regression model Model_LF1 using Y_2 as the dependent variable, which includes the interaction term $(x_{0is} \times x_{1is})$ and the Y_1 variable averaged across the samples. We also show that Model_LF0 has less biased parameters compared with the case of Model_orig if we did not have the linked file and estimated the model on file B only.

4.4 Results

Table 3 presents the average of the decision matrices, recall and precision for the 10% perturbation and Table 4 the average of the decision matrices, recall and precision for the 20% perturbation for strategies A, B, B1, C and C1 with medium and small R^2.

From Table 3, it can be seen that when the perturbation is equal to 10%, the use of the propensity score stratification, particularly strategy B1, slightly improves the record linkage in terms of recall compared with strategy A but the precision slightly decreases. The use of prediction stratification (both strategies C and C1) did not provide much improvement to precision and recall compared with strategy A when R^2 is small. However, for a medium R^2, strategy C and

Table 3. *Decision matrix for the three strategies: 10% perturbation.*

	A		*B*		*B1*		*C (medium R^2)*	
Decision	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*	*M*
NL	76 183	35	76 174	34	76 247	30	76 182	5
L	4	1 215	11	1 216	6	1 220	4	1 245
Recall	0.972		0.973		0.976		0.995	
Precision	0.997		0.991		0.995		0.998	

Table 3. *(Continued)*

	C (small R^2)		*C1 (medium R^2)*		*C1 (small R^2)*	
Decision	*NM*	*M*	*NM*	*M*	*NM*	*M*
NL	76 180	33	76 246	3	76 247	31
L	5	1 217	6	1 247	5	1 219
Recall	0.974		0.998		0.975	
Precision	0.996		0.995		0.996	

Table 4. *Decision matrix for the three strategies: 20% perturbation.*

	A		*B*		*B1*		*C (medium R^2)*		*C (small R^2)*		*C1 (medium R^2)*		*C1 (small R^2)*	
Decision	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*	*M*
NL	76 161	51	76 163	49	76 193	48	76 188	24	76 151	45	76 168	14	76 198	40
L	51	1 199	49	1 201	48	1 202	22	1 226	45	1 205	20	1 236	42	1 210
Recall	0.959		0.961		0.962		0.981		0.964		0.989		0.968	
Precision	0.959		0.961		0.962		0.983		0.964		0.984		0.966	

Table 5. *Average correlation coefficients ρ and average absolute relative bias (in parentheses) across samples for all strategies on data with small and large R^2: 10% perturbation.*

	Small R^2						*Medium R^2*					
Correlations	*ρ*	*A*	*B*	*B1*	*C*	*C1*	*ρ*	*A*	*B*	*B1*	*C*	*C1*
$\boldsymbol{Y_1}, \boldsymbol{Y_2}$	0.166	0.165 (0.003)	0.164 (0.009)	0.166 (0.001)	0.166 (0.002)	0.167 (0.007)	0.409	0.411 (0.006)	0.405 (0.011)	0.406 (0.011)	0.408 (0.003)	0.408 (0.002)
$\boldsymbol{Y_2}, \boldsymbol{X_1}$	0.000	0.002 (0.002)	0.002 (0.002)	0.000 (0.001)	0.002 (0.001)	0.000 (0.001)	−0.214	−0.213 (0.020)	−0.213 (0.028)	−0.213 (0.027)	−0.212 (0.021)	−0.213 (0.010)
$\boldsymbol{Y_2}, \boldsymbol{X_2}$	0.201	0.201 (0.001)	0.200 (0.006)	0.198 (0.014)	0.202 (0.003)	0.199 (0.010)	0.131	0.128 (0.004)	0.127 (0.011)	0.127 (0.011)	0.128 (0.003)	0.130 (0.002)
$\boldsymbol{Y_2}, \boldsymbol{X_3}$	0.299	0.300 (0.000)	0.298 (0.007)	0.298 (0.005)	0.299 (0.001)	0.298 (0.004)	0.676	0.673 (0.003)	0.668 (0.009)	0.667 (0.008)	0.675 (0.001)	0.675 (0.001)

strategy C1 both outperform all other strategies particularly in the recall measure. There is no marked difference between strategy C and strategy C1 for the 10% perturbation. In the case of a 20% level perturbation shown in Table 4, as expected, the measures of recall and precision are generally smaller than those obtained from the 10% perturbation. We see more clearly under the 20% perturbation that the prediction stratification of strategies C and C1 outperform other approaches with respect to recall and precision when the R^2 is medium, with strategy C1 having the highest recall and precision. There is also a slight improvement in strategies C and C1 when R^2 is small under the 20% perturbation.

Table 5 presents the average correlation coefficients across the samples of Y_1 and Y_2 and Y_2 with the variables X_1, X_2, X_3 according to strategies A, B and C and modifications B1 and C1, using the data based on the small and medium R^2 compared with the correlations in the population for the 10% perturbation. Table 6 presents the same findings for the 20% perturbation. We also include in the tables the average absolute relative bias as defined in (9) in parentheses.

From Table 5, it can be seen that the correlation coefficients are similar across all strategies with smaller absolute relative bias under the 10% perturbation compared with the 20% perturbation in Table 6. Under the 10% perturbation in Table 5, there are little differences between the strategies for both small and medium R^2. Strategy B shows slightly more bias in the correlations of Y_2 and X_2. There is higher bias for all strategies in the correlation of Y_2 and X_1 with strategy C1 showing the smallest bias. Under the 20% perturbation in Table 6, there are little differences between the strategies for small R^2 except for strategy C1, which outperforms all other strategies. This also holds when R^2 is medium. In summary, similar to the conclusions found on the recall and precision measures, strategy C1 with prediction stratification using Model_C1 with the interaction term and having a medium R^2 outperforms the other linkage approaches under the correct model. Strategy C on Model_C0 also performed well compared with other strategies A, B and B1.

In Tables 7 and 8, we show the results of the estimation of three regression models. We show the two models in the linked file: Model_LF0: $y_{2i} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$ and Model_LF1: $y_{2is} = \beta_1 x_{0is} + \beta_2 x_{1is} + \beta_3 x_{2is} + \beta_4 x_{3is} + \beta_5 y_{1is} + \beta_6 (x_{0is} \times x_{1is}) + e_i$. Recall that the variables Y_2 and X_0 were not part of the record linkage strategies, X_0, X_1, X_2, X_3 and Y_1 are taken from file A whereas Y_2 is taken from file B, as described in the simulation steps. We also present the results of model Model_orig: $y_{2i} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$ estimated on file B only.

The averages of the regression coefficient estimates across the samples are shown in Table 7 for the 10% perturbation with the average absolute relative bias in (9) in parentheses. Similarly, the averages of the regression coefficient estimates and the average absolute relative bias are shown in Table 8 for the 20% perturbation.

In Table 7 for the 10% perturbation, when R^2 is small and the model of analysis is Model_LF0 (without an interaction), there are mixed results across the strategies. For the intercept, strategy A has the smallest bias for this case but when R^2 is medium, we see slightly smaller bias on the intercept with strategies C and C1. The bias is more pronounced in Table 8 under the 20% perturbation with strategies C and C1 having overall the smallest bias in all

Table 6. *Average correlation coefficients and average absolute relative bias (in parentheses) across samples for all strategies on data with small and large R^2: 20% perturbation.*

	Small R^2					*Medium R^2*	
Correlations	*ρ*	*A*	*B*	*B1*	*C*	*C1*	*ρ*
$\boldsymbol{Y_1, Y_2}$	0.166	0.159 (0.038)	0.160 (0.035)	0.160 (0.033)	0.166 (0.014)	0.166 (0.000)	0.409
$\boldsymbol{Y_2, X_1}$	0.000	0.002 (0.007)	0.002 (0.005)	−0.001 (0.001)	0.001 (0.001)	−0.001 (0.001)	−0.214
$\boldsymbol{Y_2, X_2}$	0.201	0.196 (0.025)	0.196 (0.024)	0.192 (0.046)	0.197 (0.018)	0.196 (0.001)	0.131
$\boldsymbol{Y_2, X_3}$	0.299	0.287 (0.042)	0.288 (0.040)	0.288 (0.039)	0.287 (0.042)	0.291 (0.030)	0.676

Table 6. *(Continued)*

	Medium R^2				
Correlations	*A*	*B*	*B1*	*C*	*C1*
$\boldsymbol{Y_1, Y_2}$	0.390 (0.044)	0.393 (0.040)	0.392 (0.043)	0.405 (0.010)	0.406 (0.008)
$\boldsymbol{Y_2, X_1}$	−0.210 (0.004)	−0.212 (0.003)	−0.214 (0.004)	−0.213 (0.002)	−0.214 (0.001)
$\boldsymbol{Y_2, X_2}$	0.124 (0.054)	0.126 (0.052)	0.124 (0.040)	0.129 (0.036)	0.129 (0.026)
$\boldsymbol{Y_2, X_3}$	0.647 (0.042)	0.650 (0.040)	0.648 (0.021)	0.670 (0.020)	0.673 (0.014)

Table 7. *Regression model on linked file LF (Model_LF0) and on file B only (Model_orig) $y_{2i} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$ and regression model on linked file LF (Model_LF1) $y_{2is} = \beta_1 x_{0is} + \beta_2 x_{1is} + \beta_3 x_{2is} + \beta_4 x_{3is} + \beta_5 y_{1is} + \beta_6 (x_{0is} \times x_{1is}) + e_i$ coefficient estimates averaged across the samples with absolute relative bias (in parentheses): 10% perturbation.*

	Small R^2						*Medium R^2*					
	True	*Strategy*					*True*	*Strategy*				
	value	*A*	*B*	*B1*	*C*	*C1*	*value*	*A*	*B*	*B1*	*C*	*C1*
						Model_LF0						
β_1	−0.012	−0.011	−0.008	−0.010	−0.010	−0.010	−0.903	−0.899	−0.894	−0.898	−0.900	−0.898
		(0.133)	(0.333)	(0.200)	(0.168)	(0.165)		(0.003)	(0.009)	(0.005)	(0.002)	(0.005)
β_2	0.404	0.402	0.401	0.402	0.402	0.402	0.353	0.355	0.352	0.353	0.353	0.353
		(0.006)	(0.007)	(0.006)	(0.005)	(0.004)		(0.004)	(0.001)	(0.002)	(0.001)	(0.003)
β_3	0.179	0.180	0.179	0.180	0.180	0.180	0.548	0.545	0.547	0.548	0.549	0.548
		(0.004)	(0.000)	(0.002)	(0.003)	(0.002)		(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
						Model_orig						
β_1	−0.012	−0.029 (0.198)					−0.903	−0.746 (0.173)				
β_2	0.404	0.387 (0.042)					0.353	0.358 (0.010)				
β_3	0.179	0.169 (0.056)					0.548	0.506 (0.079)				
						Model_LF1 (with interaction)						
β_1	−4.593	−1.335	−5.851	−4.523	−3.763	−5.290	−4.456	−3.928	−4.792	−4.600	−4.575	−4.437
		(0.709)	(0.300)	(0.090)	(0.181)	(0.152)		(1.881)	(0.237)	(0.172)	(0.222)	(0.169)
β_2	−0.055	−0.051	−0.050	−0.053	−0.050	−0.050	−0.946	−0.940	−0.942	−0.945	−0.936	−0.937
		(0.061)	(0.025)	(0.019)	(0.020)	(0.018)		(0.009)	(0.005)	(0.000)	(0.010)	(0.010)
β_3	0.406	0.400	0.399	0.400	0.400	0.400	0.358	0.355	0.356	0.356	0.355	0.355
		(0.013)	(0.016)	(0.014)	(0.014)	(0.016)		(0.007)	(0.005)	(0.005)	(0.002)	(0.003)
β_4	0.180	0.180	0.179	0.180	0.180	0.180	0.549	0.545	0.546	0.547	0.547	0.547
		(0.002)	(0.001)	(0.000)	(0.000)	(0.000)		(0.003)	(0.002)	(0.001)	(0.005)	(0.003)
β_5	0.005	0.011	0.009	0.007	0.012	0.012	0.005	0.011	0.009	0.007	0.009	0.008
		(1.189)	(0.102)	(0.008)	(0.106)	(0.014)		(0.205)	(0.091)	(0.022)	(0.020)	(0.021)
β_6	0.047	0.045	0.040	0.042	0.045	0.044	0.047	0.042	0.044	0.045	0.044	0.043
		(0.058)	(0.060)	(0.030)	(0.059)	(0.068)		(0.121)	(0.096)	(0.003)	(0.002)	(0.008)

[Correction added on 26 December 2022, after first online publication: format of table 7 has been corrected to improve clarity.]

Table 8. *Regression model on linked file LF (Model_LF0) and on file B only (Model_orig)* $y_{2i} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$ *and regression model on linked file LF (Model_LF1)* $y_{2is} = \beta_1 x_{0is} + \beta_2 x_{1is} + \beta_3 x_{2is} + \beta_4 x_{3is} + \beta_5 y_{1is} + \beta_6 (x_{0is} \times x_{1is}) + e_i$ *coefficient estimates averaged across the samples with absolute relative bias (in parentheses): 20% perturbation.*

	Small R^2						*Medium R^2*					
		Strategy						*Strategy*				
	True value	*A*	*B*	*B1*	*C*	*C1*	*True value*	*A*	*B*	*B1*	*C*	*C1*
						Model_LF0						
β_1	−0.012	0.006 (1.500)	0.001 (1.136)	−0.001 (0.800)	−0.004 (0.702)	−0.006 (0.600)	−0.903	−0.868 (0.038)	−0.879 (0.036)	−0.881 (0.034)	−0.900 (0.009)	−0.900 (0.007)
β_2	0.404	0.400 (0.009)	0.398 (0.014)	0.398 (0.013)	0.403 (0.001)	0.404 (0.000)	0.353	0.362 (0.024)	0.362 (0.023)	0.362 (0.024)	0.355 (0.009)	0.357 (0.007)
β_3	0.179	0.174 (0.028)	0.176 (0.015)	0.176 (0.013)	0.178 (0.013)	0.177 (0.013)	0.548	0.534 (0.027)	0.535 (0.024)	0.535 (0.023)	0.543 (0.004)	0.543 (0.003)
						Model_orig						
β_1	−0.012	0.050 (5.224)					−0.903	−0.639 (0.291)				
β_2	0.404	0.367 (0.093)					0.353	0.300 (0.039)				
β_3	0.179	0.166 (0.073)					0.548	0.401 (0.218)				
						Model_LF1 (with interaction)						
β_1	−4.593	−2.300 (0.599)	−3.211 (0.430)	−3.994 (0.399	−4.000 (0.100)	−4.011 (0.087)	−4.456	−2.989 (0.601)	−3.800 (0.398)	−4.001 (0.060)	−4.300 (0.009)	−4.129 (0.010)
β_2	−0.055	−0.012 (0.770)	−0.015 (0.724)	−0.043 (0.753)	−0.050 (0.465)	−0.051 (0.470)	−0.946	−0.894 (0.050)	−0.899 (0.041)	−0.900 (0.030)	−0.942 (0.014)	−0.943 (0.012)
β_3	0.406	0.394 (0.039)	0.399 (0.030)	0.399 (0.029)	0.400 (0.020)	0.402 (0.016)	0.358	0.299 (0.049)	0.300 (0.041)	0.305 (0.031)	0.340 (0.023)	0.349 (0.019)
β_4	0.180	0.176 (0.024)	0.177 (0.022)	0.179 (0.019)	0.180 (0.015)	0.180 (0.012)	0.549	0.459 (0.100)	0.499 (0.071)	0.500 (0.061)	0.540 (0.008)	0.543 (0.005)
β_5	0.005	0.011 (0.998)	0.009 (0.400)	0.007 (0.298)	0.006 (0.007)	0.005 (0.004)	0.005	0.001 (0.079)	0.002 (0.008)	0.002 (0.008)	0.004 (0.003)	0.004 (0.002)
β_6	0.047	0.002 (0.963)	0.025 (0.906)	0.036 (0.200)	0.040 (0.100)	0.041 (0.006)	0.047	0.038 (0.059)	0.039 (0.041)	0.039 (0.037)	0.042 (0.007)	0.043 (0.006)

[Correction added on 26 December 2022, after first online publication: format of table 8 has been corrected to improve clarity.]

regression coefficients of Model_LF0 under both small and medium R^2. Note that the fact that strategy B1 (Model_B1) and strategy C1 (Model_C1), where the predictions for the record linkage included an interaction term, did not make much difference to the overall bias in the model of analysis Model_LF0, thus adding a variable in the prediction model with respect to the model of analysis, did not substantially increase bias. This result was found in additional work not presented here, which showed that adding a variable in the prediction model did not change the results of the bias for an analysis model that did not include the additional variable.

Regarding model LF with interaction and the Y_1 variable (Model_LF1), we can see that strategy A introduces a large bias in the model parameter estimates for both small and medium R^2 under both the 10% perturbation in Table 7 and 20% perturbation in Table 8. Under the 10% perturbation in Table 7, we see smaller biases for strategy B1 when R^2 is small and similar biases for strategies B1 and C1 when R^2 is medium. However, under the 20% perturbation in Table 8, it is clear that strategies C and C1 have overall the smallest biases although strategy B1 is outperforming strategies A and B. Thus, we see that a propensity score model with an interaction term, strategy B1, outperforms strategies A and B for both 10% and 20% perturbation levels and adding an interaction term in the propensity score model increases the discriminatory power of the propensity scores when used in the record linkage procedure. In addition, it is interesting to note that generating the prediction using Model_C0 without an interaction term where the analysis model Model_LF1 includes the interaction term caused more bias in strategies B and C compared with their counterpart strategies B1 and C1, thus showing that the analysis model is impacted if the interaction term is not included in the prediction model. This result was found in additional work not presented here, which showed that omitting a variable in the prediction model can change the results of the bias for an analysis model that includes that variable.

Considering now the performances of model $y_{2i} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$ (Model_orig) estimated on file B only, this clearly introduces large bias in the model parameter estimates due to the perturbed independent variables. Hence, there is a clear advantage to carrying out the record linkage and producing a linked file for further analyses. This is consistent across all scenarios investigated in Tables 7 and 8. The absolute relative bias, as expected, increases under the 20% perturbation case.

In summary, similar to the conclusions found on the recall and precision measures and correlations, as the perturbation (error) rates increase, strategies C and C1 with prediction stratification having a medium R^2 outperform the other linkage strategies on the correct model in the case of the 20% perturbation level, but there is clear evidence for using strategy B1 under the 10% perturbation.

5 Application Based on Real Data

In this section, we present an application based on real data from a 1991 Israel Income Survey in order to evaluate the strategies described in Sections 2 and 3 and studied via the model-based simulation study in Section 4.

5.1 The Data and the Generation of the Files

From the original file with $n = 3\,841$ individuals, we create file A and file B. We use as matching variables: age, education, name, month of birth and locality. Similar to the simulation study, we follow the same technique of perturbation for the variables in file B according to all possible agreement patterns described in Section 4 at both the 10% and 20% perturbation levels. Note that for 5 variables, there are 32 agreement patterns. File A also contains the variable of income and this is not perturbed. Although the data is a real dataset, we generate an additional variable *Y* in file B correlated to the natural logarithm of the income for the sole purpose of evaluating the record linkage strategies in the linked file (LF): $y_i = N(1\,100,\ 150)(1\ -\ 0.99)\log(income_i)$ for $i =,\ 1\ldots,\ n_B$. This is to replicate the situation of the simulation study where variable Y_2 was present in file B only and was attached to file A as a result of the record linkage process.

In addition, file A includes variables: region, sex and an interaction term between employed (binary) and first digit of the occupation. This first digit of the occupation variable in the interaction term is also separately perturbed in file B according to the 10% and 20% perturbation levels. These variables will be used to carry out a linear regression model to predict income in file B that will be used for the record linkage strategy C1 to match to the income variable in file A together with other matching variables and the interaction term. Region is used as a blocking variable for all three record linkage strategies and is not perturbed.

To evaluate the record linkage strategies on the linked file (LF), we carry out a regression model with the response variable *Y* coming from file B and covariates age, education, log(income), sex, household size, marital status and the interaction term between employed and first digit of occupation from file A. The estimates of the coefficients from this regression model are compared with the coefficients obtained on the same model in the original data, and we calculate the absolute relative bias in 9 as a measure of quality.

We show in Table 9 a summary of the variables that are used in the application with the type of variable and the number of categories.

5.2 Results

Strategies A, B1 and C1 (with the interaction term) are applied here as in the simulation study and details are discussed below. In strategy A, we use the matching variables: age, education, name, month of birth and locality.

Similar to the simulation study, we estimate the propensity scores for strategy B1 by stacking file A and file B and run a logistic regression model where the dependent variable is $I = 1$ for those records in file A and $I = 0$ for those records in file B. The independent variables used as covariates are age, locality, education, interaction between employed and first digit of the occupation and sex of the respondent. Then, the record linkage is performed with the same matching variables as in strategy A plus the estimated propensity scores categorised in 20 quantiles.

In strategy C1, we first run a linear regression model in file A, where the response variable is the natural logarithm of the income. We choose the same independent variables as in the

Table 9. *Variables in the application.*

Variable	*Type*	*Number of categories*
Age	Continuous	70
Education (years of)	Continuous	25
Employment status	Categorical	2
First digit of occupation	Ordinal	10
Household size	Continuous	10
log income	Continuous	Mean: 8.72
Locality	Categorical	385
Marital status	Categorical	4
Month of birth	Categorical	12
Name	Categorical	1 498
Region	Categorical	25
Sex	Categorical	2
Y	Continuous	Mean: 19.64

propensity score model: age, locality, education, interaction between employed and first digit of the occupation and sex of the respondent. Then, based on the estimated coefficients of the model, the predictions for income are obtained in file B. The income in file A and the predicted income in file B are categorised into 20 quantiles. Record linkage is then carried out with the matching variables in strategy A plus the income (predicted income) stratification. We note that in this linear regression model for predicting income, the R^2 coefficient is 0.32, which is similar to the one used for the simulation study for the case of small R^2. Similar to the simulation study, all linkage strategies were compared using the one-to-one forced matching (van der Laan 2018), normalised matching weights on a scale [0,1] and the same threshold.

Table 10 shows the decision matrix, recall and precision for the three record linkage strategies for the 10% perturbation, and similarly, Table 11 shows the decision matrix, recall and precision for the 20% perturbation.

As expected, when the degree of perturbation is small at 10%, the quality of record linkage improves compared with the perturbation at 20%. Under the 10% perturbation in Table 10, there are not much differences in the precision and recall measures between the three record linkage strategies because the type I and type II errors are generally small. Nevertheless, we see an improvement in the recall and precision measures in strategy C1 compared with the other strategies, which included the prediction stratification as a matching variable. When the degree of perturbation increases to 20%, Table 11 shows that strategy C1 provides better results, higher recall and precision, compared with the other strategies. The improvement is as expected given the small R^2 similar to the findings in the simulation study.

To evaluate the quality of the final linked file LF, Table 12 shows the beta coefficients obtained from the regression model where the response variable Y is from file B and covariates age, education, log(income), sex, household size, marital status and interaction between employed and first digit of occupation from file A. Table 12 shows the results for both 10% and 20% perturbation levels under the three linkage strategies. Note that the R^2 coefficient for the model of analysis is 0.17. Table 12 also includes the absolute relative bias in (9) in parentheses.

From Table 12 and in line with the simulation study results, it can be seen that when the perturbation level is equal to 10%, the results of the linkage strategies show that strategy C1 has

Table 10. *Decision matrix for the three record linkage strategies: 10% perturbation.*

	Strategy A		*Strategy B1*		*Strategy C1*	
Decision	*NM*	*M*	*NM*	*M*	*NM*	*M*
NL	2 615 325	19	2 615 235	17	2 615 338	9
L	19	3 822	12	3 824	9	3 832
Recall	0.995		0.996		0.998	
Precision	0.995		0.997		0.998	

Table 11. *Decision matrix for the three record linkage strategies: 20% perturbation.*

	Strategy A		*Strategy B1*		*Strategy C1*	
Decision	*NM*	*M*	*NM*	*M*	*NM*	*M*
NL	2 614 670	674	2 614 669	675	2 615 059	285
L	674	3 167	675	3 166	641	3 200
Recall	0.825		0.824		0.918	
Precision	0.825		0.824		0.833	

Table 12. *Beta coefficients of the regression model on the linked file (LF) with Y as the response variable with the absolute relative bias in parentheses: 10% and 20% perturbation.*

		Perturbation level					
		10%			*20%*		
Coefficients	*True*	*A*	*B1*	*C1*	*A*	*B1*	*C1*
Age	−0.001	−0.001 (0.000)	−0.001 (0.000)	−0.001 (0.000)	−0.002 (0.100)	−0.001 (0.000)	−0.001 (0.000)
Education	−0.011	−0.011 (0.000)	−0.001 (0.909)	−0.008 (0.273)	−0.007 (0.364)	−0.007 (0.364)	−0.008 (0.273)
Log income	1.078	1.036 (0.039)	1.047 (0.029)	1.055 (0.021)	0.800 (0.258)	0.913 (0.153)	0.937 (0.131)
Household size	0.010	0.008 (0.200)	0.010 (0.000)	0.008 (0.200)	0.077 (6.700)	0.053 (4.300)	0.053 (4.300)
Sex	−0.065	−0.082 (0.262)	−0.086 (0.323)	−0.080 (0.231)	−0.136 (1.092)	−0.079 (0.215)	−0.063 (0.031)
Employed	0.224	0.221 (0.013)	0.222 (0.009)	0.223 (0.004)	0.230 (0.027)	0.235 (0.049)	0.226 (0.009)
Occupation 1st digit	0.006	0.007 (0.167)	0.007 (0.167)	0.006 (0.000)	−0.003 (1.500)	0.002 (0.667)	0.007 (0.167)
Employed × occupation (1st digit)	−0.103	−0.092 (0.107)	−0.099 (0.039)	−0.100 (0.029)	−0.140 (0.359)	−0.138 (0.340)	−0.136 (0.320)
Marital status 1	10.362	10.642 (0.027)	10.633 (0.026)	10.587 (0.022)	12.702 (0.226)	11.398 (0.100)	11.003 (0.062)
Marital status 2	10.482	10.731 (0.024)	10.707 (0.021)	10.699 (0.021)	12.679 (0.210)	11.228 (0.071)	11.001 (0.050)
Marital status 3	10.385	10.593 (0.020)	10.575 (0.018)	10.511 (0.012)	12.615 (0.215)	11.274 (0.086)	11.012 (0.060)
Marital status 4	10.348	10.659 (0.030)	10.538 (0.018)	10.511 (0.016)	12.589 (0.217)	11.159 (0.078)	11.000 (0.063)

generally smaller biases. Exceptions are for the education coefficient in strategy A and the household size in strategy B1. When the perturbation is equal to 20%, strategy C1 clearly outperforms the other linkage strategies and provides a smaller absolute relative bias in the coefficients for all variables. In addition, at this level of perturbation, strategy B1 propensity score stratification had smaller absolute relative bias in all variables compared with strategy A.

This application confirms the overall conclusions seen in the simulation study that for high perturbation levels, including linear predictions in the matching variables to carry out probabilistic record linkage improves the performance of the record linkage and subsequent regression models on the linked file that involve these variables.

6 Conclusions

The objectives of this study were to investigate whether the traditional F&S probabilistic record linkage approach can be improved by including predictions estimated from statistical models in the matching variables. This is motivated by two assumptions: (1) that the power of the decision theory for classifying record pairs into matches and non-matches could be increased by including not only direct matching variables but also variables that account for second-order correlation structures in their data; (2) that subsequent statistical analyses on the linked file will have better performance if these correlation structures are included in the linkage process. Almost all applications of record linkage are aimed to enhance the data at hand to be able to carry out statistical modelling such as regression models or multivariate analysis to address specific research questions. In this framework, it is important that relationships between variables in the linked data, for example, correlations and associations, are preserved. Indeed, we showed in the simulation study, that the regression analysis on the linked file produces better estimates (in terms of absolute relative bias) than analysis performed on file B only with perturbed predictors.

We proposed two strategies to improve the traditional F&S probabilistic record linkage: adding an additional matching variable of propensity score stratification (strategy B) and the stratification of predictions estimated from a linear regression model (strategy C). In order to evaluate whether these strategies improve (and to what extent) the traditional record linkage where only direct matching variables are used (strategy A), we conducted a simulation study under two levels of perturbation (10% and 20%) to simulate typical errors in matching variables due to spelling mistakes, transpositions, missing data and so forth. We also considered how the coefficient of determination R^2 for the linear predictions in strategy C impacts on the performances of these enhanced strategies. We note here that we also calculated the pseudo-R^2 for the logistic regression models for calculating the propensity scores in strategy B. The pseudo-R^2 were all very small, showing that there was little difference in discriminating between file A and file B, and hence, we saw that there was hardly any improvement in strategy B of adding a propensity score stratification compared with strategy A of just using the original matching variables in the record linkage. We analysed our results in terms of the quality of the classification of pairs into matches/non-matches based on a common threshold and a measure of the relative absolute bias of correlation estimates and regression model parameter estimates computed on the final linked file (LF).

We found that when R^2 is medium for the regression model used to calculate predictions in strategy C and particularly under the 20% perturbation, including prediction stratification in the matching variables improves the quality of record linkage (number of true links that are matched correctly) and the quality of subsequent modelling and correlation structures in the linked file (LF). When the perturbation level is 10%, the traditional record linkage in strategy

A returns good results and the improvements obtained by the use of prediction stratification are more modest. Therefore, we recommend the use of model predictions in strategy C when the level of perturbation is larger and the R^2 increases. Under strategy B with propensity score stratification, we saw some modest improvements compared with strategy A using the original matching variables in the application and 10% perturbation, but in general, strategy A and strategy B performed similarly. Users can benefit from the use of strategy B in case of smaller levels of perturbation.

We also experimented with adding and omitting an additional predictor (an interaction variable) to the prediction models for both strategies B and C (labelled strategies B1 and C1, respectively) and examined the impact when the regression model of analysis on the linked file also included the interaction variable or not. We found that if the prediction model includes an interaction term, it made little difference to the bias of regression parameters on a smaller model of analysis in the linked file (Model_LF0). However, biases were larger when the prediction model did not include the interaction term (strategies B and C) but was included in the model of analysis on the linked file (Model_LF1). This case clearly showed that strategies B1 and C1 outperformed strategies B and C. Moreover, the use of an interaction term in the prediction models helps in providing better quality model estimates in the linked file (LF) as it is shown in the simulation. This is something record linkage users should consider, with the aim that the linked file may be used for further multivariate analysis. We also add that strategies B and B1 might perform better if there are dependencies between the matching variables as the propensity score can replace those variables as a combination and better approximate the F&S model.

Future work will take into account other types of models, such as flexible non-linear models to generate predictions. Our strategies can be extended to these cases too, and further work will aim to evaluate them.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 730998 (InGRID-2 Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy).

References

Abrahams, C. and Davy, K. (2002). Linking HES maternity records with ONS birth records. *Health Stat. Q.*, **13**. 22–30.

Armstrong, J. and Mayda, J. (1992). Estimation of record linkage models using dependent data. In *Proceedings of the Section on Survey Research Methodology*, 853–858. American Statistical Association.

Avoundjian, T., Dombrowski, J. C., Golden, M. R., Hughes, J. P., Guthrie, B. L., Baseman, J., and Sadinle, M. (2020). Comparing methods for record linkage for public health action: Matching algorithm validation study. *JMIR Public Health Surveill.*, **6**(2) e15917.

Binette, O. & Steorts, R.C. (2020). (Almost) All of entity resolution. arXiv preprint arXiv:2008.04443. https://arxiv.org/abs/2008.04443

Breiman, L., Freidman, J., Olshen, R., Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Caliendo, M. & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. *Discussion Paper Series IZA*. IZA DP No. 1588.

Chambers, R. (2009). Regression analysis of probability-linked data. Research Series, Official Statistics.

Chambers, R., and Kim, G. (2016). Secondary analysis of linked data. In Harron, K., Goldstein H., and Dibben, C. (**2016**). *Methodological Developments in Data Linkage*. Wiley.

Chen, B., Shrivastava, A., and Steorts, R. C. (2018). Unique entity estimation with application to the Syrian conflict. *Ann. Appl. Stat.*, **12**(2), 1039–1067.

Chipperfield, J., Hansen, N. & Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *Int. Stat. Rev.*, **86**(2), 219–236.

Chipperfield J.O., and Chambers R. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *J. Off. Stat.*, **31**(3), 397–414.

Christen, P. (2008a) Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: *ACM SIGKDD*, 151–159. Las Vegas.

Christen, P. (2008b). Automatic training example selection for scalable unsupervised record linkage. In: *PAKDD*, Springer LNAI, Vol. **5012**, 511–518. Osaka.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications: Springer-Verlag Berlin Heidelberg.

Clark, S. J., Halter, M., Porter, A., Smith, H. C., Brand, M., Fothergill, R., Lindridge, J., McTigue, M., and Snooks, H. (2019). Using deterministic record linkage to link ambulance and emergency department data: Is it possible without patient identifiers? A case study from the UK. *Int. J. Popul. Data Sci.*, **4**,1:20.

Cochinwala, M., Kurien, V., Lalk, G., Shasha, D. (2001). Efficient data reconciliation. *Inform. Sci.* **137**(1–4), 1–15.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 2003 International Conference on Information Integration on the Web*, 73–78. AAAI Press.

Daggy, J., Xu, H., Hui, S., and Grannis, S. (2014). Evaluating latent class models with conditional dependence in record linkage. *Stat. Med.*, **33(**24), 4250–4265.

Dehejia, R. H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Revaluating the evaluation of training programs. *J. Am. Stat. Assoc.*, **94**(448), 1053–1062.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.: Ser. B* , **39**(1), 1–22.

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.

Dunn, H.L. (1946). Record linkage. *Am. J. Public Health Nations Health*, **36**(12), 1412–1416.

Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., and Carpenter, W. R. (2014). *Linking Data for Health Services Research: A Framework and Instructional Guide*. Rockville, MD: Agency for Healthcare Research and Quality.

Fellegi I. P. and Sunter, A. B. (1969). A theory for record linkage. *J. Am. Stat. Assoc.*, **64**:1183–1210.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001). On Bayesian record linkage. *Res. Off. Stat.*, **4**(1):185–198.

Goldstein, H. & Harron, K. (2016). *Record linkage*. John Wiley & Sons.

Grannis, S. J., Overhage, J. M., Hui, S., and McDonald, C. J. (2003). Analysis of a probabilistic record linkage technique without human review. *AMIA Annual Symposium Proceedings*, 259–263**.**

Grannis, S.J., Overhage, J. M., and McDonald, C. J. (2002). Analysis of identifier performance using a deterministic linkage algorithm. *AMIA Annual Symposium Proceedings*, 305–309.

Harron, K., Goldstein H., and Dibben, C. (2016). *Methodological Developments in Data Linkage*. Wiley.

Herzog, T., Scheuren, F. and Winkler, W. (2007). *Data Quality and Record Linkage Techniques*. Springer Verlag.

Hitt, L., and Frei, F. (2002). Do better customers utilize electronic distribution channels? The case of PC banking *Manag. Sci.*, **48**, No. 6, 732–748.

Kim, G., and Chambers, R. (2012a). Regression analysis under incomplete linkage. *Comput. Stat. Data Anal.*, **56**, 2756–2770.

Kim, G., and Chambers, R. (2012b). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, **66** (1), 64–79.

Kum, H, and Masterson, T.N. (2010). Statistical matching using propensity scores: Theory and application to the analysis of the distribution of income and wealth. *J. Econ. Soc. Meas.*, **35** (3–4), 177–196.

Larsen, M. D. and Rubin, D. B. (2001). Iterative automated record linkage using mixture models, *J. Am. Stat. Assoc.,* **96**(453), 32–41.

Maso, L., C. Braga, and S. Franceschi. (2001). Methodology used for software for automated linkage in Italy (SALI). *J. Biomed. Inform.*, **34**(6), 387–395.

McVeigh, B.S., Spahn, B.T. & Murray, J.S. (2019). Scaling Bayesian probabilistic record linkage with post-hoc blocking: An application to the California Great Registers. arXiv:1905.05337.

Mears, G. D., Rosamond, W. D., Lohmeier, C., Murphy, C., O'Brien, E., Asimos, A. W., Brice J. H. (2010). A link to improve stroke patient care: A successful linkage between a statewide emergency medical services data system and a stroke registry. *Acad. Emerg. Med.*, **17**(12), 1398–1404.

Moriarity, C. & Scheuren, F. (2003). A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.*, **21**(1), 65–73.

Newcombe, H. B. (1965). The study of mutation and selection in human populations. *Eugen. Rev.*, **57**(3), 109–125.

Newcombe, H. B.; Kennedy, J.M., Axford, S.J. and James, A. P. (1959). Automatic linkage of vital records. *Science*, **130** (3381), 954–959.

Newcombe, H. B. and Rhynas, P. O. W. (1962). Child spacing following stillbirth and infant death. *Eugen. Q.*, **9**(1), 25–35.

Newcombe, H. B. and Tavendale, O. G. (1965). Effects of father's age on the risk of child handicap or death. *Obstet. Gynecol. Surv.*, **20**(4), 655–656.

Perkins, S. M., Tu, W. Underhill, M. G. Zhou, X. Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research, *Pharmacoepidemiol. Drug Saf.*, **9**, 93–101.

Prasad, K., Faruquie, T., Joshi, S., Chaturvedi, S., Subramaniam, L., Mohania, M. (2009). Data cleansing techniques for large enterprise datasets. In: *SRII Global Conference*, 135–144. San Jose, USA.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Roos, LL; Wajda A. (1991). Record linkage strategies. Part I: Estimating information and evaluating approaches. *Methods Inf. Med.* **30** (2), 117–123.

Rosenbaum P.R., and Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.* **4** (1), 87–94.

Sadinle, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.*, **8**(4), 2404–2434.

Sadosky, P., Shrivastava, A., Price, M. & Steorts, R.C. (2015). Blocking methods applied to casualty records from the Syrian conflict. arXiv e-prints, 1–25.

Sarawagi, S. (2008). Information extraction. *Found. Trends Databases* **1**(3), 261–377.

Scheuren, F. and Winkler, W. (1993). Regression analysis of computer files that are computer matched. *Surv. Methodol.*, **19**(1), 39–58.

Scheuren, F. and Winkler, W. (1997). Regression analysis of computer files that are computer matched—Part II. *Surv. Methodol.*, **23**(2), 157–165.

Scholtus, S., Shlomo, N. and De Waal T. (2022). Correcting for linkage errors in contingency tables—A cautionary tale. *J. Stat. Plan. Infer.*, **18,** 122–137.

Shlomo, N. (2019). Overview of data linkage methods for policy design and evaluation in N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*. Springer.

Smith, D. & Shlomo, N. (2014). Report for the data without boundaries project. Technical report, University of Manchester. Available at http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf

Steorts, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.*, **10(**4), 849–875.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2014). SMERED: A Bayesian approach to graphical record linkage and de-duplication. *J. Mach. Learn. Res.*, **33**, 922–930.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *J. Am. Stat. Assoc.*, **111**(516), 1660–1672.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Surv. Methodol.*, **19**(1), 31–38.

Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A., and Reitsma, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *J. Clin. Epidemiol.*, **64**(5), 565–572.

van der Laan, J. (2018). reclin: Record linkage toolkit. R package version 0.1.1. https://CRAN.R-project.org/package=reclin

Winkler W.E. (1988). Using the EM algorithm for weight computation in the Fellegi Sunter model of record linkage. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association, **671**.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 354–359.

Winkler, W. E. (1992). Comparative analysis of record linkage decision rules. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association*,* 829–834.

Winkler, W. E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association*,* 274–279.

Winkler, W.E. (2006). Overview of record linkage and current research directions. Research Report Series, RRS2006-02. Available at https://www.census.gov/library/working-papers/2006/adrm/rrs2006-02.html [accessed 17072022]

Xu, H., Li, X., and Grannis, S. (2021). A simple two-step procedure using the Fellegi–Sunter model for frequency-based record linkage. *J. Appl. Stat.*, **49**(11), 2789–2804.

Xu, H., Li, X., Shen, C., Hui, S. L., and Grannis, S. (2019). Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? *J. Appl. Stat.*, **13**(3), 1753–1790.

Zhu, V., Overhage, M. J. Egg, J. Downs, S. M. Grannis, S. J. (2009). An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. *J. Am. Med. Inform. Assoc.*, **16**(5), 738–745.

[*Received March 2021; accepted November 2022*]