|  |  |  |
| --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2022-2024 | FG-AI4H-S-018-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | Geneva, 3-5 July 2023 |
| **DOCUMENT** |
| **Source:** | TG-Outbreaks Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Outbreaks) |
| **Contact:** | Auss Abbood TG-Outbreaks Robert Koch Institute, Germany | E-mail: abbooda@rki.de |
| **Contact:** | Alexander UllrichTG-Outbreaks Robert Koch Institute, Germany | E-mail: ullricha@rki.de |
| **Contact:** | Khahlil LouisyTG-Outbreaks Institute for Technology and Global Health, USA | E-mail: klouisy@hks.harvard.edu |
| **Contact:** | Alexander Radunsky  Institute for Technology and Global Health, USA; UT Southwestern Medical Center, USA  | E-mail: Alexander.Radunsky@UTSouthwestern.edu |

|  |  |
| --- | --- |
| **Abstract:** | 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 (and follows the template structure defined in document FGAI4H-J-105). The creation of this TDD is an ongoing iterative process until it is approved by the Focus Group on AI for Health (FG-AI4H) as deliverable DEL10.10. This draft will be a continuous input- and output document. |
| **Change notes:** | *Topic Driver: Please list the changes of the current TDD version in comparison to earlier versions*. *This can include content updates in specific sections, additional or completed sections, updates on subtopics, etc.* **Version 1 (submitted as FGAI4H-S- to meeting S in Geneva)*** Merged former TG Outbreaks TDD with former TG Sanitation TDD
* Updated ethical considerations
* Updated regulatory considerations
* Generated more labelled data for benchmarking setup

.. |

Contributors

|  |  |
| --- | --- |
| 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 FuchsÖsterreichische 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 MarquitanmTOMADY, 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, MasanjaIfakara 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 AbboodRobert Koch Institute  |   |
| Stéphane Ghozzi WHO Hub |   |
| Bryan KimKorean 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  |   |

*The table of contents/list of tables/ list of figures are generated automatically by MS Word provided that the correct WinWord styles are used (Heading 1, Heading 2, etc.). They can be updated with a right click when the cursor is over the table/list, then “Update table”. Please familiarize yourself on how to properly use MS word formatting with “apply styles”, in particular when adding new (sub)-sections, figures and tables.*

**CONTENTS**

|  Page |
| --- |
| [1 Introduction 7](#_Toc138859858)[2 About the FG-AI4H topic group on outbreak detection for public health 8](#_Toc138859859)[2.1 Documentation 8](#_Toc138859860)[2.2 Status of this topic group 9](#_Toc138859861)[2.2.1 Status update for meeting S 9](#_Toc138859862)[2.2.2 Status update for meeting J 10](#_Toc138859863)[2.2.3 Status update for meeting M 10](#_Toc138859864)[2.2.4 Status update for meeting N 10](#_Toc138859865)[2.2.5 Status update for meeting O 10](#_Toc138859866)[2.2.6 Status update for meeting S 11](#_Toc138859867)[2.3 Topic group participation 11](#_Toc138859868)[3 Topic description 11](#_Toc138859869)[3.1 Definition of the AI task 12](#_Toc138859870)[Community and public data collection in eThekwini 12](#_Toc138859871)[Summary of the solution for sanitation 12](#_Toc138859872)[3.2 Current gold standard 15](#_Toc138859873)[3.3 Relevance and impact of an AI solution 15](#_Toc138859874)[3.4 Existing AI solutions 16](#_Toc138859875)[3.5 Subtopic 16](#_Toc138859876)[4 Ethical considerations 17](#_Toc138859877)[4.1 Privacy 19](#_Toc138859878)[4.2 Fairness 19](#_Toc138859879)[5 Existing work on benchmarking 20](#_Toc138859880)[5.1.1 Publications on benchmarking systems 20](#_Toc138859881)[5.1.2 Benchmarking by AI developers 21](#_Toc138859882)[5.1.3 Relevant existing benchmarking frameworks 21](#_Toc138859883)[6 Benchmarking by the topic group 22](#_Toc138859884)[6.1.1 Benchmarking version [Y] 22](#_Toc138859885)[6.1.2 Benchmarking version [X] 35](#_Toc138859886)[6.2 Subtopic [B] 35](#_Toc138859887)[7 Overall discussion of the benchmarking 35](#_Toc138859888)[8 Regulatory considerations 36](#_Toc138859889)[8.1 Existing applicable regulatory frameworks 36](#_Toc138859890)[8.2 Regulatory features to be reported by benchmarking participants 38](#_Toc138859891)[8.3 Regulatory requirements for the benchmarking systems 38](#_Toc138859892)[8.4 Regulatory approach for the topic group 38](#_Toc138859893)[9 References 39](#_Toc138859894)[Annex A: Glossary 39](#_Toc138859895)[Annex B: Declaration of conflict of interests 41](#_Toc138859896) |

**List of Tables**

|  Page |
| --- |
| [Table 1: Topic group output documents 7](#_Toc55950345) |

**List of Figures**

|  Page |
| --- |
| [Figure 1: Example of a figure 25](#_Toc55950354) |

FG-AI4H Topic Description Document

Topic group-Outbreaks

# Introduction

*Topic Driver: Add a short (half page) introduction to the topic. The introduction should provide a general overview of the addressed health topic and basic information about the AI task, including the input data and the output of the AI. The objective and expected impact of the benchmarking should also be described. More detailed information about the topic will appear in section 1.3.*

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.10.10 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.10.10 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

# 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.

*[Topic Driver: If there are modifications of the topic group driver, please indicate any changes here.]*

## 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 10.10 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)  |
| 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

Select the following link:

* <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B493339B7-4AF1-4875-9896-0281BF762280%7D&file=TDD-FGAI4H-J-105.docx&action=default>
* https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/\_layouts/15/WopiFrame.aspx?sourcedoc=%7B5F88E95B-9516-4ADD-A7D2-1585774574DD%7D&file=FGAI4H-I-013\_TDD.docx&action=default&CT=1611174658729&OR=DocLibClassicUI

## Status of this topic group

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

### Status update for meeting S

*Topic Driver: Please insert a one-page summary of the work since the last focus group meeting. This can include:*

* 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

### 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-)

### 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.

### 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.

### 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.

### 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.

## 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, each topic group can create an *individual mailing list:*

* 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>

# 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.

*Topic Driver: Topic groups typically begin* ***without*** *subtopics. Please write a few lines indicating future subtopics that might become relevant. Once you have defined subtopics, their focus/mandate should be explained in this section.*

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.

## Definition of the AI task

*This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is not about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to* [*DEL03*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7997F2C1-5A1D-4409-B2A0-CBC4E9CE8CDA%7D&file=DEL03.docx&action=default) *“AI requirements specifications,” which describes the functional, behavioural, and operational aspects of an AI system.*

* What is the AI doing?
* What kind of AI task is implemented (e.g., classification, prediction, clustering, or segmentation task)?
* Which input data are fed into the AI model?
* Which output is generated?

## Community and public data collection in eThekwini

There wereopportunities 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 fecal 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.

## Current gold standard

*This section provides a description of the established gold standard of the addressed health topic.*

* How is the task currently solved without AI?
* Do any issues occur with the current gold standard? Does it have limitations?
* Are there any numbers describing the performance of the current state of the art?

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].

## Relevance and impact of an AI solution

*This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.*

* Why is solving the addressed task with AI relevant?
* Which impact of deploying such systems is expected (e.g., impact on the health system, overall health system cost, life expectancy, or gross domestic product)?
* Why is benchmarking for this topic important (e.g., does it provide stakeholders with numbers for decision-making; does it simplify regulation, build trust, or facilitate adoption)?

## Existing AI solutions

*This section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain details of the operations, limitations, robustness, and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered in chapter 6.*

* Description of the general status and the maturity of AI systems for the health topic of your TG (e.g., exclusively prototypes, applications, and validated medical devices)
* Which are the currently known AI systems and their inputs, outputs, key features, target user groups, and intended use (if not discussed before)? This can also be provided as a table.
* What are the common features found in most AI solutions that might be benchmarked?
* What are the relevant metadata dimensions characterizing the AI systems in this field and with relevance for reporting (e.g., systems supporting offline functions, availability in certain languages, and the capability to process data in a specific format)?
* Description of existing AI systems and their scope, robustness, and other dimensions.

## 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).

# 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](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*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.

* What are the ethical implications of applying the AI model in real-world scenarios?
* What are the ethical implications of introducing benchmarking (having the benchmarking in place itself has some ethical risks; e.g., if the test data are not representative for a use case, the data might create the illusion of safety and put people at risk)?
* What are the ethical implications of collecting the data for benchmarking (e.g., how is misuse of data addressed, is there the need for an ethics board approval for clinical data, is there the need for consent management for sharing patient data, and what are the considerations about data ownership/data custodianship)?
* What risks face individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground?
* How is the privacy of personal health information protected (e.g., in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?
* How is ensured that benchmarking data are representative and that an AI offers the same performance and fairness (e.g., can the same performance in high, low-, and middle-income countries be guaranteed; are differences in race, sex, and minority ethnic populations captured; are considerations about biases, when implementing the same AI application in a different context included; is there a review and clearance of ‘inclusion and exclusion criteria’ for test data)?
* What are your experiences and learnings from addressing ethics in your TG?

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](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*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 [11]. 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 [12]. 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 [13] 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 [14]. Other biases can occur in this context and can undermine the correct ML functioning [15]. 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.

## 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.

## 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** [16]: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* [17], 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.

# 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 …

### Publications on benchmarking systems

While a representative comparable benchmarking for outbreak detection does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL07 *“AI for health evaluation considerations,”* [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“AI4H evaluation process description,”* [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification*,*”* [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM),”* and [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default) *“Clinical Evaluation of AI for health”*.

* What is the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
* State what are the most relevant approaches used in literature?
* Which scores and metrics have been used?
* How were test data collected?
* How did the AI system perform and how did it compare the current gold standard? Is the performance of the AI system equal across less represented groups? Can it be compared to other systems with a similar benchmarking performance and the same clinically meaningful endpoint (addressing comparative efficacy)?
* How can the utility of the AI system be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)?
* Have there been clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings?
* What are the most relevant gaps in the literature (what is missing concerning AI benchmarking)?

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.

### 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.

* What are the most relevant learnings from the benchmarking by AI developers in this field (e.g., ask the members of your topic group what they want to share on their benchmarking experiences)?
* Which scores and metrics have been used?
* How did they approach the acquisition of test data?

### 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 [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“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).

* Which benchmarking platforms could be used for this topic group (e.g., EvalAI, AIcrowd, Kaggle, and CodaLab)?
* Are the benchmarking assessment platforms discussed, used, or endorsed by FG-AI4H an option?
* Are there important features in this topic group that require special attention?
* Is the reporting flexible enough to answer the questions stakeholders want to get answered by the benchmarking?
* What are the relative advantages and disadvantages of these diverse solutions?

# 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: [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default) *“Data specification”* (introduction to deliverables 5.1-5.6), [DEL05\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B19830259-F63B-42D4-A408-48C854D6C124%7D&file=DEL05_1.docx&action=default)*“Data requirements”* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL05\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B25141F77-E59A-45F1-B081-185C2194FE67%7D&file=DEL05_2.docx&action=default) *“Data acquisition”*, [DEL05\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B05D8938E-BC2A-4A62-BCB0-1FD46AA72235%7D&file=DEL05_3.docx&action=default) *“Data annotation specification”*, [DEL05\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF267A95C-4C5B-4D63-A135-58AF487C3AD3%7D&file=DEL05_4.docx&action=default) *“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), [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) *“Data handling”* (which outlines how data will be handled once they are accepted), [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) *“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](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF5967277-90C8-4252-A0B9-43A5692F35E2%7D&file=DEL06.docx&action=default) *“AI training best practices specification”* (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL07](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B47E77197-F87B-49F4-80B3-2DD949A5F185%7D&file=DEL07.docx&action=default)*“AI for health evaluation considerations”* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“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), [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification”* (which specifies how an AI can and should be tested *in silico*), [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM)”* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default)*“Clinical Evaluation of AI for health”* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL09](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3E940987-8D75-44B8-85E4-F0E475964F15%7D&file=DEL09.docx&action=default) *“AI for health applications and platforms”* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL09\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1A2EC8D5-53CA-4C8C-9B09-B61CA6F428C5%7D&file=DEL09_1.docx&action=default) *“Mobile based AI applications,”* and [DEL09\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3B5A31DE-D3B1-4EC1-A261-2C2E19F73810%7D&file=DEL09_2.docx&action=default) *“Cloud-based AI applications”* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

*Topic driver: Please refer to the above comments concerning subtopics.*

The benchmarking of [YOUR TOPIC] 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.

* Which benchmarking iterations have been implemented thus far?
* What important new features are introduced with each iteration?
* What are the next planned iterations and which features are they going to add?

### Benchmarking version [Y]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [Y] (latest version, chronologically reversed order).

#### Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [Y].

* What is the overall scope of this benchmarking iteration (e.g., performing a first benchmarking, adding benchmarking for multi-morbidity, or introducing synthetic-data-based robustness scoring)?
* What features have been added to the benchmarking in this iteration?

#### Benchmarking methods

This section provides details about the methods of the benchmarking version [Y]. 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.

##### Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

* How does the architecture look?
* What are the most relevant components and what are they doing?
* How do the components interact on a high level?
* What underlying technologies and frameworks have been used?
* How does the hosted AI model get the required environment to execute correctly? What is the technology used (e.g., Docker/Kubernetes)?

##### Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

* How do benchmarking data access the system?
* Where and how (data format) are the data, the responses, and reports of the system stored?
* How are the inputs and the expected outputs separated?
* How are the data sent to the AI systems?
* Are the data entries versioned?
* How does the lifecycle for the data look?

##### Safe and secure system operation and hosting

*From a technical point of view, the benchmarking process is not particularly complex. It is more about agreeing on something in the topic group with potentially many competitors and implementing the benchmarking in a way that cannot be compromised. This section describes how the benchmarking system, the benchmarking data, the results, and the reports are protected against manipulation, data leakage, or data loss. Topic groups that use ready-made software might be able to refer to the corresponding materials of the manufacturers of the benchmarking system.*

This section addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage, or data loss.

In the case of a manufactured data source (vs. self-generated data), it is possible to refer to the manufacturer’s prescriptions.

* Based on the architecture, where is the benchmarking vulnerable to risk and how have these risks been mitigated (e.g., did you use a threat modelling approach)? A discussion could include:
* Could someone access the benchmarking data before the actual benchmarking process to gain an advantage?
* What safety control measures were taken to manage risks to the operating environment?
* Could someone have changed the AI results stored in the database (your own and/or that of competitors)?
* Could someone attack the connection between the benchmarking and the AI (e.g., to make the benchmarking result look worse)?
* How is the hosting system itself protected against attacks?
* How are the data protected against data loss (e.g., what is the backup strategy)?
* What mechanisms are in place to ensure that proprietary AI models, algorithms and trade-secrets of benchmarking participants are fully protected?
* How is it ensured that the correct version of the benchmarking software and the AIs are tested?
* How are automatic updates conducted (e.g., of the operating system)?
* How and where is the benchmarking hosted and who has access to the system and the data (e.g., virtual machines, storage, and computing resources, configurational settings)?
* How is the system’s stability monitored during benchmarking and how are attacks or issues detected?
* How are issues (e.g., with a certain AI) documented or logged?
* In case of offline benchmarking, how are the submitted AIs protected against leakage of intellectual property?

##### Benchmarking process

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

* How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
* How do possible participants learn about an upcoming benchmarking?
* How can one apply for participation?
* What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine learning technology used, company size, company location)?
* Are there any contracts or legal documents to be signed?
* Are there inclusion or exclusion criteria to be considered?
* How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
* How can participants test their interface (e.g., is there a test dataset in case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
* Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
* If there are problems with an AI, how are problems resolved (e.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)?
* How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?
* In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?

#### AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of [YOUR TOPIC]. 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.

* What are the general data types that are fed in the AI model?
* How exactly are they encoded? For instance, discuss:
	+ The exact data format with all fields and metadata (including examples or links to examples)
	+ Ontologies and terminologies
	+ Resolution and data value ranges (e.g., sizes, resolutions, and compressions)
	+ Data size and data dimensionality

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.

#### 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.

* What are the general data output types returned by the AI and what is the nature of the output (e.g., classification, detection, segmentation, or prediction)?
	+ How exactly are they encoded? Discuss points like:
		- The exact data format with all fields and metadata (including examples or links to examples)
		- Ontologies and terminologies
* What types of errors should the AI generate if something is defective?

#### Test data label/annotation structure

*Topic driver: Please describe how the expected AI outputs are encoded in the benchmarking test data. Please note that it is essential that the AIs never access the expected outputs to prevent cheating. The topic group should carefully discuss whether more detailed labelling is needed. Depending on the topic, it might make sense to separate between the best possible output of the AI given the input data and the correct disease (that might be known but cannot be derived from the input data alone). Sometimes it is also helpful to encode acceptable other results or results that can be clearly ruled out given the evidence. This provides a much more detailed benchmarking with more fine-grained metrics and expressive reports than the often too simplistic leader boards of many AI competitions.*

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.

* What are the general label types (e.g., expected results, acceptable results, correct results, and impossible results)?
* How exactly are they encoded? Discuss points like:
	+ The exact data format with all fields and metadata (including examples or links to examples)
	+ Ontologies and terminologies
* How are additional metadata about labelling encoded (e.g., author, data, pre-reviewing details, dates, and tools)?
* How and where are the labels embedded in the input data set (including an example; e.g., are there separate files or is it an embedded section in the input data that is removed before sending to the AI)?

#### Scores and metrics

*Topic drivers: This section describes the scores and metrics that are used for benchmarking. It includes details about the testing of the AI model and its effectiveness, performance, transparency, etc. Please note that this is only the description of the scores and metrics actually used in* ***this*** *benchmarking iteration. A general description of the state of the art of scores and metrics and how they have been used in previous work is provided in section 3.*

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.

* Who are the stakeholders and what decisions should be supported by the scores and metrics of the benchmarking?
* What general criteria have been applied for selecting scores and metrics?
* What scores and metrics have been chosen/defined for robustness?
* What scores and metrics have been chosen/defined for medical performance?
* What scores and metrics have been chosen/defined for non-medical performance?
	+ Metrics for technical performance tracking (e.g., monitoring and reporting when the performance accuracy of the model drops below a predefined threshold level as a function of time; computational efficiency rating, response times, memory consumption)
* What scores and metrics have been chosen/defined for model explainability?
* Describe for each aspect
	+ The exact definition/formula of the score based on the labels and the AI output data structures defined in the previous sections and how they are aggregated/accumulated over the whole dataset (e.g., for a single test set entry, the result might be the probability of the expected correct class which is then aggregated to the average probability of the correct class)
	+ Does it use some kind of approach for correcting dataset bias (e.g., the test dataset usually has a different distribution compared to the distribution of a condition in a real-world scenario. For estimating the real-world performance, metrics need to compensate this difference.)
	+ What are the origins of these scores and metrics?
	+ Why were they chosen?
	+ What are the known advantages and disadvantages?
	+ How easily can the results be compared between or among AI solutions?
	+ Can the results from benchmarking iterations be easily compared or does it depend too much on the dataset (e.g., how reproducible are the results)?
* How does this consider the general guidance of WG-DAISAM in [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) “Data and artificial intelligence assessment methods (DAISAM)”?
* Have there been any relevant changes compared to previous benchmarking iterations? If so, why?

When we want to measure the performance of an algorithm, we might look for criteria such as usefulness, cost, sensitivity, representativeness, timeliness, simplicity, flexibility, and acceptability. These are measures that include not only the statistical algorithms but also the more general criterions for public health systems. Common measures for the comparison of statistical algorithms are (more closely described in our review [How to benchmark disease outbreak detection algorithms: A review](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/outbreaks/review_benchmark_outbreaks.pdf); located at our 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-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.

#### 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.

* How does the overall dataset acquisition and annotation process look?
* How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
* Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
* How was the dataset documented and which metadata were collected?
	+ Where were the data acquired?
	+ Were they collected in an ethical-conform way?
	+ Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
	+ Do the data contain ‘sensitive information’ (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
	+ What kind of data anonymization or deidentification has been applied?
	+ Are the data self-contained (i.e., independent from externally linked datasets)?
	+ How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
	+ What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
* Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
* Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
* How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
* Specific data governance derived by the general data governance document (currently [F-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-F-103-DataPolicy.pdf) and the deliverables beginning with [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default))
* How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
* Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
* How was the annotation organized?
	+ How many annotators/peer reviewers were engaged?
	+ Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
	+ How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
	+ If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
	+ How was the annotation of each case documented?
	+ Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
* Were data/label update/amendment policies and/or criteria in place?
* How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
* How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
* Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
* What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)

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.

#### 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 [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) on *data handling* and [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) on *data sharing practices*).

* Which legal framework was used for data sharing?
* Was a data sharing contract signed and what was the content? Did it contain:
	+ Purpose and intended use of data
	+ Period of agreement
	+ Description of data
	+ Metadata registry
	+ Data harmonization
	+ Data update procedure
	+ Data sharing scenarios
		- Data can be shared in public repositories
		- Data are stored in local private databases (e.g., hospitals)
	+ Rules and regulation for patients’ consent
	+ Data anonymization and de-identification procedure
	+ Roles and responsibilities
		- Data provider
		- Data protection officer
		- Data controllers
		- Data processors
		- Data receivers
* Which legal framework was used for sharing the AI?
* Was a contract signed and what was the content?

#### 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.

* Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
* Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group, and ethnicity)?
* Was a baseline assessed as part of the benchmarking?
* How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous section.
* What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

#### Reporting methodology

*After the benchmarking, the next step is to describe how the results are compiled into reports that allow stakeholders to make decisions (e.g., which AI systems can be used to solve a pre-diagnosis task in an offline –field –clinic scenario in central America). For some topic groups, the report might be as simple as a classical AI competition leader board using the most relevant performance indicator. For other tasks, it could be an interactive user interface that allows stakeholders to compare the performance of the different AI systems in a designated context with existing non-AI options. For the latter, statistical issues must be carefully considered (e.g., the multiple comparisons problem). Sometimes, a hybrid of prepared reports on common aspects are generated in addition to interactive options. There is also the question of how and where the results are published and to what degree benchmarking participants can opt in or opt out of the publication of their performance.*

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

* What is the general approach for reporting results (e.g., leader board vs. drill down)?
* How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
* How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
* What additional metadata describing the AI models have been selected for reporting?
* How is the relationship between AI results, baselines, previous benchmarking iterations, and/or other benchmarking iterations communicated?
* What is the policy for sharing participant results (e.g., opt in or opt out)? Can participants share their results privately with their clients (e.g., as in a credit check scenario)?
* What is the publication strategy for the results (e.g., website, paper, and conferences)?
* Is there an online version of the results?
* Are there feedback channels through which participants can flag technical or medical issues (especially if the benchmarking data was published afterwards)?
* Are there any known limitations to the value, expressiveness, or interpretability of the reports?

#### 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.

* When was the benchmarking executed?
* Who participated in the benchmarking?
* What overall performance of the AI systems concerning medical accuracy, robustness, and technical performance (minimum, maximum, average etc.) has been achieved?
* What are the results of this benchmarking iteration for the participants (who opted in to share their results)?

#### Discussion of the benchmarking

This section discusses insights of this benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process).

* What was the general outcome of this benchmarking iteration?
* How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
* Are there real benchmarking results and interesting insights from this data?
	+ How was the performance of the AI system compared to the baseline?
	+ How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
	+ How did the results change in comparison to the last benchmarking iteration?
* Are there any technical lessons?
	+ Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
	+ How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
* Are there any lessons concerning data acquisition?
	+ Was it possible to collect enough data?
	+ Were the data as representative as needed and expected?
	+ How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
	+ Was it possible to find annotators?
	+ Was there any relevant feedback from the annotators?
	+ How long did it take to create the dataset?
* Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
	+ Are metrics missing?
	+ Do the stakeholders need different reports or additional metadata (e.g., do they need the “offline capability” included in the AI metadata so that they can have a report on the best offline system for a certain task)?
* Are there insights on the benchmarking process?
	+ How was the interest in participation?
	+ Are there reasons that someone could not join the benchmarking?
	+ What was the feedback of participants on the benchmarking processes?
	+ How did the participants learn about the benchmarking?

#### Retirement

*Topic driver: describe what happens to the benchmarking data and the submitted AI models after the benchmarking.*

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) “*AI software lifecycle specification”* (identification of standards and best practices that are relevant for the AI for health software life cycle).

* What happens with the data after the benchmarking (e.g., will they be deleted, stored for transparency, or published)?
* What happens to the submitted AI models after the benchmarking?
* Could the results be reproduced?
* Are there legal or compliance requirements to respond to data deletion requests?

### Benchmarking version [X]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [X].

*Topic driver: Provide details of previous benchmarking versions here using the same subsection structure as above.*

## Subtopic [B]

*Topic driver: If there are subtopics in your topic group, please provide the details about the benchmarking of the second subtopic [B] here using the same subsection structure as above (please refer to earlier comments – in red fonts - concerning subtopics).*

# Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 4.2.11).

* What is the overall outcome of the benchmarking thus far?
* Have there been important lessons?
* Are there any field implementation success stories?
* Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
* Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
* Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
	+ Did the AI system perform as predicted relative to the baselines?
	+ Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust, and quality of service)?
* Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?
* What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

# Regulatory considerations

*Topic Driver: This section reflects the requirements of the working group on* [***Regulatory considerations on AI for health (WG-RC)***](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) *and their various deliverables. It is* ***NOT requested to re-produce regulatory frameworks****, but to show the regulatory frameworks that have to be applied in the context of your AIs and their benchmarking (****2 pages max****).*

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. 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)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“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](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) 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

## 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

* What existing regulatory frameworks cover the type of AI in this TDD (e.g., MDR, FDA, GDPR, and ISO; maybe the systems in this topic group always require at least “MDR class 2b” or maybe they are not considered a medical device)?
* Are there any aspects to this AI system that require additional specific regulatory considerations?

Smart sanitation—the use of biosensors in toilets, sewage pipes, and septic tanks—is an emerging technology with the potential to improve individual health diagnostics as well as enhance disease surveillance at the community level. However, the implementation of this technology also presents many ethical challenges that are reflected (or should be reflected) in the regulatory framework that applies to it.

This section will review what Human Rights apply to smart sanitation technology, with particular attention to the work of technology legal scholars who have applied the Human Rights framework to other technologies with the potential for systemic surveillance. It will also review how major data privacy regulations, such as the General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), and the new Chinese Personal Information Protection Law (PIPL) applies to smart sanitation technology. Finally, it makes policy recommendations.

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 ALI/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 Software that meets the following criteria (under 21 USC § 360j(o)(1)(E))

Is not “intended to acquire, process, or analyze a medical image or a signal from an in vitro diagnostic device or signal acquisition system”

Is intended for the purpose of “displaying, analyzing, 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.

US Modifications to Software:

FDA requires premarket 501(k) submission which demonstrates that a device is safe and effective.

Framework for Modifications to AI/ML-based SaMD: (internationally harmonized 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 TPLC approach as envisioned in the Digital Health Software PreCertification )Pre-Cert) Program.)

## Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

* Which certifications and regulatory framework components of the previous section should be part of the metadata (e.g., as a table with structured selection of the points described in the previous section)?

## Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

* Which regulatory frameworks apply to the benchmarking system itself?
* Are viable solutions with the necessary certifications already available?
* Could the TG implement such a solution?

## Regulatory approach for the topic group

*Topic Driver: Please select the points relevant for your type of AI and the corresponding benchmarking systems. If your AIs and your benchmarking are not a medical device, this might be quite short.*

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations.”*

* Documentation & Transparency
	+ How will the development process of the benchmarking be documented in an effective, transparent, and traceable way?
* Risk management & Lifecycle approach
	+ How will the risk management be implemented?
	+ How is a life cycle approach throughout development and deployment of the benchmarking system structured?
* Data quality
	+ How is the test data quality ensured (e.g., the process of harmonizing data of different sources, standards, and formats into a single dataset may cause bias, missing values, outliers, and errors)?
	+ How are the corresponding processes document?
* Intended Use & Analytical and Clinical Validation
	+ How are technical and clinical validation steps (as part of the lifecycle) ensured (e.g., as proposed in the IMDRF clinical evaluation framework)?
* Data Protection & Information Privacy
	+ How is data privacy in the context of data protection regulations ensured, considering regional differences (e.g., securing large data sets against unauthorized access, collection, storage, management, transport, analysis, and destruction)? This is especially relevant if real patient data is used for the benchmarking.
* Engagement & Collaboration
	+ How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

# 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] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, “On Fairness and Calibration”.

*Topic driver: Add the bibliography here.*

*Topic driver: If you include figures in this document, please use the following MS Word format/style (otherwise the figure won’t be included in the table of figures).*

|  |  |  |
| --- | --- | --- |
| A |  | B |

Captions for figures use WinWord style "Figure\_No & title"

Figure 1: Example of a figure

Annex A:
Glossary

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

|  |  |  |
| --- | --- | --- |
| Acronym/Term | Expansion | Comment |
| 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 [YOUR TOPIC GROUP] |
| TG | Topic Group |  |
| WG | Working Group |  |
| FGAI4H | Focus Group on AI for Health |  |
| AI | Artificial intelligence |  |
| ITU | International Telecommunication Union |  |
| WHO | World Health Organization |  |
| DEL | Deliverable  |  |
| CfTGP | Call for topic group participation |  |
| AI4H  | Artificial intelligence for health |  |
| IMDRF | International Medical Device Regulators Forum |  |
| MDR | Medical Device Regulation |  |
| ISO | International Standardization Organization |  |
| GDPR | General Data Protection Regulation |  |
| FDA | Food and Drug administration |  |
| SaMD | Software as a medical device |  |
| AI-MD | AI based medical device |  |
| LMIC | Low-and middle-income countries |  |
| GDP | Gross domestic product |  |
| API | Application programming interface |  |
| IP | Intellectual property |  |
| PII | Personal identifiable information |  |
| […] |  |  |

Annex B:
Declaration of conflict of interests

In accordance with the ITU transparency rules, this section lists the conflict-of-interest declarations for everyone who contributed to this document. Please see the guidelines in FGAI4H-F-105 “ToRs for the WG-Experts and call for experts” and the respective forms (Application form & Conflict of interest form).

Company/Institution/Individual XYZ

A short explanation of the company’s area of activity and how the work on this document might benefit the company and/or harm competitors. A list of all people who contributed to this document on behalf of this company and any personal interest in this company (e.g., shares).

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