|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-E-026 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Geneva, 30 May – 1 June 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | Robert Koch Institut | | |
| **Title:** | | New topic area: Outbreak detection | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Martina Fischer Robert Koch Institut, Germany | | Tel: +49 30 18754 4485 Email: [FischerMa@rki.de](mailto:FischerMa@rki.de) |
| **Contact:** | | Bernhard Renard Robert Koch Institut, Germany | | Tel: +49 30 18754 2561 Email: [RenardB@rki.de](mailto:RenardB@rki.de) |
|  | | Alexander Ullrich  Robert Koch Institut, Germany | | Email: [ullricha@rki.de](mailto:ullricha@rki.de) |

|  |  |
| --- | --- |
| **Abstract:** | Infectious disease outbreaks pose a major risk to public health. However, early detection can prompt fast interventions to limit spread of the disease or even prevent an outbreak altogether.  According to the German infection protection law, cases of notifiable pathogen are reported and collected via a mandatory reporting system at the RKI. The available data set contains a collection of 8 million reported infectious disease cases within Germany. The data holds expert labels relating cases to specific disease outbreaks.  The aim of outbreak surveillance is to detect changes and conspicuous events within the case data in a fast and automatic manner. Both classic statistical methods and supervised learning methods have been applied for signal detection on the reported case data. To this, AI algorithms can increase the timeliness and accuracy of outbreak detection and improve the understanding of the warnings. It can particularly do so by incorporating multiple data streams with diverse properties. To achieve earlier and more comprehensive detection of notifiable and non-notifiable pathogens, the integration of real-time-surveillance data with data from the reporting system is crucial. For this task, internal syndromic surveillance sources are available and valuable external data sources (google trends, health apps) are present. |

# Overview

We present a use case on outbreak detection of emerging infectious diseases. Infectious disease outbreaks pose a major risk to public health. However, early detection can prompt fast interventions to limit spread and effects of the disease or even prevent an outbreak altogether.  
According to the German infection protection law, which specifies notifiable pathogens and information to be reported, infectious cases are reported and collected via a mandatory reporting system. This electronic reporting system, named SurvNet, was established at RKI in 2001. So far, the available data set contains a collection of 8 million reported infectious disease cases within Germany. The data contains expert labels indicating which cases can be related to specific disease outbreaks. Hence, it poses a reliable and excellent training data set.

The aim of outbreak surveillance is to detect changes and conspicuous events within the case data in a fast and automatic manner. We have applied both classic statistical methods as well as supervised learning methods for signal detection on the reported case data. To this, AI algorithms can increase the timeliness and accuracy of outbreak detection as well as improve the understanding of the warnings. It can particularly do so by incorporating multiple data streams with diverse properties. To achieve earlier and more comprehensive detection of notifiable as well as non-notifiable pathogens, the integration of real-time-surveillance data with data from the reporting system is crucial. For this task, internal syndromic surveillance sources are available and further valuable external data sources (google trends, health apps) are present.

# Relevance

Infectious disease outbreaks pose a major risk to public health. However, early detection can prompt appropriate reactions to contain the consequences or even prevent the threat altogether. Infectious diseases are a concern globally. There are many established infectious diseases that cause the death of millions of people every year and new infectious diseases are emerging. The risk and occurrence of infectious diseases is influenced by globalization, mass migration and climate change. Therefore, new and faster methods of detection and prevention have to be found to deal with these challenges. Better and faster outbreak detection can save lives, increase quality of life, and in general will benefit the health of the population.

# Impact

AI algorithms can increase the timeliness and accuracy of outbreak detection as well as improve an understanding of the warnings and the disease spread itself. It can particularly do so by incorporating multiple data sources with diverse properties. Combining real-time sources like internet searches or emergency department visits with high quality sources, e.g. lab results, can give accurate warnings ahead of previous approaches. Furthermore, it can provide additional information on the underlying causes allowing more specific actions to be taken for prevention.

Currently, outbreak detection algorithms are often parametrized, compared, and tested on small sets of data or on simulations. These simulations are a very simplistic representation of outbreaks and capture only a few aspects of the real world, therefore algorithms are benchmarked often on the task of detecting elevated levels of counts in univariate time-series. By using actual outbreak data from the German mandatory surveillance system, algorithms can be benchmarked on the actual task of detecting real world outbreak events. Furthermore, new benchmarking metrics should capture the actual costs and benefits of timely detection and false alarms, which may differ by disease and region.

# Existing work

The field of outbreak detection has several established methods, comprising mostly of regression methods on time-series, or cluster-finding approaches. Some machine learning methods, mostly Hidden Markov Models have been proposed. Several tools (surveillance package, SatScan) exist to perform outbreak detection on infectious disease data.   
Currently, outbreak detection is primarily performed on only one data source at a time and thus classic statistical models are often sufficient. However, as more and more informative internal (several different surveillance systems) and external data sources (google trends, health apps) appear, the need for more sophisticated learning approaches emerges.

Concerning benchmarking, there has been research in assessing and comparing outbreak detection methods which mostly use rather simple statistical metrics, such as accuracy or F1-score, and a few more sophisticated scores incorporating prediction uncertainty and timeliness. For the most part, simulated outbreaks are used to assess algorithms for individual time-series.  
However, using actual outbreak data from the German mandatory surveillance system, algorithms could be benchmarked on the actual task of detecting real world outbreak events, which is accessible as a gold standard from within the RKI.

# Feasibility

We have applied both classical statistical methods as well as supervised learning methods to the problem of outbreak detection. Thereby we use exactly the described dataset of lab confirmed outbreak cases from the German mandatory reporting system (see point 6). The machine learning methods use the outbreak labels. We already see first improvements in the accuracy using ML approaches compared to classic statistical approaches. E.g. when keeping the same sensitivity in outbreak detection, the false alarms are considerably decreased. This reduces the number of alarms the experts have to assess. Hence, a training set is already established, and we have some baseline methods at hand.   
The next steps would be to define a metric, that considers timeliness and importance (size of outbreak, seriousness of disease) and incorporates the prediction uncertainty. Further, AI-methods such as LSTMs will be investigated for the use in outbreak detection. The goal is then to extend these algorithms to integrate further data streams from both internal sources and external sources. Relevant internal sources are already available from an emergency department (ED) syndromic surveillance system, and external sources can either be publicly available data streams, such google trends, or data streams from health companies that collect relevant health data. Finally, an evaluation framework needs to be developed that uses the above-mentioned metric and assesses the AI-methods on the level of the entire surveillance system, i.e. for one or several diseases and all covered regions of the surveillance systems.

# Data availability

The available data contains 8 million infectious disease cases collected through the German mandatory reporting system. The data contains expert labels indicating which cases are related to specific disease outbreaks. For each case, there is information about the pathogen, demographics (age, sex), location (NUTS-3 level, 400 counties) and some other relevant information such as hospitalization, fatality, affiliation with care facilities and others. Some of this data is publicly available in an aggregated form, e.g. counts for a specific disease, by week and county. However, details and single cases are not published and most importantly, the expert outbreak labels have not been disclosed so far. All of the data is collected through the SurvNet@RKI reporting system via a webservice and stored in a structured relational SQL database. Training data sets can be provided.

Data from the national Syndromic Surveillance system, currently under development at the RKI, comprises of near real-time emergency department attendances, covering information on the person (age, gender), location, preliminary diagnosis and chief complaints, vital signs as well as diagnostic and administrative details. The data is available in an anonymized and semantically labelled data model. It comprises about 50,000 attendances per hospital and year, amounting to an approximate of 10 million single cases for the participating 40 pilot hospitals and retrospective data usually available for about five years.

Other data sources potentially accessible in near real-time (web, apps, healthcare providers, hotlines, …) are currently assessed for their use in a national Syndromic Surveillance system to support outbreak detection. The variety of data sources is likely to improve the accuracy and timeliness of this valuable Public Health tool.

# Data quality

The data is collected since 2001 at the RKI via the mandatory reporting system and undergoes constant data quality checks by data engineers and review by epidemiologists. Some of the data is publicly available and is also communicated in monthly bulletins and yearly reports. The data is used by hundreds of local health agencies on a daily basis. While a fully clean dataset is generally difficult to assure, this data is heavily and routinely checked for quality.

# Annotation/label quality

The labels ‘outbreak’ and ‘no outbreak’ are given by experts from actual outbreak investigations. Hence, the information that a case, which is labelled as an outbreak case, was actually part of an outbreak is highly trustworthy. However, it is not clear to which extent all true outbreak cases are labelled as an outbreak, since not all outbreaks are found or investigated by the local health agencies.  
An independent and external study on gastrointestinal diseases by the food safety agencies, however confirmed, that most of the larger outbreaks reported by them were also found in similar case numbers in our dataset.

# Data provenance

The mandatory reporting system is based on the German infection protection law, which specifies which diseases and which information must be reported. The data arrives pseudonymized at the RKI. The data stems from about 400 local health agencies with somewhat different regulations for each of the 16 federal states and a diverse software landscape. Data from emergency departments collected within the ESEG-project, piloting the national Syndromic Surveillance strategy, is fully anonymized and available for scientific use.

# Benchmarking

Since many of the previous approaches were time-series based, we expect further Hidden Markov Models and deep learning methods appropriate for sequential data such as Long Short Term Memory Networks (LSTM). However other methods, like multivariate Bayesian regression or all-purpose deep learning (CNN, RNN) are conceivable.

Metrics should contain cost functions for timeliness, outbreak size and false alarms, and incorporate prediction uncertainty.

# Organizer

One of the main responsibilities of the Robert Koch institute (RKI) is to detect, prevent and control the spread of infectious diseases in the population. For this purpose, the RKI has established a traditional surveillance system, collecting lab confirmed cases for around 80 infectious diseases as well as several syndromic surveillance systems using data from e.g. emergency departments and general practitioners. An automated outbreak detection system, utilizing this data, is in use for early detection of infectious disease outbreaks.  
While assessing more and more external data sources to improve RKI’s mandate to support outbreak detection through a Syndromic Surveillance System, a strategy for a comprehensive exhaustion of the full information content is under development. With the variety of data sources, types, and formats to become available in the next years, methods of AI are expected to outpace performance of traditional methods using a manual descriptive-analytic framework for near real-time outbreak detection.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_