International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.10

FG-AI4H Topic Description Document for the Topic Group on outbreak detection (TG-Outbreaks)

ITU-T

International Telecommunication Union

Summary

This topic description document (TDD) specifies a standardized benchmarking for AI in outbreak detection for public health. It covers scientific, technical, and administrative aspects relevant for setting up this benchmarking.

Keywords

Artificial intelligence; benchmarking; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; public health; epidemiology; computational methods; disease outbreak detection; early detection; emergency response; mathematical methods; patterns; spread reduction

Change Log

This document contains Version 1 of the Deliverable DEL10.10 on "*FG-AI4H Topic Description Document for the Topic Group on outbreak detection (TG-Outbreaks)*" approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editors:

Auss Abbood and Alexander Ullrich
TG-Outbreaks
Robert Koch Institute, Germany
E-mail: abbooda@rki.de, ullricha@rki.de

Khahlil Louisy
TG-Outbreaks
Institute for Technology and Global Health, USA
E-mail: klouisy@hks.harvard.edu

Alexander Radunsky
Institute for Technology and Global Health, USA; UT Southwestern Medical Center, USA
E-mail: Alexander.Radunsky@UTSouthwestern.edu

Contributors: (in alphabetical order)

Maria Carnovale
ITGH
USA
Email: carnovalemaria@outlook.com

Augusto Gesualdi
ITGH
USA
Email: augusto.gesualdi@pathcheck.org

Khahlil Louisy
ITGH
USA
Email: klouisy@hks.harvard.edu

Gokul Parameswaran
ITGH
USA
Email: gokul.parameswaran@keble.ox.ac.uk

Rebecca Perez
ITGH
USA
Email: rebecca.perez@wadham.ox.ac.uk

Alex Radunsky
ITGH
USA
Email: alex.radunsky@mail.harvard.edu

Simona Tiribelli
ITGH
USA
Email: simona.tiribelli@pathcheck.org

Reinhard Fuchs
Österreichische Agentur für Gesundheit und Ernährungssicherheit (AGES)

Ian Kopacka
Österreichische Agentur für Gesundheit und Ernährungssicherheit (AGES)

Philippe P. Verstraete
"Milan and Associates", an ethical empathetic social enterprise

Giovanna J. Gutierrez ,
"Milan and Associates", an ethical empathetic social enterprise

Elaine Nsoesie
School of Public Health, Boston University

Sophie Marquitan
mTOMADY, a project of Doctors for Madagascar

Dr. Julius Emmrich
mTOMADY, a project of Doctors for Madagascar

Dr. Samuel Knauss
mTOMADY, a project of Doctors for Madagascar

Noelson Lahiafake
mTOMADY, a project of Doctors for Madagascar

Victor Akelo
US CDC, Child health and mortality Prevention Surveillance (CHAMPS) project

M. Claire Jarashow
Los Angeles County Department of Public Health

Sharon K. Greene
NYC Department of Health and Mental Hygiene

Robert Istepanian
Imperial College

Richard Aubrey White
Norwegian Public-Health-Institut
FHI

Birgitte Freiesleben De Blasio
Norwegian Public-Health-Institut
FHI

Gunnar Rø
Norwegian Public-Health-Institut
FHI

Claudia Coipan
RIVM

Roger Antony Morbey
Public Health England; National
Infection Service

Amy FW Mikhail
Public Health England; National
Infection Service

Angela Noufaily
University of Warwick

Anette Hulth
Public Health Agency of Sweden

Pär Bjelkmar
Public Health Agency of Sweden

Henrik Källberg
Public Health Agency of Sweden

Yann Le Strat
Santé publique France (SpF), PH Fr

Céline Caserio-Schönemann
Santé publique France (SpF), PH Fr

Honorati, Masanja
Ifakara Health Institute (IHI),
Tanzania

Salim Abdullah
Ifakara Health Institute (IHI),
Tanzania

Irene Masanja
Ifakara Health Institute (IHI),
Tanzania

Nada Malou
Médecins Sans Frontières (MSF),
France

Ally Salim Jr.
Inspired Ideas, Tanzania

Meghan Hamel
Public Health Agency of Canada

David L. Buckeridge
McGill University

Auss Abbood
Robert Koch Institute

Stéphane Ghozzi
WHO Hub

Bryan Kim
Korean CDC

Azadur Rahman Sarker
Tech Valley Networks Limited

Helmi Zakariah
AIME Inc.

Meerjady Sabrina Flora
Institute of Epidemiology, Disease Control, and Research (Bangladesh)

Chawetsan Namwat
Bureau of Epidemiology, Ministry of Public Health (Thailand)

Rome Buathong
Bureau of Epidemiology, Ministry of Public Health (Thailand)

Derrick Bary Abila
One Health Fellow

Rachel Lowe
London School of Hygiene & Tropical Medicine

CONTENTS

Page

List of Tables

Page

List of Figures

Page

ITU-T FG-AI4H Deliverable 10.10

FG-AI4H Topic Description Document for the Topic Group on outbreak detection (TG-Outbreaks)

1 Introduction

Disease outbreak detection describes a process usually found in the field of epidemiology that uses mathematical and/or computational methods to find salient, unusual patterns in health-related and associated data that hint to an outbreak. A disease outbreak is an excess of cases compared to what you would expect to observe. These cases can be related to exposure to a common source (e.g. close contact with an infected person or vector, exposure to contaminated food or, breeding site of disease transmitting insects). Early detection and response to outbreaks can substantially reduce their spread. Outbreaks that spread quickly and are hard to contain can still come in predictable patterns. Accurate outbreak detection helps to detect the build-up of such a wave quickly to ensure appropriate public health response.

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 will continue to emerge. 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, trend change, and other conspicuous events within data streams, pointing to the emergence of infectious disease outbreaks, in a fast and automatic manner. To this end, AI algorithms can increase the timeliness and accuracy of outbreak detection.

Additionally, disease outbreak algorithm development happens mostly in countries with a strong research infrastructure. Such algorithms may subsequently be biased towards the environment, endemic diseases, and infrastructure of these countries. In Europe, for example, an algorithm developed in the UK (namely., Farrington's algorithm) is used across other neighbouring countries with no public benchmark assessing them. It is more common to evaluate such algorithms on expert-generated synthetic data, which may not be representative. The development of a disease outbreak detection benchmarking would help to provide a low entry into testing and using outbreak detection algorithms regardless of available resources. Not only are developments of outbreak detection algorithms unevenly funded, but systemic disadvantages in civil and public health infrastructure make some nations at greater risk of inadequate sanitation and poor public health surveillance. This increases the likelihood and likely severity of an outbreak.

Safe sanitation remains inaccessible to over 50% of the world population, contributing to nearly 1 million deaths in low- and middle-income countries (World Health Organization, 2019a). Inadequate sanitation and unsafe water supply contribute to diarrhoeal disease, which is a leading cause of global childhood mortality and morbidity. Poor sanitation is estimated to have cost $260 billion in disruption to economic productivity and healthcare costs per year from 2012 to 2015 (Hutton, 2012).

We highlight a set of public health surveillance efforts designed to use AI informed analytics to detect disease outbreaks. This topic description document specifies the standardized benchmarking for sanitation systems. It serves as deliverable No.10.23 of the ITU/WHO Focus Group on AI for Health (FG-AI4H). Safe sanitation remains inaccessible to over 50% of the world population, contributing to nearly 1 million deaths in low- and middle-income countries (World Health

Organization, 2019a). Inadequate sanitation and unsafe water supply contribute to diarrhoeal disease, which is a leading cause of global childhood mortality and morbidity. Poor sanitation is estimated to have cost $260 billion in disruption to economic productivity and healthcare costs per year from 2012 to 2015 (Hutton, 2012).

We highlight a set of public health surveillance efforts designed to use AI informed analytics to detect disease outbreaks. This topic description document specifies the standardized benchmarking for sanitation systems. It serves as deliverable No.10.23 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

2 About the FG-AI4H topic group on outbreak detection for public health

The introduction highlights the potential of a standardized benchmarking of AI systems for outbreak detection to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Outbreaks at the meeting E in E in Geneva, June 2019

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. During Meeting G in New Delhi, 14 November 2019, Stéphane Ghozzi from the Helmholtz Centre for Infection Research and Auss Abbood from the Robert Koch Institute were nominated as topic drivers for the TG-Outbreaks. During Meeting L held virtually, May 2021, TG-Sanitation was established. Khahlil Louisy and Alexander Radunsky from ITGH were nominated as co-driver for the TG-Sanitation by FG-AI4H.

Meeting N, in Berlin, TG-Outbreaks and TG-Sanitation merged into a single Topic Group with Khahlil Louisy and Alexander Radunsky from ITGH and Auss Abbood from RKI remaining co-topic drivers and with Alexander Ullrich from RKI replacing Stéphane Ghozzi.

2.1 Documentation

This document is the TDD for the TG-Outbreak. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for outbreak detection for public health. It describes the existing approaches for assessing the quality of outbreak detection with a focus on sanitation systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable "DEL 10.23 Outbreaks (TG-Outbreaks)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable (Table 1) to each FG-AI4H meeting.

Table 1: Topic Group output documents

Number	Title
FGAI4H-O-028-A01	Latest update of the Topic Description Document of the TG-Sanitation
FGAI4H-M-028-A02	Latest update of the Call for Topic Group Participation (CfTGP)

Number	Title
FGAI4H-O-028-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Sanitation

The working version of this document can be found in the official topic group SharePoint directory.

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Sanitation.aspx
- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Outbreaks.aspx

2.2 Status of this topic group

The following subsections describe the update of the collaboration within the TG-Outbreaks for the official focus group meetings.

2.2.1 Status update for meeting J

- Work on this document
- Work on the benchmarking software
- Progress with data acquisition, annotation, etc.
- Overview of the online meetings including links to meeting minutes
- Relevant insights from interactions with other working groups or topic groups
- Partners joining the topic group
- List of current partners
- Relevant next steps
- Phone meeting with interested parties (Dec 2019)
- Further acquisition of members (Jan-Feb 2020)
- Review of existence methods and metrics and in disease outbreak detection and existing approaches for benchmarking or similar endeavours. (Mar 2020)
- Survey on how disease outbreak detection is done among our members (Feb-Mar 2020)
- Implementation of a new metric to test different families of outbreak detection algorithms (July 2020-)

2.2.2 Status update for meeting M

TG-Sanitation Outreach to potential partners is ongoing. We have drafted a Call for Participation and outlined areas of expertise we would be interested in incorporating in our focus group. We have Initial TG planning and group delegation of initial TDD tasks. The topic group has researched and written preliminary drafts for portions of sections 1, 2, 3, 4 and 8 of TDD.

2.2.3 Status update for meeting N

Based on interviews, literature reviews, and questioners, TG-Outbreaks crafted a preprint and developed a software library based on said work that would allow scoring outbreak detection algorithms with different aggregation and testing strategies. Since we found that the approaches common in outbreak detection as well as the data which depends on the surveillance strategy and disease vary, we needed a method to make algorithm performance comparable in order to properly proceed with our work in TG Outbreaks.

TG-Sanitation has begun 1) community engagement planning with eThakwini communities by UKZN team and Woodco, 2) sensor and data systems design testing and fielding by Woodco. We also will assess data availability of current and historical manually sampled data from the Palmiet River system as a potential source of training data along. We will also assess potential data collection methods and sources useful to detection of diarrheal disease outbreak. We have begun to further research potential sensors in the CAB: occupation sensors, water meters, and acoustic diarrheal sensors; and in the pyrolysis plant: faecal sludge moisture content, calorific values, heavy metal content, presence and severity of pathogenic contamination.

2.2.4 Status update for meeting O

TG-Sanitation has identified potential sensors for testing by Woodco partner, associated with the community ablution block (CAB) and the pyrolysis waste treatment facility. These are currently undergoing testing in Ireland. System assessment is being planned including the collection and storage of sensor data and performance data.

2.2.5 Status update for meeting S

We concluded the merging of both TDDs. It included filling gaps in the document and adopting the former TG Outbreaks and TG Sanitation objectives under a common narrative. With the Global Initiative in mind, we have started exploring possible partners to conduct implementation work with our topic group. For the benchmarking to be richer, we started creating more challenges following a data simulation approach. Relevant next steps are reaching out and discussing needs and interest for potential collaborations with the Global Initiative. Also, to conclude our benchmarking work, we plan to submit a paper describing our work for a technical audience.

2.3 Topic Group participation

The participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding 'Call for TG participation' (CfTGP) can be found here:

- https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Sanitation.pdf

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Sanitation.aspx
- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Outbreaks.aspx

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG 'zoom' link:

- https://itu.zoom.us/my/fgai4h

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the 'Call for Topic Group participation' and this link:

- https://itu.int/go/fgai4h/join

In addition to the general FG-AI4H mailing list, the following dedicated mailing list was used:

- fgai4htgoutbreaks@lists.itu.int

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

- https://itu.int/go/fgai4h

3 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in outbreak detection and how this can help to solve a relevant 'real-world' problem.

Topic Groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG- Outbreaks currently has no subtopics. Future subtopics for outbreak detection might be introduced.

This topic group has been approaching the objective of outbreak detection from two sides: TG Sanitation focused on the feasibility and usability of an on-site waste water surveillance system in South Africa, highlighting ethical and regulatory considerations. TG Outbreaks before the merging of both groups focused on the technical aspects of outbreak detection. As a result, this document follows to narratives in describing the topic group's work.

3.1 Definition of the AI task

Community and public data collection in eThekwini

There were opportunities to focus on planning stages for data collection of health event, environmental contamination data, weather, and watershed ecological data. Woodco, an Ireland-based sensor developer, and local partners at University of KwaZulu-Natal, have previously engaged with these communities in a set of informal settlements on the outskirts of eThekwini in South Africa. Although the burden of diarrhoeal disease is high, current capacity to detect these outbreaks and intervene is severely limited.

Community engagement and understanding around health and data privacy is a critical step in using some public and community sensors and other local sources of data. The ethical and regulatory considerations of this collection effort, especially in the context of highly marginalized and systematically disadvantaged communities, must be given sufficient consideration.

The primary output of interest is the incidence of diarrhoeal disease. Data collection is planned to include case counts and other local health data, ongoing testing for waterborne pathogens in local water systems, communal ablution block sensors, and pathogen testing in the waste treatment stream before and after pyrolysis treatment. This ground data is complemented by satellite EO, GNSS data, and weather data systems. These are to be collected in compliment with local data collection to predict and prevent diarrhoeal disease outbreak.

Summary of the solution for sanitation

The AI's ultimate goal is to enable stewardship of diarrhoeal and sanitation related health problems in communities with limited sanitation infrastructure. The system currently in development by our field partners will enable the generation of several data streams, whose frequency (weekly, daily, NRT) will evolve progressively as the roll out of the project advances.

The data thus collected will be — on top of being consolidated for basic analysis — fed into an algorithm to predict outbreaks of diarrhoeal disease in the community. As such, the task is expected to be a binary prediction. The geographical resolution of the same, the prediction window, and the exact FP/FN trade-off are expected to be defined during the course of the present FG.

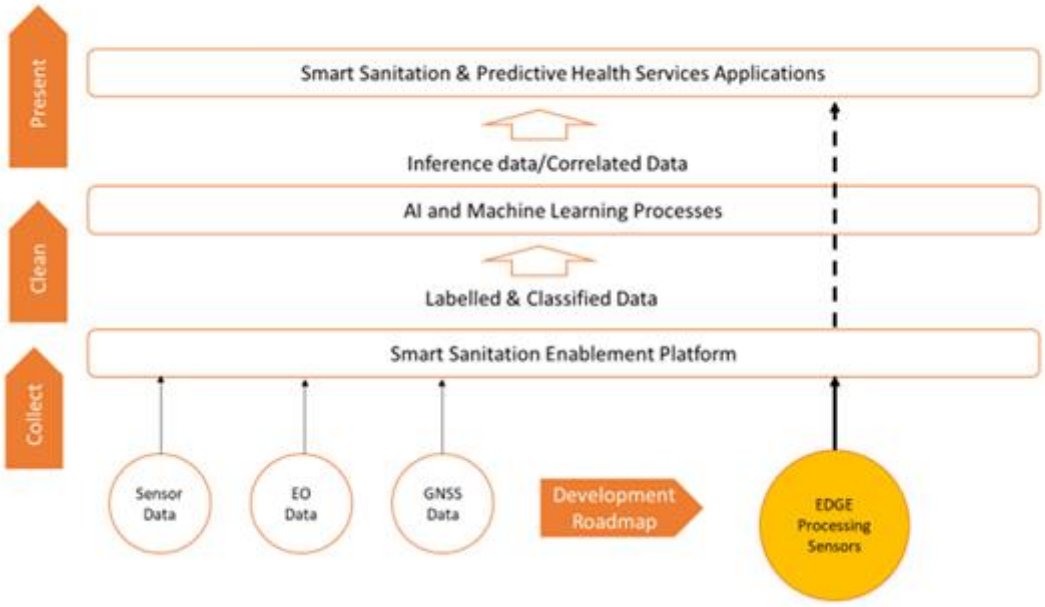

Figure 1: Solution architecture blocks

To detect signals in data streams like those produced by wastewater surveillance, there is a variety of published statistical and machine learning methods [1]–[3]. 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 during and after outbreak investigations by our experts. The main methods used by us are based on Hidden Markov Models and the improved Farrington method. We have already observed first improvements in the accuracy using ML approaches compared to classic statistical approaches [4]. In particular, keeping the same sensitivity in outbreak detection, the false alarms are considerably decreased using supervised learning. This reduces the number of alarms the experts have to assess.

Since the aforementioned approaches are time-series based, we expect the relevance of Hidden Markov Models and deep learning-based methods appropriate for sequential data such as Long Short Term Memory Networks (LSTM) or transformers to increase for the tasks of outbreak detection tasks. However, other methods like multivariate Bayesian regression or all-purpose deep learning (CNN, RNN) are conceivable, especially when variety of input modalities increases beyond the more common univariate time series.

Data streams

Disease surveillance and subsequently outbreak detection, traditionally operate on data created by medically sound diagnostic methods. Diagnostic capabilities, country-dependent disease and syndrome definitions, and the structure of the public health system influences the granularity and quality of the data sources. It can be said that a combination of different data streams is favourable as they allow to combine each other's strengths and counterbalance their weaknesses. Slow and reliable laboratory confirmed data can be combined with fast but informal information like news articles or social media activities. The COVID-19 pandemic produced and matured additional data streams such as satellite imagery to estimate deaths from dug graves, fitness tracker to track temperature and sleep disturbances indicating infections, and wastewater surveillance, allowing for a cheap, non-invasive but geographically comprehensive data stream. In TG-Outbreaks, we are piloting waste water surveillance in South Africa using different systems.

Sensors to detect presence of pathogens in faecal sludge, as well as acoustic-based diarrhoea detectors in Community Ablution Blocks (CAB's) are planned to be deployed on a pilot community in KwaZulu-Natal, South Africa. Signals from the sensors are edge-processed (using standard Raspberry Pi devices) and propagated primarily through standard LoRaWAN to central processing. These features are expected to provide small scale information about potential outbreaks. In the early stages of the project, the pathogen sensing technology will be replaced by frequent laboratory testing and manual input into the system.

Earth observation data from ESA missions Sentinel-5p (atmospheric composition) and Sentinel-3 (vegetation, water and moisture indices) provided by the European Space Agency allows the system to assess environmental and ecological changes including water chemistry, conditions at dumping sites, temperature changes. In combination with terrestrial sources for water level and turbidity at select sampling points of the basin, and weather observation data, we expect the system to capture weather patterns, water level, atmospheric conditions and land use (proxying for factors such as illegal dumping), and model their combined impact on disease propagation in the pilot communities.

Additional to the aforementioned streams, data from a sludge pyrolysis plant (including inflow / outflow measures as well as process KPIs), sanitation supply chain management data (CAB usage levels, consumables, sludge transport data) will provide a fuller picture of the state of the system, and may also be incorporated into the predictive model provided they add significant performance.

The combination of these data streams is expected to be used to identify the presence of disease-causing pathogens in water bodies in communities, and to serve as input for AI models that predict possible disease outbreaks based on those observations.

The data and findings from the analyses are published to a centralized platform that is accessible to health practitioners, equipping them with the knowledge required to make rapid decisions aimed at controlling the spread of any disease outbreaks.

The solution combines repurposed space technology to conduct ecological and environmental observations which is then combined with data from IoT sensors - acoustic in public toilets, from faecal sludge in sewage systems, and in water systems to detect the presence of disease-causing pathogens. Using these datasets, machine learning models and AI can be developed and trained to predict potential community disease outbreaks, when the conditions that are conducive to these phenomena converge. The data and results from the analyses are maintained in a global, centralized, and accessible platform with no government intervention, which is an important feature for communicating vital and valid information.

The combination of sanitation systems data and earth observation data to predict disease outcome is not currently practiced, yet we know that environmental and ecological changes may create the conditions necessary for diseases to incubate and propagate. Analysing faecal waste in community sewage systems also eliminates violating individual privacy. The availability of both ecological and faecal analysis data presents opportunities for researchers and health practitioners to utilize in their various approaches to understanding the nature of disease spread and their effects in communities.

3.2 Current gold standard

AI algorithms can increase the timeliness and accuracy of outbreak detection, and further have the potential to improve the understanding of the warnings and the disease spread itself. AI algorithms are particularly powerful in incorporating multiple data sources with diverse properties. The integration of high-quality data sources, from, e.g., mandatory reporting systems and laboratory tests, or wastewater surveillance is crucial to achieve earlier and more comprehensive detection of notifiable and non-notifiable pathogens. Different syndromic surveillance systems and valuable external data sources (google trends, health apps) can be incorporated. The gain of additional information on the underlying causes, by using explainable AI approaches, further enables for more

specific actions to be taken for prevention. More specifically, in the field of sanitation, statistical and AI methodology need to be linked with a community-wide understanding of prevention that cannot be replaced by algorithms.

Inadequate water, sanitation, and hygiene (WASH) is linked to water-borne illnesses such as cholera, intestinal worms and typhoid: diarrhoeal disease is implicated in the deaths of 297,000 children under 5 every year [5] and an economic burden estimated at over $12 billion [6]. These diseases are especially prevalent in communities with poorly developed sanitation systems and limited access to safe drinking water or toilets. Therefore, these communities face constant outbreaks of water-borne illnesses, leading to chronic malnutrition and ill-health in the local population. To mitigate the effect of these outbreaks, the WHO as well as other organisations have published clear guidelines to detect and manage outbreaks of water-related infectious diseases (WRID) [7], [8].These guidelines suggest that local health authorities constantly monitor the health of their community using a combination of makers directly assessing WRID (e.g. reports from healthcare providers) as well as more indirect markers (e.g. sale of antidiarrheal drugs, complaints of water quality, etc.). Based on these different markers, health authorities can rapidly detect and verify the outbreak of disease. Once identified, the authorities collect information about the spread of cases and generate hypotheses about the possible sources of outbreak. They then collect water or other specimens to validate their hypothesis, helping contain an outbreak.

These methods of detection and management have been successful in helping us rapidly identify the outbreak of WRIDs. For example, a recent study considering time to detection for any infectious disease outbreak in Africa from 2017 to 2019 showed WRIDs have the shortest median time to detection of just 2 days [9]. While these methods allow us to rapidly mount a response to disease outbreaks, they do not seem to allow predictive modelling of WRID outbreaks. This limitation in our current approach was highlighted in a recent CDC report where it was stated that it would be 'impossible to predict the type of contamination or illness prior to an outbreak' using our current methods [10].

3.3 Existing AI solutions

Currently, outbreak detection is performed using statistical models. Usually, input data produced by health authorities or hospitals are line lists, which are often too small for AI models. Even when using online text data like news articles or blogs, data is transformed into lines lists of numbers of documents containing certain keywords [11]. With the increase of LLMs, we can expect, however, a heavier use of AI models to, for example, analyse text data beyond keyword matching but on a semantic level.

The state is similar for sanitation-level outbreak detection. The more traditional monitoring of concentration levels of indicators for pathogen is a task that does not require AI models nor does it produce enough data for AI models to show their advantage. However, with more sensors and secondary data at hand, as described in this TDD, AI models will probably have an advantage detecting alerting changes in data that is otherwise hard to model like acoustic, weather, and pathogen concentration levels.

3.4 Subtopic

Pathogen specialization

One area of expanded focus is the application of these benchmarking tools for other developed algorithms. This expansion should include other datasets, other locations, other pathogens and other algorithms.

Further, because different pathogens are expected to behave differently, it may well be reasonable to differentiate food-borne (e.g., salmonella) and vector-borne diseases (such as Dengue). Potential differences in pathogenicity, social factors impactful to outbreak pattern, or differential impact on

the health system, may justify differentiation of benchmarking methodology, standards, algorithms, and data streams to function well.

Integrated genomic surveillance

Clearly missing in this topic group is the utilization of genomic data to aid outbreak detection. In cases where a pathogen's mutation rate and quality are well understood, outbreaks can be detected by linking genomic markers of pathogens across infected to retrace the course and potentially the source of an outbreak.

More prominently discussed due to COVID-19, is the use of routine sequencing data to detect the emergence of variants of concerns. International research efforts quickly described SARS-CoV 2's replication cycle and how the immune response in humans helps avoiding infections. This allowed bioinformaticians to model the likelihood of a new variant avoiding an immune response or due to changes in the genome responsible for the spike protein (an important physiological component for infecting a host cell).

4 Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable DEL01 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-Outbreaks.

The rapidly evolving field of AI and digital technology in the fields of public health raises a number of ethical, legal, and social concerns that ought to be considered urgently. They are discussed in deliverable DEL01 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-Outbreak.

Ethical determinations and recommendations for AI in outbreak detection must include ethical sustainability of the AI application in health, i.e., the ethical assessment of the risks and benefits raised by the introduction of the technology to address a public health crisis such as disease outbreak detection, analysis, mitigation, and communication.

Our project designed an ethical evaluation framework for the full deployment of AI in outbreak detection for an existing pathogen that can also be applied to a novel pathogen as well. Diverse datasets such as largescale standardized population level datasets, as well as publicly available GNSS data, local health system community health data, and environmental sensors were all considered in our ethical analysis of this challenging public health question. We consider the ethical implications of proposed data collection and use across these dimensions: 1) the quality of knowledge (evidence), 2) the quality of data, 3) privacy and 4) fairness. Our framework prioritizes conducting risk-assessment evaluation early in the design process. Early detection of potential problems is of high value, but perhaps just as important is a different-level risks analysis. While, indeed, benefits related to the potential of AI for social good in sanitation and outbreak detection have been clarified in section 1. (introduction), technical risks related to the specific ML model in use and the dataset collected need to be anticipated and addressed by the very first stage of the project design – and this task specifically pertains to the ethics' domain.

The ethical concerns related to the introduction of benchmarking AI in real-world outbreak detection scenarios can be related to 1) the **quality of knowledge** (evidence) that predictive ML systems can produce, i.e., the quality of correlations discovered by AI on the presence of pathogens and their relation to certain diseases' outbreak, as well as the disclosure of new potential environmental factors as specific causes of disease. But ML algorithms are probabilistic, and certainly not infallible [12]. Probabilistic algorithms are vulnerable to mistakes. Overfitting can find patterns where none exist (phenomenon also known as apophenia), and underfitting can overlook a

pattern where actually there is one [13]. In these cases, the evidence they produce is highly vulnerable to inaccuracy and without insight into training data and methods, the ability to evaluate this inaccuracy is severely limited. ML knowledge (evidence) can also be limited, as **inconclusive**: indeed, such models are probabilistic and therefore they rarely can posit causal relationships. These causal relationships are difficult to determine in almost all non-experimental conditions. Focus on non-causal indicators may distract attention from the underlying causes of a given disease, leading to focus on inaccurate or completely wrong indicators.

Beyond the ethical considerations and risks that can be raised by the model itself, other concern the **quality of data used to train the ML model and the insurgence of bias.** Indeed, algorithmic outcomes can only be as reliable as the data they are based on. The presence of bias in the input dataset or in the training dataset [14] of the ML model will produce wrong and misguided evidence. Unwanted bias can occur due to improper deployment of an algorithm. Consider transfer context bias: the problematic bias that can emerge when a functioning algorithm is used in a new environment. For example, if a research hospital's healthcare algorithm is used in a rural clinic and assumes that the same level of resources is available to the rural clinic as the research hospital, the healthcare resource allocation decisions generated by the algorithm will be inaccurate and flawed [15]. Other biases can occur in this context and can undermine the correct ML functioning [16]. Biases can emerge from an absence of sufficient representativeness of certain diseases for a model to learn the correct statistical pattern (minority bias). There are also biases depending on a lack of data of diseases related to members of protected groups; lack of data that makes an accurate prediction hard to render (missing data bias). Other biases might be due to availability of features that are less informative to render an accurate prediction; an example in healthcare ML consists of identifying melanoma from an image of a patient with dark skin may be more difficult (informativeness bias). Biases in ML's functioning can generate discriminatory knowledge which leads in turn to produce disparate impact (positive or negative) on one group of people rather than another (algorithmic discrimination and unfairness). This is specifically true when the dataset used to train ML algorithms reflect and can unintentionally exacerbate existing inequalities. Such flaws can make the evidence produced by ML **biased and misguiding**. Moreover, such knowledge is very often also opaque and therefore **inscrutable**, due the complexity of ML (models as black boxes). Indeed, very often, the probabilistic path ML develops to reach a certain prediction or decision by analysing data is not comprehensible to the human (expert) eye. This makes the detection of biases an extremely difficult task. This would also hamper public health decision-makers' validation and audit procedures of technology and the evidence it produces.

If such evidence is used – without precautionary assessment – by policy-makers and public institutions broadly to make decisions (e.g., how to allocate resources or how to implement measures to prevent the spread of certain diseases), it can lead to risks for the society at different levels. At the individual level, risks related to the previous concerns can be, for example, the wrong identification of certain disease causes in reference to a specific person or groups of people (a person or a community using public sanitation services can be wrongly identified as connected to the spread of certain disease and be blamed for that). This would cause massive or disproportionate health surveillance for certain people rather than other. This would entail privacy and autonomy infringements and also lead to phenomena of social injustice towards vulnerable groups, due to more severe profiling towards members of low-income communities (for example, as they use more public toilets).

At the society level, ethical risks related to the previous concerns can be, for example: an excessively broad data sharing between public and private entities (privacy issues); waste of funds and resources which are not directed to areas of greater need, therefore, to poorer public healthcare provision, worsening health outcomes, due to the use of inaccurate evidence; inequality in outcome due to the use on scale of biased evidence; as well as a low adoption and loss of trust on technology and public sanitation due to the use of inscrutable (or black box) ML.

Beyond the ethical implications of proposed data collection and use related to the quality of knowledge (evidence) and 2) the quality of data, we consider also the dimensions concerning data 3) privacy and 4) fairness. For the next phase of the project, specific privacy and fairness criteria to be met in order to carrying on our ethical assessment have been identified as critical aspects on which focusing further work in our TG. We specify such ethical criteria and use them to develop our ethical framework for AI in outbreak detection.

4.1 Privacy

About **privacy**: individuals' privacy is taken into account from the choice of the specific ML model to deploy for the predictive task. Highly advanced privacy-preserving technique, such as federated learning and/or split learning, will be deployed to drive ML functioning to safeguard users' privacy. Moreover, to be ethically justifiable, the project should meet the following privacy enabling factors: 1. the collection of users' data cannot be mandatory (it is always optional for the members of the communities involved accepting or not the profiling); 2. the collection of users' data requires the clear consensus of the participants (the community involved should have choice over what of their data is shared and when, as well as to be in the position to ask for removal); 3. privacy-preserving techniques deployed – as those above mentioned – should ensure that users' data is not re-identifiable. Furthermore, 4. the purpose of the data collection phase should be limited to a clearly defined scope (it can range from the sole prevention to a more influencing health-monitoring, but it needs to be declared from the beginning); 5. the scope definition and communication concern also the data collected and the correlations discovered for secondary uses and/or in combination with other/multiple data sources: these aspects should be made transparent and subject to users' and/or a public health ethics board's approval. Lastly, health data collected will be managed and stored according to the EU regulation (GDPR): as health data is labelled as "special category", its use can be limited to the sole scope of the project; this means that, for example, although datasets are anonymized, their sharing/selling with third-party entities outside the project will not be allowed.

4.2 Fairness

About **fairness**: a first step to operationalize fairness is based on choosing an ML model able to ensure at a minimum threshold three main criteria known as **distributive justice options** [17]:1) *equal outcomes*, i.e., the benefits produced from the deployment of ML models in terms of outcomes ought to be the same for protected and unprotected groups; 2) *equal performance*, i.e., performance and results of ML ought to be equally accurate for members belonging to protected and unprotected groups for such metrics as accuracy, sensitivity (*equal opportunity*), specificity (*equalized odds*), and positive predictive value (or *predictive parity* [1]); and 3) *equal allocation*, also called as *demographic parity* [18], i.e., the allocation of resources as decided by the model ought to be equal across groups and especially proportionally allocated to members of the protected group. The metric used to evaluate is the rate of positive predictions produced by ML for protected and unprotected groups. Further work on fairness in AI for sanitation and how to operationalize it will be developed in the next phase of the project.

These considerations constitute a first ethical compass to acknowledge and systematically analyse the major ethical issues connected to the use of AI for outbreak detection which underpin our ethics by design approach. In the next phase of the project, we will expand such analysis and our ethical risk assessment through the analysis of specific case studies in order to build specific guidelines for the responsible use of AI in outbreak detection along with an operationalizable ethical risk.

5 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and outbreak detection for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings

from previous benchmarking that could help to implement the benchmarking process in this topic group.

At RKI, we have running a small benchmarking setup occasionally to compare models:

– Mandatorily reported data in infections and pathogens in Germany was aggregated to weekly reported infection cases and cases being part of an outbreak

– Several outbreak detection algorithms operating on univariate data were trained on data of the past 5 years per diseases (exception may be necessary)

– Testing on was conducted on a held-out data set of, for example, a year following the training data set (the 6th year, if you like). Outbreak detection was applied to the next week under realistic conditions (prospective 1 week ahead: data available until last week)

– Models were compared using scores that are or comprised of functions using true/false positive/negative rate (TP, FP, TN, FN) like sensitivity, specificity, precision, F1 …

5.1.1 Publications on benchmarking systems

Existing work in benchmarking of outbreak detection algorithms in the literature is more closely described in our review *How to benchmark disease outbreak detection algorithms: A review,* which can be found on our ITU collaboration site.

5.1.2 Benchmarking by AI developers

All developers of AI solutions for outbreak detection implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

The most crucial insight in benchmarking outbreak detection is that (labelled) data is rare. First, the quantity of it low. Opposed to sensor and diagnostic data in medicine, which are indispensable tools of the daily work in a medical facility, surveillance of infectious diseases in a public health setting is focused on the rare and impactful diseases. Thus, by definition, we expect less data.

Second, the ground truth on outbreaks is most likely not known. In individual-level data, we tend to have less uncertainty when diagnosing a patient with a notifiable infectious disease. However, whether cases got infected by the same event or source and whether all affected by such an outbreak have been recorded by the health authorities is unknown. Only in rare occasions, outbreaks are well investigated and understood. This information, i.e., these labels are, however, not publicly available.

Thus, outbreak detection is often an unsupervised classification. The goal is to detect an anomaly. Due to the small size of available data, outbreak detection algorithms are usually more top-down, meaning, they have stronger assumption about the data generation process. This is in strong contract to AI models which will learn this process by being trained on vast amounts of data.

To bridge the lack of labels on outbreaks and low numbers of data, it is common to simulate timeseries and inject outbreaks using statistical methods. During our work we realized, that this procedure is not ideal. Our main insight was to not only use the parameters introduced in the literature but to use curve fitting to find a parameter set for the simulation models that will be close to your internal data.

As described in the document an in our internal review *How to benchmark disease outbreak detection algorithms: A review,* which can be found on our ITU collaboration site, we want to highlight how important specialized metrics are. Good performance tends to be measured much better by timeliness or the detection of prominent outbreaks rather than F1 score or accuracy which appear in more classical machine learning tasks.

5.1.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable DEL7.5 *"FG-AI4H assessment platform"* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

Given the sensitive nature of the data, it is unlikely that a benchmark will be hosted by a commercial platform. Also, most benchmarking platforms lack the possibility to ask for more qualitative features of the model. While performance of models can be well described using metrics, especially in a health setting, possibly even more so on a population-level, biases

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the TG Outbreaks AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: DEL5 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL5.1*"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL5.2 *"Data acquisition"*, DEL5.3 *"Data annotation specification"*, DEL5.4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL5.5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL5.6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL7*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL7.1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL7.2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL7.4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL7.5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL9 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL9.1 *"Mobile based AI applications,"* and DEL9.2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

The benchmarking of TG Outbreaks is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

Benchmarking in this document has been more focused on identifying the right data acquisition processes and metrics than on introducing a powerful algorithm. The task of outbreak detection is quite diverse. For example, data quality and the feasibility to achieve good results in outbreak detection will heavily depend on the disease and how data is obtained for this disease. (Public) health systems may monitor different diseases with different methods which would lead to an algorithm performing well in one setting not performing well in a different one.

6.1.1 Benchmarking version 1

This section includes all technological and operational details of the benchmarking process for the benchmarking version 1.

6.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version 1.

In this iteration, a very basic reimplementation of a benchmarking setup used in the literature is applied [19]. Later, a variant is introduced on how to obtain data for benchmarks that utilize real data that cannot be shared.

6.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version 1. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

At present, outbreak detection algorithms are commonly parametrized and benchmarked on small sets of data or on simulations. These simulations are mimicking infection counts with outbreak and capture only a few, well-known aspects of disease transmission, 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.

The architecture, the data flow, and other technical details are described within the focus group since we adhered to the internal health.aiaudit platform for our work. An example benchmark is uploaded and can be checkout at health.aiaudit.org.

6.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of TG Outbreaks. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

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 Trends, 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.

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 (EBS) 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 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 statistic can be used where closeness is evaluated given a pre-specified 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 like SaTScan. Other operate on aggregated timeseries such as CUSUM or regression models.

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 that it 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.

Even though the usage of real data might have clear disadvantages, such as being incomplete, which motivated the development of disease outbreak simulations, it is still desirable to utilize real data for the evaluation and training of disease outbreak algorithms as these are the data on which we will later apply outbreak detection 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.

6.1.1.4 AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking.

The output may be binary (outbreak or not) or a probability (like) score indicating the chance of an outbreak. It could also be a probability distribution if a Bayesian approach is used. In any case, the output will need to be created for a meaningful temporal, spatial, and demographic stratification. Most commonly, we want to make predictions for each disease, per week, and on the smallest geographical unit a country usually uses (e.g., a county).

6.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called 'labels') for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

Labels in this task usually indicate the occurrence of an outbreak. Individuals that got affected by the same outbreak may be assigned a common outbreak ID. These IDs can be assigned after experts have investigated an outbreak, conducted interviews, or even performed genomic sequencing on the pathogens that causes in infection in the affected individuals.

Labels can also be generated automatically using statistical methods that detect strong increases of case counts, bell like shapes in the data, or similar. The goal in this approach is to develop early warnings systems that are faster than the traditional surveillance systems such as the laboratory system. Detecting a strong increase in case counts in laboratory confirmed cases of influenza is well understood. Running an outbreak detection algorithm on influenza case counts is of not real advantage since the onset of an influenza wave is easily spotted event without algorithmic help. Other data sources can, however, offer a time advantage. To then save time labelling easily identifiable peaks in public health data, labels can be produced automatically.

6.1.1.6 Scores and metrics

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems.

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 the review How to benchmark disease outbreak detection algorithms: A review; located at the TG-Outbreaks collaboration site):

- 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 $1\text{-}s/D$ 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.
- Hitrate: If outbreak detection is applied forecast-based, we can calculate the number of equal signs between forecasts and real data, i.e., by looking at the sign of the difference of the last forecast to its predecessor and vice versa for the real data.

6.1.1.7 Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources, and storage.

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.

6.1.1.8 Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL5.5 on *data handling* and DEL5.6 on *data sharing practices*).

In Germany, there is no framework on how to share data. Acquiring data on notifiable diseases is regulated by a specific law. It is not like a clinical trial that has clear frameworks for data sharing of, e.g., anonymized data. In practice, it seems to be similar in other countries since there are hardly any labelled data sets on case count numbers and respective outbreak labels. A possible alternative is the aforementioned method of fitting a simulation model to non-sharable internal data. This possibility is explored at RKI right now. We achieved promising preliminary results by running a non-linear fitting algorithm (using the lmfit Python library) of labelled case count data on the simulation model described by Noufaily et al.[19]. The fitting is conducted in two steps. First, data without outbreak cases is fitted to the simulation model to find the best parameters describing the real data. To increase the chance to find a good fit, the trend is estimated from the data using a linear model and used as an initial value for the fitting routine. Second, the parameters from step on are used to run a second fit, this time, however, only the outbreak scaling factor is tuned. Outbreaks are seeded using a Markov chain. Whenever an outbreak occurs, the scaling factor determines how many more cases are observed compared to the number of endemic cases. Once the optimal parameters for the simulation models are found, data for the benchmarking can be produced.

Experiments to scale this approach have yet not been successful. In Germany, almost 100 pathogens and disease are notifiable to authorities. There are 412 counties in Germany which means that there are 41,200 timeseries that could possibly be used to curate a diverse data set for the benchmarking challenge. We have been experimenting with clustering to identify around 30 timeseries that best describe the set of timeseries to be expected in the German public health setting. However, results were not satisfactory before the deadline of submitting this TDD. Therefore, the proposed parameters for the simulation model from Noufaily et al. were used instead.

6.1.1.9 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance

of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

The baseline in this TDD is the set of timeseries proposed by Noufaily et al. [19]. It is a set of univariate timeseries. Thus, it misses to appreciate the spatio-temporal nature of infectious diseases. Also, the authors picked the parameters based on case counts observed in the UK. They may miss patterns observed for other countries and infectious diseases not notifiable or endemic in the UK. Using the simulation approach described in section 6.1.1.8, more diverse timeseries could be submitted to our benchmark.

6.1.1.10 Reporting methodology

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

The native reporting output of the health.aiaudit platform of this Focus Group is used.

6.1.1.11 Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

Data simulation were still ongoing during final submission for the TDD. Furthermore, approval to share simulated data was also not yet legally cleared. Thus, results will be shared at a later point in time.

7 Regulatory considerations

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices, though significant progress has been made within the last year, with the passing of legislation and agreements from the European Union and United States. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *"Regulatory considerations on AI for health" (WG-RC)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL2 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL2.1 *"Mapping of IMDRF essential principles to AI for health software"*, and DEL2.2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL04 identifies standards and best practices that are relevant for the "*AI software lifecycle specification."* The following sections discuss how the different regulatory aspects relate to the TG- Outbreaks

7.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for outbreak detection

The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly issued 10 guiding principles to inform the development of what they call Good Machine Learning Practice (GMLP), to help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AL/ML):

– Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle

– Good Software Engineering and Security Practices Are Implemented

– Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population

– Training Data Sets Are Independent of Test Sets

– Selected Reference Datasets Are Based Upon Best Available Methods

– Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device

– Focus Is Placed on the Performance of the Human-AI Team

– Testing Demonstrates Device Performance During Clinically Relevant Conditions

– Users Are Provided Clear, Essential Information

– Deployed Models Are Monitored for Performance and Re-training Risks Are Managed

The use of AI/ML and devices utilizing these advanced technologies may be exempt from FDA oversight under the 21st Century Cures Act which was enacted in December 2016 and which modified the Federal Food, Drug, and Cosmetic Act (FFDCA) to create the exemption. Clinical Decision Support (CDS) Software that meets the following criteria (under 21 USC § 360j(o)(1)(E)):

– Is not "intended to acquire, process, or analyse a medical image or a signal from an in vitro diagnostic device or signal acquisition system"

– Is intended for the purpose of "displaying, analysing, or printing medical information about a patient or other medical information (such as peer-reviewed clinical studies and clinical practice guidelines)"

– Is intended for the purpose of "supporting or providing recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition"

– Is intended for the purpose of "enabling such health care professional to independently review the basis for such recommendations that such software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient"

To meet the scope of this statutory CDS exemption, the software must be intended for use by a healthcare professional (HCP)—software intended for patient or consumer use is outside the scope of the exemption. For HCP applications software must be FDA approved for premarket 501(k) safety and effectiveness assessment.

This FDA Framework for Modifications to AI/ML-based SaMD is an internationally harmonized framework drawing from: The International Medical Device Regulators Forum (IMDRF) risk categorization principles, FDA's benefit-risk framework, risk management principles in the software modifications guidance, and the organization-based Total Product Life Cycle (TPLC) approach as envisioned in the Digital Health Software Pre-Certification (Pre-Cert) Program.

References

[1] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks: A review," *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 175, no. 1, pp. 49–82, 2012, doi: 10.1111/j.1467-985X.2011.00714.x.

[2] M. Yuan, N. Boston-Fisher, Y. Luo, A. Verma, and D. L. Buckeridge, "A systematic review of aberration detection algorithms used in public health surveillance," *J. Biomed. Inform.*, vol. 94, no. May 2018, p. 103181, 2019, doi: 10.1016/j.jbi.2019.103181.

[3] B. Allévius and M. Höhle, "Prospective Detection of Outbreaks," *Handb. Infect. Dis. Data Anal.*, pp. 411–436, 2019, doi: 10.1201/9781315222912-21.

[4] B. Zacher and I. Czogiel, "Supervised learning using routine surveillance data improves outbreak detection of Salmonella and Campylobacter infections in Germany," *PLOS ONE*, vol. 17, no. 5, p. e0267510, May 2022, doi: 10.1371/journal.pone.0267510.

[5] World Health Organization, "1 in 3 people globally do not have access to safe drinking water – UNICEF, WHO." https://www.who.int/news/item/18-06-2019-1-in-3-people-globally-do-not-have-access-to-safe-drinking-water-unicef-who (accessed Jun. 28, 2023).

[6] F. S. Alhamlan, A. A. Al-Qahtani, and M. N. A. Al-Ahdal, "Recommended advanced techniques for waterborne pathogen detection in developing countries," *J. Infect. Dev. Ctries.*, vol. 9, no. 02, Art. no. 02, Feb. 2015, doi: 10.3855/jidc.6101.

[7] World Health Organization. Regional Office for Europe and U. N. E. C. for Europe, *Surveillance and outbreak management of water-related infectious diseases associated with water-supply systems*. World Health Organization. Regional Office for Europe, 2019. Accessed: Jun. 28, 2023. [Online]. Available: https://apps.who.int/iris/handle/10665/329403

[8] "Waterborne Disease Outbreak Investigation Toolkit | Water, Sanitation, & Hygiene-related Emergencies & and Outbreaks | Healthy Water | CDC," Mar. 30, 2023. https://www.cdc.gov/healthywater/emergency/preparedness-resources/outbreak-response.html (accessed Jun. 28, 2023).

[9] B. Impouma *et al.*, "Measuring Timeliness of Outbreak Response in the World Health Organization African Region, 2017–2019 - Volume 26, Number 11—November 2020 - Emerging Infectious Diseases journal - CDC", doi: 10.3201/eid2611.191766.

[10] "Preparation for a Waterborne Disease Outbreak Investigation Waterborne Disease Outbreak Investigation Toolkit | CDC," Apr. 12, 2023. https://www.cdc.gov/healthywater/emergency/waterborne-disease-outbreak-investigation-toolkit/preparation.html (accessed Jun. 28, 2023).

[11] N. E. Kogan *et al.*, "An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time," *Sci. Adv.*, vol. 7, no. 10, p. eabd6989, Mar. 2021, doi: 10.1126/sciadv.abd6989.

[12] J. Morley, C. Machado, C. Burr, J. Cowls, M. Taddeo, and L. Floridi, "The Debate on the Ethics of AI in Health Care: A Reconstruction and Critical Review." Rochester, NY, Nov. 13, 2019. doi: 10.2139/ssrn.3486518.

[13] S. M. Liao, "Ethics of AI and Health Care: Towards a Substantive Human Rights Framework," *Topoi*, Apr. 2023, doi: 10.1007/s11245-023-09911-8.

[14] H. Shah, "Algorithmic accountability," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 376, no. 2128, p. 20170362, Aug. 2018, doi: 10.1098/rsta.2017.0362.

[15] D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems," pp. 4691–4697, 2017.

[16] B. Giovanola and S. Tiribelli, "Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms," *AI Soc.*, vol. 38, no. 2, pp. 549–563, Apr. 2023, doi: 10.1007/s00146-022-01455-6.

[17] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[18] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration".

[19] A. Noufaily, D. G. Enki, P. Farrington, P. Garthwaite, N. Andrews, and A. Charlett, "An improved algorithm for outbreak detection in multiple surveillance systems," *Stat. Med.*, vol. 32, no. 7, pp. 1206–1222, 2013, doi: 10.1002/sim.5595.

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
CfTGP	Call for topic group participation	
DEL	Deliverable	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MDR	Medical Device Regulation	
PII	Personal identifiable information	
SaMD	Software as a medical device	
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group Symptoms
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	