|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FGAI4H-K-011-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plen | E-meeting, 27-29 January 2021 |
| **DOCUMENT** |
| **Source:** | TG-FakeMed Topic Driver |
| **Title:** | Att.1 – TDD update (TG-FakeMed) [same as Meeting J] |
| **Purpose:** | Discussion |
| **Contact:** | Frank VerzeféTrueSpec-Africa, DRC | Email: fverzefe@gmail.com |

|  |  |
| --- | --- |
| **Abstract:** | This document “DEL10.18” is the topic description document (TDD) of the topic group *TG-FakeMed*, which is concerned with the standardized benchmarking of AI for the detection of falsified medicine. The outline will follow the template structure defined in FGAI4H-C-105. This TDD draft will be created in a joint effort by the topic group and continuously improved over the upcoming meetings until it is finally approved by the focus group. The presence of substandard and falsified medical products in countries and their use by patients threatens to undermine progress towards meeting the Sustainable Development Goals. Such products may be of poor quality, unsafe or ineffective, threatening the health of those that take them. The problem of substandard and falsified medical products continues to increase, as globalized manufacturing and distribution systems grow ever more complex. That complexity heightens the risk that production errors will occur, or that medicines will degrade between factory and consumer. Increasing demand for medicines, vaccines and other medical products in almost every country, in addition to poor supply-chain management and the growth of e-commerce also creates opportunities for falsified medicines to be introduced into the supply chain. Unfortunately, reliable information on the true public health and socioeconomic impacts of substandard and falsified medical products is sparse. A stronger evidence base is needed to help prevent, detect and respond to substandard and falsified medical products, and the public health threat they represent. The falsified and sub-standard drugs today cause, according to the University of Edinburgh (childhood pneumonia model), the deaths of 250,000 children a year. Technological innovation, more precisely AI technology is one of the most effective means of dealing with increasingly creative counterfeiters.This version of the TDD is the same as seen in Meeting J (FGAI4H-J-011-A01-R01), reproduced for easier reference as a Meeting K document. |

*The outline follows the template structure defined in FGAI4H-C-105.*

Table of Contents

[1 Introduction 3](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776057)

[1.1 Document structure 3](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776058)

[1.2 Topic description 3](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776059)

[1.2.1 Overview 3](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776060)

[1.2.2 Relevance 4](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776061)

[1.2.3 Impact 5](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776062)

[1.3 Ethical considerations 5](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776063)

[1.3.1 Relevance of data and models 5](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776064)

[1.3.2 Database 6](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776065)

[1.3.3 Data availability 6](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776066)

[1.3.4 Data quality 6](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776067)

[1.3.5 Accuracy 7](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776068)

[1.3.6 Size and severity of impact 7](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776069)

[1.4 Existing AI solutions 7](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776070)

[1.5 Existing work on benchmarking 8](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776071)

[2 AI4H topic group 8](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776072)

[2.1 AI for detection of falsified medicine (AI4DFM) objectives & deliverables 8](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776073)

[2.2 Topic group structure 8](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776074)

[3 Method 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776075)

[3.1 AI input data structure 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776076)

[3.2 AI output data structure 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776077)

[3.3 Test data labels 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776078)

[3.4 Scores & metrics 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776079)

[3.5 Undisclosed test data set collection 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776080)

[4 Benchmarking methodology and architecture 10](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776081)

[5 Reporting methodology 11](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776082)

[6 Results 11](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776083)

[7 Discussion 11](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776084)

[8 Declaration of conflict of interest 11](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776085)

[9 References 11](file:///C%3A%5CUsers%5Cbanno%5CAppData%5CLocal%5CTemp%5CTemp2_Fillers.zip%5CFillers%5CFGAI4H-J-011-A01-R01.docx#_Toc51776086)

# Introduction

The presence of substandard and falsified medical products in countries and their use by patients threatens to undermine progress towards meeting the Sustainable Development Goals. Such products may be of poor quality, unsafe or ineffective, threatening the health of those that take them.

The problem of substandard and falsified medical products continues to increase, as globalized manufacturing and distribution systems grow ever more complex. That complexity heightens the risk that production errors will occur, or that medicines will degrade between factory and consumer. Increasing demand for medicines, vaccines and other medical products in almost every country, in addition to poor supply-chain management and the growth of e-commerce also creates opportunities for falsified medicines to be introduced into the supply chain.

Unfortunately, reliable information on the true public health and socioeconomic impacts of substandard and falsified medical products is sparse. A stronger evidence base is needed to help prevent, detect and respond to substandard and falsified medical products, and the public health threat they represent.

The falsified and sub-standard drugs today cause, according to the University of Edinburgh (childhood pneumonia model), the deaths of 250,000 children a year. Technological innovation, more precisely AI technology is one of the most effective means of dealing with increasingly creative counterfeiters. This topic group on AI-based detection of falsified medicine aims to develop artificial intelligence algorithms and to collect data available on falsified drugs.

## Document structure

TBC

## Topic description

### Overview

An assessed 250,000 children die each year from falsified or poor quality drugs ([2]Nayyar et al., 2019). Based on a 10% prevalence of substandard and falsified antibiotics, the model developed by the University of Edinburgh (childhood pneumonia model) with figures reproduced in Table 1 estimates that up to 72 430 childhood pneumonia deaths can be attributed to the use of substandard and falsified antibiotics if there is reduced antibiotic activity and this increases up to 169 271 deaths if substandard and falsified antibiotics have no activity ([5] WHO, 2017a).

Table 1: Findings on excess deaths from severe pneumonia due to substandard and falsified antibiotics in hospital and community settings

| Prevalence of substandard and falsified products (%) | Number of excess deaths in most likely scenario (two-fold increase in CFR) | Number of excess deaths in alternative scenario (four-fold increase in CFR) |
| --- | --- | --- |
| 1 | 8688 | 18372 |
| 5 | 37018 | 85438 |
| 10 | 72430 | 169271 |



Figure 1: Impact of substandard and falsified medical product

### Relevance

One of the reasons for the proliferation of falsified and substandard drugs is the lack of drug analysis equipment or simply too expensive equipment for LMICs that cannot afford a price tag of EUR 60.000 per device, many healthcare facilities and pharmaceutical depots on the continent are unable to get the technology to stop counterfeited medicine.

This lack of analytical devices leads to the lack of database that stores clean data on falsified and sub-standard drugs. Lack of high-quality data from the majority of LMICs means that first estimates will depend largely on data modelling.

AI is presented as a solution to the problem and must:

– Save life by verifying the authenticity of drugs (Customs, warehouses, pharmacies, hospitals) anywhere rapid verification is required.

– To build a database: To provide information and quantify the cost and socioeconomic impact of falsified and substandard medicines and establish the potential costs.

– To save money: An affordable solution than existing test machines on the market

– To identify the unknown chemical components: By using artificial intelligence-based technology to match collected drug data to datasets in our database stored on secure servers in the cloud.

In order to address the global challenge, the system should satisfy the following objectives:

i. Visual measurement data Devices should work with an affordable smartphone

ii. Easy backup generation of database

iii. Continuous operational status follow-up measurement devices

iv. The system should be able to work offline and online if connection is available

v. The system should allow identification and quantification of drugs

vi. The system should have two app, one for non-professional and the second for health professional

vii. The system should also provide automatic reporting (weekly, monthly, yearly)

viii. The system should provide simple answers and interpretation messages for non-professional

### Impact

Detecting substandard and falsified medical products requires a keen awareness of the likely risk factors (including product shortages), a culture that promotes the rapid exchange of information, and the technology and trained personnel to follow up suspicion with appropriate action.

The nearly 1500 cases reported to the WHO GSMS over its first four years of operation provide many very graphic examples of how global changes contribute to the production and trade in medicines and other products that fail to meet quality standards. Although they represent only a fraction of the true number of suspect products in circulation, these cases are already increasing knowledge of the forces that underpin and facilitate the manufacture, sale and distribution of substandard and falsified medical products ([6] WHO, 2017b).

Substandard and falsified medical products in one country can make diseases impossible to treat even in another country that has a very well-regulated medicine market. This is because substandard medicines promote antimicrobial resistance. Antibiotics and other antimicrobial medicines are manufactured and prescribed at doses designed to destroy the pathogens that are causing illness. If a treatment course contains only a fraction of the correct dose, or if it is so badly made that the active ingredients are not released properly, then it is only likely to destroy some of the pathogens, but not all of them. The ones that survive will be the ones that have mutated enough to survive low doses of the medicine. Usually, they do not reproduce very quickly. But with all the more susceptible strains killed by the weak medicines, they have room to multiply and spread to more people. There is clear evidence that resistance to the most important antimalarial medicine, artemisinin, first appeared in a part of the world where at one point between 38 and 90% of the artemisinin medicines on the market were substandard or falsified ([3] Paul Newton et al.,2001)

Improving detection technologies in the field and the laboratory detection

Detection is difficult, in part because sophisticated laboratory equipment is expensive to acquire and run, difficult to maintain and requires trained personnel who may be in short supply. A number of field detection technologies exist, but each has advantages and disadvantages and they are not always in the hands of people who have been trained to use them effectively ([4] Pisani, 2015).

It is important to develop new tools that will help front-line health care workers record and report suspect medicines an important entry point for detection.

● Impact of our work in this period of health crisis (COVID-19)

Global operation made by Interpol sees a rise in fake medical products related to COVID-19. Compared to the week of action in 2018, this latest edition of the operation reported an increase of about 18 per cent in seizures of unauthorized antiviral medication, and an increase of more than 100 per cent in seizures of unauthorized chloroquine (an antimalarial medication), which could also be connected to the COVID-19 outbreak.

## Ethical considerations

### Relevance of data and models

The term "data" here includes all of the spectral signatures associated with the metadata, including the name and brand of the drug, the active ingredient, the excipients, and the operator, the date of measurement, its geolocation, and all other information available. The development of classification models it is based on the data available for each brand of spectrometer and the merged data. It is important to remember that the scope of validity of the models can only extend to the information entered in the database. This step brings together all of the Machine Learning work, including the identification of outliers, the selection of spectral pre-processing, the selection of modelling strategies, the optimization of hyper parameters of the selected models.

It is therefore important to think carefully about the data used to train our technology. Are the data and models appropriate to the real-life problem they are solving? It is tempting to believe causal forces are at play when we find correlation on a single dataset. Does the data capture the true variable of interest? Is it consistent across observations and over time?

Training data rarely aligns with real-life goals. For example, off-line training data is not always representative of the true environment, and real-world objectives can be difficult to encode as simple value functions. What was the original purpose in collecting the data, and how did that determine its content? There is often a trade-off between accuracy and the intelligibility of a model (Dent, 2020). More predictive but harder-to-understand models can make it difficult to know which personal characteristics determine the decision and are therefore not available for validation against human judgment.

To have a good result, the colouring of the sample must be of good quality because colour is also a physicochemical criterion used when analysing drugs.

Table 2: Data annotation

| Annotation task category | Classification & Detection |
| --- | --- |
| Data modalities relevant for the topic group | 2D & videos |
| Nature of the annotations | Class label |
| Annotation procedure | Python/pip LabelIng |
| Annotation quality criterions | Size, width, height, depth |
| Metadata relevant for data annotation | Institution and responsible or corresponding PIConstruction dates of annotation dataset |

### Database

Develop tools and systems that countries can adapt to make reporting of suspected products easier and more efficient; develop and maintain a global database relating to the discovery of substandard or falsified medicines, for use by regulatory agencies globally; analyse global data to provide evidence-based recommendations for appropriate decision-making and effective action.

### Data availability

Anonymized data will be made available by the organizers of this Topic group.

Data available:

– Data from TrueSpec Africa and other companies

– Anonymized Drugs information

### Data quality

The data from TrueSpec devices is collected in real-time. This data is structured and analysed automatically and by data scientist before being made available for further processing.

The proprietary machine learning algorithm reads the scan from the scanner, identifies the drug, assesses its quality and lets the user know whether the drug is genuine or falsified, in about 20 seconds. The information collected is a spectral signature of the drug, and once checked, the database sends back information to an app on your phone, the vast database is continually updated.

### Accuracy

In pharmacy, the cost of a wrong decision can be very high. How accurate is the algorithm and how accurate does it need to be? Do our stakeholders understand the number of people who will be subject to a missed prediction given our measure of accuracy? A model in that misses only 1% shows phenomenally good performance, but if hundreds or thousands of people are still adversely affected, that might not be acceptable. Are there human inputs that can compensate for the system’s misses and can we design for that? What about post-deployment accuracy? Accuracy in training data does not always reflect real usage. Do we have a way to measure runtime accuracy? The world is dynamic and changes with time. Is there a way to continue to assess the accuracy after release? How often does it have to be reviewed?

There is always a need to improve the performance and for that, we have the following options:

* Adjust the current algorithm.
* Get more spectrum data or improve the data.
* Improve training.
* Switch algorithms.

At this stage, we have only the last option to choose i.e. Switch algorithms. So, we are switching to the Random Forest algorithm which is:

* An ensemble algorithm.
* Fits multiple trees with subsets of data.
* It includes average tree results to improve performance and control overfitting.

If we have got an excellent accuracy with the training data with Random Forest. We also need to test with the test data: if we have a good accuracy, that means our model has learned training data too well in comparison to the test data and this is called as overfitting.

### Size and severity of impact

Think about the numbers of people affected. Of course, we want to avoid harming anyone but knowing the size or the severity of negative consequences can justify the cost of extra scrutiny. We might also be able to design methods that mitigate for them. Given an understanding of the impact, we can make better decisions about the value required by the extra effort.

## Existing AI solutions

Wider use of authentication technologies while smartphone reporting technologies straddle the territories of detection and response, other technologies deal with prevention and detection. These include track and trace technologies, which allow for the seamless tracking of products through the supply chain. This sort of authentication technology allowed the Ugandan national regulatory authority to identify falsified contraceptives. Although the falsifiers had included a greyed-out area on the fake packaging that imitated a scratch-off authentication device, it was not actually scratchable. This alerted inspectors to the likelihood that the tablets were not genuine, a fact later confirmed both by the manufacturer of the original product, and by laboratory analyses, which found no active ingredient.

These devices use a barcode that are often photocopy and do not allow the user to know the chemical composition of the drugs. Existing devices to identify the chemical components of drugs are not adapted to the financial realities of the LMICs and do not allow an automatic constitution of the database.

## Existing work on benchmarking

TBC

# AI4H topic group

## AI for detection of falsified medicine (AI4DFM) objectives & deliverables

Mapping out existing AI4DFM systems and use cases:  Collect such use cases from members and others using these 9 descriptors

* Condition
* Spectral signature modality
* AI task/problem description (e.g. Image Classification)
* General algorithm description (if shareable)
* Project goal and current stage (if shareable)
* Input structure and format
* Output structure and format
* Evaluation metrics
* Explainability and Interpretability framework
1. Domain experts (spectroscopy) evaluation of structures and descriptors
2. AI/Technologists evaluation of structures and descriptors
3. Standardization of structures for each condition and task
4. Development of rapid drug Authentication regulatory recommendations
5. Undisclosed Test Data Collection
6. Review of existing regulations on spectroscopy
7. Evaluation and Benchmarking
8. Consideration of Assessment Platform for AI systems in spectroscopy
9. Precision Evaluation
10. Uncertainty Measurement Considerations
11. Explainability and Interpretability Considerations
12. Robustness Considerations
13. Privacy and Security Considerations
14. Ethical Considerations
15. Transparency, Fairness and Trust Considerations
16. Accountability Considerations
17. Legal Considerations

## Topic group structure

1. **Pros & cons**

This is a starting point for discussion about what we are trying to achieve, what needs and priorities we have. This can be done as a full group, or by asking pairs, to work on the pros and cons of one option and report back to the group. One benefit of this tool is that it allows us to talk about the drawbacks of particular ideas in a way that isn't too personalised. We simply need to list the benefits and drawbacks of each idea and compare the results. We may find that we don't all agree about what is a pro and what is a con. Or perhaps we have very different ideas about which pros and cons are most significant.

1. **Evaluating ideas**

These tools enable us to explore ideas in much more depth and evaluate how well they will work. In most cases it is useful to remember that our options are not necessarily limited to the ones that are currently on the table. If we cannot decide between two ideas, we can find a new option that combines the benefits of both and addresses key concerns. We use these tools to get a better understanding of everyone's priorities, even if we do not end up going ahead with any of the options exactly as they are.

1. **Urgent/Important grid**

A classic time-management tool that can be applied to group prioritisation. The group ranks ideas according to their urgency and importance, illustrated in Figure 2.



Figure 2: Urgency and importance grid for idea ranking

Six thinking hats

This tool encourages a group to look at a situation from a new angle. Each 'hat' represents a different way of looking at something. The roles the 'hats' bring us a chance to thoroughly examine every option and to prioritise or choose the best one(s).

* White hat: White hatted people concentrate on the facts what information and knowledge do we know about the situation? What can we learn about the situation from this information? What info is missing? Can we plug the gap? If not, can we take it into account when discussing the situation? What can we learn from past trends?
* Green hat: Green hat people think creatively in a no-criticism, freeform thinking kind of way.
* Red hat: Red hats are the emotional input of the discussion. They allow themselves to be intuitive and act as much on hunches as fact. They are sensitive to the emotional responses of others in the group.
* Black hat: Black hatted people evaluate ideas logically, and look for reasons to be cautious.
* Yellow hat: Yellow hats should think optimistically looking for the value in every possibility. What benefits does it bring?
* Blue hat: The blue hat is worn by the facilitator(s). They concentrate on process, calling on the other hats to add in their thinking as and when it's appropriate and making sure that each option is scrutinised from all perspectives. They are neutral, helping the group achieve its task without trying to shape the decision. This tool actively seeks out the optimistic analysis, the pessimistic analysis etc., so every idea is thoroughly tested and when the decision is made, it is made on the basis of a creative and thorough process.
1. **Evaluating meetings**

Evaluation allows us to learn from our experiences. It is a regular part of our online meetings as it gives us the chance for honest feedback on the process and content of the event, allowing us to improve. Everyone who participated in the topic group- AI-based detection of falsified medicine should be encouraged to take part in its evaluation.

# Method

* Overview of the benchmarking

## AI input data structure

* possible inputs for benchmarking
* ontologies, terminologies
* data format

## AI output data structure

* outputs to benchmark
* ontologies, terminologies
* data format

## Test data labels

* label types
* ontologies, terminologies
* data format

## Scores & metrics

* which metrics & scores to use for benchmarking
* considering relation to parameters stakeholders need for decision making
* considering scores that providers use
* considering the scope providers designed their solutions for
* Considering the state of the art in RCT, statistics, AI benchmarking etc.
* considering bias transparency

## Undisclosed test data set collection

* raw data acquisition / acceptance
* test data source(s): availability, reliability,
* labelling process / acceptance
* bias documentation process
* quality control mechanisms
* discussion of the necessary size of the test data set for relevant benchmarking results
* Specific data governance derived by DEL 5.1 Data requirements and DEL 5.5. Data handling

# 4 Benchmarking methodology and architecture

* technical architecture
* hosting (IIC, etc.)
* possibility of an online benchmarking on a public test dataset
* protocol for performing the benchmarking (who does what when etc.)
* AI submission procedure including contracts, rights, IP etc. considerations

# 5 Reporting methodology

* Report publication in papers or as part of ITU documents
* Online reporting
* public leaderboards vs. private leaderboards
* Credit-Check like on approved sharing with selected stakeholders
* Report structure including an example
* Frequency of benchmarking

# 6 Results

* insert here the reports of the different benchmarking runs

# 7 Discussion

* Discussion of the insights from executing the benchmarking on
	+ external feedback on the whole topic and its benchmarking
	+ technical architecture
	+ data acquisition
	+ benchmarking process
	+ benchmarking results
	+ field implementation success stories

# 8 Declaration of conflict of interest

* by each contributor to this document

# 9 References

[1] Dent, K. (2020). *Ethical Considerations for AI Researchers*. www.aaai.org

[2] Nayyar, G. M. L., Breman, J. G., Mackey, T. K., Clark, J. P., Hajjou, M., Littrell, M., & Herrington, J. E. (2019). Falsified and substandard drugs: Stopping the pandemic. In *American Journal of Tropical Medicine and Hygiene* (Vol. 100, Issue 5, pp. 1058–1065). American Society of Tropical Medicine and Hygiene. https://doi.org/10.4269/ajtmh.18-0981

[3] Paul Newton, Stephane Proux, Michael Green, Frank Smithuis, Jan Rozendaal, Sompol Prakongpan, Kesinee Chotivanich, Mayfong Mayxay, Sornchai Looareesuwan, Jeremy Farrar, Francois Nosten, N. J. W. (2001). *GPHF | Wellcome Trust 2001: Fake Artesunate In Southeast Asia*. https://www.gphf.org/minilab/hintergrund\_fake\_artesunate.htm

[4] Pisani, D. E. (2015). *Antimicrobial resistance: What does medicine quality have to do with it?* https://amr-review.org/sites/default/files/ElizabethPisaniMedicinesQualitypaper.pdf

[5] WHO. (2017a). *A study on the public health and socioeconomic impact of substandard and falsified medical products*. https://www.who.int/medicines/regulation/ssffc/publications/se-study-sf/en/

[6] WHO. (2017b). *WHO Global Surveillance and Monitoring System for substandard and falsified medical products*. https://apps.who.int/iris/bitstream/handle/10665/326708/9789241513425-eng.pdf?ua=1

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