|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-I-018-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 7-8 May 2020 |
| **DOCUMENT** |
| **Source:** | TG-Outbreaks Topic Driver |
| **Title:** | TDD (TG-Outbreaks) |
| **Purpose:** | Discussion |
| **Contact:** | Stéphane GhozziRobert Koch InstituteGermany | Email: GhozziS@rki.de  |
| **Contact:** | Auss AbboodRobert Koch InstituteGermany | Email: AbboodA@rki.de  |

|  |  |
| --- | --- |
| **Abstract:** | This document contains an outline of a topic description document (TDD) on Outbreak detection (TG-Outbreaks). This version makes necessary references to a recently written review about benchmarking of outbreak detection algorithms. NOTE – retrieved from the Deliverables folder in the collaboration area. |

**Table of Contents**

[1 Introduction 3](#_Toc36659122)

[1.1 Document Structure 3](#_Toc36659123)

[1.2 Topic Description 3](#_Toc36659124)

[1.3 Ethical Considerations 3](#_Toc36659125)

[1.4 Existing AI Solutions 3](#_Toc36659126)

[1.5 Existing work on benchmarking 4](#_Toc36659127)

[1.5.1 Data format 4](#_Toc36659128)

[1.5.2 Metrics 5](#_Toc36659129)

[2 AI4H Topic Group 6](#_Toc36659130)

[2.1 Topic group structure 6](#_Toc36659131)

[2.1.1 Topic group participation 6](#_Toc36659132)

[2.1.2 Tools/process of TG cooperation 7](#_Toc36659133)

[2.1.3 TG interaction with WG and FG 7](#_Toc36659134)

[2.1.4 Next meetings 7](#_Toc36659135)

[2.1.5 Next steps 7](#_Toc36659136)

[3 Method 7](#_Toc36659137)

[3.1 AI Input Data Structure 7](#_Toc36659138)

[3.1.1 Available Data 7](#_Toc36659139)

[3.2 AI Output Data Structure 7](#_Toc36659140)

[3.3 Test Data Labels 8](#_Toc36659141)

[3.4 Score and Metrics 8](#_Toc36659142)

[3.5 Undisclosed Test Data Set Collection 8](#_Toc36659143)

[3.6 Benchmarking Methodology and Architecture 9](#_Toc36659144)

[3.6.1 Benchmark Tasks 9](#_Toc36659145)

[3.7 Reporting Methodology 9](#_Toc36659146)

[4 Results 9](#_Toc36659147)

[5 Discussion 9](#_Toc36659148)

[6 Declaration of Conflict of Interest 9](#_Toc36659149)

[References 10](#_Toc36659150)

DEL10.10

Topic Description Document for TG-Outbreaks

# Introduction

* Infectious disease outbreaks pose a major risk to public health
* Early detection of outbreaks can prompt fast interventions
* Case data are collected by diverse surveillance systems
* AI algorithms can be applied to detect aberrant case numbers based on these data collections
* AI algorithms have the potential to increase the timeliness and accuracy of outbreak detection

## Document Structure

This TDD will cover core topic-specific questions, including:

* Relevant definitions of the term and event of outbreaks for detection algorithms
* Label uncertainties (e.g. issue of unlabelled outbreak cases, no lab-confirmed outbreaks)
* Performance evaluation: relevant score and metrics for benchmarking detection algorithms
* Test set(s) (undisclosed) to serve as a gold standard for benchmarking (based on different pathogens, time frames, demographic groups, and regional resolution)

## Topic Description

Infectious disease outbreaks pose a major risk to public health and are of global concern. Many established infectious diseases 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, migration, and climate change. According to a World Health Organization (WHO) ranking, infectious diseases are ranked in the top 10 causes of death worldwide.

However, early detection of outbreaks can prompt fast interventions to limit spread of the disease or even prevent an outbreak altogether. Improved algorithms for outbreak detection can save lives, increase quality of life and will benefit the overall health of the world population.

The aim of outbreak detection algorithms is to detect aberrant case numbers and conspicuous events within data streams, pointing to the emergence of infectious disease outbreaks, in a fast and automatic manner. To this, AI algorithms can increase the timeliness and accuracy of outbreak detection as well as improve the understanding of the warnings.

## Ethical Considerations

Relating to medical doctors having the Declaration of Geneva and the Nuremberg Code, analogue principles and quality standards need to be established for AI applications used in health enquiries.

* Best scientific practices need to be assured
* Data protection: drawing references to an individual person has to be prevented.
	+ The integration of several and further data sources needs to assure that the personal identity is still protected.
	+ For the collection of labelled outbreak data, the levels of data aggregation needs to be defined carefully
* Balanced selection of test data: prevent discrimination of demographic groups, risk groups or even countries

## Existing AI Solutions

There is a variety of published statistical approaches and machine learning methods ((Unkel, Farrington, Garthwaite, Robertson, & Andrews, 2012) (Yuan, Boston-Fisher, Luo, Verma, & Buckeridge, 2019) (Allévius & Höhle, 2017) (Salmon, Schumacher, & Höhle, 2016)), which are used for the detection of outbreaks in given surveillance data.

At the Robert Koch Institute (RKI) we have applied both classical statistical methods as well as supervised learning methods to the problem of outbreak detection. The machine learning methods use outbreak labels, assigned from expert investigations. The main methods at hand are based on Hidden Markov Models and the improved Farrington method. We already see first improvements in the accuracy using ML approaches compared to classic statistical approaches (Zacher & Czogiel, 2019). 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.

For the future, 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.

## Existing work on benchmarking

Minimal benchmarking setup at RKI:

* Data: weekly reported infection cases and outbreaks for notifiable diseases in Germany
* training of the algorithms on data of the past 5 years
* testing on next week (prospective 1 week ahead: data available until last week)
* Scores = functions of TP, FP, TN, FN: sensitivity, specificity, precision, F1, …

Existing work in benchmarking of outbreak detection algorithms (more closely described in our review *How to benchmark disease outbreak detection algorithms: A review*)

### Data format

Outbreak detection traditionally happens as part of indicator-based surveillance (IBS). According to WHO, it is defined as the “systematic collection, monitoring, analysis, and interpretation of structured data, i.e. indicators, produced by a number of well-identified, predominantly health-based formal source''. The complementing form of surveillance to IBS is called event-based surveillance and can be understood, according to WHO, as “the organized collection, monitoring, assessment and interpretation of mainly unstructured *ad hoc* information regarding health events. Since benchmarking relies somewhat on a pre-specified data model to be able to easily run different algorithms that we will focus describing benchmarking on IBS data. EBS data lacks structure by definition and therefore, it is hard to adjust benchmarking to all possible forms EBS data can assume.

Although IBS is more structured, IBS data still comes in different shapes which might be relevant for the later use of algorithms. For example, it might be important to have a long history of data since some algorithms require data to be have been collected for at least five years. Furthermore, almost any surveillance system that reports notifiable diseases does so by providing the date of infection or report and cases numbers aggregated to weeks months, or quarter and a location of varying precision (street address, county, region, federal state,…). Our choice of algorithms, however, depends on the available granularities of the former properties. For example, to detect whether two cases are part of an outbreak, the Knox test can be applied where locations are close given a prespecified critical distance and time span. This makes it desirable to have a more exact location than using the former method. Most algorithms can incorporate spatial information given there is a meaningful metric for distance and a sufficiently strong spatial resolution such as CUSUM or regression models.

#### Simulated data

If we were to agree on a data format, we still would need to determine the source for this data. It is not, as obviously assumed, the best way to benchmark using real data from a public health institute. There are studies that use wholly simulated data, real data with simulated outbreaks and other artificial alterations of real data to assure where an outbreak is situated, and only real data where outbreak labels are known form the evaluations of epidemiologists. All these different approaches have their advantages and disadvantages.

The main motivation to evaluate outbreak detection algorithms using simulated data is this approach provides a ground truth about the outbreaks injected into the (often also simulated) endemic baseline. Since disease dynamics, such as seasonality, reporting behavior, and trends, are known, a good estimate of realistic data can be formulated. The ground truth knowledge about outbreaks might be missing in real data and therefore makes it impossible to calculate several performance scores such as specificity and sensitivity.

One approach for such a simulation is to produce a linear model that generates mean outbreak cases per week which are then used as an input for a negative binomial model to introduce some natural variance. The model parameters are chosen to mimic characteristics of timeseries for different pathogens. Outbreaks are then generated using a Markov process to selected weeks as outbreak weeks. On such outbreak weeks a realization of a Poisson distribution with mean equaling to a chosen constant is added. The added cases are distributed over the outbreak week given a lognormal distribution.

#### Real data

Even though the usage of real data might have clear disadvantages, such as being incomplete, which motivated the development of disease outbreak simulations, there are still viable approaches to still utilize real data for the evaluation and training of disease outbreak algorithms.

A straightforward approach to train/test an outbreak detection algorithm is to use real data where outbreaks are labeled by epidemiologists. Downsides of this method is that not all outbreaks are recognized by epidemiologists, sometimes only the reporting data and not the data of infection is known, or the data suffers from reporting delays which can degrade the performance of an algorithm.

Another approach is to select the 20% highest values from a time series and subtract them to create an endemic timeseries on which outbreak detection happens in form of aberration detection. Due to down-weighting of high baseline values of algorithms trained on synthetic data, one alternative is to take real data, train a generalized linear model or, given seasonality, a generalized additive model let the model detect extreme values, and then replace these values with the realization of a negative binomial distribution using a lower expected value than the removed values. This way, extreme values, considered as outbreaks, are removed and we get two timeseries, one with, and one without outbreaks/extreme values. These two timeseries of endemic and epidemic case counts are reunited with the epidemic outbreak timeseries being shifted by one year into the future, incorporating knowledge about the seasonality of the disease of interest, to create new labeled timeseries from real data.

### Metrics

When we want to measure the performance of an algorithm, we might look for criteria such as usefulness, cost, sensitivity, representativeness, timeliness, simplicity, flexibility, and acceptability. These are measures that include not only the statistical algorithms but also the more general criterions for public health systems. Common measures for the comparison of statistical algorithms are (more closely described in our review *How to benchmark disease outbreak detection algorithms: A review*):

* Sensitivity
* Specificity
* Precision
* Negative predictive value
* F1
* ROC/AUC
* ROC using a timeliness measure where we define a minimum timeliness *D* such that outbreaks must be detected within *t+D* with *t* being the time point where an outbreak started. Let *s* be the timepoint where an outbreak started, then *1-s/D* replaces the false positive rate in our ROC curve. This timeliness measure is defined to not be smaller than 0.
* ROC where we use a normalized measure to punish time elapsed since the begin of an outbreak. This might be important to compare timeseries with various time granularity. Such a method could be to count the timesteps elapsed since an outbreak, where a timestep is defined by the granularity or some other criterion.
* Instead of replacing some axis on the ROC, we can add a third dimensions such as timeliness and calculate a volume under the curve to measure the performance of an algorithm.
* Matthews Correlation Coefficient
* Scaled probability of detection (POD), where we count whether an algorithm detected a count within an outbreak as being extreme. The proportion of outbreaks detected this way is called POD.
* One extension of the POD is the Scaled POD which takes the size of the detected outbreak into account. By weighting the amount of detected outbreak with the size of the outbreaks, i.e. the amount of cases belonging to an outbreak.
* Another timeliness measure is the average time before detection. It is the sum of all detected outbreaks by an algorithm multiplied by the time elapsed since outbreak normalized by the overall number of outbreaks.
* A variation of the average time before detection that punished absolute delays in detection of an outbreak is the relative size before detection. This metric consists of the sum of detected outbreaks multiplied by the deviation of the epidemic time series from the endemic timeseries, i.e. the fraction of cases during the detection of the outbreak divided by the number of cases not part of an outbreak. This metric is then normalized by the overall number of outbreaks.

# AI4H Topic Group

## Topic group structure

Topic Groups summarize uses cases of a certain health topic or problem and similar AI benchmarking requirements. However, inside a Topic Group different Sub-topic Groups can be established to pursue different topic-specific specializations. TG-Outbreak will start without separate subtopic Groups. However, it is possible that during the process subtopics will be introduced. Possible examples are:

* Outbreak detection for specific pathogens
* Different national outbreak detection methods for different nations based on different national reporting systems
* …

### Topic group participation

The topic group on outbreak detection algorithms currently includes members from the Robert Koch Institute (National Public Health Institute Germany), involving members from different groups within the institute

* Dr Stéphane Ghozzi, Infection epidemiology, Signale team (Robert Koch Institute, Unit 31)
* Auss Abbood, Infection epidemiology, Signale team (Robert Koch Institute, Unit 31)
* Dr Alexander Ullrich, Infection epidemiology, Signale team (Robert Koch Institute, Unit 31)

### Tools/process of TG cooperation

The TG will utilize any available online collaboration tools to further its work (between FG meetings)

### TG interaction with WG and FG

TBC

* Regular attendance of the Focus Groups meetings

### Next meetings

The Focus Groups meets about every two months at changing locations. The upcoming meetings are:

* I: May in Geneva/virtual (TBD)

An up to date list can be found at the official [ITU FG AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

### Next steps

Future contributions expected for the TG will revolve around:

* Collection of labelled test data from different sources: Any data stream (case reporting systems, surveillance systems, etc.) directly linked to outbreak labels (expert/lab confirmed) is of high value. The ultimate goal is to cover outbreak data from different systems and countries.
* Provision of AI models and approaches for outbreak detection: Contributing to the development of a viable and accepted benchmarking framework
* Support on various aspects (data, methods, benchmarking, etc.) of this topic

# Method

TBC

## AI Input Data Structure

### Available Data

There are different potential data sources which can be used for outbreak detection and serve as input for the detection algorithms. Possible data input sources can be based on different surveillance systems, such as national mandatory reporting systems or syndromic surveillance systems. Further input data sources, particularly accessible in near real-time, are online sources (wikipedia, google clicks, HealthTweets, Twitter) or data from symptom-assessment apps, healthcare providers, hotlines etc. Real time data sources have a high potential of significantly improving the outbreak detection particularly in accuracy or timeliness.

## AI Output Data Structure

TBC

## Test Data Labels

German SurvNet data

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, e.g. confirmed, that most of the larger outbreaks reported by them were also found in similar case numbers in the SurvNet dataset.

Methods on how to use real data for algorithm testing is elaborated in 1.5.1.2.

We, however, did not mention the use of molecular data yet.

## Score and Metrics

Standard statistical metrics

* + sensitivity
	+ specificity
	+ precision
	+ F1

Outbreak detection specific metrics

* time-to-event characterizations (e.g. by time (days/weeks) passed, by number of occurred cases before detection)
* cost functions that interrelate false positives (false alarms) and false negatives (missed true alarms)
* cost functions for outbreak size detection and incorporated prediction uncertainty.

There are many more metrics defined in 1.5.2 which need to be further elaborated on their use for benchmarking.

As of now, we are missing metrics to deal with molecular data and spatial information.

## Undisclosed Test Data Set Collection

German SurvNet data

In Germany, data from the German mandatory reporting system, collected since 2001 at the Robert Koch Institute (RKI), contains 8 million infectious disease cases and undergoes constant data quality checks by data engineers and review by epidemiologists. The data contains expert labels indicating which cases are related to specific disease outbreaks. All of the data is collected through the national reporting system via a web service and stored in a structured relational SQL database. The data arrives pseudonymized at the RKI from about 400 local health agencies. The data holds expert labels relating cases to specific disease outbreaks. For each case, information is given on the pathogen, demographics (age, sex), location (NUTS-3 level, county) and additional features such as hospitalization, fatality, and affiliation with care facilities and others. Some data is publicly available in an aggregated form, e.g. by counts for a specific disease, by week and county. However, details and single cases are not published. Most importantly, the expert outbreak labels have not been disclosed so far. In this document this set is referred to as German SurvNet data.

## Benchmarking Methodology and Architecture

### Benchmark Tasks

At present, outbreak detection algorithms are commonly parametrized and benchmarked on small sets of data or on simulations. These simulations are very simplistic outbreak representations, which capture only few aspects and often reduce benchmarking to the task of detecting elevated case count levels. By creating solutions for using real outbreak data from mandatory surveillance system, e.g. by “sending the algorithm to the place of the data”, algorithms could be benchmarked on the actual task of detecting real world outbreak events.

The topic of outbreak detection is of national and international concern. The development of most detection algorithms is, however, naturally executed on national level. Thereby, each country relies on individual national disease surveillance systems.

To create a standardised benchmarking for output detection algorithms, the topic group aims to address all aspects, which are relevant and shared across countries.

## Reporting Methodology

TBC

# Results

TBC

# Discussion

TBC

# Declaration of Conflict of Interest

TBC

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

Robert Koch Institute (RKI)

One of the main responsibilities of the Robert Koch institute 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.

References

Allévius, & Höhle. (2017). Prospective Detection of Outbreaks. *arXiv preprint*. https://arxiv.org/abs/1711.08960.

Hossin, M., & Md Sulaiman, N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process 5.2*, pp. 1-11 (http://doi.org/10.5121/ijdkp.2015.5201).

Salmon, Schumacher, & Höhle. (2016). Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v070.i10 .

Unkel, Farrington, Garthwaite, Robertson, & Andrews. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *J Royal Statistical Society A*. https://doi.org/10.1111/j.1467-985X.2011.00714.x.

Yuan, Boston-Fisher, Luo, Verma, & Buckeridge. (2019). A systematic review of aberration detection algorithms used in public health surveillance. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2019.103181.

Zacher, B., & Czogiel, I. (2019, Feb 6). Supervised learning improves disease outbreak detection. arXix:1902.10061v1.

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