User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-639

Challenges for Signal Generation from Medical Social Media Data

Johannes DREESMAN[a], Kerstin DENECKE[b,1],
[a] *Niedersächsisches Landesgesundheitsamt, Hannover, Germany*
[b] *L3S Research Center, Hannover, Germany*

Abstract. Early detection of disease outbreaks is crucial for public health officials to react and report in time. Currently, novel approaches and sources of information are investigated to address this challenge. For example, data sources such as blogs or Twitter messages become increasingly important for epidemiologic surveillance. In traditional surveillance, statistical methods are used to interpret reported number of cases or other indicators to potential disease outbreaks. For analyzing data collected from other data sources, in particular for data extracted from unstructured text, it is still unclear whether these methods can be applied. This paper surveys existing methods for interpreting data for signal generation in public health. In particular, problems to be addressed when applying them to social media data will be summarized and future steps will be highlighted.

Keywords. Epidemic Intelligence, Signal Generation, Disease Surveillance

1. Introduction

Threats to public health, for example those, that are related to disease activity in humans and animals or bioterrorism, are under monitoring by public health officials. Factors such as globalization or climate change contribute to the fast emergence of disease activity and lead to an increased necessity to detect those threats as early as possible. For this reason, early detection of disease activity became even more important. Thus, besides indicators such as number of reported cases or drug prescriptions, new information sources (e.g., social media data from the Web) are considered and investigated for the purpose of epidemiologic surveillance.

Epidemiologic surveillance comprises the process of gathering and analyzing data related to human health and disease to early identify and characterize public health events and to be aware of disease activity in the human population. Thus, the objectives of this process are situational awareness and early event detection [1]. In general, within this process occurrence of a public health event is recognized from input data. The input data is processed by detection methods and indicators or hints to health events are identified. In the following signal generation step, several indicators are combined and analyzed. In the simplest case, the indicator frequency is compared to a predefined threshold. In more complex approaches, statistical algorithms are exploited.

By now, these algorithms have been used to analyze indicator data received through traditional reporting mechanisms such as the number of cases reported by physicians or laboratories. It is still unclear, whether they are useful to analyze also

[1] Corresponding Author.

data collected from other information sources. In particular, the challenges to be addressed when interpreting indicators gained from medical social media data have not yet been identified.

In this paper, we describe the problem of signal generation from medical social media data. Existing methods to signal generation are briefly summarized. Finally, the challenges for signal generation to be addressed when exploiting medical social media data for signal generation and epidemiologic surveillance are summarized and steps for future work are pointed out.

2. Signal Generation from Medical Social Media Data

2.1. Signal Generation from Medical Social Media Data

A signal is considered a hint to a public health event. To generate a signal, input data is processed, indicators are detected and analyzed. In traditional surveillance, values of indicators are directly collected (e.g., number of cases reported by a laboratory) and provided to the analysis methods. When considering unstructured texts such as medical social media data, these indicators need to be detected first, before generating signals. Indicators to be detected from social media data could be frequent mentions of specific symptoms or disease names. A signal is generated on the basis of indicators and thresholds. If an indicator exceeds the threshold, a signal is generated. The thresholds may be constant or variable in time. They may refer to the absolute value of the indicator or to changes of the indicator in time.

2.2. Existing Systems for Epidemiologic Surveillance

Existing systems for epidemiologic surveillance differ in the kind of sources they monitor and the kind of information they provide. Hartley et al. provide an overview on the landscape of event-based surveillance [2]. In contrast to indicator-based (or traditional) surveillance where indicators are used for surveillance (such as number of reported cases or drug prescriptions), event-based systems use additional sources of information. In this paper, the focus is on event-based systems. Two example systems are HealthMap[2] and MedISys[3] using news media and public health websites as information source. Signals are generated by using a simple threshold method. BioSense[4] exploits outpatient data along with medical laboratory test results. To this structured data, methods for signal generation as introduced in section 3 are applied. Considering social media data for this purpose has just been started and many challenges still need to be addressed. Further, statistical methods known from traditional surveillance are not yet applied in event-based surveillance. In this paper, we summarize the problems for signal generation from medical social media data.

[2] http://www.healthmap.org

[3] http://medusa.jrc.it/medisys

[4] http://www.cdc.gov/biosense

3. Methods for Signal Generation

Indicator-based surveillance systems exploit a variety of temporal and spatial methods for signal generation. The basic idea behind all methods is to search for aberrations of the observed values from the expected level. These aberrations might occur in time, in space or in combinations of both. Depending on the disease under investigation and on the reporting system, there might be a substantial variation of the indicator in time or space anyway. In the time domain, this variation is often caused by seasonal effects (e.g. seasonal influenza) and can be taken into account, if expected values are mainly generated from the same season of the past years. Incompleteness of a reporting system, i.e. data for some period of time is missing, is not such an issue for the statistical procedures, if the amount of incompleteness remains stable over time.

3.1. Methods based on Simple Thresholds

The *Bayes threshold* is calculated from the reference data of six weeks before the currently observed value, by estimating the coefficients of a negative binomial distribution from the observed data and by calculating an upper threshold with an alpha of 0,05. The set of reference data can also be extended to reference data from earlier years [7]. Furthermore, the surveillance institutes or health organizations such as Robert Koch-Institut (RKI, Germany) or Center of Disease Control (CDC) have implemented their own thresholds. The *RKI threshold* [3] is calculated from the reference data of six weeks before the currently observed value. The set of reference data can also be extended to reference data from earlier years. The *CDC threshold* is usually based on reference data from the past five years. The reference values are taken from several weeks. From the reference values the mean and the variance are calculate an upper 95% prediction limit which serves as threshold [9].

3.2. Methods based on Regression Analysis

The threshold methods described in section 3.1 suffer from the weakness that a secular trend in the indicator can not be considered. Outbreaks in the past are likely to disturb the estimation of the threshold. Finally statistical properties of the method become questionable, if the indicator is presenting very low numbers. To overcome all these problems, Farrington et al. proposed an approach based on generalized linear models, which is by now broadly applied in European countries for indicator based surveillance [8]. The approach fits a regression model to the data over several years, allowing for a secular trend. Outbreaks in the past are automatically identified and removed, and the statistical distribution fits either to rare counts or to frequent counts [8].

3.3. Methods based on Quality Control Measures

Cumulative sum or CUSUM methods [5,6] originated in quality control. They do not focus on the total aberration of an observed value from an expected value in one particular period of time, but on several consecutive periods, and sum up aberrations in one particular direction. If there is a similar tendency in consecutive weeks, the sum of aberrations increases over a particular threshold and a signal is generated.

Michael Höhle [3] provides a software package for the statistical software R with several statistical algorithms for surveillance implemented. It contains functionality to

visualize routinely collected surveillance data and provides algorithms for the statistical detection of potential outbreaks. The inputs to these algorithms are univariate or multivariate time series of case counts (number of cases and whether there was an outbreak). In the package, all time series methods mentioned before are implemented. The package is currently used by several European national health authorities.

3.4. Detection of Spatial Clusters by using the Scan-Statistic

The spatial Scan-Statistic scans the area of interest by using a circular window [4]. This window is moved over the area and the diameter is changed, such that the window covers one district or several neighboring districts. For each window, joint case density of the districts inside the window is compared with case density outside. Simulation methods are used to assess whether a difference between inside and outside the window is statistically significant. In order to adjust for varying population densities, population data of the regions can be incorporated. If no population data are available, data of another "control disease" can be used as a reference. The spatial Scan-Statistics can be calculated by the computer program SatScan which is freely available [4].

4. Challenges for Generating Signals from Social Media Data

Shmueli and Burkom summarized in [10] statistical challenges when monitoring time series. They came up with four differences that characterize the epidemiologic surveillance setting: the underlying background behavior, the nature of outbreaks, evaluation of performance and the requirements and uses of surveillance systems. When evaluating signal generation methods, three criteria are used: sensitivity, specificity and timeliness [2]. Sensitivity and specificity are well known evaluation measures.

Timeliness can be measured by subtracting the time of generation from the time of the event itself. This measure is crucial since the early detection of disease activity is a high priority and tools are only accepted by public health officials when they show a clear benefit in contrast to existing reporting and analysis processes. Signal generation methods can normally be adjusted to increase or decrease the single values for the quality measures mainly by adapting thresholds or other parameters. But, this process is very difficult since the measures depend on each other. An increase of specificity can lead to a decrease of sensitivity and so on.

Already in traditional surveillance systems, a challenge in the spatial domain is the unequal distribution of the population. In addition, a substantial spatial variation might occur due to different diagnostic or reporting behavior. The latter is much more difficult to assess and adjustments are complicating.

In the following, we focus more specifically on concrete challenges when considering social media data for epidemiological surveillance. The statistical methods described before base on certain assumptions [1]. Already indicator-based surveillance in general violates some of these assumptions (e.g., observations are auto correlated). When using social media data, even more assumptions are violated due to the peculiarities of this data. While indicators used in traditional surveillance can be considered true, social media data is extremely noisy and data is coming in every second (e.g. Twitter data).

Indicators to a public health event can be detected in various social media sources (e.g. in news articles and in blog postings). It is unclear, how to aggregate these indicators before analyzing them within the signal generation process. Further, indicators might be mentions of symptoms. But, which symptoms together make a signal to a public health event? In addition, information that is normally available in traditional surveillance could be missing in social media data (e.g. detailed information on the location or the time). It is still a challenge to find a solution how to process incomplete data with these algorithms, whether to consider them anyway or better leave them out.

5. Conclusions and Future Work

In this paper, the problem of generating signals for epidemiological surveillance form medical social media data has been characterized. Even though there are signal generation methods from indicator-based surveillance available, it is still unclear, to what extent these methods are applicable to the specific problem of signal generation form social media data. We collected the challenges to be considered.

In future, the methods need to be applied to data collected from unstructured documents and solutions for these challenges need to be found. For this purpose, a standard data set would be very helpful that could be used to evaluate the various statistical methods with respect to the evaluation measures sensitivity, specificity and timeliness. Such data set for medical social media data is still missing. This would help to test and adapt methods more easily.

Aknowledgements: This research is part of the M-Eco project funded partly under 247829 by the European Commission

References

[1] Fricker RD: Biosurveillance: Detecting, Tracking, and Mitigating the Effects of Natural Disease and Bioterrorism. *Encyclopedia of Operations Research and the Management Sciences ,* 2010

[2] Hartley DM, et al. The Landscape of International event-based Biosurveillance. *Emerging Health Threats Journal,* 3, 2009

[3] Höhle M. surveillance: An R package for the surveillance of infectious diseases, *Computational Statistics* (2007), 22(4), pp. 571-582.

[4] Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* (1995) 14:799-810.

[5] Rossi G, Lampugnani L, Marchi M. An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*,1999, 18:2111-2122.

[6] Rogerson PA, Yamada I. Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, 2004. 53/Supplement:79-85.

[7] Riebler A. Empirischer Vergleich von statistischen Methoden zur Ausbruchserkennung bei Surveillance Daten (Empirical Comparison of statistical methods for outbreak detection in surveillance data), *Bachelor thesis*, 2004

[8] Farrington P, Andrews N, Beale A, Catchpole M. A statistical algorithm for the early detection of outbreaks of infectious disease. *J. R. Statist. Soc. A*, 159, 1996: 547-563.

[9] Stroup D, Williamson G, Herndon J, Karon J. Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine*, 8, 1989: 323 – 329

[10] Shmueli G, Burkom H. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics*, Vol. 52, No. 1. (February 2010), pp. 39-51.