ISDS 2019 Conference Abstracts

A strategy of analysis of free-text E-death certificates using machine learning

Yasmine Baghdadi, Anne Gallay, Céline Caserio-Schönemann, Marie-Michèle Thiam, Anne Fouillet

Santé publique France, Saint-Maurice, France

Objective

The aim of this study is to present the syndromic groups that will be routinely monitored for the reactive mortality surveillance based on free-text medical causes of death.

Introduction

In 2004, Santé publique France, the French Public Health Agency set up a reactive all-cause mortality surveillance based on the administrative part of the death certificate, in the final objectives 1/ to detect unexpected or usual variations in mortality and 2/ to provide a first evaluation of mortality impact of events. In 2007, an Electronic Death Registration System (EDRS) was implemented, enabling electronic transmission of the medical causes of death to the agency in real- time. To date, 12% of the mortality is registered electronically. A pilot study demonstrated that these data were valuable for a reactive mortality surveillance system based on causes of death [1].

A strategy has thus been developed for the analysis in routine of the medical causes of death with the objectives of early detection of expected and unexpected outbreaks and reactive evaluation of their impact. This system will allow approaching the cause accountability when an excess death will be observed.

Methods

Mortality syndromic groups (MSG) were defined as clusters of medical causes of death (pathologies, syndromes or symptoms) that meet the objectives of the surveillance system. The causes of death are available reactively in free-text (words, terms, expressions) and with a delay of 6 to 24 months in ICD10 codes format.

We explored multiple biomedical classifications such as the Mesh, SNOMED, UMLS or ICD10 to learn from their various ways to classify diseases. Based on ICD10, we defined MSGs by a list of ICD10 codes, each codes belonging to a unique MSGs. Each MSG definition was then discussed in working group including medical and epidemiological experts.

Additionally, we used a dictionary (provided by the Epidemiology Center on Causes of Death (Inserm-CépiDc)) of each term/expression found in the death certificates since the early 2000 to enrich variety of expression of each MSG. We classified causes of death into MSGs from E-death certificates from 2012 to 2016: 1/ using the ICD10 codes assigned by Inserm-CépiDc based on rules defined by WHO in order to produce the national mortality statistics and 2/ using a linear Support Vector Machine (SVM) method to classify free-text causes of death. Then we compared the fluctuations of the weekly numbers of each MSG built by using both classification methods (ICD10 codes and the SVM classification) [2].

Results

A list of a hundred MSGs was defined, divided into 20 topics (Respiratory conditions, Digestive conditions, Infectious conditions, Cardio and Cerebrovascular conditions, General symptoms…). 60 MSGs were dedicated to alert and detection of both expected seasonal epidemics (12 MSGs) and unexpected events (42 MSGs). They contain unspecific or acute pathologies and symptoms. 40 MSG included medical causes of death related to chronic diseases and medical history.

The list of established MSGs was composed of:

- MSGs for detection of expected seasonal events such as: “Influenza”, “Low acute respiratory infection”, “Gastroenteritis”, “Chikungunya”, “Heat related death”, “Dehydration”…

- MSGs for detection of the impact of unexpected events such as: “Epilepsy”, “Choc”, “Coma”, “Unspecified fever”, “Headache”, “Suicide”, “Drugs/opioids poisoning”…
- MSGs for Chronic diseases and Medical history: “Chronic digestive diseases”, “Chronic endocrine diseases”, “Genitourinary chronic diseases”, “History of diseases”…

The weekly number of MSGs built using SVM classification was close and highly correlated to the weekly number of MSGs built using ICD10 codes (Figure 1). Seasonality and peaks were visible using both classifications. For instance, the increase of the MSG “Influenza” occurred during winter months which are known to be the circulating months of the influenza virus (Figure 1, left) [3].

For unusual and rare events such as death due to burns, we observed that the weekly numbers of MSG “Burns” were also similar using both methods. We observed (Figure 1, right) that the outbreak that occurred in September 2016 related to a major bus accident was found using ICD10 codes or SVM classification.

Conclusions

The use of free-text causes of death for reactive mortality surveillance requires the development of a strategy for the analysis of these data. Defining MSGs was essential for the implementation of automatic classification methods of the death certificates in routine.

The dynamic of MSGs using ICD10 codes or SVM classification were comparable. However, the use of ICD10 codes for reactive mortality surveillance is not an option due to the delay of availability of the codes. The uses of machine learning methods, thus, enable to harness free-text causes of death for the reactive mortality surveillance with an objective of detection and early impact assessment.

Acknowledgement

The authors thank the medical and epidemiological expert for the discussion and validation of definitions of Mortality Syndromic Groups

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 [1]. *Public Health*. 143, 85-93. PubMed https://doi.org/10.1016/j.puhe.2016.10.029
2. Baghdadi Y, Bourrée A, Robert A, Rey G, Gallay A, et al. 2018. Automatic classification of medical causes from free-text death certificates for reactive mortality surveillance in France [2]. *Int J Med Inform*. (Under review).
3. Bedford T, Riley S, Barr IG, Broor S, Chadha M, et al. 2015. Global circulation patterns of seasonal influenza viruses vary with antigenic drift [3]. *Nature*. 523(7559), 217-20. PubMed https://doi.org/10.1038/nature14460

Figure 1: Weekly numbers of MSGs “Influenza” and “Burns” using ICD10 codes (Black) and SVM classification (Red) from 2012 to 2016 in France

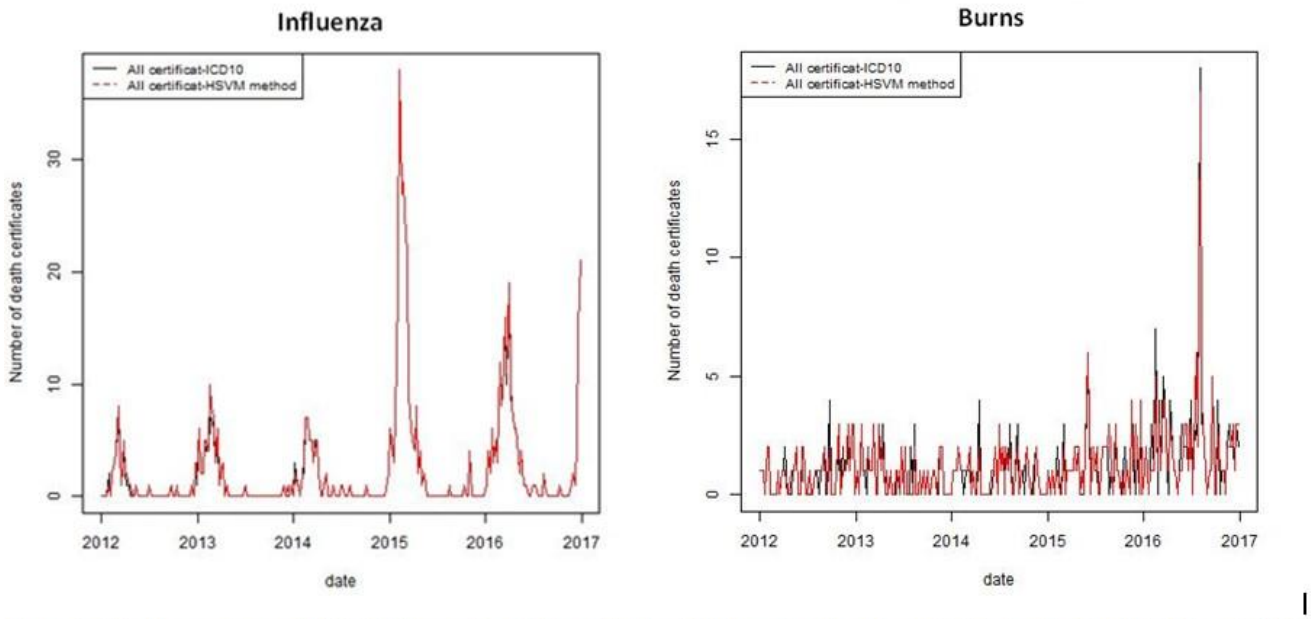

Figure 1: Weekly numbers of MSGs “Influenza” and “Burns” using ICD10 codes (Black) and SVM classification (Red) from 2012 to 2016 in France