Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system

Thanh Huan Vo, Guillaume Chauvet, André Happe, Emmanuel Oger, Stephane Paquelet, Valérie Garès

Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system

Thanh Huan Vo[a,c,*], Guillaume Chauvet[b], André Happe[d], Emmanuel Oger[d], Stéphane Paquelet[c], Valérie Garès[a]

[a] *Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*
[b] *Univ Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*
[c] *IRT b<>com, Rennes, France*
[d] *EA 7449 REPERES, France*

Abstract

Probabilistic record linkage is a process of combining data from different sources, when such data refer to common entities and identifying information is not available. A probabilistic record linkage framework that takes into account multiple non-identifying information that this is limited to simple binary comparison between matching variables has been previously proposed. An extension of this method is proposed for mixed-type comparison vectors. A mixture model for handling comparison values of low prevalence categorical matching variables, and a mixture of hurdle gamma distribution for handling comparison values of continuous matching variables have been developed. The parameters are estimated by means of the Expectation Conditional Maximization (ECM) algorithm. Through a Monte Carlo simulation study, both the posterior probability estimation for a record pair to be a match and the prediction of matched record pairs are evaluated. The simulation results indicate that the proposed methods outperform existing ones in most considered cases. The proposed methods are applied on a real dataset, to perform linkage between a registry of patients suffering from venous thromboembolism in the Brest district area (GETBO) and

*Corresponding author
Email address: vthuan.math@gmail.com (Thanh Huan Vo)

[1] A supplementary material containing detailed computations and additional simulations is available online with the article.

the French national health information system (SNDS).

Keywords: Expectation Conditional Maximization (ECM) algorithm, hurdle gamma distribution, low prevalence variables, mixture model, probabilistic record linkage.

1. Introduction

Electronic health records have become more and more prevalent in medical fields, and the ability to exchange this information can help in providing better care for patients as well as richer sources for researchers. Record linkage is a process of combining data from different sources that refer to the same entity. The process is straightforward if each record contains a unique identifier such as Social Security Number (Zhu et al., 2015). However, some large health databases may not contain such identifying information. In other cases, this information is available but may contain errors, or may not be used for record linkage due to ethical reasons. Fellegi & Sunter (1969) proposed a probabilistic framework that takes into account multiple quasi-identifiers such as name, address and postal code. It has become widely used in applications when unique identifiers are unavailable or when data contain errors (e.g. Grannis et al., 2003; Sayers et al., 2015).

The French SNDS (Système National des Données de Santé) is the national health data system including the national health insurance information (SNIIRAM: Système National d'Information InterRégimes de l'Assurance Maladie) of around 99% of the French population (Bezin et al., 2017). This data system also includes information on all health care expenses, as well as private and public hospital data collected in the medical information system (see Tuppin et al., 2017b). There is therefore an increasing demand of getting this information from SNDS, to enrich research datasets in epidemiology or public health. However, due to ethical reasons, the SNDS database is anonymous. This means that personal identifying information such as Social Security Number, Name or Address is not available. We are therefore interested in proposing a probabilis-

tic record linkage model using other variables in common represented by the so-called matching variables. They can be of various types (categorical, binary, continuous) depending on the research study. For example, the matching variables may include postal code (categorical), date of treatment (continuous) and medical diagnosis (binary).

The Fellegi-Sunter probabilistic record linkage model laid the foundation for most record linkage models until now (Christen & Winkler, 2017). Although this model is useful for many applications in sample surveys and epidemiology, it has a limitation when some matching variables are binary and with a low prevalence (e.g., medical diagnoses). In that case, the simple binary comparison method proposed by Fellegi & Sunter (1969) can not distinguish the agreement of low prevalence values, which is much more informative than the agreement of high prevalence values. Such cases are considered in Hejblum et al. (2019), who propose a Bayesian linkage framework outperforming the Fellegi-Sunter model. However, their model is restricted to binary matching variables only.

Another limitation is that most probabilistic record linkage models only make use of simple binary or categorical comparison values (see Christen, 2012) even if the matching variables are continuous. Some authors introduced continuous similarity measures for comparing string data, but then comparison values are transferred to categorical values representing different levels of agreement (e.g., Herzog et al., 2007; Sadinle, 2017; Enamorado et al., 2019), which may result in a loss of information.

In this article, we propose a new linkage model adapted from the framework of Fellegi and Sunter, which handles such situations. We aim at better taking into account the nature of matching variables (e.g., low-prevalence binary, or continuous), so as to improve the performances of record linkage. The article is organized as follows. In Section 2, we review the Fellegi-Sunter probabilistic record linkage model and some relevant problems. We then propose two comparison strategies for low prevalence binary or continuous matching variables in Section 3. An extended mixture model taking into account both categorical and continuous comparison values is also introduced in Section 3. In Section 4, we

evaluate the proposed methods through simulation studies. In Section 5, a real data application is proposed, where we perform record linkage between SNDS and the GETBO (Groupe d'Etude de la Thrombose de Bretagne Occidentale) registry. Finally, possible further research is discussed in Section 6.

2. Probabilistic record linkage

Consider two databases A and B containing n_A and n_B records respectively, and with elements in common. Following the terminology in Fellegi & Sunter (1969), each possible pair of individuals (a_i, b_j) with $a_i \in A, i = 1, \ldots, n_A$ and $b_j \in B, j = 1, \ldots, n_B$ either belongs to the set of true matched pairs

$$M = \{(a, b); a = b, a \in A, b \in B\},$$

or to the set of true unmatched pairs

$$U = \{(a, b); a \neq b, a \in A, b \in B\}.$$

Because an identifying variable is not available, other less discriminant data are used in the probabilistic record linkage procedure, such as the name, date of birth, postal code, or some diagnosis codes. This information needs to be registered in both data sets and is referred to as matching variables. The matching variables in two databases are required to have the same format (Christen, 2012).

It is supposed that there is no prior knowledge on how likely the matches are, which is often the case in practice. The strategy therefore begins by comparing K matching variables for all records $X_{A,i} = (X^1_{A,i}, \ldots, X^K_{A,i}), i = 1, \ldots, n_A$ of n_A individuals in A, with all records $X_{B,j} = (X^1_{B,j}, \ldots, X^K_{B,j}), j = 1, \ldots, n_B$ of n_B individuals in B. This leads to $n_A \times n_B$ comparison vectors γ_{ij} such that

$$\gamma_{ij} = \left\{\gamma^1_{ij}, \ldots, \gamma^k_{ij}, \ldots, \gamma^K_{ij}\right\}, \tag{1}$$

where $\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j})$ and h^k is a comparison function for the $k-$th matching variable.

Because the number of all record pairs is quadratic in the number of individuals in each database, making the comparison for all possible record pairs

is often impracticable in applications. One of the most popular methods to reduce the number of record pairs that need to be compared is blocking, in which only records from the two databases that are in a same block (i.e., sharing the same values for the blocking variables) are compared with each other. Record pairs disagreeing on the blocking variable are automatically classified as non-matches. Therefore, blocking is a trade-off between computational cost and the proportion of missed matches (matched pairs are missed because of errors in the blocking variable), see Herzog et al. (2007).

The set of all possible realizations of γ is called the comparison space and denoted by Γ. The comparison function γ^k for the k-th matching variable can be defined in different ways depending on the type of matching variable (Christen, 2012). The most common way consists in a binary comparison, i.e.

$$\gamma_{ij}^k = h^k(X_{A,i}^k, X_{B,j}^k) = \begin{cases} 1 & \text{if} \quad X_{A,i}^k = X_{B,j}^k, \\ 0 & \text{if} \quad X_{A,i}^k \neq X_{B,j}^k. \end{cases} \tag{2}$$

If there is no error in the matching data, all components of a comparison vector of a matched pair are equal to 1. However, application data usually contain errors (e.g., typographical), and some similarity measures that can take them into account have been developed in the literature for string variables (Herzog et al., 2007).

Once all candidate pairs are compared, various approaches are possible to classify the set of comparison vectors into matches and non-matches (Christen, 2012). If training data where we observe the true matched status of record pairs is available, supervised classification methods (Christen, 2008) can be used to find a classification rule. If there is no training data but some clerical review is possible, some semi-supervised approaches (e.g. Enamorado, 2018) may be applied. However, the exact knowledge of matches is rarely possible in real world situations, and the clerical review is costly. Unsupervised methods (e.g. Winkler, 1988; Mamun et al., 2016) are therefore the more common approaches. From a Bayesian perspective, Tancredi & Liseo (2011) introduced a paradigm for probabilistic record linkage, and Steorts et al. (2016) proposed a Bayesian

approach to graphical record linkage.

In the frequentist view, Fellegi & Sunter (1969) assumed that each record pair belongs to one of the two latent classes. The distribution of comparison vector γ for each pair is assumed to follow a mixture model

$$\mathbb{P}(\gamma) = \mathbb{P}(\gamma|M)\mathbb{P}(\gamma \in M) + \mathbb{P}(\gamma|U)\left[1 - \mathbb{P}(\gamma \in M)\right]. \tag{3}$$

If we do not make additional assumptions on the joint agreement pattern, the comparison vector γ may take 2^K different values, each of which corresponds to a parameter that we need to estimate. To reduce this number, some authors (Fellegi & Sunter, 1969; Winkler, 1988), have proposed to make the so-called conditional independence assumption between fields of the comparison vector. Under this assumption, we obtain:

$$\mathbb{P}\left[\gamma = (\gamma^1, \ldots, \gamma^K)|M\right] = \prod_{k=1}^{K} \mathbb{P}(\gamma^k|M), \tag{4}$$

$$\mathbb{P}\left[\gamma = (\gamma^1, \ldots, \gamma^K)|U\right] = \prod_{k=1}^{K} \mathbb{P}(\gamma^k|U). \tag{5}$$

The conditional independence assumption is common in most probabilistic record linkage models (Winkler, 1988), although it may not hold in some practical cases. For example, if some records agree on a chronic disease, they are more likely to agree on the drug used. Although the assumption is invalid in some cases, the linkage result is still quite robust, in the sense that we may have a good linkage performance even if the conditional independence assumption does not hold (Winkler, 1988; Grannis et al., 2003; Sayers et al., 2015). Some authors (e.g. Xu et al., 2019) relaxed this assumption and showed better record linkage results in some specific scenarios.

Under the conditional independence assumption, we only need to estimate $2K+1$ parameters which are the marginal probabilities of agreement for matched and unmatched pairs $m^k \equiv \mathbb{P}(\gamma^k = 1|M)$ and $u^k \equiv \mathbb{P}(\gamma^k = 1|U)$, and the overall matching probability $p_M \equiv \mathbb{P}(\gamma \in M)$. Winkler (1988) proposed to apply the expectation maximization (EM) algorithm (Dempster et al., 1977; Wu, 1983), to find the maximum likelihood estimates for the vector of parameters

$\theta \equiv \{p, m^k, u^k, k = 1, \ldots, K\}$. It has become widely used in probabilistic record linkage (Grannis et al., 2003; Christen, 2012). Once all the parameters are estimated, the record pairs may be ordered by either matching weights

$$\hat{w}_{ij} = \frac{\mathbb{P}(\gamma_{ij}|M, \hat{\theta})}{\mathbb{P}(\gamma_{ij}|U, \hat{\theta})},$$

see Fellegi & Sunter (1969); Belin & Rubin (1995), or by posterior probabilities of matching $\hat{q}_{ij} \equiv \mathbb{P}(M|\gamma_{ij}, \hat{\theta})$ (Larsen & Rubin, 2001). Then, the pairs are classified into matches, non-matches or possible matches based on two defined thresholds (Fellegi & Sunter, 1969). Because the possible matches require manual review which is sometimes not available, Grannis et al. (2003) propose to establish only a single threshold to avoid human review. Although the matching scores and the posterior probabilities produce the same ordering for record pairs (Larsen & Rubin, 2001), the posterior probabilities are preferable in our case because they may be useful for further analyses (Lahiri & Larsen, 2005; Kim & Chambers, 2012; Hof & Zwinderman, 2012; Zhang & Tuoto, 2020).

In some applications, a one-to-one matching restriction may be needed; namely, that each record in B can be matched to one and only one record in A, and conversely. One possible approach to respect a one-to-one matching is to solve a linear sum assignment problem proposed by Jaro (1989). If the optimal score is not demanded, a simple approach is to sort all candidate pairs according to their estimated posterior probabilities of matching, and to select matched pairs in a greedy approach (Christen, 2012).

3. An extension of the Fellegi-Sunter model

In this section, we extend the Fellegi-Sunter model by making better use of low prevalence categorical matching variables and of continuous variables. Two new comparison approaches and a mixture model for mixed type of comparison values are introduced.

3.1. Comparison approaches

For a categorical matching variable, it is likely that the proportions for each category are different, and accounting for these differences in a record linkage model may help to improve the linkage results. This idea was proposed by Fellegi & Sunter (1969); Winkler (1989), and is applied on a real clinical data in Zhu et al. (2009). These authors use the same model for simple agreement/disagreement comparison, but the matching weights are rescaled a posteriori, using a frequency-based correction. We introduce a new comparison approach for categorical matching variables, which differs from simple binary comparison and may naturally handle different proportions for categories.

Let X^k be a categorical matching variable taking L different values, which means that the comparison function for this variable may take up to L^2 values. For example, the comparison for a binary matching variable may lead to four possible realizations $\{(0,0),(0,1),(1,0),(1,1)\}$ and a comparison function can be defined as follows

$$h^k(0,0) = c_1, \quad h^k(0,1) = c_2, \quad h^k(1,0) = c_3 \quad \text{and} \quad h^k(1,1) = c_4, \qquad (6)$$

where c_1, c_2, c_3 and c_4 stand for four different categories. It should be noted that the values taken by the comparison function have no ordinal meaning. If this is a low prevalence binary matching variable (e.g. a rare disease) such that only 5% (say) of the values in the dataset are equal to 1, the agreement on the value "1" is much more informative than the agreement on the value "0". Our comparison approach aims at using this information while the simple agreement comparison method does not, leading to poor performance. Hejblum et al. (2019) propose a Bayesian record linkage framework making use of a similar idea, and which is efficient in case of a large number of low-prevalence binary matching variables. However, their model is designed for binary variables only.

If the number of matching variables and/or the number of categories is large, the number of parameters to be estimated is $L^2 - 1$, which may be too large in practice. This number may be reduced by assigning the same comparison value for the agreement/disagreement of categories which have a close meaning. For

instance, we may reduce the comparison values given in (6) as

$$h^k(0,0) = c_1, \quad h^k(0,1) = h^k(1,0) = c_2 \quad \text{and} \quad h^k(1,1) = c_3. \tag{7}$$

In general, the number of comparison values depends on which realizations we would like to distinguish. Suppose that we are interested in a categorical matching variable X^k with categories $1, 2, \ldots, L$. If the first category seems particularly meaningful, we may distinguish whether we have an agreement on the first category, an agreement on another category, or a disagreement. In such case, the comparison function would be defined as

$$h^k(i,j) \quad = \quad \begin{cases} c_1 & \text{if } i = j = 1, \\ c_2 & \text{if } i = j \neq 1, \\ c_3 & \text{if } i \neq j = 1, \ldots, L. \end{cases}$$

The objective of this comparison approach is to distinguish the agreement of low prevalence values from other agreements, which differs from multiple levels of agreement introduced in (Sadinle, 2017) and (Enamorado et al., 2019).

Now, let us consider the case of a continuous variable X^k. For example, date variables (e.g., admission to the hospital, or medical act) are common in medical datasets. By converting each date into a duration from a specified origin, they may be treated as continuous counting variables. Even if an individual is present in both datasets, a lag between dates is likely to appear. The simple binary comparison is therefore not appropriate. In this article, if the k^{th} matching variable is continuous, we propose to consider

$$\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j}) = d(X^k_{A,i}, X^k_{B,j}), \tag{8}$$

where d is a distance which can be used to measure the difference between two dates of events, in which case it can be interpreted as a time lag. By using the distance, the continuous comparison values γ^k of matching pairs $(X^k_{A,i}, X^k_{B,j})$ can be described as

$$\gamma^k_{ij} | (X_{A,i}, X_{B,j}) \in M = \begin{cases} 0 & \text{with probability} \quad 1 - e^k, \\ \epsilon^k_{ij} > 0 & \text{with probability} \quad e^k, \end{cases}$$

where e^k is the proportion of error, and ϵ_{ij}^k is the error term of the k^{th} matching variable among matched pairs. For example, two patients who refer to the same individual should have the same day for a medical act, up to some errors in the registration process, and the distance should therefore be equal to 0 or to a small error term ϵ_{ij}^k. Therefore, $\gamma_{ij}^k|M$ follows a hurdle distribution in which the positive part depends only on the distribution of errors. On the other hand, the distribution of $\gamma_{ij}^k|U$ depends mostly on the distribution of the k^{th} matching variable, since ϵ_{ij}^k is often small compared to the distance between records for two unmatched units.

3.2. Estimation of parameters

Let

$$\gamma_{ij} \quad = \quad \left(\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}, \gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}\right) \tag{9}$$

be a mixed type comparison vector which includes K_1 categorical comparison values $\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}$ and K_2 continuous distances $\gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}$. Following the Fellegi-Sunter framework, these comparison vectors are assumed to follow the mixture model (3).

Under the conditional independence assumption between the different fields in the comparison vector for both the matched and the unmatched sets, we have

$$\mathbb{P}(\gamma_{ij}|M) = \underbrace{\prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|M)}_{P_{ij}^{1M}} \underbrace{\prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|M)}_{P_{ij}^{2M}}, \tag{10}$$

$$\mathbb{P}(\gamma_{ij}|U) = \underbrace{\prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|U)}_{P_{ij}^{1U}} \underbrace{\prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|U)}_{P_{ij}^{2U}}, \tag{11}$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. For both equations (10) and (11), the first term in the right hand side involves K_1 categorical comparison values of the comparison vector γ_{ij}. We define

$$m_s^k = \mathbb{P}(\gamma_{ij}^k = s|M) \text{ and } u_s^k = \mathbb{P}(\gamma_{ij}^k = s|U) \text{ for } s \in S^k, \tag{12}$$

with S^k the set of all possible categorical comparison values for the k^{th} variable. Then

$$P_{ij}^{1M} = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|M) = \prod_{k=1}^{K_1} \prod_{s\in S^k} (m_s^k)^{\mathbb{1}_{\gamma_{ij}^k=s}},$$
$$P_{ij}^{1U} = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|U) = \prod_{k=1}^{K_1} \prod_{s\in S^k} (u_s^k)^{\mathbb{1}_{\gamma_{ij}^k=s}},$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$, and with $\sum_{s\in S^k} m_s^k = \sum_{s\in S^k} u_s^k = 1$.

The second part in the right hand side of equations (10) and (11) involves K_2 continuous values of the comparison vector γ. We define

$$\begin{aligned} P_{ij}^{2M} &= \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|M) \quad \text{with} \quad \mathbb{P}(\gamma_{ij}^k|M) \sim f_M^k(\phi_M^k), \\ P_{ij}^{2U} &= \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|U) \quad \text{with} \quad \mathbb{P}(\gamma_{ij}^k|U) \sim f_U^k(\phi_U^k), \end{aligned} \tag{13}$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. The distributions f_M^k and f_U^k need to be postulated, depending on the characteristics of the matching variables and on the chosen distance.

To find the maximum likelihood estimates for parameters, we apply the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) or the Expectation Conditional Maximization (ECM) algorithm (Meng & Rubin, 1993), depending on the distribution f^k. In Section 1 of the supplementary material, we present the details of the ECM algorithm, when both f_M^k and f_U^k correspond to a hurdle gamma distribution, which is used in the next part of this article.

Once all parameters are estimated by means of the EM/ECM algorithm, the posterior probabilities $q_{ij} = \mathbb{P}(M|\gamma_{ij})$ are estimated for all record pairs by the Bayes formula

$$\widehat{q}_{ij} = \frac{\widehat{p}_M \widehat{P}_{ij}^{1M} \widehat{P}_{ij}^{2M}}{\widehat{p}_M \widehat{P}_{ij}^{1M} \widehat{P}_{ij}^{2M} + (1-\widehat{p}_M)\widehat{P}_{ij}^{1U} \widehat{P}_{ij}^{2U}}. \tag{14}$$

These estimated posterior probabilities are then used to find proper matched pairs.

4. Simulation studies

In this section, our proposed approaches are evaluated and compared to other existing approaches. To facilitate interpretation, two simulation studies are performed to evaluate the properties of the proposed methods for binary and continuous variables separately. A simulation study for a combination of both categorical and continuous matching variables is presented in Section 4 of the supplementary material. All the simulations are implemented in a R program, which is available on Github repository: https://github.com/thanhhuanVO/Extending-FellegiSunter-Record-linkage.git.

4.1. Simulation designs

In the following simulations, we consider two databases A and B containing $n_A = 500$ and $n_B = 200$ individuals and K matching variables. We assume that there is no duplicate in both databases, and that all individuals in B have corresponding individuals in A. The number of individuals in both databases remains fixed in our simulations. However, different sizes are considered in additional simulations available as a supplement.

We first generate the observations in A, and a random subset of n_B units is used to obtain the database B. For $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$ let us denote by

$$X_{A,i} = \left(X^1_{A,i}, \ldots, X^K_{A,i}\right) \quad \text{and} \quad X_{B,j} = \left(X^1_{B,j}, \ldots, X^K_{B,j}\right) \tag{15}$$

the i^{th} and j^{th} individual in A and B, respectively. Without loss of generality, we assume that the first unit in B is the first unit in A, ..., the n_B^{th} unit in B is the n_B^{th} unit in A. The full comparison matrix $\gamma = \left\{\gamma^k_{ij}\right\}$ contains $n^A \times n^B = 100\,000$ lines and K columns.

Once the posterior matching probabilities are estimated for all possible record pairs, a pair is classified as a match if $\hat{q}_{ij}$ (see equation 14) is larger than a predefined threshold τ, and is classified as a non-match otherwise. The choice of the threshold depends on the objectives of the study, a higher threshold leading to a lower number of false matches.

Scenario 1: binary matching variables

Data generating process. In this scenario, each variable $X_{A,i}^k$ is first generated according to a Bernoulli distribution with parameter p^k, for $k = 1, \ldots, K$. To account for possible errors in the matching variables, the variables $X_{B,j}^k$ in database B are then obtained as

$$X_{B,j}^k = \begin{cases} X_{A,j}^k & \text{with probability} \quad 1 - e^k, \\ 1 - X_{A,j}^k & \text{with probability} \quad e^k. \end{cases} \tag{16}$$

Simulation parameters. Since the binary matching variables are less discriminant, all the methods tested require a large number K of matching variables, in order to have sufficient information for achieving acceptable linkage results. We therefore used $K \in \{30, 40, 50\}$. The probability of error is chosen as $e^k \in \{0.02, 0.04, 0.06\}$. For simplicity, the probability p^k for each Bernoulli variable is fixed to 0.2.

Methods. Once the variables in the databases were generated, we considered four possible record linkage methods: **FS**, the Fellegi-Sunter model with simple binary comparison as described in (2); **FS3**, the Fellegi-Sunter model using a comparison with 3 categories, as described in (7); **FS4**, the Fellegi-Sunter model using a comparison with 4 categories, as described in (6); **Bayesian**, the bayesian method described in Hejblum et al. (2019). With the methods **FS**, **FS3** and **FS4**, the parameters p_M, m_s^k and u_s^k (see equation 12) are estimated by means of the EM algorithm, and some initial values are required. We initialize with $1/n_A$ for p_M. The formulas to compute the initial values for m_s^k and u_s^k and the stopping criteria are given in Section 2.1 of the supplementary material. The **Bayesian** method is performed by means of the package ludic of Hejblum et al. (2019), where we used 0.01 as the discrepancy rates needed for the method.

Scenario 2: continuous matching variables

Data generating process. In this scenario, each variable X_A^k is generated according to an exponential distribution with parameter λ^k, for $k = 1, \ldots, K$.

To account for possible errors in the matching variables, the variables $X^k_{B,j}$ in database B are then obtained as

$$X^k_{B,j} = \begin{cases} X^k_{A,j} & \text{with probability} \quad 1 - e^k, \\ X^k_{A,j} + \epsilon^k_j & \text{with probability} \quad e^k, \end{cases} \tag{17}$$

where the ϵ^k_j's are iid, generated according to an exponential distribution of parameter λ^k_e.

Simulation parameters. We used $K = 3$ matching variables and $\lambda^k = 0.02$ for $k = 1, \ldots, K$. Because small lags are likely to happen in the registration process, we considered as possible proportions of errors $e^k \in \{0.1, 0.2, 0.3\}$ and $\lambda^k_e \in \{1/2, 1/3, 1/4\}$. This leads to a mean value of approximately 50 days for X^k, and a mean value of approximately $2, 3$ or 4 days for the lag value ϵ^k_j.

Methods. Once the databases were generated, we compared three possible record linkage methods: **FS**, the Fellegi-Sunter model with simple binary comparison as described in (2); **FS3**, the Fellegi-Sunter model using a comparison with 3 categories defined as follows:

$$\gamma^k_{ij} = \begin{cases} 0 & \text{if } |X^k_{B,j} - X^k_{A,i}| = 0, \\ 1 & \text{if } 0 < |X^k_{B,j} - X^k_{A,i}| \leq 3, \\ 2 & \text{if } 3 < |X^k_{B,j} - X^k_{A,i}|, \end{cases} \tag{18}$$

for $k = 1, \ldots, K$; **FS-HGa**, the Fellegi-Sunter model using the absolute distance for comparison defined as:

$$\gamma^k_{ij} = d(X^k_{A,i}, X^k_{B,j}) = |X^k_{B,j} - X^k_{A,i}|. \tag{19}$$

For the **FS-HGa** method, we used the hurdle Gamma distribution

$$f(\gamma^k; p^k_0, \alpha^k, \beta^k) = \begin{cases} p^k_0 & \text{if } \gamma^k = 0, \\ (1 - p^k_0) \dfrac{(\gamma^k)^{(\alpha^k - 1)} e^{-\gamma^k / \beta^k}}{(\beta^k)^{(\alpha^k)} \Gamma(\alpha^k)} & \text{if } \gamma^k > 0, \end{cases} \tag{20}$$

for both f_M^k, f_U^k in equation (13) where $\alpha^k, \beta^k \in \mathbb{R}^+$ and $\Gamma(\alpha^k)$ is the gamma function for $k = 1, \ldots, K$. This is the true distribution for $\gamma^k|M$ under our simulation set-up, since

$$\gamma_{j,j}^k|M = |X_{B,j}^k - X_{A,j}^k| = \begin{cases} 0 & \text{with probability} \quad 1 - e^k, \\ \epsilon_j^k & \text{with probability} \quad e^k, \end{cases} \tag{21}$$

and since ϵ_j^k follows an exponential distribution, which is a particular case of the Gamma distribution with parameters $\alpha^k = 1$ and $\beta^k = 1/\lambda^k$. On the other hand, it is more complicated to describe the true distribution of $\gamma^k|U$. For $j \neq i$, we have

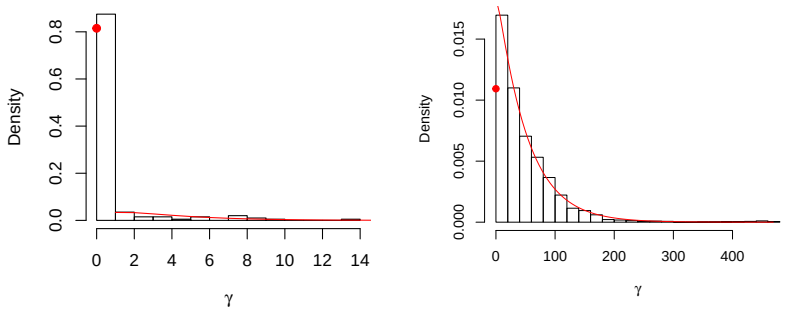
$$\gamma_{i,j}^k|U = |X_{B,j}^k - X_{A,i}^k| = \begin{cases} |X_{A,j}^k - X_{A,i}^k| & \text{with probability} \quad 1 - e^k, \\ |X_{A,j}^k - X_{A,i}^k + \epsilon_j^k| & \text{with probability} \quad e^k. \end{cases} \tag{22}$$

Since $X_{A,j}^k$ and $X_{A,i}^k$ are independent for $i \neq j$, $\gamma_{i,j}^k|U$ follows an exponential distribution with probability $1 - e^k$. With probability e^k, the distribution of $\gamma_{i,j}^k|U$ also involves that of the error ϵ^k. Since this error is typically small compared to the difference $X_{B,j}^k - X_{A,i}^k$, we may also consider that $\gamma^k|U$ approximately follows an exponential distribution.

With the **FS-HGa** method, we propose using the hurdle gamma distribution for f, since it adds more flexibility to the modeling. The histogram of $\gamma_{i,j}^k$ values on one sample is given in Figure 1a for the matched pairs, and in Figure 1b for the unmatched pairs. They decidedly indicate that the hurdle gamma distribution fits well to these values in this example. The robustness of our modeling with different families of distributions for X^k and ϵ^k is studied in Section 3.6 of the supplementary material.

While the parameters for **FS** and **FS3** are estimated by the EM algorithm, the parameters for **FS-HGa** are estimated by the ECM algorithm, which is presented in Section 1 of the supplementary material. The starting values and stopping criteria for all methods are presented in Section 3.1 of the supplementary material.

Since the matching variables are interpreted as durations (in days) in the application presented in Section 5, the generated values X_A^k and ϵ^k values are rounded to the smallest larger integer in this simulation. For example, patients may get a medical act at different times (days, hours and minutes), but the durations are registered in days only.



(a) Histogram of $\gamma|M$

(b) Histogram of $\gamma|U$

Figure 1: Histogram of the positive values of $\gamma_{i,j}^k$ for the matched pairs (left side) and the unmatched pairs (right side) and fitted gamma density estimation (red curve) when $\lambda_e^k = 1/2$ and $e^k = 0.2$

4.2. Performance criteria

All the methods tested for record linkage are evaluated by means of the True Positive Rate

$$\text{TPR}_\tau \quad = \quad \mathbb{P}\left\{q_{ij} \geq \tau | (X_{A,i}, X_{B,j}) \in M\right\},$$

and the Positive Predicted Value

$$\text{PPV}_\tau \quad = \quad \mathbb{P}\left\{(X_{A,i}, X_{B,j}) \in M | q_{ij} \geq \tau\right\}.$$

The True Positive Rate (a.k.a sensitivity or recall) is the proportion of matched pairs which are correctly identified. The Positive Predicted Value (a.k.a. precision) is the proportion of predicted matched pairs which are correctly identified.

These are the most common criteria in an imbalanced binary classification problem, which is the case when the overall set of record pairs is extremely dominated by non-matches. In this work, these criteria are estimated by means of $1,000$ independent Monte Carlo simulations. To save time, all the results are obtained by using the package simsalapar (Hofert & Mächler, 2016) for parallelizing the estimation of all combinations of simulation parameters. A server with 2 Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz with 12 cores in each has been used.

In the simulations, the EM/ECM algorithm is used for the estimation of parameters, with a convergence tolerance of 10^{-6} before reaching the maximum number of iterations (set equal to 500). We observed convergence issues for record linkage methods, more particularly for the usual Fellegi-Sunter method in the Scenario 1: that is, there are some simulations for which the tolerance value is not reached after 500 iterations. The Monte Carlo approximation of the TPR and PPV for any method is therefore obtained from the subset of simulations for which the convergence is attained for all methods. The proportion of cases for which the convergence is attained is presented for all methods in Section 2.2 and 3.2 of the supplementary material, along the results we would obtain for the TPR and PPV if the Monte Carlo iterations with convergence issues were taken into account.

The Monte Carlo approximation for the TPR and PPV are presented in Section 4.3 for the methods considered, with a threshold $\tau = 0.5$. This is the most natural threshold, since it is equivalent to classify a pair as a match if $\mathbb{P}(M|\gamma) \geq \mathbb{P}(U|\gamma)$. In practice, the choice of the threshold τ corresponds to a trade-off between TPR and PPV: a more stringent threshold may increase PPV, but decrease TPR. For a particular case of each scenario, we therefore plotted in Figures 3 and 5 the PPV-TPR curve (a.k.a. precision-recall curve) for different values of τ. Two types of curves are plotted. The "observed" curves correspond to the (theoretical) situation when the parameters are directly estimated by maximizing the full likelihood, assuming that the true status (matched/unmatched) is known for each pair. The "estimated" curves correspond to the

(practical) situation when this status is not known, and the parameters are estimated by the EM/ECM algorithm as described in Section 4.1. The difference between an observed curve and its estimated counterpart is helpful to separate the effect in parameter estimation when using the EM/ECM algorithm.

4.3. Results

Scenario 1

The Monte Carlo estimates for the TPR and the PPV are presented in Figure 2. We first note that for all the methods considered, both criteria improve when the number of matching variables K increases and/or when the probability of error e decreases, as could be expected. In terms of TPR, **FS3** is preferable, followed by **FS4**; **Bayesian** and **FS** show comparable results for $K \leq 40$, but **FS** performs better for $K = 50$. In terms of PPV, **FS4** and **Bayesian** are preferable, with almost identical results; **FS3** performs slightly worse, while **FS** performs poorly, but both methods improve as K increases. Overall, **FS3** performs better than **FS** in both TPR and PPV. As explained in Hejblum et al. (2019), **FS** has many false matches, leading to the smallest PPV. In comparison to **FS4** and **Bayesian**, **FS3** improves the TPR substantially with a slight decrease of the PPV. **FS4** and **Bayesian** show a similar behavior when $e = 0.02$. However, when the error increases, **FS4** has a better TPR with a minimal decrease in PPV as compared to **Bayesian**. In addition, we have also reported the proportion of convergence and the average execution time in Table 1 presented in Section 2.2 of the supplementary material. Generally, the **FS3** and **FS4** have a higher chance of convergence since their comparison vectors provide more information for the algorithm. However, they require a longer computation time than **FS**, since more parameters need to be estimated. In this specific scenario, the execution time of **Bayesian** is much faster than with other linkage methods, because it was performed by package ludic which is optimally designed for this specific method.

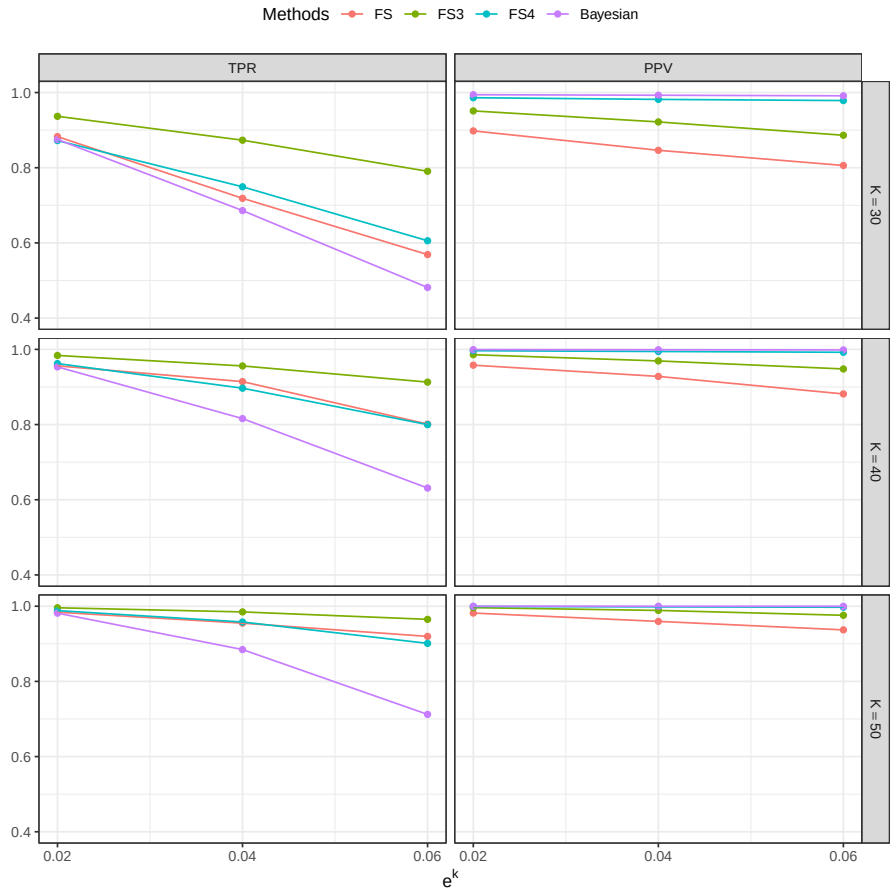


Figure 2: Monte-Carlo estimates of TPR and PPV with binary matching variables only and sample sizes $n_A = 500$ and $n_B = 200$, $p^k = 0.2$ for the parameter of the Bernoulli distribution, a number of matching variables $K \in \{30, 40, 50\}$, and a proportion of errors $e^k \in \{0.02, 0.04, 0.06\}$.

To evaluate the impact of the choice of the threshold τ in the performances of the methods, we consider the particular scenario with the parameters $K = 40$, $p^k = 0.2$ and $e = 0.04$. We plot in Figure 3 the PPV in function of the TPR for different thresholds. The Figure 3 indicates that the observed **FS4** performs better among the observed methods, while the estimated **FS3** performs better among the estimated methods.

Additional simulations with a fixed number of matching variables $K = 40$

and different values for the probability p^k were performed in Section 2.3 of the supplementary material. The results showed that all methods improve significantly when p^k rises from 0.1 to 0.3. Also, the results in Section 2.4 of the supplementary material indicate that all methods gradually improve as the ratio n_B/n_A increases.

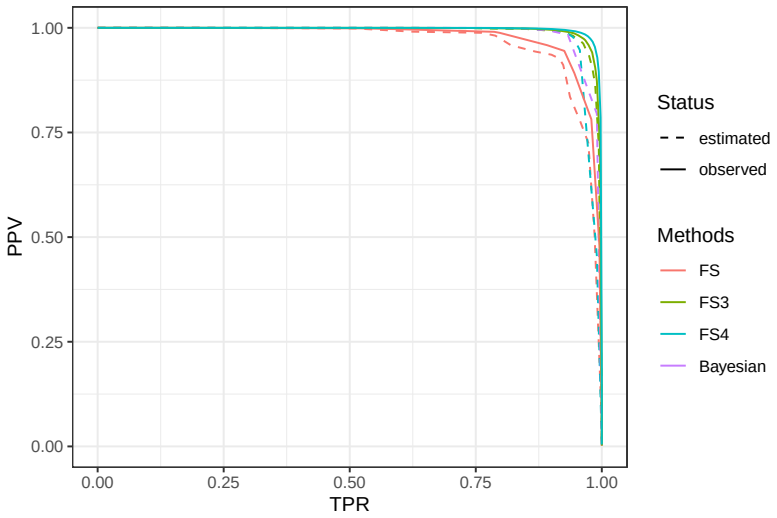


Figure 3: PPV-TPR curves for the observed/estimated version of the methods considered with binary matching variables only, with sample sizes $n_A = 500$ and $n_B = 200$, $K = 40$ matching variables, $p^k = 0.2$ for the parameter of the Bernoulli distribution, and a proportion of errors $e^k = 0.04$.

Scenario 2

The Monte Carlo estimates for the TPR and the PPV are presented in Figure 4. For each method, both the TPR and the PPV decrease as the proportion of errors e increases. We note that the slower decrease is observed for **FS-HGa**, which is also the method which gives both the best TPR and the best PPV in all cases. We also observe that for **FS-HGa** and **FS3**, both the TPR and the PPV decrease with λ_e, but the decrease is very limited for **FS-HGa**. On the other hand, **FS** is not affected by λ_e: this is likely due to the fact that **FS** only considers exact agreement/disagreement in comparison step, while **FS3**

accounts for an additional category when the time lag is no greater than 3 days. Therefore, **FS3** performs better than **FS** when the proportion of error is large ($e = 0.3$) and the mean value of the error is small ($\lambda_e = 1/2; 1/3$). We have also reported the proportion of convergence and the average execution time of each method in Table 2 of Section 3.2 in the supplementary material. Since the implementation of the ECM algorithm in **FS-HGa** has 2 maximization steps, it requires a longer computation time.

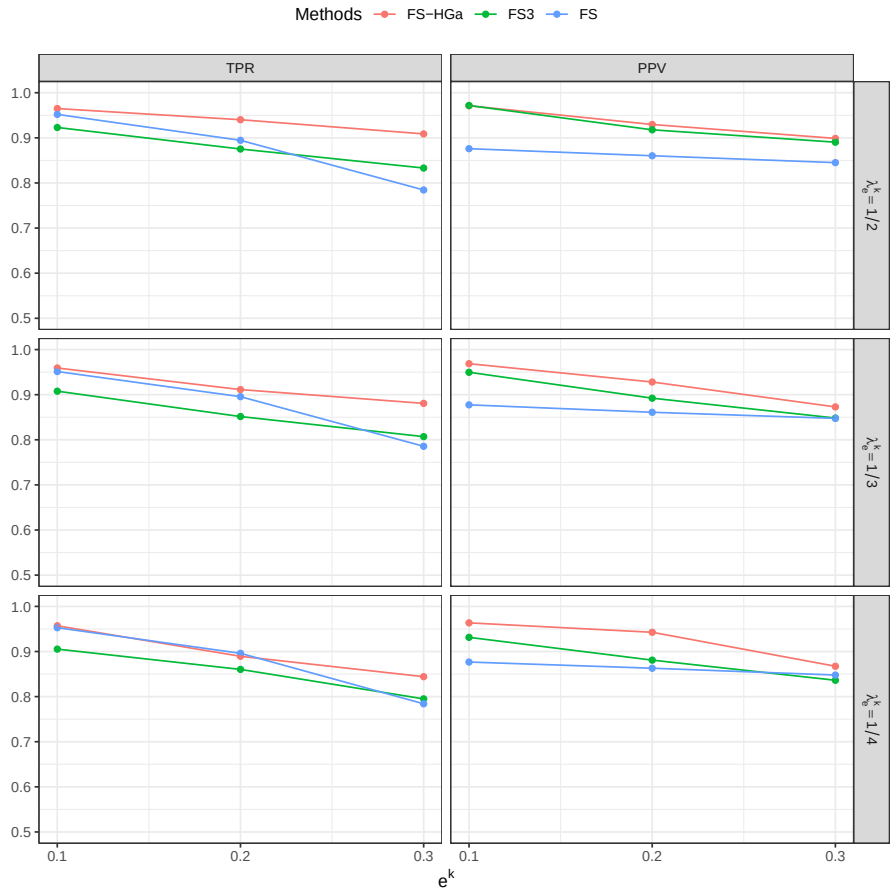


Figure 4: Monte-Carlo estimates of TPR and PPV over different simulation cases when there are only continuous matching variables with sample sizes $n_A = 500$ and $n_B = 200$, $K = 3$ matching variables, $\lambda^k = 0.02$ for the parameter of the Exponential distribution, a proportion of errors $e^k \in \{0.1, 0.2, 03\}$, and a parameter $\lambda_e^k \in \{1/2, 1/3, 1/4\}$ for the error lag.

To evaluate the impact of the threshold τ, we consider the particular scenario with the parameters $K = 3$, $\lambda^k = 0.02$, $e = 0.2$ and $\lambda_e = 1/2$. We observe in Figure 5 that the PPV-TPR curves obtained for a given estimated method and for its observed counterpart are very similar. Also, **FS-HGa** performs significantly better than **FS3** and **FS**.

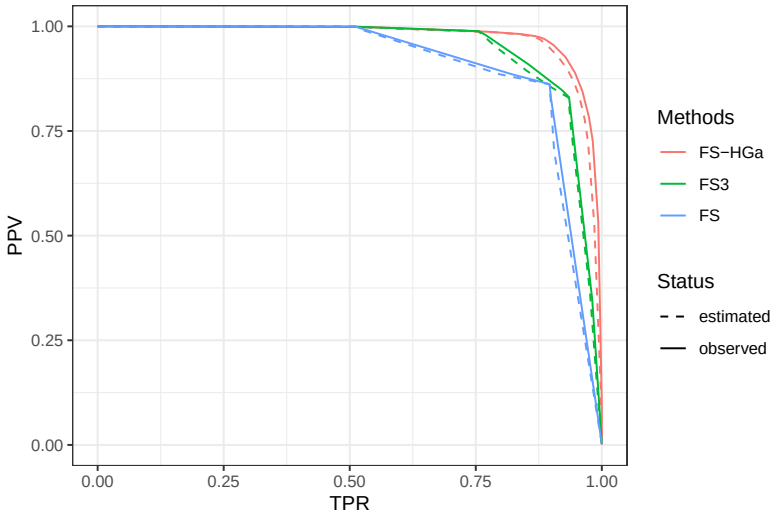


Figure 5: PPV-TPR curves for the observed/estimated version of the methods considered with continuous matching variables only, with sample sizes $n_A = 500$ and $n_B = 200$, $K = 3$ matching variables, $\lambda^k = 0.02$ for the parameter of the Exponential distribution, a proportion of errors $e = 0.2$, and a parameter $\lambda_e = 1/3$ for the error lag.

To evaluate the robustness of **FS-HGa**, we performed additional simulations presented in Section 3.6 of the supplementary material. In these simulations, X^k is generated according to a uniform distribution and ϵ according to a normal distribution. The results indicate that even when the model is misspecified, **FS-HGa** is robust and performs better than the other methods. Also, we considered different values for K and λ^k in Section 3.3 and 3.4 of the supplementary material. In general, under the fixed sample sized n_A and n_B, all methods perform better with more matching variables (larger K) and/or when the matching variables are more informative (smaller λ^k). Finally, the results in Section 3.5

of the supplementary material indicate that all methods gradually improve as the ratio n_B/n_A increases.

5. Application

5.1. Description of SNDS and GETBO databases

The French national health information system SNDS was first created mainly based on the national register of health insurance information (SNIIRAM), which is currently one of the largest claims database in the world (Bezin et al., 2017). The SNDS includes information such as socio-demographic data, real-life use of drugs, chronic medical conditions (ICD10 codes), date and duration of hospital admissions. These databases are therefore of major interest, and their study has already led to several useful findings (e.g. Tuppin et al., 2017a,b). Because of this interest, there is an increasing demand for using this database to enrich existing cohorts or medical registers. However, most of the time, no common identifier is available in the database. Our objective is therefore to link the de-identified GETBO database to the SNDS, when no common individual identifier is available.

The GETBO database results from a data management process of the raw data of the GETBO registry. It is built as a list of documented cases of venous thromboembolism (VTE) recorded between 2013 and 2015 in Brest metropolitan area (Delluc et al., 2016). A given patient may have several events, and the database contains $1,404$ VTE events concerning $1,332$ distinct patients. For each documented case, the diagnostic or therapeutic medical acts were recorded with their type and the precise date, as well as the demographic information for each patient (date of birth, gender, residency code). Linked data consisting of VTE cases from GETBO and corresponding valuable health information from SNDS are used to build a prediction model, which can identify symptomatic VTE early for French people (Noboa et al., 2006; Delluc et al., 2016).

In this application, the so-called SNDS database results from a data extraction process of the raw data from SNDS, including the health insurance

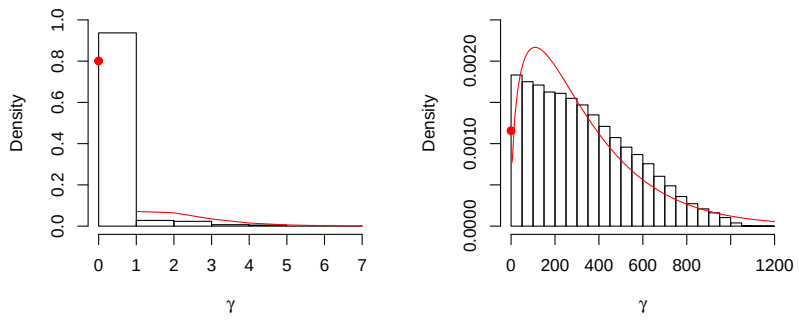
data from SNIIRAM and the national hospital discharge databases. The complete extraction was designed to select patients living in the Brest area, and having at least one care reimbursement between 2013 and 2015. It concerned $369,695$ distinct individuals. We selected patients having, during the studied period, at least one medical act either prescribed for diagnosis purposes of VTE (echodoppler, scintigraphy, tomoscintigraphy and angiography), or for therapeutic purposes (vena cava filter and thrombolysis) that were supposed to be recorded in the GETBO registry. This led to a list of $48,102$ timestamped medical acts concerning $32,382$ distinct patients with all the related demographic information (date of birth, gender, residency code). This database is expected to contain all medical acts in the GETBO database.

5.2. Probabilistic record linkage process

Since some VTE events in GETBO can relate to several medical acts, we first restructure this database such that each row contains only one medical act. This results in a new GETBO database with $1,919$ medical acts associated to $1,332$ patients. There are 6 available matching variables: year of birth, month of birth, residency code, gender, type and date of medical act. A full Cartesian product of the GETBO and SNDS databases requires computing $1,919 \times 48,102 = 92,307,738$ comparison vectors. Therefore, we need to choose a blocking variable to reduce computational time. A good blocking variable should have high quality, and multiple categories distributed as uniformly as possible (Herzog et al., 2007). The gender variable has only two categories and is therefore not very successful in reducing the dimension of the comparison space. Besides, the year of birth is not uniformly distributed and the residency code is likely to change due to moves, for example. Therefore, the month of birth seems the more reasonable choice. It should also be noted that only records with the same type of medical acts should be compared. By employing this scheme, there remains $4,308,847$ candidate pairs that need to be compared in terms of year of birth, residency code, gender and date of medical act.

We use the simple binary comparison function (2) for the year of birth and

residency code variable. For the gender variable, since there is an imbalance between male and female in SNDS database (36.6% compared to 63.4%), we choose (7) as the comparison function. Finally, we choose the absolute distance (19) for the dates of medical acts variable. The comparison step results in a set of $4,308,847$ mixed-type comparison vectors. They are fitted by our proposed extension of Fellegi-Sunter model for mixed-type data, denoted by **FS-ext**. The ECM algorithm is applied to estimate all the model parameters. It stopped after 5 iterations when the relative difference of log-likelihood values of two successive steps was less than 10^{-7}. Once all parameters are estimated, we compute the estimated posterior probabilities of matching (14) for all record pairs of medical acts. Finally, we define a threshold $\tau = 0.5$, and a pair with a greater estimated posterior probability is predicted as a match.



(a) Histogram of $\gamma|\hat{M}$ (b) Histogram of $\gamma|\hat{U}$

Figure 6: Histogram of the comparison values for dates of medical acts of predicted matched pairs (a) and unmatched pairs (b), and the fitted distribution (red line) of our model.

In Figure 6, we present two histograms of comparison values of the dates of medical acts for our predicted matched/unmatched pairs. The red line is the hurdle gamma distribution fitted by our model. Figure 6a indicates that there are more than 90% predicted matches with the same dates, and the others have 1 to 5 days in difference between dates.

5.3. Results

The observation unit is a medical act, and the matching variables are therefore observed at this level. On the other hand, the outcomes are needed at the patient level. We therefore report the application results in two steps. In the first one, we identify record pairs which refer to the same medical act, by applying the record linkage method on the observed data. In the second one, an ad-hoc procedure is performed to get corresponding pairs of patients from the pairs of medical acts.

Firstly, after performing the linkage method on the two databases of medical acts, we get $1,810$ pairs of medical acts that have estimated posterior matching probabilities no smaller than a threshold of 0.5. It is required that one patient in GETBO may be linked to one patient only in SNDS, and conversely. Therefore, if different pairs of medical acts lead to more than two candidates for one patient, we only keep the pairs of medical acts with the highest estimated probabilities, and suppress the others. Eventually, there remains $1,627$ pairs of medical act predicted as matches. The distribution of their (estimated) posterior matching probabilities is presented in Table 1. Among the predicted matched pairs, $1,410/1,627 = 87\%$ have estimated posterior probabilities larger than 0.9.

$\hat{q}$	$(0.5, 0.6]$	$(0.6, 0.7]$	$(0.7, 0.8]$	$(0.8, 0.9]$	$(0.9, 0.95]$	$(0.95, 1]$
Number of pairs of medical acts	4	9	18	186	188	1,222

Table 1: Frequency distribution table of estimated posterior probability of matching for predicted matched pairs of medical acts

From the $1,627$ pairs of medical acts predicted as matches, we obtain $1,146$ corresponding pairs of patients, since one patient may have several medical acts. Among them, 13 patients in GETBO have two different matched candidates in SNDS with the same probability. A random choice between two SNDS candidates is made for these patients. We also consider two different approaches

for linking the two databases: the Fellegi-Sunter model **FS** with a binary comparison (2), and the deterministic method. Under the latter, a pair of medical acts is classified as a match if both records share the same type of medical act, month, year of birth, gender, residency code, while the date of medical act is compared with a tolerance of 3 days. Some manual review is required for pairs that link an individual in the database to more than two individuals in another database. We compare the three approaches in terms of predicted matched pairs of patients.

We summarize the linkage results of the different methods in Table 2. As could be expected, the set of predicted matched pairs of patients obtained under both **FS-ext** and **FS** include all the pairs identified by the deterministic record linkage. All 867 pairs predicted as matches by the deterministic method have a very high average posterior probability of matching for both **FS-ext** ($\bar{\hat{q}}_{\text{FS-ext}} = 0.993$) and **FS** ($\bar{\hat{q}}_{\text{FS-ext}} = 0.996$). Among the 247 remaining pairs which are classified as a match by **FS**, $245/247 \approx 99.2\%$ are also identified by **FS-ext**. Besides, 34 additional pairs of patients are identified as matches by **FS-ext**, with a high average probability ($\bar{\hat{q}}_{\text{FS-ext}} = 0.868$). From a look at the data, these pairs are not predicted by **FS** because they often correspond to a difference of 1 to 5 days in the date of medical acts. Consequently, the proposed method **FS-ext** predicts $1,146$ matched pairs for $1,332$ patients in GETBO, which represents 86% of the patients. On the other hand, the deterministic and **FS** only account for 65% and 83.6% respectively.

	Classified as a match by					
	FS-ext	**FS**	**Deterministic method**	Number of pairs of patients	$\overline{\hat{q}}_{\text{FS-ext}}$	$\overline{\hat{q}}_{\text{FS}}$
	X	X	X	867	0.993 (0.003)	0.996 (0)
	X	X		245	0.900 (0.045)	0.911 (0)
	X			34	0.868 (0.136)	
		X		2		0.911 (0)
Total	1146	1114	867			

Table 2: Comparison of three different record linkage methods with the number of pairs, the average estimated posterior probability of matching $\overline{\hat{q}}$ and the standard deviation (in parentheses)

6. Discussion

In this article, we proposed two comparison approaches for low prevalence categorical and continuous matching variables. The proposed comparison functions aim to make a more extensive use of the matching variables in the comparison vectors. We propose an extension of the Fellegi-Sunter probabilistic record linkage model, for comparison vectors containing both categorical and continuous comparison values. This model allows for using a variety of comparison functions, which can reflect matching data more accurately. We also suggest the use of a mixture of hurdle gamma distributions, for modeling the absolute difference between continuous variables such as dates. This distribution has never been formerly considered in the record linkage literature. In practice, the distribution for comparison values of continuous matching variables should be considered and validated a posteriori.

The simulation studies show that our proposed model outperforms the simple model with binary comparison in all the scenarios considered. For categorical matching variables, in Scenario 1, we have showed that the proposed model is more efficient than the standard model, especially when there are low prevalence

values. However, if the frequencies of the different categories of a matching variable are similar, then there is not much difference between our approach and the standard one. In that case, the model with binary comparison should be considered due to its simplicity. For continuous matching variables, in Scenario 2, the proposed mixture of hurdle gamma distributions performs better than the standard model, and is robust to some misspecification of the distribution of the comparison function (see Section 3.6 in the supplementary material). However, our evaluation remains specific to the fact that we are dealing with continuous time variables, which may be naturally modelled by Gamma distributions. A similar approach could be pursued for other types of matching variables (e.g., string variables), but would require a different modelling for the similarity measure between strings.

We also conducted a simulation with mixed-type data in Section 4 of the supplementary material. Consistently with the previous simulation results, the proposed model has a better performance than the standard model. In the application on real data, the performance is also better. We obtain a larger number of patients matched between the SNDS and the GETBO datasets, with high matching probabilities.

In practice, the matching variables that can be used for record linkage may include missing data. Also, dates of events may be censored. It would be of great practical interest to develop a joint modeling for record linkage and handling of missing values, to improve the performance of the record linkage process in this case. This is an important matter for further research. In a different approach, Copas & Hilton (1990) described a hit-miss model for record linkage which can accommodate the frequency distribution and missing values of the matching variables. However, this approach is not as commonly used in practice as the Fellegi-Sunter model due to its specific context (Goldstein et al., 2017). Besides, we did not consider matching variables varying over time. A study of Li et al. (2011) suggests that considering matching variables along with their time stamp (if applicable) may improve matching quality.

A problem of most probabilistic record linkage models lies in the imbalance

between matched and non-matched pairs in the set of all comparison vectors, which may cause bias in parameter estimation. Blocking methods have been introduced to reduce the number of non-matched pairs, along with the computational cost. However, some true matched pairs may be overlooked if the blocking variable contains errors. Recently, Fortini (2020) introduced a robust approach where the EM algorithm is modified to obtain unbiased estimates of parameters in this context. However, this approach is designed for binary comparison values only.

The construction of the complete likelihood function rests on the assumption that the comparison vectors are independent. Such assumption may not be valid in practice, especially when there are matching restrictions, such that each record in a database can be linked to only one record in another database. Lee et al. (2020) recently proposed a maximum entropy classification for record linkage which overcomes this assumption.

References

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

Bezin, J., Duong, M., Lassalle, R., Droz, C., Pariente, A., Blin, P., & Moore, N. (2017). The national healthcare system claims databases in france, sniiram and egb: Powerful tools for pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety*, *26*, 954–962.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '08 (p. 151–159). Association for Computing Machinery.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Link-*

age, Entity Resolution, and Duplicate Detection. Springer Publishing Company, Incorporated.

Christen, P., & Winkler, W. E. (2017). Record linkage. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1066–1075). Springer US.

Copas, J. B., & Hilton, F. J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *153*, 287–320.

Delluc, A., Tromeur, C., Ven, F., Gouillou, M., Paleiron, N., Bressollette, L., Nonent, M., Salaun, P., Lacut, K., Leroyer, G., C.and Le Gal, Couturaud, F., Mottier, D., & study group, E. (2016). Current incidence of venous thromboembolism and comparison with 1998: a community-based study in western france. *Thromb Haemost*, *116*, 967–974.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Enamorado, T. (2018). Active learning for probabilistic record linkage. *Social Science Research Network (SSRN)*, .

Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, *113*, 353–371.

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

Fortini, M. (2020). An improved fellegi-sunter framework for probabilistic record linkage between large data sets. *Journal of Official Statistics*, *36*, 803–825.

Goldstein, H., Harron, K., & Cortina-Borja, M. (2017). A scaling approach to record linkage. *Statistics in Medicine*, *36*, 2514–2521.

Grannis, S., Overhage, J., Hui, S., & McDonald, C. (2003). Analysis of a probabilistic record linkage technique without human review. *Annual Symposium proceedings. AMIA Symposium*, (pp. 259–63).

Hejblum, B., Weber, G., Liao, K., Palmer, N., Churchill, S., Shadick, N., Szolovits, P., Murphy, S., Kohane, I., & Cai, T. (2019). Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific Data*, *6*.

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

Hof, M., & Zwinderman, A. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in medicine*, *31*, 4231–4242.

Hofert, M., & Mächler, M. (2016). Parallel and other simulations in r made easy: An end-to-end study. *Journal of Statistical Software*, *69*, 1–44. doi:10.18637/jss.v069.i04.

Jaro, M. (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.

Kim, G., & Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis*, *56*, 2756 – 2770.

Lahiri, P., & Larsen, M. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, *100*, 222–230.

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

Lee, D., Zhang, L., & Kim, J. (2020). Maximum entropy classification for record linkage. arXiv:2009.14797.

Li, P., Dong, X., Maurino, A., & Srivastava, D. (2011). Linking temporal records. In *Proceedings of the VLDB Endowment* (pp. 956–967). volume 4.

Mamun, A., Aseltine, R., & Rajasekaran, S. (2016). Efficient record linkage algorithms using complete linkage clustering. *PLOS ONE*, *11*, 1–21.

Meng, X., & Rubin, D. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, *80*, 267–278.

Noboa, S., Mottier, D., Oger, E., & GROUP, T. E.-G. S. (2006). Estimation of a potentially preventable fraction of venous thromboembolism: a community-based prospective study. *Journal of Thrombosis and Haemostasis*, *4*, 2720–2722.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, *112*, 600–612.

Sayers, A., Ben-Shlomo, Y., Blom, A., & Steele, F. (2015). Probabilistic record linkage. *International journal of epidemiology*, *45*, 954–964.

Steorts, R., Hall, R., & Fienberg, S. (2016). A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, *111*, 1660–1672.

Tancredi, A., & Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, *5*, 1553 – 1585.

Tuppin, P., Pestel, L., Samson, S., Cuerq, A., Rivière, S., Tala, S., Denis, P., Drouin, J., Gissot, C., Gastaldi-Ménager, C., & Fagot-Campagna, A. (2017a). Poids humain et économique des cancers en france en 2014, les données du sniiram. *Bulletin du Cancer*, *104*, 524 – 537.

Tuppin, P., Rudant, J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., Roquefeuil, L., Maura, G., Caillol, H., Tajahmady, A., Coste, J., Gissot, C., Weill, A., & Fagot-Campagna, A. (2017b). Value of a national administrative

database to guide public decisions: From the système national d'information interrégimes de l'assurance maladie (sniiram) to the système national des données de santé (snds) in france. *Revue d'Épidémiologie et de Santé Publique*, *65 Suppl 4*, 146–167.

Winkler, W. (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* (pp. 667–671).

Winkler, W. (1989). Frequency-based matching in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 778–783).

Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, *11*, 95–103.

Xu, H., Li, X., Shen, C., Hui, S., & Grannis, S. (2019). Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? *The Annals of Applied Statistics*, *13*, 1753–1790.

Zhang, L., & Tuoto, T. (2020). Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *100*, 222–230.

Zhu, V. J., 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. *Journal of the American Medical Informatics Association*, *16*, 738–745.

Zhu, Y., Matsuyama, Y., Ohashi, Y., & Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? a simulation study. *Journal of Biomedical Informatics*, *56*, 80 – 86.