ISDS 2019 Conference Abstracts

Performance of machine learning method to classify free-text medical causes of death

Yasmine Baghdadi[1], Alix Bourrée[1, 2], Aude Robert[3], Grégoire Rey[3], Anne Gallay[1], Pierre Zweigenbaum[2], Cyril Grouin[2], Anne Fouillet[1]

[1]Santé publique France, Saint-Maurice, France, [2]CNRS-LIMSI, Orsay, France, [3]Inserm-CépiDc, Le Kremlin-Bicêtre, France

Objective

This study aims to implement and evaluate two automatic classification methods of free-text medical causes of death into Mortality Syndromic Groups (MSGs) in order to be used for reactive mortality surveillance.

Introduction

Mortality is an indicator of the severity of the impact of an event on the population. In France mortality surveillance is part of the syndromic surveillance system SurSaUD and is carried out by Santé publique France, the French public health agency. The set-up of an Electronic Death Registration System (EDRS) in 2007 enabled to receive in real-time medical causes of death in free-text format. This data source was considered as reactive and valuable to implement a reactive mortality surveillance system using medical causes of death [1].

The reactive mortality surveillance system is based on the monitoring of Mortality Syndromic Groups (MSGs). An MSG is defined as a cluster of medical causes of death (pathologies, syndromes, symptoms) that meet the objectives of early detection and impact assessment of events [2].

Since causes of death are entered in free-text format, their automatic classifications into MSGs require the use of natural language processing methods. We observe a constant increase in the use of these methods to classify medical information and for health surveillance over the last two decades [3].

Methods

Data consisted of the medical part of electronic death certificates received in routine by Santé publique France from 2012 to 2016. We split the dataset into training and test sets. Among each set, a subset of certificates was selected by a random sampling without replacement. Two annotators manually assigned MSGs to each death certificates in all subsets. Discordances were discussed and corrected if necessary. The agreement rate between the two annotators was 0.90 on the test set. Final annotated subsets represent the ground truth against which the methods tested were evaluated. The final evaluation was performed on the test set of 1,000 death certificates while the classifiers were trained on 3500 death certificates.

Two classification methods were implemented: a rule-based method and a supervised machine learning method. The rule-based method was based on four processing steps: applying standardization rules, splitting of medical expression using delimiters, spelling correction and dictionary projection. The supervised machine learning method was set up using a linear Support Vector Machine (SVM) classifier. We trained a multi-label classifier using the one- versus-all strategy. We implemented two models: one based on surface features (SVM model) and the other, a hybrid model, combining surface features and features obtained by the rule-based method. Surface features were bags-of-word unigrams and bigrams and of character trigrams.

The rule-based method and the two supervised machine learning models were evaluated using the three evaluation measures: precision (Positive Predictive Value), recall (Sensitivity) and F-measure (P/R/Fm). The study focused on the classification performance of MSGs defined for the reactive detection of outbreaks and are composed of unspecific or acute pathologies, or general symptoms (related to pain, fever, cognitive disorder…). Only the 40 MSGs mentioned at least 3 times in the test set were considered in this study, they belonged to 13 topics (Respiratory conditions, Cardio and cerebrovascular conditions, Infectious diseases, Digestive conditions…).

Results

With the rule-based method, among the 40 MSGs, 24 obtained a P/R/Fm over 0.90. They belonged mainly to the topics *Cardio and Cerebrovascular conditions* (5 MSGs), *Respiratory conditions* (6), *and General symptoms* (5). Four MSGs obtained P/R/Fm below 0.85 belonging to the topics *Infectious conditions* (2), *Blood condition* (1) and *Unspecified causes of death* (1).

The hybrid model obtained P/R/Fm over 0.90 for 25 MSGs. Among them, 21 were the same as the rule-base method. Performance of the rule-based method and the hybrid model were over 0.95 for the same 13 MSGs. The hybrid model obtained P/R/Fm below 0.85 for 4 MSGs also belonging to the same topics as those of the rule-based method.

The SVM model had lower classification performance than the two other models.

Conclusions

For syndromic mortality surveillance both precision and recall are important for all MSGs. Indeed to meet the objective of a reactive detection of events, high precision is needed to limit false alarms. To measure the impact of an event, the surveillance system should have high recall, to avoid an underestimation of this impact. This is especially true for rarer diseases. The results showed that the rule-based method and the hybrid model are the most effective to classify causes of death into MSGs. For some MSGs with less than 5 mentions in the test set (7%), these results must be qualified. Also, to improve classification performance for MSGs with performance below 0.90, and to confirm these results further analysis must be conducted. The results suggest the relevance of these methods to set up a reactive mortality surveillance system for detection and alert based on free-text causes of death. Such a system will provide useful information to health authorities regarding the causes of death during an event, helping them to adapt counter and prevention measures.

Acknowledgement

The authors thank Thomas Lavergne for his valuable advice during this study. The authors also thank the IT department of LIMSI for computer support and help in computer setup.

References

1. Lassalle M, Caserio-Schönemann C, Gallay A, Rey G, Fouillet A. 2017. Pertinence of electronic death certificates for real-time surveillance and alert, France, 2012–2014. *Public Health*. 143, 85-93. PubMed https://doi.org/10.1016/j.puhe.2016.10.029

2. Baghdadi Y, Gallay A, Caserio-Schönemann C, Thiam M-M, Fouillet A. 2018. Towards real-time mortality surveillance by medical causes of death: A strategy of analysis for alert. *Rev Epidemiol Sante Publique*. 66, S402. https://doi.org/10.1016/j.respe.2018.05.453

3. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, et al. 2018. Clinical information extraction applications: A literature review. *J Biomed Inform*. 77, 34-49. PubMed https://doi.org/10.1016/j.jbi.2017.11.011