A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-571

Retrieving the Vital Status of Patients with Cancer Using Online Obituaries

Emmanuelle SYLVESTRE [abe,1], Guillaume BOUZILLE [abcd], Mathias BRETON [e], Marc CUGGIA [abcd], Boris CAMPILLO-GIMENEZ [abe]
[a] *INSERM, U1099, Rennes, F-35000, France*
[b] *Université de Rennes 1, LTSI, Rennes, F-35000, France*
[c] *CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France*
[d] *CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France*
[e] *Centre Eugène Marquis, Rennes, F-35000, France*

Abstract. The aim of this study was to develop a methodology to link mortality data from Internet sources with administrative data from electronic health records and to assess the performance of different record linkage methods. We extracted the electronic health records of all adult patients hospitalized at Rennes comprehensive cancer center between January 1, 2010 and December 31, 2015 and separated them in two groups (training and test set). We also extracted all available online obituaries from the most exhaustive French funeral home website using web scraping techniques. We used and evaluated three different algorithms (deterministic, approximate deterministic and probabilistic) to link the patients' records with online obituaries. We optimized the algorithms using the training set and then evaluated them in the test set. The overall precision was between 98 and 100%. The three classification algorithms performed better for men than women. The probabilistic classification decreased the number of manual reviews, but slightly increased the number of false negatives. To address the problem of long delays in the publication or sharing of mortality data, online obituary data could be considered for real-time surveillance of mortality in patients with cancer because they are easily available and time-efficient.

Keywords. Digital epidemiology, web mining, vital status, Medical Record Linkage

Introduction

Cancer is among the leading causes of death worldwide and cancer-related statistics are closely monitored in many countries. Several key statistics such as incidence, prevalence, mortality, survival and type of cancer, are used to assess the impact of cancer in the general population.[1] To collect these metrics, two types of data sources are generally used: cancer registries and mortality databases. Population-based cancer registries collect data on all new cases of cancer that occur in a well-defined population[2] (regional or nationwide). Mortality data come from different sources such as civil registration systems[3] or reimbursement claims. [4]. However, given the sensitive nature of these data, these databases are not available for public use or for

[1] Corresponding Author, Faculté de médecine, Université Rennes 1, 2 Avenue du Professeur Léon Bernard 35043 Rennes Cedex 9, France; E-mail: emmasyl@gmail.com.

routine use, and investigators must obtain specific approval to access them for clinical research purposes. Moreover, cancer surveillance statistics are usually national rather than local, and are available with a few years of delay because in most cases, they are manually reported and need to be reviewed before publication. Mining web content, such as web queries or social media, allows near real-time digital surveillance. Studies using Internet based-data identified trends that are comparable to those obtained using established indicator-based surveillance methods.[5] However, most studies based on web data have focused on reproducing mortality trends,[6] or on comparing patient-reported outcomes with traditional health records.[7].However, cancer studies involving death rates and/or survival rates use traditional sources of data such as national cancer registries[8,9] or reimbursement claims to link their medical record to mortality statistics. Nevertheless, it has been shown that obituaries might provide reasonably reliable mortality data that could be used to generate study hypotheses for future epidemiological studies.[10] Therefore, we hypothesized that it should be possible to match medical records to death announcements to calculate mortality data, although these data sources are incomplete and do not entirely overlap. The aim of this study was to develop a methodology to link mortality data from online obituaries at funeral home websites to administrative data from electronic health records and to assess the performance of different record linkage methods.

1. Material and Methods

1.1. Data Sources

We used two data sources: the patients' electronic health records (EHR) from the Rennes Comprehensive Cancer Center (Centre Eugène Marquis, CEM) database, and obituaries from French funeral home websites. We extracted ten fields from the CEM database that were necessary for the linkage: First Name, Middle Name, Maiden Name, Last Name, Date of Birth, Birthplace, Last known address, Zip Code, Sex, and Hospitalization Dates. For patient with multiple hospitalizations, we only kept the most recent stay, and considered it as the last known date when the patient was alive**.** We then randomly divided the patients included in the CEM database in two datasets: training set (82% of all patients) and test set (18% of all patients). Obituaries were extracted from the most exhaustive French funeral home website: www.avisdedeces.net using an online third-party service for web scraping[11] (www.import.io). For confidentiality and safety reasons, we downloaded the scraping result files and matched them locally with the EHR data. The following data were extracted from the obituaries: First Name, Middle Name, Maiden Name, Last Name, Date of Death, Age of Death, City and Zip Code.

1.2. Linking model

Our approach included four steps (Figure 1): i) data cleaning and standardization; ii) indexing and creation of candidate record pairs; iii) comparison of candidate record pairs; and iv) classification of the candidate record pairs into matches and non matches. The first step was carried out to improve data quality and transform both databases in a common standardized form. To match two databases that contain *n* and *m* records, *m* x *n* comparisons are required. As both databases were quite large, the aim of the indexing step was to reduce the number of record pairs that were compared and accelerate the linkage process. For this, we used a block search and applied the following blocking key: Sex (male/female) concatenated with the first three letters of the Last Name. Record comparison was based on similarity. For each candidate record pair generated in Step 2, we compared several record fields and used different approximate string comparison functions, depending on the field type. For each field, a similarity score ranging from 0 (total dissimilarity) to 1 (exact match) was generated. In-between scores corresponded to some degree of similarity. If a field was empty, its similarity score was arbitrarily noted as “0”. In the final step, we assessed whether each candidate record pair belonged to the same person and classified the record pairs into matches and non-matches, based on their overall similarity score. As women, but not men, can have a maiden name (and therefore a supplementary record field), we separated our record pairs by sex before their classification. We optimized our matching approach in the training set, and then evaluated them using a random sample from the test set. We reviewed manually the classification algorithm with one medical expert. Uncertain pairs were reviewed again by two other investigators

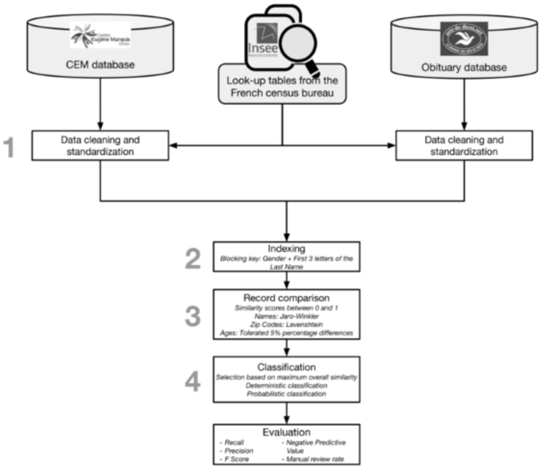
CEM database
Insee
Look-up tables from the French census bureau
Obituary database
1
Data cleaning and standardization
Data cleaning and standardization
2
Indexing
Blocking key: Gender + First 3 letters of the Last Name
3
Record comparison
Similarity scores between 0 and 1
Names: Jaro-Winkler
Zip Codes: Levenshtein
Ages: Tolerated 5% percentage differences
4
Classification
Selection based on maximum overall similarity
Deterministic classification
Probabilistic classification
Evaluation
- Recall
- Precision
- F Score
- Negative Predictive Value
- Manual review rate

Figure 1. Study design.

1.3. Ethics approval

This study was approved by National Commission for Informatics and Freedom (CNIL, Commission Nationale Informatique et Libertés) who agreed to give a waiver of consent for the study. However patients were collectively informed of the study on the hospital website and in the patients’ book.

2. Results

The CEM database included 74,257 patients hospitalized between January 1, 2010 and December 31, 2015. The training set had 60,672 patients (82% of the CEM database) and the test set 13,585 patients (18% of the CEM database). The obituary database included 1,885,816 subjects deceased between January 1, 2010 and December 31, 2015.

The overall precision was between 98 and 100%. For women, the three algorithms displayed the same Positive Predictive Value (PPV) (100%). For men, the exact deterministic algorithm showed the best PPV (100%). Overall, the three classification algorithms performed better for men than for women. The Negative Predictive Value (NPV) was around 99% regardless of the sex. In both cases, the probabilistic classification decreased the number of manual reviews, but slightly increased the number of false negatives (Table 1).

Table 1. Algorithm Performance

	Women			Men		
	Exact deterministic	**Approximate deterministic**	**Probabilistic**	**Exact deterministic**	**Approximate deterministic**	**Probabilistic**
PPV	100 %	100 %	100 %	100%	98%	98%
Manual review		34 %	9 %		9%	4%
NPV		100%	99 %		99%	99%

PPV: Positive Predictive Value. NPV: Negative Predictive Value.

3. Discussion

All classification approaches showed very high precision (between 98 and 100%) and NPV (around 99%), which are highly sought in identity matching. The approximate deterministic and probabilistic approaches outperformed the exact deterministic method, especially for men (who had fewer classification fields than women). The probabilistic method did not improve precision and recall for the single-threshold classification, but greatly reduced the number of manual reviews in the dual threshold situation, especially for women (from 34.3% to 8.9%).

Our study has several limitations. First, obituaries came from a unique digital source (funeral homes). However, it was rather exhaustive because we estimated that it covered more than half of the absolute number of deaths nation-wide and 86% in Brittany where most of our patients lived (data not shown). Second, we chose to adjust our thresholds to obtain the highest possible specificity. Previous studies [12] showed that for record linkage, it is the most pertinent choice, although it automatically increases the false negative results. Women with a missing name field (missing maiden name or married name) were particularly affected by this methodological choice. As we gave a “0” similarity score also to empty fields, we did not differentiate between missing and totally dissimilar data, and these women were classified as non-matches, although they were matches for all the other fields. Finally, we only used a limited number of methods for record linkage. Most record linkage studies use several combinations of blocking variables.[13] We might have missed potential record pairs because we only had one combination for our blocking variable, but we were limited by the available information: inferred sex and last name were the most reliable attributes, while other combinations were risky.

Most studies using mortality data from external sources have tried to reproduce mortality trends or used more conventional sources for record linkage.[14] This study moves one step further by linking obituary data to real patient records. Despite the limitations and the relatively low number of variables, we still managed to match accurately a good number of records. Furthermore, our source is almost real-time, whereas most mortality data from conventional sources are only available on an annual basis.[1] Therefore, this method can be applied for real-time monitoring of death rates of patients with cancer without the usual delay of the traditional methods.

This system cannot make the same claims of completeness as official mortality registries, but it could become a supplemental and reliable source of information for routine vital status surveillance.

4. Conclusion

Our study demonstrated that online obituary data could be considered for real-time surveillance of mortality in patients with cancer. This information is easily available and time-efficient and addresses the problem of long delays in the publication or sharing of mortality data. Next, we would like to compare our results to the mortality data from the National Directory of Identification of Natural Persons (the French gold standard) to confirm the accuracy of our findings.

References

[1] Jemal A, Ward EM, Johnson CJ, *et al.* Annual Report to the Nation on the Status of Cancer, 1975-2014, Featuring Survival. *J Natl Cancer Inst* 2017;**109**. doi:10.1093/jnci/djx030

[2] DosSantos Silva I. *Cancer epidemiology: principles and methods*. Lyon: IARC 1999.

[3] Mahapatra P, Shibuya K, Lopez AD, *et al.* Civil registration systems and vital statistics: successes and missed opportunities. *Lancet Lond Engl* 2007;**370**:1653–63. doi:10.1016/S0140-6736(07)61308-7

[4] Moulis G, Lapeyre-Mestre M, Palmaro A, *et al.* French health insurance databases: What interest for medical research? *Rev Med Interne* 2015;**36**:411–7. doi:10.1016/j.revmed.2014.11.009

[5] Velasco E, Agheneza T, Denecke K, *et al.* Social media and internet-based data in global systems for public health surveillance: a systematic review. *Milbank Q* 2014;**92**:7–33. doi:10.1111/1468-0009.12038

[6] Tourassi G, Yoon H-J, Xu S, *et al.* The utility of web mining for epidemiological research: studying the association between parity and cancer risk. *J Am Med Inform Assoc JAMIA* Published Online First: 27 November 2015. doi:10.1093/jamia/ocv141

[7] Eichler GS, Cochin E, Han J, *et al.* Exploring Concordance of Patient-Reported Information on PatientsLikeMe and Medical Claims Data at the Patient Level. *J Med Internet Res* 2016;**18**:e110. doi:10.2196/jmir.5130

[8] Peres SV, Latorre M do RD de O, Tanaka LF, *et al.* Quality and completeness improvement of the Population-based Cancer Registry of São Paulo: linkage technique use. *Rev Bras Epidemiol Braz J Epidemiol* 2016;**19**:753–65. doi:10.1590/1980-5497201600040006

[9] Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. *J Am Med Inform Assoc JAMIA* 1997;**4**:233–7.

[10] Mpinga EK, Delley V, Jeannot E, *et al.* Testing an Unconventional Mortality Information Source in the Canton of Geneva Switzerland. *Glob J Health Sci* 2014;**6**:1.

[11] Glez-Peña D, Lourenço A, López-Fernández H, *et al.* Web scraping technologies in an API world. *Brief Bioinform* 2014;**15**:788–97. doi:10.1093/bib/bbt026

[12] Fonseca MGP, Coeli CM, Lucena F de F de A, *et al.* Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database. *Cad Saúde Pública* 2010;**26**:1431–8. doi:10.1590/S0102-311X2010000700022

[13] Winkler WE, Yancey WE, Porter EH. Fast record linkage of very large files in support of decennial and administrative records projects. In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 2010. 2120–2130.

[14] Granbichler CA, Oberaigner W, Kuchukhidze G, *et al.* Decrease in mortality of adult epilepsy patients since 1980: lessons learned from a hospital-based cohort. *Eur J Neurol* 2017;**24**:667–72. doi:10.1111/ene.13267