INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2022-2024

FGAI4H-P-018-A01

ITU-T Focus Group on AI for Health

Original: English

WG(s):	Plen	Helsinki, 20-22 September 2022

DOCUMENT

Source:	TG-Outbreaks Topic Driver
Title:	Att.1 – TDD update (TG-Outbreaks) [same as Meeting K]
Purpose:	Discussion

Contact: Auss Abbood, Stephane Ghozzi
Robert Koch Institute, Helmholtz HZI, Germany
Email: ghozzis@rki.de, stephane.ghozzi@helmholtz-hzi.de

Abstract: This topic description document (TDD) specifies a standardized benchmarking for AI-based outbreak detection algorithms. It covers all 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 No. 10.10. This draft will be a continuous input- and output document.

This version of the TDD is the same as seen in Meeting K (FGAI4H-K-018-A01), reproduced for easier reference as a Meeting N document.

Change notes:

Version 1 (submitted as FGAI4H-J-013 to meeting J in location 'Online')

- Adopted new TDD-template

..

Contributors

Reinhard Fuchs, Österreichische Agentur für Gesundheit und Ernährungssicherheit (AGES)

Ian Kopacka, Österreichische Agentur für Gesundheit und Ernährungssicherheit (AGES)

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

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

Elaine Nsoesie, School of Public Health, Boston University

Sophie Marquitan, mTOMADY, a project of Doctors for Madagascar

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

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

Noelson Lahiafake, mTOMADY, a project of Doctors for Madagascar

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

M. Claire Jarashow, Los Angeles County Department of Public Health

Sharon K. Greene, NYC Department of Health and Mental Hygiene

Robert Istepanian, Imperial College

Richard Aubrey White, Norwegian Public-Health-Institut FHI

Birgitte Freiesleben De Blasio, Norwegian Public-Health-Institut FHI

Gunnar Rø, Norwegian Public-Health-Institut FHI

Claudia Coipan, RIVM

Roger Antony Morbey, Public Health England; National Infection Service

Amy FW Mikhail, Public Health England; National Infection Service

Angela Noufaily, University of Warwick

Anette Hulth, Public Health Agency of Sweden

Pär Bjelkmar, Public Health Agency of Sweden

Henrik Källberg, Public Health Agency of Sweden

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

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

Honorati, Masanja, Ifakara Health Institute (IHI), Tanzania

Salim Abdullah, Ifakara Health Institute (IHI), Tanzania

Irene Masanja, Ifakara Health Institute (IHI), Tanzania

Nada Malou, Médecins Sans Frontières (MSF), France

Ally Salim Jr., Inspired Ideas, Tanzania

Meghan Hamel, Public Health Agency of Canada

David L. Buckeridge, McGill University

Auss Abbood, Robert Koch Institute

Stéphane Ghozzi, Helmholtz Centre for Infection Research

Bryan Kim, Korean CDC

Azadur Rahman Sarker, Tech Valley Networks Limited

Helmi Zakariah, AIME Inc.

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

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

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

Derrick Bary Abila, One Health Fellow

Rachel Lowe, London School of Hygiene & Tropical Medicine

CONTENTS

Page

1 Introduction 6

2 About the FG-AI4H topic group on outbreak detection 6
2.1 Documentation 7
2.2 Status of this topic group 7
2.2.1 Status update for meeting J 7
2.2.2 Status update for meeting [MEETING LETTER] 8
2.3 Topic group participation 8

3 Topic description 8
3.1 Subtopic 9
3.1.1 Definition of the AI task 9
3.1.2 Current gold standard 9
3.1.3 Relevance and impact of an AI solution 9
3.1.4 Existing AI solutions 10

4 Ethical considerations 11

5 Existing work on benchmarking 11
5.1 Publications on benchmarking systems 12
5.1.1 Benchmarking by AI developers 12
5.1.2 Relevant existing benchmarking frameworks 13

6 Benchmarking by the topic group 13
6.1.1 Benchmarking version [Y] 14
6.1.2 Benchmarking version [X] 25

7 Overall discussion of the benchmarking 25

8 Regulatory considerations 26
8.1 Existing applicable regulatory frameworks 26
8.2 Regulatory features to be reported by benchmarking participants 27
8.3 Regulatory requirements for the benchmarking systems 27
8.4 Regulatory approach for the topic group 27

9 References 28

Annex A: Glossary 28

Annex B: Declaration of conflict of interests 29

List of Tables

Page

Table 1 :Topic group output documents 7

List of Figures

Page

Figure 1: Example of a figure ..... **Error! Bookmark not defined.**

FG-AI4H Topic Description Document

Topic group-Outbreaks

1 Introduction

This topic description document specifies the standardized benchmarking for outbreak detection systems. It serves as deliverable No. 013 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

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, in contrast to endemic infections, has infected cases that can be related to a source (bad foodstuff, breeding site of disease transmitting insects, a sick person in close contact with many others, etc.) and usually needs to be contained to stop the spread of this disease.

Infectious disease outbreaks pose a major risk to public health and are of global concern. Many established infectious diseases cause the death of millions of people every year and new infectious diseases are emerging. The risk and occurrence of infectious diseases is influenced by globalization, migration, and climate change. According to a World Health Organization (WHO) ranking, infectious diseases are ranked in the top 10 causes of death worldwide.

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

The aim of outbreak detection algorithms is to detect aberrant case numbers and conspicuous events within data streams, pointing to the emergence of infectious disease outbreaks, in a fast and automatic manner. To this 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 are mostly biased towards the environment, diseases, and infrastructure of these countries. In the EU, e.g., algorithms developed for one country (e.g., Farrington) are used across other neighbouring countries with no public benchmark assessing them. 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.

2 About the FG-AI4H topic group on outbreak detection

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 Geneva, 1 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 FG-AI4H meeting G in New Dehli, 14 November 2019 Stéphane Ghozzi from the Helmholtz Centre for Infection Research and Auss Abbood from the Robert Koch Institute was nominated as topic driver for the TG-Outbreaks.

2.1 Documentation

This document is the TDD for the TG-Outbreaks. 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. It describes the existing approaches for assessing the quality of outbreak detection 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 10.013 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-J-013-A01	Latest update of the Topic Description Document of the TG-Outbreaks
FGAI4H-I-013-A02	Latest update of the Call for Topic Group Participation (CfTGP)
FGAI4H-I-013-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Outbreaks

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-Outbreaks.aspx

Select the following link:

- 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

2.2 Status of this topic group

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

2.2.1 Status update for meeting J

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

- List of current partners
- Relevant next steps

- Phone meeting with interested parties (Dec 2019)
- Further acquisition of members (Jan-Feb 2020)
- Review of existence methods and metrics and in disease outbreak detection and existing approaches for benchmarking or similar endeavours. (Mar 2020)
- Survey on how disease outbreak detection is done among our members (Feb-Mar 2020)
- Implementation of a new metric to test different families of outbreak detection algorithms (July 2020-)

2.2.2 Status update for meeting [MEETING LETTER]

[…]

2.3 Topic group participation

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

- Needs to be uploaded

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-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:*

- gai4htgoutbreaks@lists.itu.int

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

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

3 Topic description

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

Topic groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups

can be established within one topic group to pursue different topic-specific fields of expertise. The TG-Outbreaks currently has no subtopics.

Future subtopics for different disease groups might be introduced in TG-Outbreaks.

3.1 Subtopic

Although there are no subtopics yet, there would be two senseful subtopics. We propose the distinction into different groups of infectious disease such as for food-borne (e.g., salmonella) and vector-borne diseases (such as Dengue) since they typically have different impact on the health system, outbreak pattern, and thus require different data streams and algorithms to operate well. This distinction is not exhaustive and further differentiation between infectious disease can be done. However, in past TG-meetings we concluded that this distinction might be the first to consider.

3.1.1 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 *"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?

3.1.2 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 an 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, with data from real-time-surveillance systems 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.

3.1.3 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)?

The majority of outbreak detection algorithms today are bound by the type of data they require, and sometimes even by the disease they need to detect as of today. To find a more holistic approach would also mean to compare all these different, well-established methods to a new, probably AI-driven approach. Only then, public health agents would trust such an algorithm. Making comparisons between methods that require different data types and focus on different diseases is a non-trivial task that can as of now best be realized by an extensive benchmarking.

Another impact might be that AI developers who usually lack expert knowledge in epidemiology have a gentler entry into the field of outbreak detection since benchmarking introduces the relevant metrics and data types and narrow down the error margin.

3.1.4 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.

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

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

For the future, since many of the previous approaches were time-series based, we expect further Hidden Markov Models and deep learning methods appropriate for sequential data such as Long

Short Term Memory Networks (LSTM). However other methods, like multivariate Bayesian regression or all-purpose deep learning (CNN, RNN) are conceivable.

4 Ethical considerations

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

- 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?

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

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

5 Existing work on benchmarking

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

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

Minimal benchmarking setup at RKI:

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

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

5.1 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 *"AI4H evaluation process description,"* DEL07_2 *"AI technical test specification,"* DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL07_4 *"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)?

5.1.1 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?

5.1.2 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 *"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?

6 Benchmarking by the topic group

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

It reflects the considerations of various deliverables: DEL05 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL05_1 *"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL05_2 *"Data acquisition"*, DEL05_3 *"Data annotation specification"*, DEL05_4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL05_5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL05_6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL07 *"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 *"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 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL07_4 *"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL07_5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL09 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL09_1 *"Mobile based AI applications,"* and DEL09_2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

The benchmarking of outbreak detection 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?

6.1.1 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).

6.1.1.1 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?

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

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

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

6.1.1.2.1 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)?

6.1.1.2.2 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?

6.1.1.2.3 Safe and secure system operation and hosting

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?

6.1.1.2.4 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?

6.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of outbreak detection. 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 clicks, HealthTweets, Twitter) or data from symptom-assessment apps, healthcare providers, hotlines etc. Real time data sources have a high potential of significantly improving the outbreak detection particularly in accuracy or timeliness.

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

Although IBS is more structured, IBS data still comes in different shapes which might be relevant for the later use of algorithms. For example, it might be important to have a long history of data since some algorithms require data to have been collected for at least five years. Furthermore, almost any surveillance system that reports notifiable diseases does so by providing the date of infection or report and cases numbers aggregated to weeks months, or quarter and a location of varying precision (street address, county, region, federal state…). Our choice of algorithms, however, depends on the available granularities of the former properties. For example, to detect whether two cases are part of an outbreak, the Knox test can be applied where locations are close given a 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 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 this approach provides a ground truth about the outbreaks injected into the (often also simulated) endemic baseline. Since disease dynamics, such as seasonality, reporting behavior, and trends, are known, a good estimate of realistic data can be formulated. The ground truth knowledge about outbreaks might be missing in real data and therefore makes it impossible to calculate several performance scores such as specificity and sensitivity.

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

chosen constant is added. The added cases are distributed over the outbreak week given a lognormal distribution.

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

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

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

6.1.1.4 AI output data structure

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

- 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?

6.1.1.5 Test data label/annotation structure

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

- 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)?

6.1.1.6 Scores and metrics

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

- 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 “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*):

- 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.

6.1.1.7 Test dataset acquisition

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

- 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 and the deliverables beginning with DEL05)
- 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.

6.1.1.8 Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL05_5 on *data handling* and DEL05_6 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?

6.1.1.9 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

- 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)?

6.1.1.10 Reporting methodology

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

- 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?

6.1.1.11 Result

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

- 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)?

6.1.1.12 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?

6.1.1.13 Retirement

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 "*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?

6.1.2 Benchmarking version [X]

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

7 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 6.1.1.12).

- 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)?

8 Regulatory considerations

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

The deliverables with relevance for regulatory considerations are DEL02 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL02_1 *"Mapping of IMDRF essential principles to AI for health software",* and DEL02_2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL04 identifies standards and best practices that are relevant for the "*AI software lifecycle specification."* The following sections discuss how the different regulatory aspects relate to the TG-Outbreaks

8.1 Existing applicable regulatory frameworks

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

- 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?

8.2 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)?

8.3 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?

8.4 Regulatory approach for the topic group

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 *"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?

9 References

(To be transferred from our review *How to benchmark disease outbreak detection algorithms: A review*)

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 Outbreaks
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).

_____