JAMIA Open, 00(0), 2021, 1–5
doi: 10.1093/jamiaopen/ooab005
Application Notes

Application Notes

Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning

Sebastien Cossin [1,2], Serigne Diouf,[1,2] Romain Griffier,[1,2] Philippine Le Barrois d'Orgeval,[1,2] Gayo Diallo,[2] and Vianney Jouhet [1,2]

[1]CHU de Bordeaux, Pôle de Santé Publique, Service d'information Médicale, Informatique et Archivistique Médicales (IAM), Bordeaux F-33000, France and [2]Inserm, Bordeaux Population Health Research Center, Team ERIAS, University of Bordeaux, UMR 1219, Bordeaux F-33000, France

Corresponding Author: Sebastien Cossin, Service D'information Médicale, Informatique et Archivistique Médicales, Place Amélie Raba-Léon, Bordeaux F-33076 France (sebastien.cossin@chu-bordeaux.fr).

Received 15 December 2020; Revised 22 January 2021; Editorial Decision 25 January 2021; Accepted 2 February 2021

ABSTRACT

Introduction: Vital status is of central importance to hospital clinical research. However, hospital information systems record only in-hospital death information. Recently, the French government released a publicly available dataset containing death-certificate data for over 25 million individuals. The objective of this study was to link French death certificates to the Bordeaux University Hospital records to complete the vital status information.
Materials and Methods: Our linkage strategy was composed of a search engine to reduce the number of comparisons and machine-learning algorithms. The overall pipeline was evaluated by assembling a file containing 3,565 in-hospital deaths and 15,000 alive persons.
Results: The recall and precision of our linkage strategy were 97.5% and 99.97% for the upper threshold and 99.4% and 98.9% for the lower threshold, respectively.
Conclusion: In this study, we demonstrated the feasibility of accurately linking hospital records with death certificates using a search engine and machine learning.

Key words: medical record linkage, death certificates, search engine, information storage and retrieval, supervised machine learning

INTRODUCTION

Ascertainment of the vital status of individuals is of central importance to epidemiological studies that monitor mortality as an end point.[1] Death-related information is recorded in the hospital information system only if the death occurred at the hospital; otherwise, such information is often missing.[2]

On December 5, 2019, the French task force for Open Data made publicly available the death certificates established since 1970. This dataset, referred to as the French Death Master File (DMF), enables completion of patient vital status information in the internal registration system of hospitals.

In this context, Bordeaux University Hospital launched an initiative aiming to complete vital status information by identifying extra-hospital deaths with the French DMF dataset. We report the approach used and the results achieved. In particular, we used the information retrieval (IR) approach with machine learning (ML) to perform record linkage.

MATERIALS AND METHODS

Background on Record Linkage

There are two main types of record linkage: deterministic and probabilistic. The former is based on rules and the latter on weights or

scores. The deterministic method consists of performing linking based on exact match agreement, on several matching variables. Records are compared using a set of one or more matching variables (identifiers) that are common to both records, to be compared.[3] The probabilistic linkage involves calculating a score or probability between two records. In the well-known Fellegi-Sunter method, each variable is assigned a weight derived from two conditional probabilities: the *m*-probability (probability that an identifier agrees that given records belong to the same individual) and the *u*-probability (probability that an identifier agrees that given records belong to different individuals).[4] For each pair of records, the overall match weight is derived by calculating the ratio $\log_2(m/u)$ for each variable, and summing across all variables. Pairs with high scores have higher probabilities of being true matches.

In probabilistic approaches, two thresholds are typically chosen. Pairs with weights above the upper threshold are classified as links; pairs with weights below the lower threshold are classified as non-links; and those in the middle are inspected further (e.g., by clerical review).[5,6] To reduce the number of comparisons between data sources, blocking strategies can determine which records are considered potential matches.[5,7]

ML approaches have also been used for record linkage.[8] The Fellegi-Sunter model incorporates an independence assumption of each variable whereas ML algorithms can handle highly correlated variables. ML models can also take advantage of more high-level features, such as string distances and time differences.[9] Supervised ML models require training data for which the matching status is known.[10] In general this process is time-consuming and impractical for large administrative data sources.[11]

Common evaluation metrics to evaluate record linkage are recall (*aka* sensitivity) and precision (*aka* positive predictive value). Recall is the number of true matches correctly identified by an algorithm over the total number of true matches in the gold standard. Precision is the number of true matches correctly identified over the total number of matches outputted.

Description of the Dataset

The Bordeaux hospital registration system contains administrative information on 2.2 million patients since 2005 and a total of 54,892 in-hospital deaths.

The French DMF contains mortality data for over 25 million deceased individuals since 1970. Approximately 8.8 million death certificates were produced since 2005. Table 1 shows the common attributes of the two data sources.

The hospital registers two last names, the birth name and the used surname, whereas death certificates contain only one last name. There are also differences concerning the first name. French people can have one or more first names; one is used in daily life and the others solely for official documents.[12] The hospital registers only the first name used in daily life whereas death certificates contain all of the first names.

Overview of the Record-Linkage Strategy

Figure 1 presents the pipeline developed in this study.

Our approach was to combine a blocking strategy based on the Elasticsearch™ search engine with a ML strategy to classify match candidates. To train the ML algorithms, we applied a deterministic approach to the 2005 to 2018 data to create automatically a gold standard. The overall pipeline was evaluated in 2019. The steps in the process are detailed below. Additional technical details and a step by step example are provided in SupplementalFile 1.

Step 1: Blocking strategy

Elasticsearch™ is a distributed, document-oriented search engine, built on top of Apache Lucene[1]. Elasticsearch™ was chosen for its scalability, its speed performance and its scoring metric based on the popular TF-IDF (term frequency–inverse document frequency). The French DMF dataset was indexed in Elasticsearch™ version 7.6.1. A query was sent to Elasticsearch™ to find candidate death certificates for a given hospital record. The blocking strategy consisted in selecting only the first N search results and submitting them to the pairwise comparison in step 2. The number of Elasticsearch™ results, N, was set with the gold standard created in step 2.

Step 2: Machine learning

Gold Standard

The goal of the ML step is to compare one hospital record with one candidate death certificate and to output a match probability. To train a ML algorithm, a gold standard containing both true and false matches must be created. We used a deterministic approach to find true matches. To identify them correctly, we leveraged the fact that in-hospital deaths have two more matching variables: the date and the department of death. Seven common fields were available: last name, first name, date of birth, zip code, gender, date of death, and department of death. Keeping the date and department of death as blocking variables, the matching strategy was relaxed to capture differences in one of the remaining fields. Clerical review when only one variable mismatched made us confident that differences were due to errors in one of the two data sources and no records were falsely linked.

These pairs of true matches were next used to identify incorrect matches. The difficulty in building a gold standard is to find examples in the gray zone,[13]where pairs of records do not have an exact match agreement, for the algorithm to learn the importance of each variable. For each true match, a search query was sent to Elasticsearch™ and the incorrect match with the highest score was retained. By doing so, we generated a balanced dataset, that is, there were equal numbers of samples from the positive and negative classes.

The optimal number of results *N* in step 1 was estimated by recording the rank of the correct death certificate for each true match.

Features

The gold standard was converted to a feature matrix to train the ML algorithms. Table 1 lists the features created for each attribute. Several supervised MLmodels were compared. A random Forest model and a fully connected neural network obtained the best performance and were kept to predict whether a pair of records were a true or a false match.

Step 3: Thresholds

An upper and a lower threshold were set at the end of the pipeline evaluation to maximize precision and recall, respectively.

Evaluation of the Pipeline

The overall pipeline was evaluated by assembling a file in which the death status of the individuals included was known beforehand. In-hospital deaths in 2019 were included to evaluate the sensitivity.

1 https://lucene.apache.org/

Table 1. Shared attributes of the hospital and French DMF datasets. Missing values (percentages) are indicated in parentheses. ML features were calculated by comparing the attribute values of hospital records and those of death certificates. String distance methods based on edits (Damerau-Levenshtein, Hamming, Levenshtein, optimal string alignment), qgrams (q-gram, cosine, Jaccard distance), phonetics (soundex), and heuristic metrics (Jaro, Jaro-Winkler) were calculated for the first name and the last name

Hospital $N = 2.2$M (percent missing)	French DMF $N = 8.8$M (percent missing)	Methods of comparison (features)
Last names (0%)	Last name (0%)	String distances
First Name (0%)	First names (0%)	String distances
Birth date (0%)	Birth date (0.66%)	Equal or not for the date, year, month and day
Gender (0%)	Gender (0%)	Equal or not
Birth location (39%)	Birth location (0.54%)	Equal or not
Birth country (6.9%)	Birth country (0.02%)	Equal or not
Last registered patient address (1.7%)	Death location (0.11%)	Equal or not for department and region of death
Last visit date (0%)	Date of death (0%)	Time difference in days

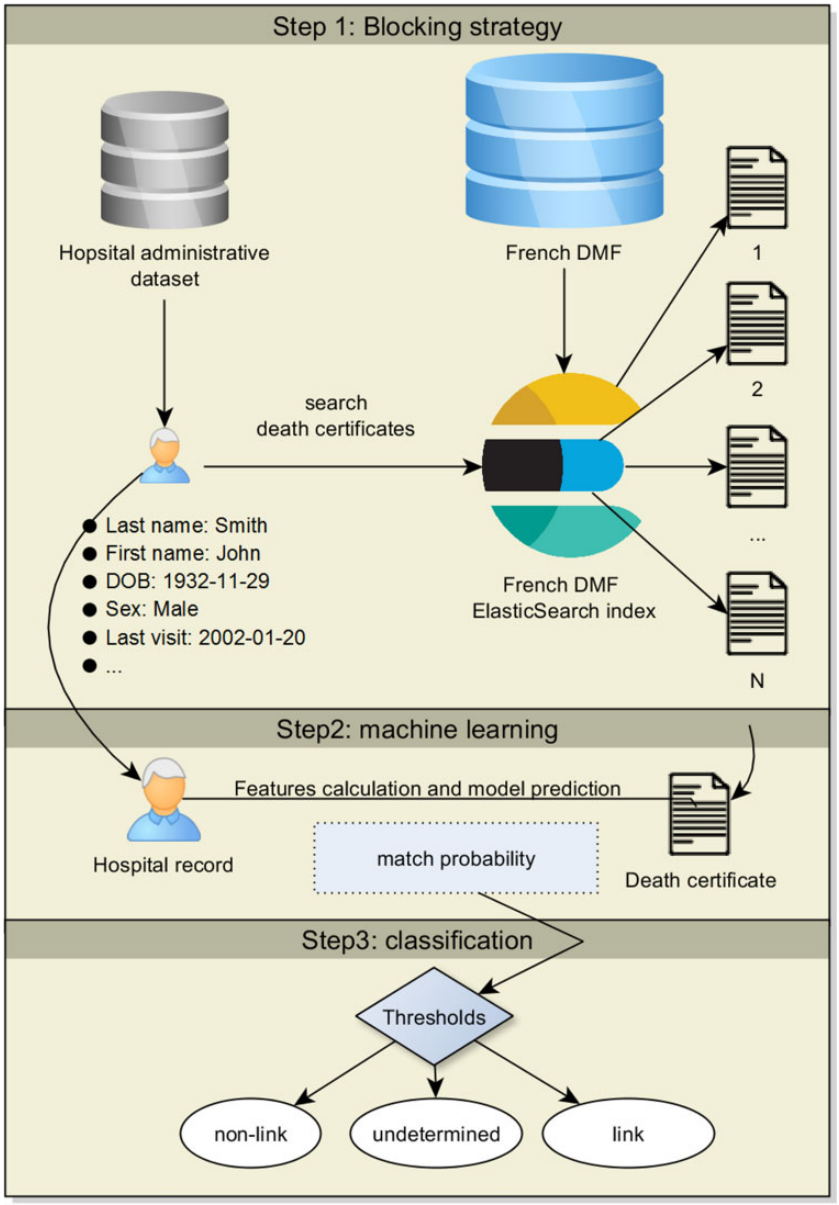

Figure 1. Overview of the record-linkage strategy. In the first step, a query is sent to Elasticsearch for each hospital record to retrieve a limited number of N candidate death certificates. In the second step, ML models predict the match probability of a hospital record and a candidate certificate. In the third step, the pair is classified as a nonlink, undetermined, or link according to the upper and lower thresholds.

The date of death was used to validate automatically if two linked records were true or false positives. If there was a discrepancy between the two dates of death, a manual review was performed to check whether it was a classification error or a data source error. Specificity was evaluated by randomly selecting 15,000 patients who attended the hospital in 2020, that is, patients known not to have died in 2019.

RESULTS

Between 2005 and 2018, 44,689 in-hospital deaths were registered in the Bordeaux University Hospital information system. With the deterministic approach 44,127 (98.7%) deaths were successfully linked to a death certificate. Of the 44,127 true matches described above, 98.9% were ranked first by Elasticsearch™ and only 0.2% were ranked after the 10th result. A decrease in sensitivity of 0.2% by the search strategy was deemed acceptable and the value of N was set to 10.

Overall Pipeline Evaluation

All the 3,565 in-hospital deaths in 2019 were successfully linked to a death certificate, demonstrating the completeness of the French DMF. Based on the model's predictions, the upper threshold was set at 0.95 to maximize precision and the lower threshold at 0.4 to maximize recall. Recall and precision were 97.5% and 99.97%, respectively, for the upper threshold and 99.4% and 98.9%, respectively, for the lower threshold.

Estimation of Outpatient Deaths

By applying the overall pipeline to the 2.2 million hospital records, 207,507 records were linked to a death certificate with a probability over the upper threshold and 29,152 had probabilities between the two thresholds, thus requiring manual validation.

In comparison, an exact match query in a relational database on the last name, first name, date of birth, and gender matched only 200,824 pairs of records. In terms of performance, it took approximately 4 minutes and 30 seconds to search and classify 1,000 hospital records on a virtual machine with Intel Core i7-8650U @1.90GH x 8 CPUs without parallelization. The source code, including the feature matrix and the trained models, is available[2].

DISCUSSION

In this study, we demonstrated the feasibility of accurately linking French hospital records with an open dataset to complete the vital-status information.

The mortality information in the Bordeaux University Hospital information system was found to be incomplete. This is unsurprising because only in-patient deaths were recorded. We found a 1:4 ratio of in-patient and out-patient deaths.

We proposed a reproducible pipeline based on a search engine and ML. Elasticsearch™ is horizontally scalable and fast. The blocking strategy based on its relevancy score gave satisfactory results—a death certificate, if it existed, appeared in 99.8% of the cases in the top 10 results. A recent Brazilian study also applied a blocking strategy with Apache Lucene to reduce the number of comparisons in the subsequent steps.[14]

We explored the utility of leveraging existing in-hospital deaths to build automatically a gold standard. A gold standard is mandatory for supervised learning but manual review of records is time-consuming, expensive, and can be error prone.[7]Matching on death date was also adopted by Newman *et al*. to check the accuracy of other variables.[15] Combined with other variables, death date is highly discriminative and the probability of falsely linking a pair of records is low.

Limitations

This study has various limitations. First, the gold standard was created automatically by a deterministic approach and the maximum difference between two records was limited to one field only. A better-quality gold standard could have been obtained by using a common identifier between the two data sources, such as the social security number. This number is not publicly available and requires authorizations that are difficult to obtain.[16]

Second, we didn't compare the performance of our approach with other linkage tools. However, not every tool was designed to cope with huge datasets which complicates comparisons.[14]

Third, traditional counterparts of given names in various languages were not taken into account, which decreased recall. Because Elasticsearch™ offers the functionality of adding synonyms, it would be feasible to broaden the search queries by providing a resource of equivalent first names.

Fourth, this was a single-center study. The excellent concordance of in-hospital deaths with the French DMF dataset may not generalize to other French regions.

CONCLUSION

The record linkage pipeline based on the search engine Elasticsearch™ and an ML strategy provides satisfactory results and could be further improved. The mortality information in our hospital information system was incomplete because only in-hospital deaths were recorded. French mortality data can enable French hospitals to add vital-status information to their records systems.

FUNDING STATEMENT

This study is part of the DRUGS-SAFER research program, funded by the French Medicines Agency (Agence Nationale de Sécurité du Médicament et des Produits de Santé, ANSM). This publication represents the views of the authors and does not necessarily represent the opinion of the French Medicines Agency.

AUTHOR CONTRIBUTIONS

SD and SC developed the pipeline and conducted the analyses. SC wrote the manuscript. RG and PLBO contributed to the evaluation of the pipeline. GD and VJ supervised and were actively involved in writing the design of the study, analysis and interpretation of the results. All authors reviewed, provided input, and accepted the submitted version. All authors duly contributed to this manuscript and they all certify that they sufficiently participated and are responsible for this work.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

2 https://github.com/scossin/record_linkage_insee

DATA AVAILABILITY

Training and test set data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.1ns1rn8sj

ACKNOWLEDGMENTS

The authors wish to thank Etalab and INSEE for publishing French death records as open data.

REFERENCES

1. Curb JD, Ford CE, Pressel S, *et al*. Ascertainment of vital status through the National Death Index and the Social Security Administration. *Am J Epidemiol*. 1985; 121(5):754–766.
2. Jones B, Vawdrey DK. Measuring mortality information in clinical data warehouses. *AMIA Jt Summits Transl Sci Proc*. 2015; 2015:450–455.
3. Doidge JC, Harron K. Demystifying probabilistic linkage.*Int J Popul Data Sci*. 2018; 3(1):410.
4. Fellegi IP, Sunter AB. A theory for record linkage. *JAm Stat Assoc*. 1969; 64(328):1183–1210.
5. Harron K, Dibben C, Boyd J, *et al*. Challenges in administrative data linkage for research. *Big DataSoc*. 2017; 4(2):205395171774567.
6. Grannis SJ, Overhage JM, Hui S, *et al*. Analysis of a probabilistic record linkage technique without human review. *AMIA Annu Symp Proc*. 2003; 2003:259–263.
7. Gu L, Baxter R, Vickers D, *et al*. Record Linkage: Current Practice and Future Directions. *Technical Report 03/83,* CSIRO Mathematical and Information Sciences. Canberra, Australia; 2003.
8. Goldstein H, Harron K, Cortina-Borja M. A scaling approach to record linkage.*Statist Med*. 2017; 36(16):2514–2521.
9. Wilson DR. Beyond probabilistic record linkage: using neural networks and complex features to improve genealogical record linkage. In: The 2011 International Joint Conference on Neural Networks; 2011: 9–14; San Jose, CA, USA.
10. Hejblum BP, Weber GM, Liao KP, *et al*. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Sci Data*2019; 6(1):180298.
11. Pita R, Mendonça E, Reis S, *et al*. A machine learning trainable model to assess the accuracy of probabilistic record linkageIn: Bellatreche L, Chakravarthy S, eds. *Big Data Analytics and Knowledge Discovery*. Cham:-Springer International Publishing; 2017:214–227. doi:10.1007/978-3-319-64283-3_16
12. French name. Wikipedia. 2020. https://en.wikipedia.org/w/index.php?title=French_name&oldid=960384153 (accessed 31 Aug 2020).
13. Capuani L, Bierrenbach AL, Abreu F, *et al*. Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database.*Cad Saúde Pública*2014; 30(8):1623–1632.
14. Barbosa GCG, Ali MS, Araujo B, *et al*.. CIDACS-RL: a novel indexing search and scoring-based record linkage system for huge datasets with high accuracy and scalability. *BMC Med Inform Decis Mak* 2020; 20 (1): 289.
15. Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. *J Am Med Inform Assoc*1997; 4(3):233–237.
16. Guesdon M, Benzenine E, Gadouche K, *et al*. Securizing data linkage in French public statistics. *BMC Med Inform Decis Mak*2016; 16(1): 129.