A Probabilistic Record Linkage Model for Survival Data

Michel H. Hof, Anita C. Ravelli, and Aeilko H. Zwinderman

1. Introduction

Computerized record linkage is often used to aggregate information from multiple existing data files from different sources. In different data files, records that belong to the same individual have been registered. With a record linkage procedure, we determine whether a pair of records belong to the same individual (match) or the records belong to different individuals (nonmatch). Using this information, we can combine the files and use this data to answer new research questions without the need of gathering additional data (Gill et al. 1993; Newcombe 1988).

Ideally, each file contains the same unique identifier, which is guaranteed to be unique for each individual (e.g., patient number). Unfortunately, due to privacy regulations or practical reasons, such a unique identifier is often not available and matches and non-matches cannot be distinguished from each other. In many situations, however, partially identifying variables may have been registered in all files. Examples of partially identifying variables are surname, zip-code, initials, date of birth, and gender. Usually, partially identifying variables have a limited number of unique values and may be contaminated with errors (e.g., typing errors). This means that the combined strength of the partially identifying variables is often not high enough to perfectly distinguish matches from non-matches, causing imperfect linkage. Importantly, if imperfect linkage is not properly dealt with, further analyses of the data could be biased and inefficient (Hof and Zwinderman 2012; Scheuren and Winkler 1993; Baldi et al. 2010).

This article was motivated by data from the Perinatal Registry Netherlands (PRN), containing information about 96% of all pregnancies in the Netherlands (Meray et al. 2007). We were interested in the effects of certain characteristics of the first delivery on the time to a second delivery. A problem of the PRN registry is that all pregnancies had been registered individually and no unique identifier was available to determine which pregnancies belonged to the same woman. To deal with imperfect linkage in a time-to-event setting, where we have a file with baseline characteristics and a file with event times, we developed a joint model in which the record linkage procedure and the time-to-event analysis are performed simultaneously. R code and example data are available as online supplemental material.

CONTACT

Michael H. Hof michelhofl@gmail.com Department of Clinical Epidemiology and Biostatistics 1100 DE, AMC / Universiteit van Amsterdam, Amsterdam, The Netherlands.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2017 Michel H. Hof, Anita C. Ravelli, and Aeilko H. Zwinderman. Published with license by Taylor & Francis. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Goldstein, Harron, and Wade (2012) proposed an imputation method to deal with imperfect linkage in the regression analysis. Although their method can deal with time-to-event data, their method requires that the linkage is complete, that is, all individuals with baseline measurements in A experience the outcome of interest and have a true match in B. If this assumption does not hold, which is likely to happen in real data, the imputation method requires external information about the mortality rate among the individuals (Goldstein, Harron, and Wade (2012), page 3485).

Since external information on the quality of the linked data is often not available, it is necessary to obtain this information by taking, for instance, an audit sample and checking the true matching status (Chambers 2009). Note that this requires extra resources, which lowers the attractiveness of combining files to answer new research questions. As an alternative, we propose a joint model in which we simultaneously estimate the record linkage model and the time-to-event regression model. The main advantage of our joint model is that it does not require additional external information.

2. Problem Description

Consider that we have two files A and B containing respectively n and m records. Let \( A = \{a_i : i = 1, \ldots, n\} \) be the file containing baseline measurements from n individuals, where each individual \( i \) is described by the record \( a_i = (x_i, r_i^A, \epsilon_i^A) \) containing the vector \( x_i = (x_{i1}, \ldots, x_{ip}) \) of \( p \) baseline measurements obtained at time \( t_i^A \) and the values of \( q \) partially identifying variables \( r_i^A = (r_{i1}^A, \ldots, r_{iq}^A) \). A partially identifying variable contains some personal information which cannot be used to perfectly distinguish individuals. Examples of such variables are gender, date of birth, and initials.

File \( B = \{b_j : j = 1, \ldots, m\} \) contains \( m \) events, where each record \( b_j = (t_j^B, r_j^B) \) contains the event time \( t_j^B \) and the same \( q \) partially identifying variables as \( A \), denoted as \( r_j^B = (r_{j1}^B, \ldots, r_{jq}^B) \). Examples of such files are death registries, hospital information systems, or healthcare (insurance) databases.

File \( B \) is obtained at time \( t_i^B \). Importantly, B is obtained after all the baseline measurements are gathered, that is, \( t_i^B > \max\{r_{i1}^A, \ldots, r_{iq}^A\} \).

Our main interest is to investigate the associations between the baseline measurements and the time from \( t_i^B \) onwards to occurrence of a particular event (e.g., death). Let \( T_i \) be the time at which individual \( i \) with a record in A experiences the event of interest, where \( T_i > t_i^B \). We assume that, if individual \( i \) experienced the event of interest before \( t_i^B \) (i.e., \( T_i \leq t_i^B \)), the event of individual \( i \) has been registered in B. In addition, if \( T_i > t_i^B \), individual \( i \) has no registered event in B.

To reconstruct the complete data, it is necessary to determine which baseline measurements from A and events from B belong to the same individual. Let \( \{(a_i, b_j) : i = 1, \ldots, n, \text{ and } j = 1, \ldots, m\} \) be the Cartesian product of A and B consisting of all \( n \times m \) possible pairs of baseline measurements from A and events from B. In addition, let \( d = \{d_{ij} : i = 1, \ldots, n, \text{ and } j = 1, \ldots, m\} \) be the set of \( n \times m \) corresponding matching indicators, where \( d_{ij} \) is a binary indicator which describes whether baseline measurement \( a_i \) and event \( b_j \) belong together (i.e., a match and \( d_{ij} = 1 \)) or not (i.e., a nonmatch and \( d_{ij} = 0 \)).

If a unique identifier, such as a citizen service number, for each individual is present in both A and B, it is possible to determine \( d \) and to calculate the follow-up times of all individuals with baseline measurements in A. For instance, consider individual \( i \) who experienced the event at \( T_i = t_i^B \) meaning that baseline measurement \( a_i \) and event \( b_j \) are a match (\( d_{ij} = 1 \)). We treat \( t_i^B \) as the start of the follow-up period and the follow-up time of individual \( i \) is \( t_i^B - t_i^B \). On the contrary, if individual \( i \) did not experience an event before \( t_i^B \), this individual was followed until time \( t_i^B = t_i^B - t_i^B \) without experiencing the event and was assumed to be right-censored.

In many situations, a unique identifier is not available and we have to rely on the partially identifying variables to determine whether a baseline measurement \( a_i \) and event time \( b_j \) are a match or a nonmatch. For each pair of observations \((a_i, b_j)\), the values of the partially identifying variables are compared to each other. In this article, we compare the values of the \( q \) partially identifying variables with a binary comparison function, that is,

\[
g_{ijk} = \begin{cases} 1 & \text{if } r_{ik}^A = r_{jk}^B \\ 0 & \text{if } r_{ik}^A \neq r_{jk}^B \end{cases}
\]

leading to the comparison vector \( g_{ij} = (g_{ij1}, \ldots, g_{ijq}) \). Note that it is also possible to use other comparison functions based on, for instance, string comparator metrics (Winkler 1990; DuVall, Kerber, and Thomas 2010; Porter and Winkler 1997). To which extent the comparison vectors can distinguish matches and non-matches depends on the combined strength of the partially identifying variables, where the strength of each variable is determined by the amount of (typing) errors, the number of unique values, and the distribution over these values. The more unique values the partially identifying variable has, the stronger the discriminative strength of the variable. For instance, gender (2 unique values) is a weaker partially identifying variable than the first initial (26 unique values) when the values of both variables are uniformly distributed in both A and B.

To determine whether a record pair is a match or not, record linkage is necessary (Fellegi and Sunter 1969). Basically, there are two types of record linkage strategies (Tromp et al. 2011). In the deterministic approach, all partially identifying variables or a subset of partially identifying variables has to agree to consider a record pair as a match. In this article, we focus on the probabilistic approach, where for each record pair we calculate the probability of a match given the comparison vector.

3. Probabilistic Record Linkage

3.1. Framework for Record Linkage

Fellegi and Sunter (1969) developed a framework for record linkage in which they treat the set of all \( n \times m \) record pairs from A and B as the union of two disjoint sets of matches M and non-matches U, where

\[ M = \{(a_i, b_j) : d_{ij} = 1, i = 1, \ldots, n, \text{ and } j = 1, \ldots, m\}, \]

\[ U = \{(a_i, b_j) : d_{ij} = 0, i = 1, \ldots, n, \text{ and } j = 1, \ldots, m\}. \]
and

\[ U = \{(a_i, b_i) : d_{ij} = 0, i = 1, \ldots, n, \text{ and } j = 1, \ldots, m\}. \]

**Remark 3.1.** If all partially identifying variables are free of errors, only the comparison vector \((1, 1, \ldots, 1)\) is observed for all record pairs in \(M\). Because of errors in the partially identifying variables, often of typographical nature, we could observe other comparison vector configurations in \(M\) (e.g., \((0, 1, \ldots, 1)\) or \((1, 0, \ldots, 1)\)). Analogously, if all partially identifying variables are unique for each individual we only observe the comparison vector \((0, 0, \ldots, 0)\) for all record pairs in \(U\). However, because the partially identifying variables have a limited number of unique values, we also observe other comparison vectors (e.g., \((1, 0, \ldots, 0)\) or \((0, 1, \ldots, 0)\)) in \(U\).

In many situations, the partially identifying variables are assumed to disagree completely at random (DCAR) in matches \(M\) and agree completely at random (ACAR) in non-matches \(U\), which Fellegi and Sunter (1969) and Schürle (2005) refer to as conditional independence. Under DCAR and ACAR, we have

\[ p(g_{ij} | d_{ij}) = \begin{cases} 
\prod_{k=1}^{m} \mu_k^{d_{ij}} (1 - \mu_k)^{1 - d_{ij}} & \text{if } d_{ij} = 1, \\
\prod_{k=1}^{m} \nu_k^{d_{ij}} (1 - \nu_k)^{1 - d_{ij}} & \text{if } d_{ij} = 0,
\end{cases} \]

where \(\mu_k\) and \(\nu_k\) are the probabilities that the \(k\)th partially identifying variable agrees in, respectively, \(M\) and \(U\). The DCAR assumption implies that disagreements of the partially identifying variables in \(M\) are uncorrelated. The ACAR assumption implies that an accidental agreement of a partially identifying variable in \(U\) happens completely at random. For instance, an accidental agreement on gender is independent of an accidental agreement on surname.

Because \(d\) is not observed, the correlation between the comparison outcomes in \(M\) and \(U\) cannot be measured. Therefore, it is not possible to test the ACAR and DCAR assumptions. If we suspect that the ACAR and DCAR assumptions do not hold, we can extend (2) by adding correlations between the comparison variables in \(M\) and \(U\) (Larsen and Rubin 2001; Winkler 1993; Schürle 2005). Another possible extension of (2) is the introduction of extra classes (e.g., potential nonmatch and likely nonmatch classes) (Scheuren and Winkler 1997; Larsen and Rubin 2001).

### 3.2. Estimation of the Matching Indicators

Since the set of matching indicators \(d\) is not observed, we need to rely on the partially identifying variables to determine which record pairs are matches and nonmatches. Because each individual has one baseline measurement in \(A\) and can experience the outcome of interest only once, not all configurations of \(d\) are possible. Each baseline measurement from \(A\) can form a match with at most one record from \(B\) and vice versa, that is, \(\sum_{j=1}^{m} d_{ij} \leq 1\) and \(\sum_{i=1}^{n} d_{ij} \leq 1\), leading to a one-to-one matching structure.

Following graph theory, Sadinle (2016) showed that a one-to-one matching structure is equivalent to a bipartite matching. The files \(A\) and \(B\) can be considered as two disjoint sets of nodes. Because of the data structure, each node from \(A\) can maximally have one possible edge that connects it to a node in \(B\) and vice versa. Therefore, all edges are pairwise disjoint and the corresponding graph is a bipartite matching.

Let \(g = \{g_{ij} : i = 1, \ldots, n, \text{ and } j = 1, \ldots, m\}\) be the set of \(n \times m\) comparison vectors. To estimate \(d\), we treat the matching indicators as missing data by considering the marginal distribution of the comparison vectors over all configurations of \(d\). Under a one-to-one matching structure, the marginal distribution of the comparison vectors is

\[ p(g) = \sum_{d \in D} p(g, d), \]

where \(p(g, d)\) is the joint probability distribution of the comparison vectors \(G\) and the matching indicators \(D\). In addition, the set \(D\) contains all possible matching indicator configurations, that is,

\[ D = \{d : \sum_{i=1}^{n} d_{ij} \leq 1, \sum_{j=1}^{m} d_{ij} \leq 1\}. \]

Using Bayes’ theorem, the most likely matching indicator configuration can be determined by calculating the probability \(p(D = d|g)\) for each possible matching indicator configuration \(d \in D\). Sadinle (2016), Gutman, Afendulis, and Zaslavsky (2013), and Tancredi and Liseo (2011) proposed methods based on Bayesian inference to estimate \(d\) under the one-to-one matching structure. The problem of imposing the one-to-one matching structure is that the size of \(D\), given by

\[ 1 + \sum_{k=1}^{n} \binom{n}{k} \prod_{l=0}^{k-1} (m - l), \]

is huge (Hof and Zwinderman 2015). Even in record linkage problems with really small files, for instance, \(n = 5\) and \(m = 30\), \(D\) already contains about \(20.6 \times 10^6\) matching indicator configurations. Consequently, these methods quickly become very computationally expensive. As a solution, Jaro (1989), Larsen and Rubin (2001), Winkler (1990), among others, proposed to simplify the model by assuming independence between all \(n \times m\) record pairs, leading to the traditional mixture model for record linkage

\[ \prod_{i=1}^{n} \prod_{j=1}^{m} \left[p(g_{ij} | d_{ij} = 1) p(d_{ij} = 1) + p(g_{ij} | d_{ij} = 0) p(d_{ij} = 0)\right], \]

where \(p(d_{ij} = 1)\) is the prevalence of matches in all \(n \times m\) record pairs.

A consequence of assuming independence between all record pairs is that the decision to declare a record pair a match or a nonmatch only depends on the information retrieved from a record pair and not on the complete data (Tancredi and Liseo 2011). Therefore, if the partially identifying variables are not strong enough, it is possible to assign multiple matches to each record from \(A\) (i.e., \(\sum_{j=1}^{m} d_{ij} > 1\)) and to each record from \(B\) (i.e., \(\sum_{i=1}^{n} d_{ij} > 1\)). To solve the problem of multiple matches, Jaro (1989) proposed a post hoc procedure to introduce the one-to-one matching structure. The optimal \(d\) that satisfies the one-to-one matching structure can be found by solving the following linear assignment problem:

\[ \max_{d} \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} \log \left[ \frac{p(g_{ij} | d_{ij} = 1)}{p(g_{ij} | d_{ij} = 0)} \right], \]
An alternative method to impose the one-to-one matching structure after fitting (4) was proposed by Zardetto and Scannapieco (2013), who used a genetic algorithm to search for the best-fitting set of matching indicators satisfying the one-to-one matching structure.

### 4. Two-Stage Approach for Survival Analysis and Probabilistic Record Linkage

After the set of matching indicators has been estimated, it is possible to fit a time-to-event model to the linked data. Let $T_i = \tilde{T}_i - t_i^0$ be a nonnegative continuous random variable representing the true follow-up time of individual $i$, that is, time from baseline measurement $t_i^0$ until the event of interest $\tilde{T}_i$. We assume that $T_i$ can be described by a proportional hazards model. Conditionally on the baseline measurements $x_i$, the hazard for individual $i$ at time $t$ is

$$
h(t|x_i; \alpha, \beta) = \begin{cases} 
    h_0(t|\alpha) \exp(x_i'\beta) & \text{if } t \geq 0, \\
    0 & \text{otherwise,}
\end{cases}
$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters and $h_0(t|\alpha)$ is a baseline hazard function with parameter vector $\alpha$. Evaluation of the hazard function at $t < 0$ gives a hazard of zero since $p(T_i < 0|x_i) = 0$. Throughout this article, we assume a parametric baseline hazard function. The baseline hazard function could be parameterized by, for instance, a constant term and linear trend over time or more flexible parametric functions based on spline specifications.

If no match was found in $B$ for individual $i$, that is, $\sum_{j=1}^{n} d_{ij} = 0$, this individual is considered to be right-censored at $t_i^0$. Analogously, if a match is found with the $j$th event time from $B$, that is, $d_{ij} = 1$, individual $i$ experienced the event of interest at $t_i^j$. Using the estimated matching indicators, the time-to-event model is defined as

$$\mathcal{L}(\alpha, \beta) = \prod_{i=1}^{n} \left\{ S(t_i|x_i; \alpha, \beta)^{\sum_{j=1}^{m} d_{ij}=0} \right. $$

$$\times \left. \prod_{j=1}^{m} f(t_{ij}|x_i; \alpha, \beta)^{d_{ij}=1\sum_{j=1}^{m} d_{ij}=1} \right\},
$$

where $S(t|x_i; \alpha, \beta) = \exp(- \int_{0}^{t} h(u|x_i; \alpha, \beta)du)$ and $f(t|x_i; \alpha, \beta) = h(t|x_i; \alpha, \beta)S(t|x_i; \alpha, \beta)$ are respectively the survival function and density function of $T_i$ evaluated at time $t$.

With time-to-event data, it is likely that the probability of a match depends on the covariates (e.g., healthier individuals are less likely to be admitted to the hospital) and the follow-up time (e.g., the longer an individual is followed in time, the higher the probability of an event). If we treat the record linkage and the time-to-event analysis as separate tasks, it is necessary to allow the probability of a match in the record linkage model to depend on the time-to-event data and the covariates. However, it is not straightforward to add the time-to-event data and the covariates in the record linkage model. Importantly, as can be seen in the simulation in Section 6, substantial bias could be introduced in the time-to-event model if the probability of a match in the record linkage model is only allowed to depend on the comparison vectors.

### 5. Joint Model for Survival Analysis and Probabilistic Record Linkage

Unlike the two-stage model, where we have the risk of misspecifying the probability of a match in the record linkage model, we propose to write the probability distribution of the matching indicators in terms of time-to-event probabilities. Since we are primarily interested in the parameters from the time-to-event model, we consider the matching indicators to be nuisance variables and treat them as missing data. By marginalizing out the matching indicators from the joint likelihood, both the record linkage submodel and the time-to-event submodel are estimated simultaneously. The main advantage of this approach is, and in contrast to the two-stage approach, that the probability of a match depends on both the time-to-event data and the covariates.

Let $\{T_i^r, T_i, G_i, D_i, X_i\}$ be the random variables of individual $i = 1, \ldots, n$, where $T_i^r$ is the maximum follow-up time, $T_i = \{T_i^j; j = 1, \ldots, m\}$ is the set of possible follow-up times, $G_i = [G_{ij}; j = 1, \ldots, m]$ is the set of comparison vectors, and $D_i = [D_{ij}; j = 1, \ldots, m]$ is the set of matching indicators. Ideally, we would impose a one-to-one matching structure to the data. However, similarly to the marginal approach for record linkage from (3), this would require the evaluation of an extremely large number of potential matching indicator configurations.

To reduce the computational burden we propose to assume independence between the observations from different individuals, that is, $\{T_i^r, T_i, G_i, D_i, X_i\} \perp \{T_j^r, T_j, G_j, D_j, X_j\}$ for $i \neq j$. Each individual can experience the outcome of interest maximally once (i.e., $\sum_{j=1}^{m} d_{ij} \leq 1$). Therefore, the set of all possible matching indicator configurations for individual $i$ is $D_i = \{D_i^m, D_i^m\}$, where the subset $D_i^m$ contains all possible vectors $d_i$ if individual $i$ has one match in $B$ and the subset $D_i^m$ represents the vectors that are possible if individual $i$ has no match in $B$, that is,

$$D_i^m = \left\{ d_i : \sum_{j=1}^{m} d_{ij} = 1 \right\} \quad \text{and} \quad D_i^m = \left\{ d_i : \sum_{j=1}^{m} d_{ij} = 0 \right\}.$$

As a consequence of assuming independence between the observations from different individuals, the decision to declare a record pair a match or a nonmatch only depends on the data from a particular individual. Therefore, it is possible to assign multiple potential matches to each record from file $B$ leading to a one-to-many matching structure. If we would be specifically interested in determining the matching indicators, this could be a problem (Tancredi and Liseo 2011). However, our main interest is the estimation of parameters from the time-to-event submodel. We conjecture that, given that the partially identifying
variables are of decent quality (i.e., relatively low level of errors), the assumption of independence between the observations from different individuals introduces ignorable bias in the parameter estimates of the joint model and therefore does not hamper our analysis. Results from our simulation study in Section 6 confirm this.

5.1. Time-to-Event Submodel

To show that the probability distribution of the matching indicators can be written in terms of time-to-event probabilities, recall from Section 2 that the event time $T_i$ of individual $i$ is registered in $B$ if individual $i$ experienced the event of interest before $T^B_i$. In other words, record $a_i$ has a match in $B$ when $T_i \leq T^B_i$, or equivalently, when the follow-up time $T_i \leq t^*_i$. Therefore, the probability that individual $i$ has a match in $B$ can be written as

$$p \left( d_i \in D^n_i \right) = p \left( \sum_{j=1}^{m} d_{ij} = 1 \bigg| x_i \right) = p \left( T_i \leq t^*_i \bigg| x_i \right) = 1 - S \left( t^*_i \bigg| x_i \right).$$

(8)

Analogously,

$$p \left( d_i \in D^n_i \right) = p \left( \sum_{j=1}^{m} d_{ij} = 0 \bigg| x_i \right) = p \left( T_i > t^*_i \bigg| x_i \right) = S \left( t^*_i \bigg| x_i \right).$$

If record $a_i$ has a match in $B$, individual $i$ experienced the event of interest at one of the event times in the set $t_i$ (i.e., $T_i \leq t^i_i$). For instance, if record pair $(a_i, b_j)$ is a match ($d_{iz} = 1$), individual $i$ experienced the event of interest at the follow-up time $T_i = t_{iz}$. The probability that $(a_i, b_j)$ is a match, given $\sum_{j=1}^{m} d_{ij} = 1$, is equivalent to the probability that individual $i$ experienced the event at follow-up time $t_{iz}$ and not at the other possible follow-up times from $t_i$, that is,

$$p \left( d_{iz} = 1 \bigg| \sum_{j=1}^{m} d_{ij} = 1, x_i \right) = \frac{f \left( t_{iz} \bigg| x_i \right)}{\sum_{j=1}^{m} f \left( t_{ij} \bigg| x_i \right)}.$$  

(9)

5.2. Record Linkage Submodel

Similarly to the traditional mixture model for record linkage from (4), the comparison vectors are assumed to be independent from each other and all observed and unobserved variables given the matching indicators of individual $i$. Consider that record $i$ from A has no match in B. The only possible vector $d_i \in D^n_i$ is a vector of zeros. Therefore, conditionally on $d_i \in D^n_i$, the probability of observing $g_i$ is

$$p \left( g_i \bigg| \sum_{j=1}^{m} d_{ij} = 0 \right) = \prod_{j=1}^{m} p \left( g_{ij} | d_{ij} = 0 \right).$$

(10)

Given that record $i$ from A has a match in B, there are $m$ possible matching indicator configurations possible. Consider the vector $d_i \in D^m_i$ with $d_{iz} = 1$, that is, record $i$ from A is a match with record $z$ from B and not with the other records from B. Conditionally on $d_i$, we can write the probability of observing $g_i$ as

$$p \left( g_i \bigg| d_{iz} = 1, \sum_{j=1}^{m} d_{ij} = 1 \right) = p \left( g_{iz} | d_{iz} = 1 \right) \times \prod_{j \neq z}^{m} \frac{1 - S \left( t_{ij} \bigg| x_i \right)}{S \left( t_{ij} \bigg| x_i \right)}.$$  

(11)

where $I()$ is the indicator function. The assumption of conditional independence between the comparison vectors and other variables requires stable partially identifying variables that do not change over time. Initials, date of birth, or gender are examples of such variables. If we use more unstable partially identifying variables, such as place of residence or marital status, it might be possible that the conditional independence in matches does not hold. The probability that place of residence agrees in matches may decrease with a larger follow-up time because some persons may have moved to different addresses. In addition, the probability that marital status agrees in nonmatches could depend on the age of the individual since younger individuals are less likely to be married. If we suspect associations between the values of the comparison outcomes and other variables, we can add these associations to the probabilistic record linkage submodel by extending (2).

In some situations, a variable can be used as baseline measurement in the time-to-event submodel and as partially identifying variable in the probabilistic record linkage submodel (e.g., gender). Because we use the observed values from A as baseline measurements in the time-to-event submodel, the variable is assumed to be without (typing) errors in A. The partially identifying variable could still disagree in matches, but these disagreements are only caused by errors in B.

5.3. Joint Likelihood

For the joint likelihood specification, we assume that the comparison vectors are independent from each other and all observed and unobserved variables given the matching indicators. If we suspect that the comparison vectors are associated with some observed variables, the record linkage submodel can be extended by modifying (2). The density function of the variables $(T^i_i, T_j, G_i, D_i, X_i)$ from individual $i = 1, \ldots, m$ is formed by combining the time-to-event submodel and record linkage submodel, that is,

$$p \left( \left( t^i_i, t^j_j, g_i, d_i; x_i; \theta \right) \right)$$

$$= \left\{ \frac{S(t^i_i \bigg| x_i; \alpha, \beta) p \left( \sum_{j=1}^{m} d_{ij} = 0; \mu, v \right)}{\sum_{j=1}^{m} f \left( t^i_i \bigg| x_i; \alpha, \beta \right)} \right\}^{\sum_{j=1}^{m} d_{ij} = 1}$$

$$\times \left\{ 1 - S(t^i_i \bigg| x_i; \alpha, \beta) \prod_{z=1}^{m} p \left( g_{iz} = 1, \sum_{j=1}^{m} d_{ij} = 1; \mu, v \right) \right\}$$

$$\times \frac{f \left( t^i_i \bigg| x_i; \alpha, \beta \right)}{\sum_{j=1}^{m} f \left( t^i_i \bigg| x_i; \alpha, \beta \right)}.$$  

(12)
where \( \theta = \{ \mu, \nu, \beta, \alpha \} \) is the set of all model parameters. The vectors \( \mu \) and \( \nu \) contain the parameters from the probabilistic record linkage submodel. In addition, \( \beta \) is the vector of regression parameters and \( \alpha \) is the vector the parameters regarding the baseline hazard function from the time-to-event submodel. Since the matching indicators are considered to be nuisance variables, we marginalize them out from (12) leading to

\[
\sum_{d_i \in D_i} p(t_i', t_i, g_i, d_i | x_i; \theta) \\
= p \left( g_i | \sum_{j=1}^{m} d_{ij} = 0; \nu \right) S(t_i' | x_i; \alpha, \beta) \\
+ \left[ 1 - S(t_i' | x_i; \alpha, \beta) \right] \sum_{z=1}^{m} p \left( g_i | d_{iz} = 1, \sum_{j=1}^{m} d_{ij} = 1; \mu, \nu \right) \\
\times \left( \sum_{j=1}^{m} f(t_{ij} | x_i; \alpha, \beta) \right). 
\]

Parameter estimates \( \hat{\theta}_n \) are obtained by maximizing the likelihood, that is,

\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log \left( \sum_{d_i \in D_i} p(t_i', t_i, g_i, d_i | x_i; \theta) \right),
\]

where \( \Theta \) is the parameter space. Since the joint model is a mixture of \( m + 1 \) components of finite measure products, identifiability of our joint model follows from Sec. 5 in Allman, Matias, and Rhodes (2009). We propose to estimate the variance of \( \hat{\theta}_n \) with the inverse Hessian evaluated at \( \hat{\theta}_n \). This variance estimate gave us good results in terms of coverages of the 95% confidence intervals in our simulations (see Section 6). However, our variance estimate should only be considered as a good approximation. A rigorous proof of asymptotic normality remains challenging.

We implemented the joint model in the statistical program R (R Core Team 2014), where we used numerical approximation techniques to maximize the log-likelihood and obtain estimates of the Hessian. The code of the joint model and example data are available in the online supplementary materials.

5.5. Reducing the Computational Burden

If \( A \) and \( B \) contain a large number of records, evaluating the likelihood from (13) is computationally expensive. Because we treat the matching indicator vector \( d_i \) as unobserved, we have to evaluate all \( m + 1 \) possible matching indicator vectors \( d_i \in D_i \) for each record \( a_i \). The computational costs to evaluate (13) can be reduced substantially when one or more of the \( q \) partially identifying variables are registered without errors in \( B \) or \( B \) files. In the record linkage literature, these variables are often referred to as blocking variables (Lahiri and Larsen 2005; Winkel 1993).

Consider that the \( k \)th \((1 \leq k \leq q)\) partially identifying variable is registered without errors, that is, \( p(g_{ijk} = 1 | d_{ij} = 1) = 1 \), which implies that

\[
p \left( g_i | d_{iz} = 1, \sum_{j=1}^{m} d_{ij} = 1; \mu, \nu \right) = 0 \quad \text{if} \quad g_{izk} = 0.
\]

The sum in (13) over all possible matching indicator vectors \( D_1 = \{ D_1^m, D_1^p \} \) becomes a sum over the smaller set \( D_1 = \{ D_1^{m^p} \} \), where \( D_1^{m^p} = \{ d_i : d_{iz} = 1, \sum_{j=1}^{m^p} d_{ij} = 1, g_{izk} = 1 \} \). The size of \( D_1^{m^p} \) depends on the number of error-free partially identifying variables and their combined strength to discriminate matches from nonmatches.

6. Simulation Study

6.1. Setup

In the simulations, file \( A \) contained baseline measurements of \( n = 750 \) individuals and file \( B \) contained \( m = 1000 \) event times. For each individual \( i \), the baseline measurements in \( A \) were obtained at time \( t_i^A = 0 \). All individuals were at risk of experiencing the event until \( t_i^B = 3 \). The baseline measurement \( a_i = (x_i, r_i) \) from individual \( i \) consisted of three baseline characteristics \( x_i = \{ x_{ip} : p = 1, \ldots, 3 \} \) and five partially identifying variables \( r_i = \{ r_{ik}^1 : k = 1, \ldots, 5 \} \). The baseline characteristics were sampled from a multivariate normal distribution with means zero, variances one, and covariances zero.

File \( B \) was generated as follows. First, file \( B \) was generated without matches by creating a file with uncorrelated records. For each record \( b_i = (t_i^A, r_i^B) \), the event time \( t_i^A \) was sampled from a continuous uniform distribution over the interval \([0, t_i^B]\) and the partially identifying variables \( r_i^B = \{ r_{ik}^1 : k = 1, \ldots, 5 \} \) were sampled from similar distributions as the partially identifying variables in \( A \). Second, we added the matches to file \( B \). For each individual \( i \) with a baseline measurement in \( A \), a follow-up time \( T_i \) was sampled from a Poisson process with intensity \( h_i(t) = h_{0i}(t) \exp(x_i^\beta) \), where \( h_{0i}(t) = 0.1 + 0.04t - 0.02t^2 \) and \( \beta = (0.25, -0.16, 0.41) \). Individual \( i \) had a record (i.e., a match) in \( B \) if \( T_i \leq t_i^B \). This record was created by changing the values of record \( b_i \) to represent the characteristics of individual \( i \), that is, \( r_i^B = r_i^B \) and \( t_i^A = T_i \). Note that when \( T_i > t_i^B \), individual \( i \) had no record in \( B \).

With respect to the distribution of the partially identifying variables located in both \( A \) and \( B \), we considered the following two scenarios. In the first scenario, we generated partially identifying variables that satisfied the ACAR and DCAR assumptions. In this scenario, we compared the joint model with the two-stage method described in more detail in Section 4. In the second scenario, we generated partially identifying variables that violated the ACAR or DCAR assumptions. In this scenario, we were particularly interested in the performance of the joint model under mis specification of the probabilistic record linkage submodel. The scenarios are described in more detail in Section 6.2.

For each model in both scenarios, the bias, mean squared error, and the coverages of the 95% confidence interval of the regression parameter estimates \( \hat{\beta} \) were reported. In addition, the estimated baseline hazards were reported. All results were obtained by repeating the simulation 500 times. A computer with 8 Intel Xeon (2.40 GHz) processors was used to fit a joint
model to the data. We reported the computer run-time for the joint models in both scenarios.

6.2. Scenarios

6.2.1. Scenario 1: Uncorrelated Partially Identifying Variables with Uncorrelated Errors (ACAR and DCAR)

In this scenario, the five partially identifying variables were sampled from independent discrete uniform distributions with 6, 7, 4, 8, and 2 unique values, meaning that they agree completely at random in nonmatches. In addition, we introduced respectively 10%, 5%, 2%, 0%, and 5% uncorrelated errors in the partially identifying variables. Errors were introduced by changing the value of the partially identifying variable to one of the other possible values. Given our parameter choices, we expected that \( n(1 - \exp(-\int_0^t h_0(u) du)) \approx 265 \) individuals experienced the event before \( t \) (i.e., 265 true matches). However, because the partially identifying variables contained errors, approximately 54 (20.4%) of these true matches had a comparison vector in which one or more partially identifying variables disagreed. In addition, because the partially identifying variables had a limited number of unique values, approximately 279 nonmatches were expected to have a comparison vector in which all the partially identifying variables agreed. Note that this was a substantial proportion of the record pairs in which all the partially identifying variables agreed; from the approximately 490 record pairs in which all the partially identifying variables agreed, about 279 (56.9%) were nonmatches.

The performance of the joint model was compared to a two-stage procedure to estimate the parameters from the time-to-event submodel. In the first stage, the matching indicators were estimated using the record linkage model mixture model from (4). We used the post hoc procedure from (5) to guarantee one-to-one matching. In the second stage, the estimated matching indicators were used in the time-to-event model from (7). Similarly to the joint model, a parametric proportional hazards model was fitted to the data. In both models, we correctly assumed ACAR and DCAR for the record linkage (sub)model and we used a natural cubic spline to estimate the log baseline hazard.

6.2.2. Scenario 2a: Correlated Partially Identifying Variables with Uncorrelated Errors (Not ACAR and DCAR)

In this scenario, the values of the first and second partially identifying variable were (positively) correlated by sampling them from a bivariate normal distribution with means zero, variances one, and covariance 0.3. Both partially identifying variables were converted into discrete variables with 10 equally likely values by using the quantile function of the standard normal distribution. Because the values of these partially identifying variables were correlated, the ACAR assumption did not hold. The remaining three partially identifying variables were sampled from independent discrete uniform distributions with 4, 8, and 2 unique values. In addition, we introduced respectively 10%, 5%, 2%, 0%, and 5% uncorrelated errors in the partially identifying variables implying that the partially identifying variables disagreed completely at random in matches.

Two joint models were fitted to the simulated data. In the first joint model, we captured the correlation between the comparison outcomes in nonmatches of the first two partially identifying variables. Instead of (2b), we used

\[
p(\mathbf{g}|d_{ij} = 0) = \prod_{k=1}^{5} \left( 1 - v_k \right)^{\mathbf{g}_{ijk}} \left( 1 - v_k \right)^{(1 - \mathbf{g}_{ijk})},
\]

where \( v_0 = 0.3, v_1 = -1.25 \) to simulate the data. The other four partially identifying variables contained 5%, 2%, 0%, and 5% uncorrelated errors.

Two joint models were fitted to the simulated data, where in the first model we replaced \( \mu_1 \) from (2a) with \( \mu_1(t_{ij}) \) from (14) to capture the relation between time and the probability that the first partially identifying variable agreed in matches. In the second joint model, we wrongly assumed DCAR and used (2a) in the record linkage parts of the joint likelihood regarding the matches. In both joint models, a natural cubic spline was used to estimate the log baseline hazard.

6.2.3. Scenario 2b: Uncorrelated Partially Identifying Variables with Errors that Depend on Time (ACAR and not DCAR)

The five partially identifying variables were sampled from independent discrete uniform distributions with 30, 7, 4, 8, and 2 unique values, meaning that all partially identifying variables agreed completely at random in nonmatches. Although the first partially identifying variable had a large number of unique values and appeared to be a relatively strong identifier, it contained a large amount of error. In addition, this amount increased with time meaning that the DCAR assumption did not hold. The probability that the first partially identifying variable agreed was

\[
\mu_1(t_{ij}) = p(g_{ij1} = 1|d_{ij} = 1, t_{ij}) = \frac{\exp(\phi_0 - \phi_1 t_{ij})}{1 + \exp(\phi_0 - \phi_1 t_{ij})},
\]

where we used \( \phi_0 = 3, \phi_1 = -1.25 \) to simulate the data. The other four partially identifying variables contained 5%, 2%, 0%, and 5% uncorrelated errors.

Two joint models were fitted to the simulated data, where in the first model we replaced \( \mu_1 \) from (2a) with \( \mu_1(t_{ij}) \) from (14) to capture the relation between time and the probability that the first partially identifying variable agreed in matches. In the second joint model, we wrongly assumed DCAR and used (2a) in the record linkage parts of the joint likelihood regarding the matches. In both joint models, a natural cubic spline was used to estimate the log baseline hazard.

6.3. Results

The simulation results regarding the regression parameters are summarized in Table 1. In addition, the estimated baseline hazard curves are given in Figure 1. The computational run-times of the joint models were comparable in both scenarios. On average, the regression parameters were estimated in about 18 minutes. In addition, estimating the Hessian matrix took about 9 minutes.

Scenario 1. The joint model gave unbiased regression parameters estimates with good coverages of the 95% confidence interval. With respect to the baseline hazard, the joint model gave slightly overestimated estimates. However, the coverage
of the 95% confidence interval was relatively good; on average 0.92.

With the two-stage method, the regression parameters estimates were all biased toward zero. The baseline hazard was accurately estimated and had a lower mean squared error than the joint model. The variance estimate of the baseline hazard, however, was too low which led to a poor coverage of the 95% confidence interval; on average 0.82.

**Scenario 2.** In Scenario 2, the joint model gave unbiased regression parameters estimates with good coverages of the 95% confidence interval, even if the probabilistic record linkage submodel was misspecified. Similarly to Scenario 1, the baseline hazard was slightly overestimated with a correctly specified probabilistic record linkage submodel. However, the coverage of the 95% confidence interval was relatively good; on average 0.92 in Scenario 2a and 0.93 in Scenario 2b.

Misspecifying the probabilistic record linkage submodel led to biased baseline hazard estimates. In Scenario 2a, wrongly assuming DCAR resulted in an overestimation of the baseline hazard. This can be explained by the fact that the identifying strength of the first two partially identifying variables was overestimated, leading to an overestimation of the number of matches (or events).

Wrongly assuming ACAR in Scenario 2b also led to biased baseline hazard estimates. In Scenario 2b, the amount of error in the first variable increased over time. Assuming ACAR, this trend was ignored which resulted in an overestimation of the baseline hazard in the beginning of the follow-up period (i.e., too many matches with short follow-up) and an underestimation at the end of the follow-up period (i.e., too few matches with long follow-up).

**7. Application: Analysis of the Time to Second Delivery Using the Perinatal Registry Netherlands**

**7.1. Data Description**

The proposed joint model was fitted to data from the Perinatal Registry Netherlands (PRN), which is a large nationwide registry containing information from pregnancies with a duration of at least 22 weeks (Méray et al. 2007). We were interested in the probability of a second delivery given the characteristics of the mother (age and social economic status) at the first delivery and characteristics of the first delivery (gender of the child, pregnancy induced with assisted reproductive technology, and pre-term birth defined as gestational age < 37 weeks). In this analysis, we focused on mothers between 20 and 40 years of age. We assumed a proportional hazards model for the time to second delivery.

File A contained the baseline measurements from \( n = 79830 \) mothers with a first delivery in 1999. The baseline characteristics are given in Table 2. File B contained the data from \( m = 696023 \) second deliveries from 1999 until 2010. We assumed that, when a mother with a delivery in 1999 had a second delivery between 1999 and 2010, the second delivery had been registered in B.

To distinguish matches from nonmatches, we relied on the partially identifying variables “birth date of the mother” and current “place of residence” described by a four digit zip-code, which had been registered for all first and second deliveries without missing values. In addition, because the date of the previous delivery was registered for 361871 (52.0 %) second deliveries in B, a third partially identifying variable “date of delivery” was created by comparing the date of the previous delivery from the second deliveries in B to the date of delivery from the first deliveries in A.
Figure 1. Simulation results regarding the estimated baseline hazard function from the time-to-event (sub)model. In the figures regarding the estimated baseline hazards, the continuous line is the median estimated baseline hazard, the light gray area represents 95% of the estimated baseline hazards, and the dashed line is the true baseline hazard.

With respect to the probabilistic record linkage submodel, we assumed that all partially identifying variables agreed completely at random in non-matches. In addition, we assumed that the birth date of the mother and the date of delivery disagreed completely at random in matches. The probability that place of residence disagreed might increase with the time between the first and second delivery. Moreover, mothers with less economic stability (i.e., low social economic status) could move more often. Therefore, the probability of a disagreement of place of residence in matches depended on maternal characteristics and the time between deliveries.

Let birth date of the mother, place of residence, and date of delivery be respectively the first, second and third partially identified variables. In the probabilistic record linkage submodel, we assumed that all partially identifying variables agreed completely at random in non-matches. In addition, we assumed that the birth date of the mother and the date of delivery disagreed completely at random in matches. The probability that place of residence disagreed might increase with the time between the first and second delivery. Moreover, mothers with less economic stability (i.e., low social economic status) could move more often. Therefore, the probability of a disagreement of place of residence in matches depended on maternal characteristics and the time between deliveries.

Let birth date of the mother, place of residence, and date of delivery be respectively the first, second and third partially identified variables.

### Table 2. Baseline characteristics of the 79830 first deliveries from file A.

| Age of the mother (years) | Gender (girl) |
|---------------------------|---------------|
| <25 years                 | Yes 38600 (48.4%) |
| 25–30 years               | Yes 3381 (4.2%) |
| 30–35 years               | Yes 6611 (8.3%) |
| ≥35 years                 | Yes 19143 (24.0%) |
| Low social economic status| Yes 27548 (34.5%) |
identifying variable. We used a binary comparison function to compare the values of the partially identifying variables (see (1)) and used the following probabilistic record linkage submodel

\[ p(g_{ij}|d_{ij} = 1, z_{ij}; \mu, \phi) = \mu_{ij}^{\phi_{ij}} (1 - \mu_{ij})^{1-\phi_{ij}} \times \frac{\exp[z_{ij}^T \Phi_{ij}]}{1 + \exp[z_{ij}^T \Phi_{ij}]} \times [\mu_{ij}^{\phi_{ij}} (1 - \mu_{ij})^{1-\phi_{ij}}]^{w_{ij}}, \]

\[ p(g_{ij}|d_{ij} = 0; \psi) = \prod_{k=1}^{2} \psi_{ij}^{g_{ijk} (1 - \psi_{ij})^{1-g_{ijk}}}, \]

where the vector \( z_{ij} \) contained the time between deliveries, the age of the mother at the first delivery, and the social economic status of the mother at the first delivery. The indicator \( w_{ij3} \) described for record pair \( ij \) whether the comparison value of date of delivery was available (\( w_{ij3} = 1 \)) or missing (\( w_{ij3} = 0 \)). We assumed that the date of the previous delivery registered in \( B \) was missing completely at random.

Let \( t \) be the time since the first delivery in weeks. Because a delivery only had been registered in the PRN registry after a pregnancy of at least 22 weeks, the risk of experiencing a second delivery is zero if \( t \leq 22 \). For our model, this meant that the baseline hazard \( h_{0i}(t|\alpha) = 0 \) if \( t \leq 22 \). The log baseline hazard was parametrized with a natural cubic spline. Since we suspected a highly nonlinear association between the age of the mother and the probability of a second delivery, a cubic B-spline was used to model the effect of the age of the mother. The effect of age was fixed to be zero at the age of 30 years. A computer cluster with 40 Intel Xeon (2.40 GHz) processors was used to fit the joint model to the data.

### 7.2. Results

Estimating the regression parameters of the joint model took about 17.8 hours. The estimation of the Hessian matrix took approximately 10.5 hours. The estimated parameters of the probabilistic record linkage submodel are given in Table 3. As expected, the probability that place of residence agreed in matches decreased with more time between first and second delivery. The probability that place of residence agreed in a match consisting of a mother between 25 and 30 years of age at her first delivery with a normal social economic status and a second delivery which occurred 0–2 years later was about 0.78. This probability decreased to 0.65, 0.67, 0.55, and 0.50 when her second delivery happened respectively 2–4, 4–6, 6–8, or 6–10 years later.

The probability of an agreement was slightly higher for younger mothers (20–25 years of age) and substantially higher for older mothers (35–40 years of age). In addition, the probability of an agreement of place of residence in matches was lower for mothers with a low social economic status.

The effects of the characteristics of the first delivery on the probability of a second delivery are summarized in Table 3 and Figure 2. Mothers with a low social economic status at the first delivery had a lower probability of a second delivery. In addition, the use of assisted reproductive technology to achieve the pregnancy that led to the first delivery decreased the probability of

### Table 3. Estimated joint model parameters. The standard errors were estimated by taking the square roots of the diagonal elements of the Hessian. In addition, the \( p \)-values for the log hazard ratio \( \hat{\beta}_g \) were computed assuming that \( Z = \frac{\hat{\beta}_g}{\sqrt{\text{var}(\hat{\beta}_g)}} \) followed asymptotically a standard normal distribution.

| Parameter | Estimate | Standard error | Z-value | p-value |
|-----------|----------|----------------|---------|---------|
| \( \phi_{\text{Intercept}} \) | 1.245 | 0.034 | | |
| \( \phi_{\text{Age mother 20-25 years}} \) | 0.412 | 0.036 | | |
| \( \phi_{\text{Age mother 30-35 years}} \) | 0.064 | 0.026 | | |
| \( \phi_{\text{Age mother 35-40 years}} \) | 1.240 | 0.072 | | |
| \( \phi_{\text{Low social economic status}} \) | -0.531 | 0.027 | | |
| \( \phi_{\text{Birth of the mother}} \) | 0.953 | | | |
| \( \phi_{\text{Date of first delivery}} \) | 0.880 | | | |
| \( \phi_{\text{Place of residence}} \) | 0.00054 | | | |
| \( \phi_{\text{Birth of the mother}} \) | 0.00016 | | | |
| \( \phi_{\text{Date of first delivery}} \) | 0.00025 | | | |

### Figure 2. Estimated effect of the age of the mother at the first delivery on the probability of a second delivery. The estimated effect, given by the solid line, was fixed at zero at 30 years of age. The dotted lines, representing the 95% confidence interval, were obtained as follows. Consider that the effect of the age of the mother is described by \( g(t; \xi) = w(t) \xi \), where \( w(t) \) is the basis vector of the spline function describing the nonlinear association, \( t \) is the age of the mother, and \( \xi \) is the vector of spline function parameters. Assuming that \( \xi \) follows asymptotically a normal distribution, the standard error of the baseline hazard is \( \text{se}(g(t; \xi)) = \sqrt{\text{diag}(w(t) \Sigma w(t)^T)} \), where \( \Sigma = \text{var}(\xi) \) is the variance-covariance matrix of \( \xi \). The 95% confidence interval was obtained by \( g(t; \xi \pm 1.96 \times \text{se}(g(t; \xi))) \).
a second delivery. No associations were found between the gender (i.e., a girl) and pre-term birth of the first child and having a second delivery. The relation between the age of the mother at the first delivery had a parabolic shape with a maximum around 30 years of age.

The estimated baseline hazard is given in Figure 3. The baseline hazard was fixed to zero until 22 weeks after the first delivery. After this time, the baseline hazard quickly increased and peaked at approximately 3.6 years after the first delivery. After this peak, the baseline hazard decreased quickly and at approximately 6 years after the first delivery, the hazard was almost zero.

8. Discussion

We proposed a joint model for the analysis of time-to-event or survival data obtained from multiple files. More specifically, we considered the situation in which baseline measurements for a set of individuals have been registered in one file and event data in the other file. In the absence of a unique identifier, we used the information from partially identifying variables to determine which records from both files belonged to the same individual. With simulations, we illustrated the performance of the joint model under misspecification of the probabilistic record linkage submodel. The results showed that the regression parameter estimates from the time-to-event submodel were robust to misspecification of the probabilistic record linkage submodel. The baseline hazard estimate, however, was affected by misspecification of the probabilistic record linkage submodel.

Further research should be focused on extending the joint model to more complex types of time-to-event data (e.g., competing risks data and left-truncated data) and to record linkage problems with more than two files. Dealing with more than two files leads to a huge increase in the number of possible matching indicator vectors. Starting points to reduce the computational burden can be found in the Bayesian record linkage literature, in which a number of methods have been proposed to deal with more than two files (Sadinle 2014; Steorts, Hall, and Fienberg 2016; Sadinle and Fienberg 2013).

Supplementary Materials

R code and example data: Implementation of the joint model in R (assuming DCAR/ACAR). In addition, two example files are provided. (zipped file)

Funding

The authors gratefully acknowledge the Perinatal Registry Netherlands for the pregnancy data, which is part of research proposal PRN 11.44.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009), “Identifiability of Parameters in Latent Structure Models With Many Observed Variables,” The Annals of Statistics, 37, 3099–3132. [1509]
Baldi, I., Ponti, A., Zanetti, R., Ciccone, G., Merletti, F., and Gregori, D. (2010), “The Impact of Record-Linkage Bias in the Cox Model,” Journal of Evaluation in Clinical Practice, 16, 92–96. [1504]
Chambers, R. (2009), Regression Analysis of Probability-Linked Data, Official Statistics Research Series, Vol 4. Official Statistics Research Series, Vol 4. [1504,1505]
DuVall, S. L., Kerber, R. A., and Thomas, A. (2010), “Extending the Fellegi-Sunter Probabilistic Record Linkage Method for Approximate Field Comparators,” Journal of Biomedical Informatics, 43, 24–30. [1505]
Fellegi, I. P., and Sunter, A. B. (1969), “A Theory for Record Linkage,” Journal of the American Statistical Association, 64, 1183–1210. [1505,1506]
Gill, L., Goldacre, M., Simmons, H., Betley, G., and Griffith, M. (1993), “Computerised Linking of Medical Records: Methodological Guidelines,” Journal of Epidemiology and Community Health, 316–319. [1504]
Goldstein, H., Harron, K., and Wade, A. (2012), “The Analysis of Record-Linked Data Using Multiple Imputation With Data Value Prior,” Statistics in Medicine, 31, 3481–93. [1505]
Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013), “A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs,” Journal of the American Statistical Association, 108, 34–47. [1506]
Hof, M. H. P., and Zwinderman, A. H. (2012), “Methods for Analyzing Data From Probabilistic Linkage Strategies Based on Partially Identifying Variables,” Statistics in Medicine, 31, 4231–42. [1504]
Hof, M. H. P., and Zwinderman, A. H. (2015), “A Mixture Model for the Analysis of Data Derived From Record Linkage,” Statistics in Medicine, 34, 74–92. [1506]
Jaro, M. A. (1989), “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” Journal of the American Statistical Association, 84, 414–420. [1506]
Kim, G., and Chambers, R. (2009), “Regression Analysis Under Incomplete Linkage,” Working Paper no. 17-09. [1504]
Kim, G., and Chambers, R. (2012), “Regression Analysis Under Incomplete Linkage,” Computational Statistics & Data Analysis, 56, 2756–2770. [1504]
Lahiri, P., and Larsen, M. D. (2005), “Regression Analysis With Linked Data,” Journal of the American Statistical Association, 100, 222–230. [1504,1509]
Larsen, M. D., and Rubin, D. B. (2001), “Iterative Automated Record Linkage Using Mixture Models,” Journal of the American Statistical Association, 96, 32–41. [1506]
Méray, N., Reitsma, J. B., Ravelli, A. C. J., and Bonsel, G. J. (2007), “Probabilistic Record Linkage is a Valid and Transparent Tool to Combine Databases Without a Patient Identification Number,” Journal of Clinical Epidemiology, 60, 883.e1–883.e11. [1504,1511]
Newcombe, H. B. (1988), Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business, Oxford, UK: Oxford Medical Publications, Oxford University Press. [1504]
Porter, E. H., and Winkler, W. E. (1997), “Approximate String Comparison and Its Effect on an Advanced Record Linkage System,” in Advanced

Figure 3. Estimated baseline hazard function. The baseline hazard is fixed to zero until the dotted line at 22 weeks after the first delivery.
R Core Team (2014), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [1509]

Sadinle, M. (2014), "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach," *Annals of Applied Statistics*, 8, 2404–2434. [1514]

Sadinle, M., and Fienberg, S. E. (2013), "A Generalized Fellegi-sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems," *Journal of the American Statistical Association*, 108, 385–397. [1504]

Scheuren, F., and Winkler, W. E. (1993), "Regression Analysis of Data Files That are Computer Matched – Part I," *Survey Methodology*, 19, 39–58. [1504]

Scheuren, F., and Winkler, W. E. (1997), "Regression Analysis of Data Files That are Computer Matched. – Part II," *Survey Methodology*, 23, 126–138. [1504,1506]

Schürle, J. (2005), "A Method for Consideration of Conditional Dependencies in the Fellegi and Sunter Model of Record Linkage," *Statistical Papers*, 46, 433–449. [1506]

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016), "A Bayesian Approach to Graphical Record Linkage and Deduplication," *Journal of the American Statistical Association*, 111, 1660–1672. [1514]

Tancredi, A., and Liseo, B. (2011), "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems," *The Annals of Applied Statistics*, 5, 1553–1585. [1506,1507]

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," *Journal of Clinical Epidemiology*, 64, 565–72. [1505]

Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research*, pp. 354–359. [1505,1506]

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1, 274–279. [1506,1509]

Zardetto, D., and Scannapieco, M. (2013), "Optimal Decision Rules for Constrained Record Linkage: An Evolutionary Approach," in *Statistical Models for Data Analysis*, eds. P. Giudici, S. Ingrassia, and M. Vichi, Studies in Classification, Data Analysis, and Knowledge Organization, New York: Springer, pp. 389–397. [1507]