Heliyon 8 (2022) e12311

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Implementation and validation of a probabilistic linkage method for population databases without identification variables

Amado D. Quezada-Sánchez [a], Iván Espín-Arellano [a], Evangelina Morales-Carmona [a], Diana Molina-Vélez [a], Lina Sofía Palacio-Mejía [b], Edgar Leonel González-González [a], Mariana Alvarez Aceves [b], Juan Eugenio Hernández-Ávila [a,*]

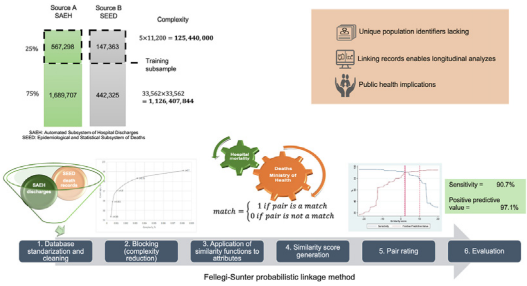
[a] *Center for Evaluation and Surveys Research, National Institute of Public Health. Av. Universidad 655 Col. Sta. María Ahuacatitlán. C.P. 62100. Cuernavaca, Morelos, Mexico*

[b] *Conacyt - National Institute of Public Health, Av. Universidad 655 Col., Sta. María Ahuacatitlán, C.P. 62100 Cuernavaca, Morelos, Mexico*

HIGHLIGHTS

- Data from different health information systems cannot be easily linked when standardized ID codes are not available.
- Linking records from the same unit of analysis makes it possible to perform key epidemiological analysis.
- We evaluated the performance of a blocking approach based on trigrams and the EM algorithm for probabilistic linkage.
- Our blocking achieved 95.76% pairs completeness and a 99.9996% complexity reduction in the validation sample.
- After classification in validation sample, we achieved a sensitivity of 90.72% and a positive predictive value of 97.10%.

GRAPHICAL ABSTRACT


Source A
SAEH
Source B
SEED
Complexity
25%
567,298
147,363
5×11,200 = 125,440,000
Training subsample
75%
1,689,707
442,325
33,562×33,562 = 1,126,407,844
SAEH: Automated Subsystem of Hospital Discharges
SEED: Epidemiological and Statistical Subsystem of Deaths
Unique population identifiers lacking
Linking records enables longitudinal analyzes
Public health implications
SAEH discharges
SEED death records
Hospital mortality
Deaths Ministry of Health
match = { 1 if pair is a match; 0 if pair is not a match
Sensitivity = 90.7%
Positive predictive value = 97.1%
1. Database standarization and cleaning
2. Blocking (complexity reduction)
3. Application of similarity functions to attributes
4. Similarity score generation
5. Pair rating
6. Evaluation
Fellegi-Sunter probabilistic linkage method

ARTICLE INFO

Keywords:
Record linkage
Blocking
Probability
Algorithm
Information systems
Hospital discharge
Mortality

ABSTRACT

Linking records of the same person from different sources makes it possible to build administrative cohorts and perform longitudinal analyzes, as an alternative to traditional cohort studies, and have important practical implications in producing knowledge in public health. We implemented the Fellegi-Sunter probabilistic linkage method to a sample of records from the Mexican Automated System for Hospital Discharges and the Statistical and Epidemiological System for Deaths and evaluated its performance. The records in each source were randomly divided into a training sample (25%) and a validation sample (75%). We evaluated different types of blocking in terms of complexity reduction and pairs completeness, and record linkage in terms of sensitivity and positive predictive value. In the validation sample, a blocking scheme based on trigrams of the full name achieved 95.76% pairs completeness and 99.9996% complexity reduction. After pairs classification, we achieved a sensitivity of 90.72% and a positive predictive value of 97.10% in the validation sample. Both values were about one percentage point higher than that obtained in the automatic classification without clerical review of potential pairs. We concluded that the linkage algorithm achieved a good performance in terms of sensitivity and positive predictive value and can be used to build administrative cohorts for the epidemiological analysis of populations with records in health information systems.

[*] Corresponding author.
E-mail address: juan_eugenio@insp.mx (J.E. Hernández-Ávila).

https://doi.org/10.1016/j.heliyon.2022.e12311
Received 8 November 2022; Received in revised form 6 December 2022; Accepted 6 December 2022

1. Introduction

The information systems and administrative records of the health sector offer valuable information at the individual level. However, this information cannot be easily integrated due to poor homogeneity and lack of unique identifiers and standards for the registration of personal data. The ability to link records of the same person from different sources would make it easier to build administrative cohorts and carry out longitudinal analysis. These designs are valuable for the healthcare field not only for their contribution to the development of knowledge, but also because of their practical implications.

Within epidemiological study designs, cohort studies are useful to approximate causal inferences between risk factors and health outcomes [1]. However, their high cost, the long times between the start and the occurrence of the events of interest, as well as the loss of study subjects, limit their implementation. In this sense, the construction of administrative cohorts by linking personal records between different health information systems is a complementary alternative to traditional cohort studies. The analysis of administrative cohorts has contributed to the progress of knowledge in health in different areas. For instance, it has been used to analyze the risk of dementia in hospitalized patients with diabetes, the risk of pneumonia in people with severe mental illness, the relationship between maternal alcohol consumption and child protection, among other applications [2, 3, 4].

To generate an administrative cohort from different data sources, a method for linking the records corresponding to the same individual is needed and must consider inaccuracies in the recording of personal information that could emerge in practice. Deterministic linkage methods have the disadvantage of requiring perfect match on the set of selected attributes. When the identifier is not unique and there are mistakes in the individual identifiers in the different data sources, it is not possible to link the records because the group of auxiliary variables or attributes do not match exactly. In situations such as confusion between homonyms, use of abbreviations, or -input errors, among others, a higher number of records can be linked using a probabilistic approach in which phonetic equivalences and similarity functions are applied to each attribute.

The Fellegi-Sunter probabilistic linkage model [5] has proven to be very useful for identifying the same person in different administrative records. For each pair of records, a similarity score is generated based on a ratio that expresses the probability of a matching result given that the pair of records is from the same individual, relative to the probability of the same matching result when the pair does not belong to the same individual. The similarity score is used to classify pairs and recognize those with a high probability of belonging to the same person. This process requires the selection of a cutoff point for the score, similarly to the approach in which diagnostic tests are carried out to detect subpopulations with a certain disease or condition [6].

Although the Fellegi-Sunter Model and probabilistic linkage methods have been implemented in tools such as Link Plus [7], there are practical limitations to use them with large databases. There is currently no software available to probabilistically link records from different information systems and data sources that is optimized for use in Spanish language, and that can be used efficiently with large databases.

In this paper, we present the implementation of a probabilistic linkage algorithm in Stata [8] based on the Fellegi-Sunter methodology, applied to databases of two information systems in Spanish. Then, we assess its performance in terms of sensitivity and positive predictive value. Additionally, we evaluate the pairs completeness of different types of blocking to reduce the number of full comparisons between the two sources of information.

2. Methodology

2.1. Data

We used 2014 data from the Automated Subsystem of Hospital Discharges (SAEH, acronym in Spanish) [9] and the Epidemiological and Statistical Subsystem of Deaths (SEED, acronym in Spanish) [10]. The administration of the SAEH and the SEED are under the responsibility of the General Directorate of Health Information, through the Directorate of Information on Health Needs and Population, of the Ministry of Health of Mexico (MoH). SEED information is mostly collected in the Civil Registry offices, and then captured in the health jurisdictions of the State Health Services (SESA) and in some hospitals of the MoH. It covers the entire population that inhabited or was present in the national territory at the time of death. SAEH records contain information about the care provided during a patient's stay in hospitalization units managed by the MoH and SESA, covering populations without social security, and those affiliated to the National Health Protection Commission (Seguro Popular), this is approximately 50% of Mexican population (people working in the informal economic sector and their families).

SAEH data included 2,257,005 records of which 64,923 were death discharges. SEED included 589,688 records, corresponding to all deaths registered in the year in Mexico.

Access to the databases with personal information was possible through a collaboration and confidentiality agreement between the National Institute of Public Health and the General Directorate of Health Information from the MoH.

2.2. Reference standard and study population design

We applied the linkage model using personal attributes such as name, paternal surname, maternal surname, year of birth, sex, and the entity and municipality of residence codes. To evaluate the performance of the model, we used the death certificate folio (DCF) present for a subset of records in both databases as a reference standard. We identified matching cases with the same DCF and person name, and then we reviewed cases with different names to rule them out. After this review, we found a total of 44,762 people who met the matching condition in both sources. For the linkage model implementation, the remaining records of hospital discharges and death records were included in our analyses. The performance evaluation was carried out only for the set of pairs for which its status (true/false pair) could be determined through the reference standard. Both the reference standard and additional records were randomly assigned to a training and a validation group (25% and 75%, respectively). Table 1 shows the partition of all possible pairs of the training group according to their source and condition of inclusion/non-inclusion in the evaluation of the model.

Once the unsupervised linkage was carried out on the training sample, partitions A–D were used to assess its performance. This information and the analysis of the classification errors informed the linkage that was carried out in the validation sample, which was also evaluated.

2.3. Variable preparation

We removed special and duplicated characters, numbers, unnecessary spaces, and gender articles in name and surname variables. We replaced all characters with uppercase and completed name and surname abbreviations. Furthermore, we unified in names and surnames, those characters with a high frequency of cognitive errors and equivalent or similar phonetics in Spanish. Then, we replaced with a null value all the records with invalid or missing information. After this procedure, 0.1% records in SAEH and 0.2% in SEED were eliminated. We also eliminated all records with ages less than 1 year (6% SAEH, 4% SEED) due to the lack of name information in most of them.

Table 1. Partition of the total set of pairs of the training group.

Partition	Source (number of records)		Included in performance evaluation	Number of possible pairs	Number of true pairs
	SAEH	SEED			
A	Reference standard (11,200)	Reference standard (11,200)	Yes	125,440,000	11,200
B	Reference standard (11,200)	Deaths, additional to standard (136,163)	Yes	1,525,025,600	0
C	Discharges alive (550,577)	Reference standard (11,200)	Yes	6,166,462,400	0
D	Discharges death, additional to standard (5,521)	Reference standard (11,200)	Yes	61,835,200	0
E	Discharges alive (550,577)	Deaths, additional to standard (136,163)	No	74,968,216,051	unknown
F	Discharges death, additional to standard (5,521)	Deaths, additional to standard (136,163)	No	751,755,923	unknown
Total	**567,298**	**147,363**		**83,598,735,174**	

In SAEH, the variable year of birth was not available, so we calculated it by subtracting the age from the year of admission to the medical unit. We eliminated records without year of birth or enough data to calculate it (0.01% SAEH, 0.9% SEED). In both sources, place of residence was specified using state and municipality codes.

2.4. Blocking

The blocking phase consists of reducing the number of full comparisons, by filtering records pairs that show similarity in one of the attributes. We evaluated the performance of different blocking types in relation to their pairs completeness (percentage of total correct pairs included in the analysis database) and complexity reduction (percentage of total possible pairs not included in the analysis database). We assessed different types of blocking applied to the name attribute that included the Soundex and NYSIIS phonetic encodings [11], as well as a similarity function from trigrams applied to the full name. Pairs that had a value equal to or greater than a defined cutoff point in the proportion of trigrams that matched were filtered. More specifically, we used the similarity function $trigrams_{common}/min(trigrams_{stringA}\ , trigrams_{stringB})$ where $trigrams_{common}$ is the number of trigrams present in two compared full name strings and $min(trigrams_{stringA}\ , trigrams_{stringB})$ is the number of trigrams from the full name string with the lowest number of characters. Before applying the function, a leading and ending special character was added to all string chains. We evaluated the performance of different cutoff points (i.e. a sequence in steps of 0.05 units, and then we evaluated an intermediate point between the two best results) in the training group in terms of pairs completeness and complexity reduction and selected the cutoff point that minimized the loss function shown in Eq. (1), in which the weights $w_1 = 50$ and $w_2 = 0.01$ were determined empirically to facilitate the cutoff selection that provided balance between maximizing pair completeness and minimizing complexity.

$$loss = w_1 \times (100 - \%\ complexity\ reduction) + w_2 \times (100 - \%\ pairs\ completeness) \tag{1}$$

In the case of sex, year of birth, and residence codes, the blocking consisted of pairs that showed a perfect match in such attributes. We performed blockings for each of those attributes and assessed its performance.

There are other approaches to pairs filtering [12]. Recent developments include the use of double-embedding record linkage and meta-blocking techniques [13, 14].

2.5. Linking algorithm

We applied the Expectation-Maximization algorithm to the Fellegi-Sunter model [5] to calculate the similarity score for classification. The algorithm is based on the likelihood function [15] and alternates between the expectation (E) and maximization (M) steps, until finding stability in the estimation of the similarity probabilities. More details of the probabilistic linkage method and the EM algorithm can be found in Herzog et al. [16].

Let $\widehat{p}$ be the estimated proportion of true matching pairs; $i = 1, 2, .., K$ the index of the attributes; K the total number of attributes; $j = 1, 2, .., N$ the index of the compared pair; N the total number of pairs; γ_i^j an indicator variable (0/1) that was set to 1 if the pair j coincides in the attribute i; m_i the probability of matching for the attribute i given that the pair is true (true status means that the information coming from the two records belongs to the same subject); and u_i the probability of matching for the attribute i given that the pair is incorrect (the information from the two records belongs to different subjects). Steps E and M are described in the following equations:

Step E – Expectation

$$\widehat{g}_j = \frac{\widehat{p}\prod_{i=1}^{K} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j}}{\widehat{p}\prod_{i=1}^{K} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j} + (1-\widehat{p})\prod_{i=1}^{K} u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j}} \tag{2}$$

Step M – Maximization

$$\widehat{m}_i = \frac{\sum_{j=1}^{N}\widehat{g}_j\gamma_i^j}{\sum_{j=1}^{N}\widehat{g}_j}\ \widehat{u}_i = \frac{\sum_{j=1}^{N}\left(1-\widehat{g}_j\right)\gamma_i^j}{\sum_{j=1}^{N}\left(1-\widehat{g}_j\right)} \tag{3}$$

$$\widehat{p} = \frac{\sum_{j=1}^{N}\widehat{g}_j}{N} \tag{4}$$

To start the algorithm, it is necessary to provide a priori values for the probabilities m_i, u_i, and the proportion $\widehat{p}$. We set $m_i = 0.95$ and $u_i = 0.20$ for all attributes, except for sex, whose a priori probability u was set to 0.5 (the probability that the sex of two different people coincide is approximately 0.5). These a priori probabilities can be set based on previous studies or they can be guessed, the EM algorithm is not particularly sensitive to starting values, however, it is important to set a priori m probabilities higher than their corresponding u probabilities [16, 17]. For the a priori probability $\widehat{p}$, we divided the number of unique records in the source with fewer records by the total pairs in the blocking.

We established as a tolerance criterion to determine the number of iterations (t), that the maximum difference between estimated probabilities of subsequent iterations did not exceed the value of 10^{-5}, that is, $max\ \{|m_{i,t} - m_{i,t-1}|, |u_{i,t} - u_{i,t-1}|\}\langle 10^{-5}$. Once we had calculated the similarity probabilities, we computed the weights for each attribute. If the attribute i matches, $(\gamma_i = 1)$, we assigned the weight

$$w_{match,i} = log_2\left(\frac{m_i}{u_i}\right) \tag{5}$$

while if it does not match $(\gamma_i = 0)$ we assigned the weight

$$w_{unmatch,i} = log_2\left(\frac{1-m_i}{1-u_i}\right) \tag{6}$$

The similarity score results from adding the weights assigned to all the attributes that are compared.

$$score = log_2(R) = \sum\nolimits_{i=1}^{K}(w_{match,i} \times \gamma_i + w_{unmatch,i} \times (1-\gamma_i)) \tag{7}$$

This score corresponds to the logarithm base 2 of the ratio R (Eq. (8)) under the assumption of conditional independence where γ is a vector of comparison results on the attributes, Γ is the set of possible comparison results, M is the set of pairs that are correct, U is the set of pairs that are incorrect, and form a partition of the total pairs: $M \cup U = N$

$$R = \frac{P(\gamma \varepsilon \Gamma | r \varepsilon M)}{P(\gamma \varepsilon \Gamma | r \varepsilon U)} \tag{8}$$

2.6. Attributes

As already mentioned, the attributes used for linking were seven: name, paternal surname, maternal surname, sex, year of birth, and entity and municipality of residence codes. For the text variables, we applied the Levenshtein similarity function [12] and classified an attribute as coincident when it was equal or greater than 0.9. Additionally, we applied the Dice bigram similarity function to the full concatenated name with an extra special character at the beginning and the end of the text string. Name, paternal surname, and maternal surname were each considered concordant when the Dice similarity score was equal or greater than 0.9. For year of birth, pairs with the same value or differing by one year were considered concordant.

2.7. Unsupervised pair classification

We applied the proposed probabilistic linkage method to the set of pairs filtered through a trigram-based blocking procedure. Using visual support from a histogram of similarity scores and inspection of attribute data from pairs in a region in which their classification status was more difficult to establish, we selected two cutoff points for the similarity score that resulted in three categories: 1) pairs classified as true 2) pairs that required review 3) pairs classified as false. After clerical review, we classified the pairs in the second category as true or false. The tasks of setting up the unsupervised cutoffs and perform the clerical review were performed by two of the authors without knowing the results of the supervised classification.

2.8. Unsupervised classification performance assessment

Using the group of pairs whose true status is known (partition A–D, Table 1), we assessed the performance of unsupervised classification in terms of sensitivity (percentage of true pairs classified as pairs) and positive predictive value (percentage of linked pairs that are correct pairs). The percentage of pairs completeness achieved with the blocking step represents the maximum value that sensitivity can reach.

2.9. Supervised pair classification

Using the set of pairs for which their true status is known, we selected the similarity score cutoff that maximized the sum of sensitivity and positive predictive value, resulting in a two-category classification of the pairs.

2.10. Analysis in the validation group

Starting from the results of the supervised classification and the performance of the unsupervised classification in the training group, we analyzed misclassification errors and the selected cutoff points for classification into three categories. This analysis informed the decisions to classify pairs in the validation group. Once the classification was carried out, we assessed its performance in terms of sensitivity and positive predictive value.

2.11. Comparison of characteristics between the reference standard and observations not in the reference standard

We compared the distribution of sex, state of residence marginalization [18], and year of birth in the reference standard with their corresponding distribution in the rest of observations from SAEH excluding live discharges and from SEED excluding observations with social security. These exclusions were performed to make distributions more clearly comparable since SAEH covers population without social security and includes both live and death discharges.

2.12. Implementation in SQL-server and Stata software

We performed all blocking procedures in SQL-Server 2012 [19] and the EM algorithm based on the Fellegi-Sunter model was programmed in Stata MP v.15 [8]. A code file for both programs is available as Supplementary Material.

3. Results

3.1. Blocking

In the training group, the different blocking approaches applied to the full name showed contrasting differences in their performance. Pairs completeness achieved by Soundex coding was 94.58% with a complexity reduction of 98.22%, while pairs completeness achieved by the NYSIIS coding was 90.72% with a complexity reduction of 99.79%. Pairs completeness and complexity reduction of the trigram blocking depended on the cutoff point for the trigram similarity function (Figure 1). According to these results, the blocking that minimized the loss function (Eq. (1)) corresponded to the cutoff point of 0.825 for the trigram similarity function, which reached 95.91% pairs completeness and a complexity reduction of 99.9994%. Trigram blocking allowed to achieve better results in terms of pairs completeness and complexity reduction compared to blocking based on phonetic encodings. It is important to note that when the total number of possible pairs is in the millions, a seemingly small improvement in complexity reduction as a percentage may correspond to a considerable drop in the size of the number of pairs filtered.

For the blocking by year of birth, we obtained 46.7% pairs completeness and 99.3916% complexity reduction. Meanwhile, blocking using residence codes (entity + municipality) achieved 84.62% pairs completeness and 99.6583% complexity reduction. These results showed that blocking on the full name using the trigram similarity function presented the best performance in terms of pairs completeness and complexity reduction. Although it is possible to filter pairs through the unions of individual blockings, full name trigram blocking simultaneously achieved relatively high pairs completeness and substantially reduced complexity.

In the validation sample, we achieved 95.76% pairs completeness and 99.9996% complexity reduction with the cutoff point selected for the trigram similarity function of the full name in the training sample.

3.2. Linkage with the EM algorithm in the training group

The total number of pairs to be compared after blocking in the training sample was 508,794. We observed 124 attribute similarity patterns out of a total of $2^7 = 128$ possible patterns. The linkage algorithm met the tolerance criterion in iteration 16. The attribute with the highest m probability was state of residence, while the one with the lowest m probability was year of birth (Table 2). We observed the highest u probability for the attribute of sex and the lowest u probability for the attribute of municipality of residence. The attributes that showed the greatest capacity for discrimination were the state and municipality

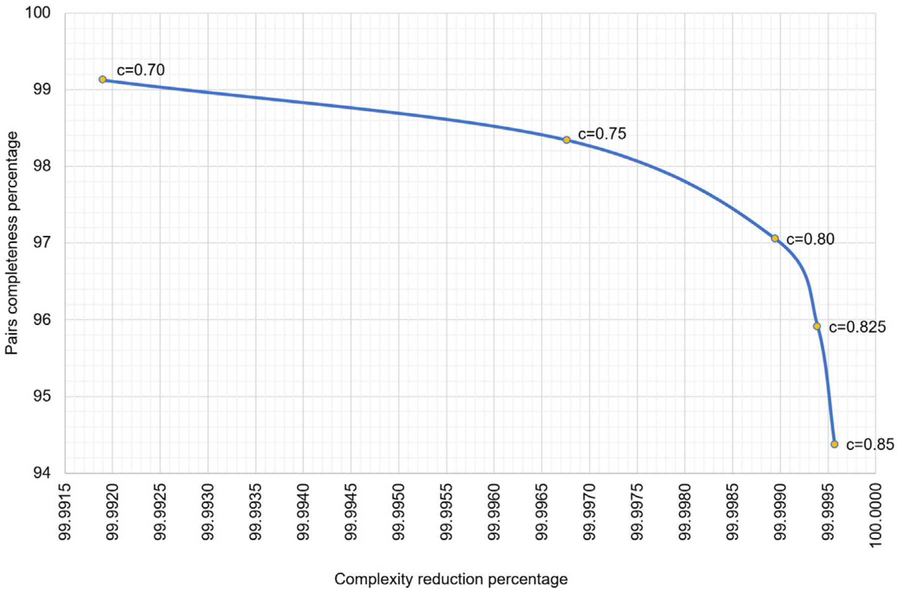


Figure 1. Pairs completeness and complexity reduction performance of trigram blockings with different similarity thresholds in the training group. c: cutoff point of the trigram similarity function.

codes. It should be noted that these results followed a pairs filtering step based on the full name. After obtaining the posterior m and u probabilities, we calculated the weights for the similarity score (Eq. (7)).

3.3. Supervised binary classification in the training group

The optimized cutoff point for a supervised classification in two categories was score = 7.90, with a sensitivity of 89.7% and a positive predictive value of 99.1%.

3.4. Unsupervised classification in the training group

We classified pairs with a similarity score greater than or equal to 9.82 as matching, and pairs with a similarity score less than or equal to 2.79 as mismatched. We performed clerical review of pairs with a score greater than 2.79 but less than 9.82. With this classification, we achieved a sensitivity of 90.2% and a positive predictive value of 99.1%.

Figure 2 shows the sensitivity and positive predictive value depending on each cutoff point. The dashed reference lines mark the cutoff points established in the unsupervised classification into three categories (non-coinciding, potentially coincident, coincident) and the solid reference line indicates the binary supervised classification (mismatched, matched).

Table 2. Similarity score weights, m and u probabilities, for each attribute after applying the EM algorithm to the Fellegi-Sunter model in the training group.

Attribute	Prior		Posterior		Weights	
	m	u	m	u	w_{match}	$w_{unmatch}$
Name	0.95	0.20	0.909	0.361	1.33	-2.81
Surname (paternal)	0.95	0.20	0.984	0.492	1.00	-4.96
Surname (maternal)	0.95	0.20	0.959	0.338	1.51	-4.03
Year of birth	0.95	0.20	0.807	0.019	5.40	-2.35
Sex	0.95	0.50	0.956	0.854	0.16	-1.73
State	0.95	0.20	0.998	0.064	3.96	-9.15
Municipality	0.95	0.20	0.851	0.012	6.10	-2.73

Estimated through the EM algorithm applied to the Fellegi-Sunter model of probabilistic linkage.
Prior p probability = 0.112, posterior p probability = 0.025.

3.5. Applying the EM algorithm in the validation group

Once we applied the blocking, we obtained a total of 4,677,152 pairs for comparison and observed 127 similarity patterns. The linkage algorithm met the tolerance criterion at iteration 19. The probabilities m and u were similar to those obtained in the training sample (Table 3).

3.6. Application of the cutoff point chosen for binary classification in the training to the validation group

With the cutoff point of 7.90 we obtained from training for a binary classification, we achieved a sensitivity of 89.86% and a predictive value of 96.18% in the validation group.

3.7. Unsupervised classification in the validation group

We classified pairs with a similarity score greater than or equal to 9.39 as matching, pairs with a similarity score less than or equal to 1.88 as mismatched and performed clerical review of pairs with a score greater than 1.88 but less than 9.39. With this classification

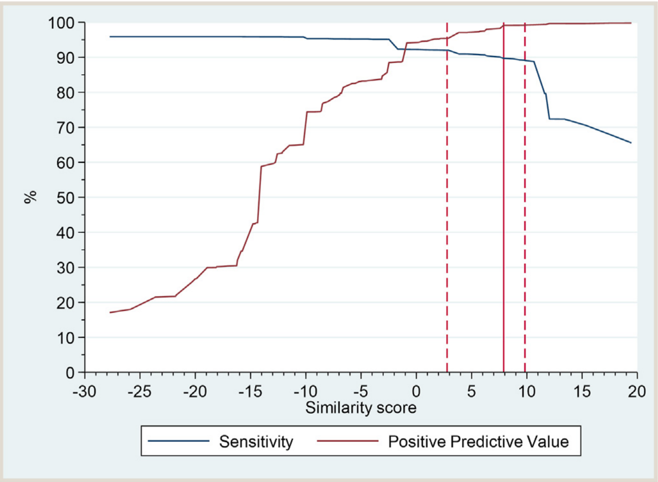


Figure 2. Sensitivity and positive predictive value in the training group and cutoff points for classification.

Table 3. Probabilities *m*, *u* and weights for the attributes in the linkage of the validation group.

Attribute	Prior		Posterior		Weights	
	m	*u*	*m*	*u*	w_{match}	$w_{unmatch}$
Name	0.95	0.20	0.869	0.353	1.30	-2.30
Surname (paternal)	0.95	0.20	0.975	0.481	1.02	-4.35
Surname (maternal)	0.95	0.20	0.936	0.319	1.55	-3.41
Year of birth	0.95	0.20	0.729	0.017	5.39	-1.86
Sex	0.95	0.50	0.945	0.859	0.14	-1.37
State	0.95	0.20	0.998	0.063	4.00	-9.27
Municipality	0.95	0.20	0.824	0.012	6.06	-2.49

Estimated through the EM algorithm applied to the Fellegi-Sunter model of probabilistic linkage.
Prior p probability = 0.051, posterior p probability = 0.017.

we achieved a sensitivity of 90.72% and a positive predictive value of 97.10%.

3.8. Comparison of characteristics between the reference standard and the rest of observations from SAEH excluding live discharges and SEED excluding observations with social security

SAEH death discharges not included in the reference standard had a higher percentage of males, a higher percentage living in states with high or very high marginalization and higher values of year of birth compared to the reference standard, whereas SEED observations without social security and not included in the reference standard showed a slightly higher percentage of males, a higher percentage living in states with high or very high marginalization, and lower values of year of birth compared to the reference standard (Table 4).

4. Discussion

In this paper we present the performance of different forms of blocking and the implementation in Stata of the Fellegi-Sunter model to probabilistically link individual records from SAEH and SEED, two health information systems in Mexico.

Table 4. Distribution of sex, state of residence marginalization and year of birth in the reference standard and the rest of observations from SAEH excluding live discharges and SEED excluding observations with social security.

	SAEH		SEED	
	Reference standard	Other death discharges	Reference standard	Other without social security
Observations	44762	10800	44762	285714
Sex				
Male	52.9	60.3	54.5	57.5
Female	47.1	39.7	45.3	42.2
Other response	0.0	0.0	0.2	0.2
State marginalization				
Very low	13.8	17.5	13.8	12.2
Low	36.4	31.2	35.8	27.8
Middle	20.2	14.2	20.3	16.7
High	20.7	26.4	21.0	28.1
Very high	9.0	10.7	9.1	15.1
Year of birth				
Mean	1955	1961	1954	1949
Std. deviation	20.8	23.2	20.7	22.2
25th percentile	1939	1942	1938	1931
Median	1953	1958	1952	1944
75th percentile	1968	1977	1967	1963

The linkage of records from different systems is a key procedure that allows researchers to perform longitudinal statistical and epidemiological analyses and uncover subpopulations with special health conditions and comorbidities [2, 3, 4]. For instance, record linkage has been recently used to compute heritability estimates for 500 disease phenotypes. In this application, emergency contact data was mined, and the identified relationships were shown to be consistent with genetically derived relatedness [20].

However, record linkage implementation presents technical challenges, such as handling large databases and the presence of variations and errors in capturing ID information. One of the main difficulties in record linking is the large number of possible comparisons between candidate pairs. One of the strategies to reduce the number of such pairs to be compared is the use of filtering techniques (blocking) that limit the comparisons to a subset of the pairs that present similarity in an attribute or group of attributes. We applied trigram-based blocking using the full name, which is more demanding in terms of the number of operations and time consuming compared to other alternatives based on phonetic similarities but had much better performance in terms of coverage (~96%) and complexity reduction (99.9994%).

4.1. Comparison with other methodologies

The linking of records with two sources of information can be seen as a classification problem: separating the pairs that belong to the same entity (e.g., subject, health center, etc.) from the rest of the pairs. The different alternatives of this analysis can be classified into supervised and unsupervised methods [21]. Supervised techniques require knowledge of the true status of a subset of pairs to train the model, after which the results can be applied to additional records. A disadvantage of this type of methodology is that the parameters of the models are specific to the training sample. Even when the attributes are the same, the model parameters could be different in other databases or for other populations. In contrast, unsupervised techniques, such as the one used in this work, do not require a gold standard or the correct state of the pairs for parameter estimation. Although the parameters of the Fellegi-Sunter model are estimated in an unsupervised way, if a reference standard is available in a subset of the data, it is possible to optimize the cutoff point for pair classification.

In this work, we use a reference standard to estimate the classification performance in terms of sensitivity and positive predictive value 1) when a cutoff point is optimized in the training sample and then applied to the validation sample 2) without establishing cutoff points informed by the reference standard. In the first case, we reached a sensitivity of 89.86% and a positive predictive value of 96.18% in the validation group. It should be noted that this type of classification is automatic, as pairs are assigned to matched or unmatched status without performing a clerical review of potential pairs. On the other hand, in the second case, we established an intermediate category of potential pairs and reviewed one by one. This achieved a sensitivity and positive predictive value about 1 percentage point higher compared to automatic classification (90.72% and 97.10%, respectively). This percentage point, in terms of number of pairs, represents a considerable amount when working with large databases. Therefore, depending on the application, the cost associated with reviewing the intermediate category of pairs, might be justified. Another advantage of reviewing a category of potential pairs, is the possibility of simultaneously improving sensitivity and positive predictive value; with an automatic binary classification there is a trade-off between sensitivity and positive predictive value when the classification cutoff point is changed.

Other record linkage methods have been developed apart from the probabilistic approach based on the Fellegi-Sunter model. These include a sorted-neighborhood approach, Bayesian methods, distance-based techniques, methodologies from machine-learning and a double-embedding scheme [14, 22].

4.2. Considerations for choosing the cutoff point

The selection of the similarity cutoff depends on the relative importance given to sensitivity and positive predictive value. Lowering the cutoff will result in higher sensitivity (more true pairs will be captured), but at the cost of classifying more false pairs as correct. On the other hand, if the cutoff point is raised, a higher positive predictive value will be obtained with a decrease in sensitivity. For example, the deterministic classification method, by requiring perfect matching on all attributes, often results in a very high positive predictive value along with an important reduction in sensitivity. If the purpose of the linkage is an epidemiological analysis by estimating statistical models for the populations of interest, a probabilistic linkage may be the most appropriate. On the other hand, if the application requires using information at the individual level, it is desirable to obtain a very high positive predictive value; in this case, a deterministic linkage might be recommended [23]. It is also possible to incorporate misclassification costs for cutoff optimization [24, 25].

4.3. Strengths and limitations

The implementation of the EM algorithm for probabilistic linkage in Stata has the advantage that the maximum number of pairs to compare is only limited by the memory available in the computer. Comparisons of the different forms of blocking allowed us to identify the most efficient alternatives in terms of reducing complexity while simultaneously maintaining high pairs completeness.

The availability of a reference standard allowed us to assess the performance of the linkage algorithm in a subset of data. However, records not included in the reference standard showed a higher percentage of observations from states with high and very high state marginalization, those records may present a higher level of linkage difficulty if data quality is negatively related with state marginalization.

Sometimes only a limited set of records in both data sources contain a standardized identifier; in such situations, the cutoff point could be optimized in these sub-samples to inform the process of parameter estimation and selection of similarity score cutoff points when using all the data. If the estimated probabilities m and u are similar to those obtained with the sub-samples with standardized identifiers, it might be useful to consider the optimized cutoff point together with the analysis of the distribution of the linkage similarity score to select cutoff points.

The importance of the database preparation processes before linking should also be considered. In this paper, we document the performance of the linking method after applying cleanup and standardization rules for text in the Spanish language. As previously mentioned, we obtained superior performance with trigram-based blocking compared to phonetic code-based blocking. It should be noted that these phonetic codes are not specifically designed for Spanish, although their use in practice has been consolidated over time. Edit distance and q-grams can also be applied with other languages or even with alpha-numeric variables.

4.4. Ethical considerations

Because of record linking, characteristics, diagnoses, and some other data about an individual can be identified. In this sense, the link itself has additional ethical implications to those considered in the management of information systems separately. Whenever these methods are applied, it should be disclosed that confidentiality of individual information will be preserved, that research will follow high-quality guidelines, and that the risks are minimal [23]. Additionally, researchers must work in strict adherence to the laws and legal regulations that govern each country [26]. Personal information should exclusively be used for the linking process, and the records should be anonymized once the linked database has been generated.

4.5. Applications and future works

Some extensions to the Fellegi-Sunter linkage model have been documented in the literature [27, 28]. In future works, we propose to apply an extension to the model for fractional comparison results and to apply the model under a Bayesian approach, which will be especially useful when linkages are routinely applied with the same information systems. The construction of a catalog of names and surnames with all their possible spelling variants in Mexican Spanish is also proposed to map these to a generic form before the application of the linking algorithm.

5. Conclusions

The algorithm for the probabilistic linkage of records based on the Fellegi-Sunter methodology achieved a good performance in terms of sensitivity and positive predictive value. It could be used to build administrative cohorts for the epidemiological analysis of populations using the records available in the health information systems.

Declarations

Author contribution statement

Amado D. Quezada-Sánchez: Conceived and designed the experiments; Performed the experiments; analyzed and interpreted the data; Contributed analysis tools and data; Wrote the paper.

Iván Espín-Arellano: Performed the experiments; Analyzed and interpreted the data; Contributed analysis tools and data; Wrote the paper.

Evangelina Morales-Carmona, Diana Molina-Vélez: Analyzed and interpreted the data; Contributed analysis tools and data; Wrote the paper.

Lina Sofía Palacio-Mejía, Juan Eugenio Hernández-Ávila: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed analysis tools and data; Wrote the paper.

Edgar Leonel González-González, Mariana Alvarez-Aceves: Analyzed and interpreted the data; Wrote the paper.

Funding statement

Dr. Mariana Alvarez-Aceves was supported by Consejo Nacional de Ciencia y Tecnología [321008].

Data availability statement

The data that has been used is confidential.

Declaration of interest's statement

The authors declare no competing interests.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e12311.

References

[1] I. Dohoo, W. Martin, H. Stryhn, Methods in Epidemiologic Research, VER Inc., 2012.

[2] K. Smolina, C.J. Wotton, M.J. Goldacre, Risk of dementia in patients hospitalized with type 1 and type 2 diabetes in England, 1998–2011: a retrospective national record linkage cohort study, Diabetologia 58 (5) (2015) 942–950.

[3] O.O. Seminog, M.J. Goldacre, Risk of pneumonia and pneumococcal disease in people with severe mental illness: English record linkage studies, Thorax 68 (2013) 171–176.

[4] K. Hafekost, D. Lawrence, C. O'Leary, C. Bower, M. O'Donnell, J. Semmens, S.R. Zubrick, Maternal alcohol use disorder and subsequent child protection contact: a record-linkage population cohort study, Child Abuse Negl. 72 (2017) 206–214.

[5] I.P. Fellegi, A.B. Sunter, A theory for record linkage, J. Am. Stat. Assoc. 40 (1969) 1183–1210.

[6] M.H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, Clin. Chem. 39 (4) (1993) 561–577.

[7] Registry PlusTM. Link Plus Version 2.0. Centers for Disease Control and Prevention.

[8] StataCorp, Stata Statistical Software: Release 15, StataCorp LLC, College Station, TX, 2017.

[9] Dirección General de Información en Salud, Subsistema Automatizado de Egresos Hospitalarios-SAEH, Secretaría de Salud, México, 2014.

[10] Dirección General Información en Salud, Data from: Subsistema Epidemiológico y Estadístico de Defunciones (SEED), Secretaría de Salud, México, 2014.

[11] D. Kaur, K. Navjot, A review: an efficient review of phonetics algorithms, Int. J. Comput. Sci. Eng. Technol. 4 (5) (2013) 506–508.

[12] P. Christen, Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection, Springer, 2012, pp. 101–127.

[13] G. Papadakis, G. Koutrika, T. Palpanas, W. Nejdl, Meta-blocking: taking entity resolution to the next level, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1946–1960.

[14] N. Adly, Efficient record linkage using a double embedding scheme, in: Conference Proceedings of the 2009 International Conference on Data Mining (DMIN). Las Vegas July 13-16, 2009, pp. 274–281.

[15] Y. Pawitan, In All Likelihood: Statistical Modeling and Inference Using Likelihood, Oxford Science Publications, 2001.

[16] T.N. Herzog, F.J. Scheuren, W.E. Winkler, Estimating the parameters of the fellegi-sunter record linkage model, in: En: Data Quality and Record Linkage Techniques, Springer, 2007, pp. 93–106.

[17] M.A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, J. Am. Stat. Assoc. 84 (406) (1989) 414–420.

[18] Consejo Nacional de Población (CONAPO), Índice de marginación por entidad federativa y municipio, Available from, https://www.gob.mx/conapo/documentos/indice-de-marginacion-2015-284579, 2015. Accessed: October 27, 2022.

[19] Microsoft, SQL Server, 2012.

[20] F.C. Polubriaginof, R. Vanguri, K. Quinnies, G.M. Belbin, A. Yahi, H. Salmasian, T. Lorberbaum, V. Nwankwo, L. Li, M.M. Shervey, P. Glowe, Disease heritability inferred from familial relationships reported in medical records, Cell 173 (7) (2018) 1692–1704.

[21] M. Sriyar, A. Borg, The record linkage package: detecting errors in data, R J. 2 (2) (2010) 61–67.

[22] S.B. Dusetzina, S. Tyree, A.M. Meyer, A. Meyer, L. Green, W.R. Carpenter, Linking Data for Health Services Research: A Framework and Instructional Guide (Prepared by the University of North Carolina at Chapel Hill under Contract No. 290-2010-000141.) AHRQ Publication No.14-EHC033-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2014. Available from: https://www.ncbi.nlm.nih.gov/books/NBK253313/. Accessed: October 28, 2022.

[23] D.E. Clark, Practical Introduction to record linkage for injury research, Inj. Prev. 10 (2004) 186–191.

[24] M. López-Ratón, M.X. Rodriguez-Álvarez, C. Cardaso-Suárez, F. Gude-Sampedro, Optimal Cutpoints: an R package for selecting optimal cutpoints in diagnostic tests, J. Stat. Software 61 (8) (2014) 1–36.

[25] K. Mendoza-Herrera, A.D. Quezada, A. Pedroza-Tobías, C. Hernández-Alcaraz, J. Fromow-Guerra, S. Barquera, A diabetic retinopathy screening tool for low-income adults in Mexico, Prev. Chronic Dis. 14 (2017), 170157.

[26] Departamento de derecho internacional, AG/RES. 2842 (XLIV-O/14) Acceso a la Información Pública y Protección de Datos Personales, Availabe at, https://www.oas.org/es/sla/ddi/proteccion_datos_personales_Resoluciones_Asamblea_General.asp.

[27] S.L. DuVall, R.A. Kerber, A. Thomas, Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators, J. Biomed. Inf. 43 (1) (2010) 24–30.

[28] T.C. Ong, M.V. Mannino, L.M. Schilling, M.G. Kahn, Improving record linkage performance in the presence of missing linkage data, J. Biomed. Inf. 52 (2014) 43–54.