ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Focus Group Deliverable

(09/2023)

Focus Group on Artificial Intelligence for Health
(FG-AI4H)

FG-AI4H DEL10.10

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

ITU-T FG-AI4H Deliverable

DEL10.10 – FG-AI4H Topic Description Document for the Topic Group on outbreak detection (TG-Outbreaks)

Summary

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

Its primary purpose is to specify a standardized benchmarking framework for artificial intelligence algorithms used in public health for detecting disease outbreaks. The document covers various essential aspects, including the definition of the AI task, ethical and regulatory considerations, existing benchmarking work, and the proposed benchmarking methodology of the topic group, aiming to create a basis for evaluating and comparing AI solutions in this critical area of public health.

Keywords

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

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

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
E-mail: carnovalemaria@outlook.com

Augusto Gesualdi
ITGH,
USA
E-mail: augusto.gesualdi@pathcheck.org

Khahlil Louisy
ITGH,
USA
E-mail: klouisy@hks.harvard.edu

Gokul Parameswaran
ITGH,
USA
E-mail: gokul.parameswaran@keble.ox.ac.uk

Rebecca Perez
ITGH,
USA
E-mail: rebecca.perez@wadham.ox.ac.uk

Alex Radunsky
ITGH,
USA
E-mail: alex.radunsky@mail.harvard.edu

Simona Tiribelli
ITGH,
USA
E-mail: 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

Table of Contents

ITU-T FG-AI4H Deliverable

DEL10.10 – 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 or computational methods to find salient, unusual patterns in health-related and associated data that indicate an outbreak. A disease outbreak is an excess of case numbers compared to that expected. 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 continue to emerge. The risk and occurrence of infectious diseases is influenced by globalization, migration and climate change. According to the World Health Organization (WHO), infectious diseases are ranked in the top 10 causes of death worldwide[1].

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, artificial intelligence (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 United Kingdom (UK) (namely the Farrington algorithm) [1] 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 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 also systemic disadvantages in civil and public health infrastructure place 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.[7] 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 by the World Bank to have cost $260 billion in disruption to economic productivity and healthcare costs per year from 2012 to 2015[2].

---

1 https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (visited 2025-04-02).

2 https://blogs.worldbank.org/en/water/what-costs-the-world-260-billion-each-year (visited 2025-04-02).

A set of public health surveillance efforts designed to use AI informed analytics to detect disease outbreaks is highlighted. This Topic Description Document (TDD) specifies standardized benchmarking for sanitation systems. It serves as deliverable DEL10.10 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

Clause 1 highlights the potential of a standardized benchmarking of AI systems for outbreak detection to help solve 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 meeting E in Geneva, Jun 2019.

FG-AI4H assigns a topic driver (similar to a moderator) to each Topic Group (TG) who coordinates the collaboration of all TG members on the TDD. During Meeting G in New Delhi, 14 Nov 2019, Stéphane Ghozzi from the Helmholtz Centre for Infection Research and Auss Abbood from the Robert Koch Institute (RKI) 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-drivers for the TG-Sanitation by FG-AI4H.

At meeting N, in Berlin, TG-Outbreaks and TG-Sanitation merged into a single TG 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 deliverable is the TDD for TG-Outbreaks. It introduces the health topic including the AI task, outlines its relevance and the potential impact that 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 to be relevant for setting up 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 clauses for all versions of benchmarking. Finally, it summarizes the results of the TG 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 TG 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 10.10 Outbreaks (TG-Outbreaks)". The TG 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)
FGAI4H-O-028-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Sanitation

The working version of this deliverable can be found in the official TG SharePoint directories.

- 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

Clauses 2.2.1 to 2.2.5 describe the update of the collaboration within TG-Outbreaks for the official focus group meetings.

2.2.1 Status update for meeting J

- Work on this deliverable
- Work on benchmarking software
- Progress with data acquisition, annotation, etc.
- Overview of online meetings including links to meeting minutes
- Relevant insights from interactions with other working groups or TGs
- Partners joining the TG
- 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 onwards)

2.2.2 Status update for meeting M

TG-Sanitation Outreach to potential partners is ongoing. A Call for Participation has been drafted and areas of expertise of interest outlined for incorporation into the focus group. Initial TG planning and group delegation of initial TDD tasks were done. The TG has researched and written preliminary drafts for portions of clauses 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 drafted 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 it was found that the approaches common in outbreak detection as well as the data that depend on the surveillance strategy and disease vary, a method was needed to make algorithm performance comparable in order to properly proceed with work in TG-Outbreaks.

TG-Sanitation has begun 1) community engagement planning with eThakwini communities by the University of KwaZulu Natal team and Woodco (an Ireland-based sensor developer); 2) sensor and data systems design testing and fielding by Woodco. Availability will also be assessed of current and historical manually sampled data from the Palmiet River system as a potential source of training data. Potential data collection methods and sources useful to detection of diarrheal disease outbreak will also be assessed. Further research has begun into potential sensors in a communal ablution block (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, associated with the 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, exploration of possible partners to conduct implementation work with this TG has been started. For benchmarking to be richer, creation of more challenges following a data simulation approach was started. Relevant next steps are reaching out and discussing needs and interest for potential collaborations with the Global Initiative. Also, to conclude benchmarking work, submission of a paper describing our work for a technical audience is planned.

2.3 Topic Group participation

- Participation in both FG-AI4H and in a TG is generally open to anyone (with a free ITU account). For this TG, applicants respond to the following CfTGP: https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Sanitation.pdf

Each TG also has a corresponding subpage on the ITU collaboration site: 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 TG, 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 CfTGP 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 clause 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.

TGs 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 a TG to pursue different topic-specific fields of expertise. The TG-Outbreaks currently has no subtopics. Future subtopics for outbreak detection might be introduced.

This TG has been approaching the objective of outbreak detection from two sides: TG-Sanitation focused on the feasibility and usability of an on-site wastewater 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 deliverable follows two narratives in describing TG 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 and local partners at the 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, CAB sensors and pathogen testing in the waste treatment stream before and after pyrolysis treatment. This ground data is complemented by satellite Earth observation (EO) and global navigation satellite system (GNSS) data and weather data systems. These are to be collected to complement local data to predict and prevent diarrhoeal disease outbreaks.

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, near-real time) 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 false positive/false negative (FP/FN) trade-off are expected to be determined during the course of the present FG. See Figure 1.

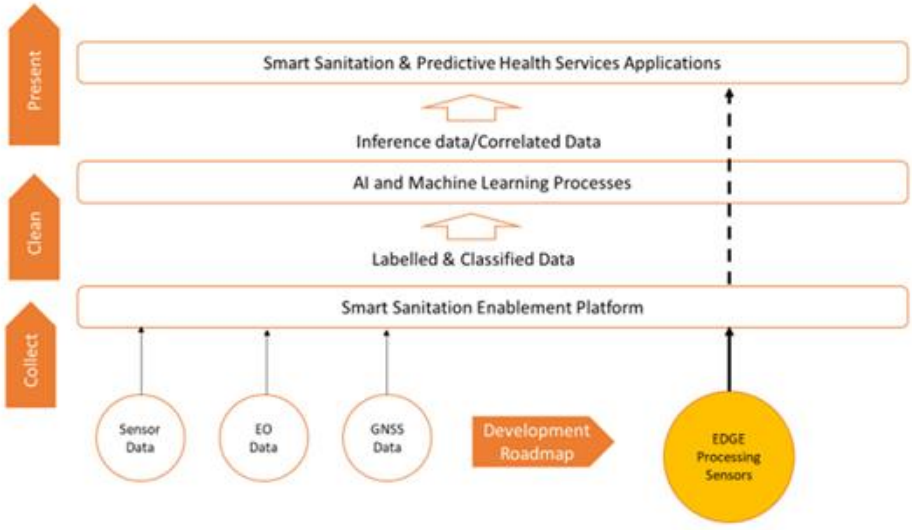

Figure 1 – Solution architecture blocks

To detect signals in data streams like those produced by wastewater surveillance, there is a variety of published statistical and machine-learning (ML) methods [1-3]. The RKI applies both classical statistical as well as supervised learning methods to the problem of outbreak detection. These ML methods use outbreak labels, assigned during and after outbreak investigations by our experts. The main methods used by RKI are based on hidden Markov models and the improved Farrington method. First improvements have already been observed in the accuracy using ML approaches compared to classic statistical approaches [4]. In particular, by maintaining the same sensitivity in outbreak detection, the number of false alarms is considerably decreased using supervised learning, reducing the need for expert assessment.

Since the aforementioned approaches are time-series based, it is expected that the relevance of hidden Markov models and deep learning-based methods appropriate for sequential data, such as long short term memory (LSTM) networks or transformers, to increase for outbreak detection tasks. However, other methods like multivariate Bayesian regression or all-purpose deep learning (convolutional neural network, recurrent neural network) 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, as well as the structure of the public health system influence the granularity and quality of data sources. It can be said that a combination of different data streams is favourable as it allows their strengths to be combined and weaknesses to be counterbalanced. 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 graves dug, 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, wastewater surveillance in South Africa is being piloted using different systems.

Sensors to detect presence of pathogens in faecal sludge, as well as acoustic-based diarrhoea detectors in CABs are planned for deployment in a pilot community in KwaZulu-Natal, South Africa. Signals from the sensors are edge processed (using a commercially available miniaturized computer) and propagated primarily through standard a long-range wide area network to central processing. These features are expected to provide small-scale information about potential outbreaks. In the early stages of the project, pathogen-sensing technology will be replaced by frequent laboratory testing and manual input into the system.

EO data from European Space Agency missions Sentinel-5p (atmospheric composition) and Sentinel-3 (vegetation, water and moisture indices) 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, the system is expected 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 or outflow measures, as well as process key performance indicators), 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 on 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 that are then combined with data from Internet of things 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, ML 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 data from sanitation systems and EO to predict disease outcome is not currently practised, yet certain environmental and ecological changes are known to create the conditions necessary for diseases to incubate and propagate. Analysing faecal waste in community sewage systems also eliminates violation of individual privacy. The availability of both ecological and faecal analysis data presents utilization opportunities for researchers and health practitioners in their various approaches to understand 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 understanding of warning indications and disease spread itself. AI algorithms are particularly powerful in incorporating multiple data sources with diverse properties. The integration of high-quality data sources, e.g., from 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 (e.g., web trends, health apps) can be incorporated. The gain of additional information on the underlying causes, by using explainable AI approaches, further enables 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 is linked to water-borne illnesses such as cholera, intestinal worms and typhoid: diarrhoeal disease is implicated in the deaths of 297 000 children aged under 5 years 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 organizations have published clear guidelines to detect and manage outbreaks of water-related infectious diseases (WRIDs) [7][8].These guidelines suggest that local health authorities constantly monitor the health of their community using a combination of makers directly assessing WRIDs (e.g., reports from healthcare providers) as well as more indirect markers (e.g., sale of antidiarrheal drugs, complaints of water quality). 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 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 the

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, outbreaks are detected 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 large language models, heavier use can be expected of AI models, e.g., to analyse text data beyond keyword matching but on a semantic level.

The situation 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 in 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 (e.g., 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 TG is the utilization of genomic data to aid outbreak detection. If the mutation rate and quality of a pathogen are well understood, outbreaks can be detected by linking genomic markers of pathogens amongst those 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 concern. International research efforts quickly described the replication cycle of SARS-CoV 2 and how the immune response in humans helps avoid infection. 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 DEL1 (see clause 6), developed by the working group on Ethical considerations on AI4H (WG-Ethics). This clause refers to DEL1 and should reflect the ethical considerations of the TG-Outbreaks.

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, e.g., disease outbreak detection, analysis, mitigation and communication.

This project designed an ethical evaluation framework for the full deployment of AI in outbreak detection for an existing pathogen that can 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 the ethical analysis of this challenging public health question. The ethical implications are considered 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. The framework prioritizes risk assessment in the design process. Early detection of potential problems is of high value, but perhaps just as important is analysis of risks at different levels. While, indeed, benefits related to the potential of AI for social good in sanitation and outbreak detection have been clarified in clause 1, 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 ethical 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 disease outbreaks, 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]. Overfitting can find patterns where none exist (a phenomenon 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, another concern is 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 on which they are based. 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 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 is the identification of melanoma from an image of a patient with dark skin, which may be more difficult than one with light colour (informativeness bias). Biases in ML functioning can generate discriminatory knowledge that 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 to 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-maker validation and audit procedures of technology and the evidence it produces.

If such evidence is used – without precautionary assessment – by policymakers 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 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 erroneously identified as connected to the spread of a certain disease and be blamed for that). This would cause massive or disproportionate health surveillance for certain people rather than others. 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 (e.g., because they use more public toilets).

At the society level, ethical risks related to the previous concerns can be, for example: excessively broad data sharing between public and private entities (privacy issues); waste of funds and resources that are not directed to areas of greater need, leading therefore to: poorer public healthcare provision and 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 in 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 the quality of data, dimensions are also considered of data privacy and fairness. For the next phase of the project, the specific privacy and fairness criteria that need to be met for ethical assessment have been identified as critical aspects on which to focus further work in this TG. Such criteria are specified and used to develop our ethical framework for AI in outbreak detection.

4.1 Privacy

Individual privacy is taken into account from the choice of the specific ML model to be deployed for the predictive task. Highly advanced privacy-preserving techniques, such as federated or split learning, will be deployed to drive ML functioning to safeguard user privacy. Moreover, to be ethically justifiable, the project should meet the following privacy enabling factors: 1. the collection of user data cannot be mandatory (it is always optional for the members of the communities involved accepting or not the profiling); 2. the collection of user data requires the clear consensus of participants (the community involved should have choice over which of their data is shared and when, as well as having the right to ask for removal); 3. privacy-preserving techniques deployed – like those previously mentioned – should ensure that user 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 or in combination with other or multiple data sources – these aspects should be made transparent and subject to user or a public health ethics board approval. Lastly, health data collected will be managed and stored according to the European Union (EU) General Data Protection Regulation: 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 or selling with third-party entities outside the project is not allowed.

4.2 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); 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 that underpin the ethics by design approach. In the next phase of the project, such analysis and ethical risk assessment will be expanded 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 clause focuses on 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 TG.

RKI has been running a small benchmarking setup occasionally to compare models as follows.

– Mandatorily reported data in infections and pathogens in Germany were aggregated to weekly numbers of infection cases reported 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 disease (exception may be necessary).

– Continuing testing was conducted on a protracted data set derived from, for example, a given year (e.g., the sixth) following the training data set. 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 or false positive or negative rates (TP, FP, TN, FN) like sensitivity, specificity, precision, F_1-score.

5.1.1 Publications on benchmarking systems

Existing work in benchmarking of outbreak detection algorithms in the literature is more closely described in *How to benchmark disease outbreak detection algorithms: A review*; located on the TG-Outbreaks collaboration site.

5.1.2 Benchmarking by AI developers

All developers of AI solutions for outbreak detection implemented internal benchmarking systems to assess performance. This clause outlines the insights and learning from this work of relevance for benchmarking in this TG.

The most crucial insight in benchmarking outbreak detection is that (labelled) data are rare. First, their quantity is 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 rare and impactful diseases. Thus, by definition, fewer data are expected.

Second, the ground truth on outbreaks is most likely not known. In individual-level data, there tends to be less uncertainty when diagnosing a patient with a notifiable infectious disease. However, whether cases were 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 on rare occasions are outbreaks 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 number of available data, outbreak detection algorithms are usually more top-down, meaning, they have stronger assumptions about the data generation process. This is in stark contrast to AI models that 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 time series and inject outbreaks using statistical methods. During this work it was realized that this procedure is not ideal. The main insight was to use not only the parameters introduced in the literature, but also curve fitting to find a parameter set for the simulation models that will be close to the internal data.

As described in *How to benchmark disease outbreak detection algorithms: A review*; located at the TG-Outbreaks collaboration site, it is desired 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 F_1-score or accuracy that appear in more classical ML 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 clause reflects on the different existing options that are relevant to this TG and includes considerations of using the assessment platform described in DEL7.5 (see clause 6), which explores implementation options that can be used to evaluate AI for health for the different TGs).

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 are likely to still be present.

6 Benchmarking by the Topic Group

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

Ethics

- DEL1 (2022): *Ethics and governance of artificial intelligence for health*

Regulatory

- DEL2 (2022): *Regulatory considerations on artificial intelligence for health*
- DEL2.1: *Mapping of IMDRF essential principles to AI for health software* (Pre-published)
- DEL2.2 (2022): *Good practices for health applications of machine learning: Considerations for manufacturers and regulators*

Technical

- DEL0.1 (2022): *Common unified terms in artificial intelligence for health*
- DEL3: *AI4H requirement specifications* (Pre-published)
- DEL4: *AI software life cycle specification* (Pre-published)
- DEL5.1: *Data requirements* (Pre-published)
- DEL5.3: *Data annotation specification* (Pre-published)
- DEL5.4: *Training and test data specification* (Pre-published)
- DEL5.5: *Data handling* (Pre-published)

- DEL6: *AI training best practices specification* (Pre-published)
- DEL7: *Artificial intelligence for health evaluation considerations* (Pre-published)
- DEL7.2: *Artificial intelligence technical test specification* (Pre-published)

Clinical evaluation and use cases

- DEL7.4: *Clinical evaluation of AI for health* (Published)
- DEL10: *AI4H use cases: Topic Description Documents* (Pre-published)
- DEL10.2: *FG-AI4H Topic Description Document for the Topic Group on AI-based dermatology (TG-Derma)* (Pre-published)
- DEL10.4: *FG-AI4H Topic Description Document for the Topic Group on falls among the elderly (TG-Falls)* (Pre-published)
- DEL10.6: *FG-AI4H Topic Description Document for the Topic Group on malaria detection (TG-Malaria)* (Pre-published)
- DEL10.7: *FG-AI4H Topic Description Document for the Topic Group on maternal and child health (TG-MCH)* (Pre-published)
- DEL10.8: *FG-AI4H Topic Description Document for the Topic Group on neurological disorders (TG-Neuro)* (Pre-published)
- DEL10.9: *FG-AI4H Topic Description Document for the Topic Group on AI for ophthalmology (TG-Ophthalmo)* (Pre-published)
- DEL10.10: *FG-AI4H Topic Description Document for the Topic Group on outbreak detection (TG-Outbreaks)* (Pre-published)
- DEL10.12: *FG-AI4H Topic Description Document for the Topic Group on AI for radiology (TG-Radiology)* (Pre-published)
- DEL10.14: *FG-AI4H Topic Description Document for the Topic Group on symptom assessment (TG-Symptom)* (Pre-published)
- DEL10.15: *FG-AI4H Topic Description Document for the Topic Group on tuberculosis (TG-TB)* (Pre-published)
- DEL10.17: *FG-AI4H Topic Description Document for the Topic Group on dental diagnostics and digital dentistry (TG-Dental)* (Pre-published)
- DEL10.20*: FG-AI4H Topic Description Document for the Topic Group on AI for endoscopy (TG-Endoscopy)* (Pre-published)
- DEL10.21: *FG-AI4H Topic Description Document for the Topic Group on musculoskeletal medicine (TG-MSK)* (Pre-published)
- DEL10.23: *FG-AI4H Topic Description Document for the Topic Group on AI for traditional medicine (TG-TM)* (Pre-published)
- DEL10.24: *FG-AI4H Topic Description Document for the Topic Group on AI-based point-of care diagnostics (TG-POC)* (Pre-published)
- TG-Dental Output 1: *Artificial intelligence in dental research: A checklist for authors and reviewers* (Pre-published)
- TG-Dental Output 2: *Artificial intelligence for oral and dental healthcare: Core education curriculum* (Pre-published)
- TG-Dental Output 3: *Ethical considerations on artificial intelligence in dentistry: A framework and checklist* (Pre-published)
- DT4HE Output 1: *Guidance on AI and digital technologies for COVID health emergency* (Pre-published)

The benchmarking of TG-Outbreaks will be further developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This clause outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to clause 6.1.1, where the actual benchmarking methodology for each version is described.

Benchmarking in this deliverable is 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 doing so in another.

6.1.1 Benchmarking version 1

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

6.1.1.1 Overview

This clause 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 clause provides details about the methods of 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 mimic 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. Most detection algorithms are, however, naturally developed at the national level. Thereby, each country relies on individual national disease surveillance systems.

To create standardized benchmarking for output detection algorithms, this TG aims to address all aspects that are relevant and shared across countries.

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

6.1.1.3 AI input data structure for the benchmarking

This clause describes the input data provided to 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 clause addressing this topic. Therefore, the description needs to be complete and precise. This clause 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 that 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 (such as Wikipedia, Google Trends, HealthTweets, X) or data from symptom-assessment apps, healthcare providers, hotlines etc. Real time data sources have high potential significantly to improve 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 easily run different algorithms that are in focus, describing benchmarking on IBS data. EBS data lack structure by definition and therefore, it is hard to adjust benchmarking to all possible forms they can assume.

Although more structured, IBS data still come in different shapes that 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 5 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, etc.). The choice of algorithms here, 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. Others, such as CUSUM or regression models, operate on aggregated time series.

If a data format were agreed upon, it would still be necessary to determine the source of these 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 ascertain 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 behaviour 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 time series for different pathogens. Outbreaks are then generated using a Markov process to selected weeks as outbreak weeks. In such outbreak weeks, a realization of a Poisson distribution with a mean equal to a chosen constant is added. The added cases are distributed over the outbreak week given a log-normal 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 outbreak detection algorithms will later be applied.

A straightforward approach to train or test an outbreak detection algorithm is to use real data where outbreaks are labelled by epidemiologists. Downsides of this method are that not all outbreaks are recognized by epidemiologists, sometimes only the reporting data and not the data of infection is known, or the data are subject to reporting delays that 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 time series on which outbreak detection happens in the 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 them with the realization of a negative binomial distribution using a lower expected value than those removed. This way, extreme values, considered as outliers, are removed and we get two time series, one with and one without outliers or /extreme values. These two time series of endemic and epidemic case counts are reunited with the epidemic outbreak time series being shifted by 1 year into the future, incorporating knowledge about the seasonality of the disease of interest, to create new labelled time series from real data.

6.1.1.4 AI output data structure

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

While AI systems can only receive the input data described in previous clauses, the benchmarking system needs to know the expected correct answers (sometimes called labels) for each element of the input data, so that it can compare the expected with the actual AI output. Since this is only needed for benchmarking, it is encoded separately. The details are described in clause 6.1.1.6.

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

Labels can also be generated automatically using statistical methods that detect strong increases of case counts, bell-like curves in the data, or similar. The goal in this approach is to develop early warning systems that are faster than traditional surveillance, such as the laboratory system. Detecting a strong increase in counts in laboratory-confirmed cases of influenza is well understood. Running an outbreak detection algorithm on influenza case counts is not a real advantage, since the onset of an influenza wave is easily spotted, even without algorithmic help. Other data sources can, however, offer a time advantage. To then save time by 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 clause describes the scores and metrics used to measure the performance, robustness and general characteristics of submitted AI systems.

When measurement of the performance of an algorithm is desired, criteria might be sought, such as usefulness, cost, sensitivity, representativeness, timeliness, simplicity, flexibility and acceptability. These are measures that include not only statistical algorithms, but also the more general criteria for

public health systems. Common measures for the comparison of statistical algorithms (more closely described in [20]) are as follows.

- Sensitivity.
- Specificity.
- Precision.
- Negative predictive value.
- F_1-score.
- Receiver operating characteristic (ROC)/area under curve.
- ROC using a timeliness measure where a minimum timeliness D is specified such that outbreaks must be detected within $t + D$ with t being the time point when an outbreak starts. Let s be the time point where an outbreak starts, then $1 - s/D$ replaces the false positive rate in our ROC curve. This timeliness measure is specified to be not less than 0.
- ROC where a normalized measure is used to correct time elapsed since the start of an outbreak. This might be important to compare time series with various time granularity measures. Such a method could be used to count the time steps elapsed since an outbreak, where a time step is determined by granularity or some other criterion.
- Instead of replacing an axis on the ROC, we can add a third dimension 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 the proportion is calculated of counts detected by an algorithm within an outbreak as being extreme.
- One extension of the POD is the scaled POD, which takes the size of the outbreak detected into account by weighting its amount with its size, i.e. the number of cases belonging to an outbreak.
- Another timeliness measure is the average time before detection. It is the sum of all outbreaks detected by an algorithm multiplied by the time elapsed since an outbreak, normalized by the overall number of outbreaks.
- A variation of the average time before detection that corrected 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 time series, 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.
- Hit rate: If forecast-based outbreak detection is applied, the number of equals signs between forecasts and real data can be calculated, 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.

Data from the German mandatory reporting system, collected since 2001 at the RKI, contains 8 million infectious disease cases and undergoes constant data quality checks by data engineers and review by epidemiologists. The data contain expert labels indicating which cases are related to specific disease outbreaks. All data are collected via a web service and stored in a structured relational structured query language database. The data arrives pseudonymized from about 400 local health agencies. For each case, information is given on the pathogen, demographics (age, sex), location (Nomenclature of Territorial Units for Statistics-3 level, county) and additional features such as

hospitalization, fatality and affiliation with care facilities and others. Some data are 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 deliverable, this set is referred to as German SurvNet data.

6.1.1.8 Data sharing policies

This clause 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 and DEL5.6 in clause 6).

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, e.g., of anonymized data. In practice, the situation seems to be similar in other countries since there are hardly any labelled data sets on case count numbers and corresponding outbreak labels. A possible alternative is the aforementioned method of fitting a simulation model to non-sharable internal data. This possibility is under exploration at RKI at the time of publication. Promising preliminary results have been achieved 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 one 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 not yet been successful. In Germany, almost 100 pathogens and diseases are notifiable to authorities. There are 412 counties in Germany, which means that there are 41 200 time series 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 time series that best describe the set of time series to be expected in the German public health setting. However, results were not satisfactory before the submission deadline of this TDD. Therefore, the proposed parameters for the simulation model from Noufaily et al. [19] 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 time series proposed by Noufaily et al. [19]. It is a set of univariate time series. Thus, it fails to take into account the spatiotemporal nature of infectious diseases. Also, the parameters were chosen 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 clause 0, more diverse time series could be submitted to the benchmark chosen.

6.1.1.10 Reporting methodology

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

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

6.1.1.11 Result

This clause gives an overview of the results from runs of this benchmarking version of this topic. Even if this TG prefers an interactive drill-down rather than a leaderboard, a context of common interest is chosen to give some examples.

Data simulation was still ongoing during final submission for the TDD. Furthermore, approval to share simulated data has also not yet been legally cleared. Thus, results will be shared after publication.

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 by the EU and USA. 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 requirements that consider these challenges.

The deliverables with relevance for regulatory considerations (see clause 6) are DEL2, DEL2.1, and DEL2.2 (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). DEL4 identifies relevant standards and best practices for AI software lifecycle specification. Clause 7.1 discusses how the different regulatory aspects relate to TG-Outbreaks.

7.1 Existing applicable regulatory frameworks

Most AI systems that are part of the FG-AI4H benchmarking process can be classified as software as medical device (SaMD) and are 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. This section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for outbreak detection.

The US Food and Drug Administration (FDA), Health Canada, and the UK Medicines and Healthcare products Regulatory Agency have jointly issued 10 guiding principles to inform the development of what they call good machine-learning practice, to help promote safe, effective and high-quality medical devices that use AI and ML (AI-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 with 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[3] and which modified the *Federal Food, Drug, and Cosmetic Act* to create the exemption. Clinical decision support (CDS) software that meets the following criteria (under 21 USC § 360j(o)(1)(E)[4]):

– 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[5] is an internationally harmonized framework drawing from: the International Medical Device Regulators Forum (IMDRF) risk categorization principles, FDA benefit-risk framework, risk management principles in the software modifications guidance and the organization-based total product life cycle approach as envisioned in the Digital Health Software Pre-Certification (Pre-Cert) Program[6].

3 https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act (visited 2025-04-02).

4 https://uscode.house.gov/view.xhtml?req=granuleid:USC-1994-title21-section360j&num=0&edition=1994 (visited 2025-04-02).

5 https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (visited 2025-04-02).

6 https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-pilot-program (visited 2025-04-02).

References

[1] Unkel, S., Farrington, C.P., Garthwaite, P.H., Robertson, C., Andrews N. (2011). *Statistical methods for the prospective detection of infectious disease outbreaks*: A review. *J. R. Stat. Soc. Ser. A Stat. Soc.* **175**(1), pp. 49-82. DOI: 10.1111/j.1467-985X.2011.00714.x.

[2] Yuan, M. Boston-Fisher, N., Luo, Y., Verma, A., Buckeridge, D.L. (2019). *A systematic review of aberration detection algorithms used in public health surveillance*. *J. Biomed. Inform.* **94**, 103181. DOI: 10.1016/j.jbi.2019.103181.

[3] Allévius, B., Höhle, M. (2019). *Prospective detection of outbreaks*. In: Held, L., Hens, N., O'Neill, P., Wallinga, J. (editors). *Handb. Infect. Dis. Data Anal.*, pp. 411-36. DOI: 10.1201/9781315222912-21.

[4] Zacher, B., Czogiel, I. (2022). *Supervised learning using routine surveillance data improves outbreak detection of Salmonella and Campylobacter infections in Germany*. *PLoS One*, **17**, e0267510. DOI: 10.1371/journal.pone.0267510.

[5] World Health Organization (2019). *1 in 3 people globally do not have access to safe drinking water – UNICEF, WHO*. Geneva: World Health Organization. Available [viewed 2024-05-20] at: https://www.who.int/news/item/18-06-2019-1-in-3-people-globally-do-not-have-access-to-safe-drinking-water-unicef-who

[6] Alhamlan, F.S., Al-Qahtani, A.A., Al-Ahdal, M.N. (2015). *Recommended advanced techniques for waterborne pathogen detection in developing countries*. *J. Infect. Dev. Ctries.* **9**, pp. 128-35. DOI: 10.3855/jidc.6101.

[7] World Health Organization-Regional Office for Europe, United Nations Economic Commission for Europe (2019). *Surveillance and outbreak management of water-related infectious diseases associated with water-supply systems*. Geneva: World Health Organization-Regional Office for Europe, pp. 110. Available [viewed 2024-05-20] at: https://apps.who.int/iris/handle/10665/329403

[8] CDC (2023). *Waterborne disease outbreak investigation toolkit*. Atlanta, GA: Centers for Disease Control and Prevention, 36 pp. Available [viewed 2024-05-20] at: https://www.cdc.gov/healthywater/emergency/preparedness-resources/outbreak-response.html

[9] Impouma, B., Roelens, M., Williams, G.S., Flahault, A., Codeço, C.T., Moussana, F., Farham, B., Hamblion, E.L., Mboussou, F., Keiser O. (2020). *Measuring timeliness of outbreak response in the World Health Organization African Region*, 2017–2019. *Emerg. Infect. Dis*. **26**, pp. 2255-64 DOI: 10.3201/eid2611.191766.

[10] CDC (2022). *Preparation for a waterborne disease outbreak investigation*. Atlanta, GA: Centres for Disease Control and Prevention, 4 pp. Available [viewed 2024-05-20] at: https://www.cdc.gov/healthywater/emergency/waterborne-disease-outbreak-investigation-toolkit/preparation.html

[11] Kogan, N.E., Clemente, L., Liautaud, P., Kaashoek, J., Link, N.B., Nguyen, A.T, Lu, F.S., Huybers, P., Resch, B., Havas, C., Petutschnig, A., Davis, J., Chinazzi, M., Mustafa, B., Hanage, W.P., Vespignani, A., Santillana, M. (2021). *An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time*. *Sci. Adv.* **7**, No. eabd6989, DOI: 10.1126/sciadv.abd6989.

[12] Morley, J., Machado, C., Burr, C., Cowls, J., Taddeo, M., Floridi, L. (2019). *The debate on the ethics of AI in health care: A reconstruction and critical review*. Centre for Digital Ethics, Research paper series. DOI: 10.2139/ssrn.3486518.

[13] Liao, S.M. (2023). *Ethics of AI and health care: Towards a substantive human rights framework*. *Topoi*. **42**, pp. 857-66. DOI: 10.1007/s11245-023-09911-8.

[14] Shah, H. (2018). *Algorithmic accountability*. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **376**, 20170362. DOI: 10.1098/rsta.2017.0362.

[15] Danks, D., London, A.J. (2017). *Algorithmic bias in autonomous systems*. In: *Proc. 26th Int. Joint Conf. on Artificial Intelligence; AI and autonomy track*, pp. 4691-7. International Joint Conference on Artificial Intelligence Organization. DOI: 10.24963/ijcai.2017/654.

[16] Giovanola, B., Tiribelli, S. (2023). *Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms*. *AI Soc.* **38**, pp. 549-63. DOI: 10.1007/s00146-022-01455-6.

[17] Rajkomar, A., Dean, J., Kohane, I, (2019). *Machine learning in medicine*. *N. Engl. J. Med.* **380**, pp. 1347-58. DOI: 10.1056/NEJMra1814259.

[18] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q. (2017). *On fairness and calibration*. In: *Proc. 31st Int. Conf. Neural Information Processing Systems*, pp. 5684-93. Red Hook, NY: Curran Associates.

[19] A. Noufaily, A., Enki, D.G., Farrington, P., Garthwaite, P., Andrews, N., Charlett, A. (2013). *An improved algorithm for outbreak detection in multiple surveillance systems*. *Stat. Med.* **32**, pp. 1206-1222. DOI: 10.1002/sim.5595.

[20] Abbood, A, Ghozzi, S. (2020). *How to benchmark disease outbreak detection algorithms*: A review. Geneva: International Telecommunication Union. Available at: https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/outbreaks/review_benchmark_outbreaks.pdf [viewed 2024-05-20].

Annex A

Glossary

Table A.1 lists all the relevant abbreviations, acronyms and uncommon terms used in this deliverable.

Table A.1

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
CAB	Community Ablution Block	
CDS	Clinical Decision Support	
CfTGP	Call for Topic Group participation	
DEL	Deliverable	
EBS	Event-Based Surveillance	
EO	Earth Observation	
FDA	Food and Drug Administration	
FGAI4H	Focus Group on AI for Health	
FN	False Negative	
FP	False Positive	
GNSS	Global Navigation Satellite System	
HCP	Healthcare Professional	
IBS	Indicator-Based Surveillance	
ID	Identifier	
IMDRF	International Medical Device Regulators Forum	
ITU	International Telecommunication Union	
ML	Machine Learning	
POD	Probability Of Detection	
RKI	Robert Koch Institute	
ROC	Receiver Operating Characteristic	
SaMD	Software as a Medical Device	
TDD	Topic Description Document	
TG	Topic Group	
TN	True Negative	
TP	True Positive	
WG	Working Group	
WHO	World Health Organization	
WRID	Water-Related Infectious Disease	