|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-L-012-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 19-21 May 2021 |
| **DOCUMENT** |
| **Source:** | TG-Falls Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Falls) |
| **Purpose:** | Discussion |
| **Contact:** | Pierpaolo PalumboTG-Falls Topic DriverUniversity of BolognaItaly | Tel: +39 3402378412E-Mail: pierpaolo.palumbo@unibo.it |
| **Contact:** | Inês SousaAssociação Fraunhofer Portugal Research – Fraunhofer AICOSPortugal | Tel: +351 220 430 326Email: ines.sousa@fraunhofer.pt  |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based prevention of falls among the elderly. 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. 10.04. This draft will be a continuous input- and output document. |

|  |  |
| --- | --- |
| **Change notes:** | Version 3 (submitted as FGAI4H-L-012-A01 to meeting L)* Update of sections Topic description, Ethical considerations, and Existing work on benchmarking systems.
* Draft schema of the version 0 of the benchmarking platform

Version 2 (submitted as FGAI4H-K-012-A01 to meeting K)* Provisional draft
* Updated accorded to the new template

Version 1 (submitted as FG-AI4H-J-012-A01 to meeting J) |

Contributors

|  |  |
| --- | --- |
| Pierpaolo PalumboUniversity of BolognaItaly | Tel: +39 3402378412Email: Pierpaolo.palumbo@unibo.it  |
| Inês Sousa Associação Fraunhofer Portugal Research – Fraunhofer AICOSPortugal | Tel: +351 220 430 326Email: ines.sousa@fraunhofer.pt |
| Barry GreeneKinesis Health Technologies Ltd.Ireland | Email: barry.greene@kinesis.ie |
| Kimberley S. van SchootenUniversity of New South Wales, Sydney, NSWAustralia. | Email: k.vanschooten@neura.edu.au |

**CONTENTS**

|  Page |
| --- |
| [1 Introduction 5](#_Toc71913813)[2 About the FG-AI4H topic group on Falls among the elderly 5](#_Toc71913814)[2.1 Documentation 6](#_Toc71913815)[2.2 Status of this topic group 6](#_Toc71913816)[2.2.1 Status update for meeting B (Lausanne) 6](#_Toc71913817)[2.2.2 Status update for meeting C (New York) 6](#_Toc71913818)[2.2.3 Status update for meeting D (Shanghai) 7](#_Toc71913819)[2.2.4 Status update for meeting E (Geneva) 7](#_Toc71913820)[2.2.5 Status update for meeting F (Zanzibar) 7](#_Toc71913821)[2.2.6 Status update for meeting G (New Delhi) 7](#_Toc71913822)[2.2.7 Status update for meeting H (Brasilia) 7](#_Toc71913823)[2.2.8 Status update for meeting J (E-meeting) 8](#_Toc71913824)[2.2.9 Status update for meeting K (E-meeting) 8](#_Toc71913825)[2.2.10 Status update for meeting L (E-meeting) 9](#_Toc71913826)[2.3 Topic group participation 9](#_Toc71913827)[3 Topic description 10](#_Toc71913828)[3.1.1 Definition of the AI task 10](#_Toc71913829)[3.1.2 Current gold standard 10](#_Toc71913830)[3.1.3 Relevance and impact of an AI solution 11](#_Toc71913831)[3.1.4 Existing AI solutions 11](#_Toc71913832)[4 Ethical considerations 12](#_Toc71913833)[5 Existing work on benchmarking 13](#_Toc71913834)[5.1 Publications on benchmarking systems 13](#_Toc71913835)[5.2 Benchmarking by AI developers 15](#_Toc71913836)[5.3 Relevant existing benchmarking frameworks 15](#_Toc71913837)[6 Benchmarking by the topic group 16](#_Toc71913838)[6.1 Subtopic Fall prediction 16](#_Toc71913839)[6.1.1 Benchmarking version 0 16](#_Toc71913840)[6.1.2 Benchmarking version [X] 28](#_Toc71913841)[7 Overall discussion of the benchmarking 28](#_Toc71913842)[8 Regulatory considerations 29](#_Toc71913843)[8.1 Existing applicable regulatory frameworks 29](#_Toc71913844)[8.2 Regulatory features to be reported by benchmarking participants 30](#_Toc71913845)[8.3 Regulatory requirements for the benchmarking systems 30](#_Toc71913846)[8.4 Regulatory approach for the topic group 30](#_Toc71913847)[9 References 31](#_Toc71913848)[Annex A: Glossary 34](#_Toc71913849)[Annex B: Declaration of conflict of interests 35](#_Toc71913850) |

**List of Tables**

[Table 1: Topic group output documents 6](#_Toc71913851)

[Table 2: Datasets available/considered so far 22](#_Toc71913852)

**List of Figures**

|  Page |
| --- |
| [Figure 1: Example of a figure 33](#_Toc71913853) |

FG-AI4H Topic Description Document

Topic group-Falls among the elderly

# Introduction

Falls are one of the most common health problems in the elderly population. About a third of community-dwelling adults aged 65 years or older fall each year [1], and these events represent more than 50% of the hospitalizations due to lesions in this age group. Falls are also considered one of the main causes for loss of independence and institutionalization. In 10% of cases falls result in fractures, thus contributing to significant increases in morbidity and mortality. Direct health care costs associated with this phenomenon are high, reaching yearly costs of 25 billion euros in the European Union and 31 billion dollars in the United States of America [2].

Falls have a multifactorial origin, however most of the fall risk factors are amendable by implementing falls prevention programs based on improving strength and balance and modifying behaviours [3]. Nevertheless, fall risk screenings and the implementation of such falls’ prevention programs are rarely part of the community-dwelling elder’s routine. Traditional clinical scales for fall risk assessment are the Morse Fall Scale [4], the Berg Balance Scale [5], and the Performance Oriented Assessment of Mobility Problems in Elderly Patients [6]. Different other multifactorial predictive models have been proposed and only few have been validated [7]. Despite being recommended by international health bodies, such as the National Institute for Health and Care excellence (NICE)[[1]](#footnote-1), multifactorial fall risk screening is still not widespread in the clinical practice. One of the reasons for this shortcoming is the difficulty in combining the multiple parameters evaluated in a meaningful scale that is able to discriminate those who are more likely to fall in a period of time following the assessment.

Systems based on Artificial Intelligence (AI) techniques identify individuals at high risk of falling. These systems leverage input information on personal risk factors for falls and/or signal recordings containing information on personal motor and balance capabilities. Their output is generally an indicator that expresses the individual risk of falling within a given time period and/or indications for reducing the risk.

A platform for standardized benchmarking of these systems would allow to consistently evaluate their predictive accuracy and their efficacy in preventing falls.

This topic description document specifies the standardized benchmarking for systems to prevent falls among the elderly. It serves as deliverable No. 10.4 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

# About the FG-AI4H topic group on Falls among the elderly

The introduction highlights the potential of a standardized benchmarking of AI systems for preventing falls among the elderly 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-Falls (Falls among the elderly) at the meeting A in Geneva, Switzerland, on 25-27 September 2018.

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 A in Geneva, Switzerland, on 25-27 September 2018, Inês Sousa from Associação Fraunhofer Portugal Research – Fraunhofer AICOS was nominated as topic driver for the TG-Falls.

In December 2020, Inês Sousa announced her absence for the period December 2020 – September 2021. Pierpaolo Palumbo was asked to drive the topic group for this period.

## Documentation

This document is the TDD for the TG-Falls. 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 preventing falls among the elderly. It describes the existing approaches for assessing the quality of fall prevention 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 10.4 Falls among the elderly (TG-Falls).” 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-K-012-A03 | The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting K |
| FGAI4H-K-012-A01 | Latest update of the Topic Description Document of the TF-Falls for meeting K |
| FGAI4H-J-012-A01 | Latest update of the Topic Description Document of the TG-Falls for meeting J |
| FGAI4H-H-012-A02 | Latest update of the Call for Topic Group Participation (CfTGP) for meeting H |
| FGAI4H-H-012-A03 | The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting H |

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-Falls.aspx

## Status of this topic group

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

### Status update for meeting B (Lausanne)

First submission was provided in response to the ITU-T FG-AI4H's call for proposals on use cases and data A‑102. The document was presented remotely.

|  |  |  |
| --- | --- | --- |
| FGAI4H-C-014 | Lausanne, 22-25 January 2019 | Status Report of: Reducing risk of falling among elderly |

### Status update for meeting C (New York)

The topic group description was refined and presented remotely.

### Status update for meeting D (Shanghai)

Inês Sousa participated remotely in the Shanghai meeting and provided an update on the progress of the topic “Standardized benchmarking of AI to prevent falls among the elderly”.

Main points:

– There were no contacts or manifestations of interest from other research groups regarding this topic;

– It was suggested that some of the groups that have been actively publishing in this area could be contacted;

– It was mentioned that the possibility of enlarging the scope of the topic to include fall detection datasets could also be considered, despite the unavailability of Fraunhofer AICOS to provide a dataset.

### Status update for meeting E (Geneva)

The Personal Health Systems Laboratory from University of Bologna joined the Topic Group following the manifestation of interest sent by Pierpaolo Palumbo, biomedical engineer, working on algorithms for health risk assessment, with a focus on fall risk in community-dwelling older adults and lower-limb amputees. He is a post-doctoral fellow at the Personal Health Systems Laboratory, headed by Prof. Lorenzo Chiari, at the University of Bologna.

### Status update for meeting F (Zanzibar)

Following the suggestion from the Personal Health Systems Laboratory, a list of longitudinal studies on ageing with data on falls has been drafted. A draft letter was created inviting these studies to share the data of the new waves with our consortium for benchmarking the algorithms.

### Status update for meeting G (New Delhi)

Update on contacts with longitudinal studies on ageing with data on falls and groups that have been actively publishing in this area, results in demonstrations of interest to join the Focus Group from Kim van Schooten, PhD, Human Frontier Science Program Postdoctoral Fellow, Conjoint Senior Lecturer, UNSW Medicine, UNSW Ageing Futures Institute, and, Barry Greene from Kinesis, Ireland.

A paper was published [8].

### Status update for meeting H (Brasilia)

Demonstrations of potential interest from two groups relative to two epidemiological studies about ageing with data on falls: InCHIANTI[[2]](#footnote-2) and TILDA[[3]](#footnote-3).

Questions:

* The InCHIANTI dataset is generally shared with interested researchers under formal agreements with a non-disclosure clause. The InCHIANTI board keep track of the researchers that have accessed the different versions of their dataset. Furthermore, the waves that could be available for the benchmarking activities of the Focus Group have been shared with a relatively small number of persons. Could the benchmarking framework accept these data (excluding from the benchmark the models coming from researchers that have accessed the data)?
* The different datasets are mostly similar but slightly different in terms of available variables. Should we keep all useful variables from both datasets or should we restrict the datasets to the variables that are present in both datasets?
* Because of their design, neither FallSensing nor InCHIANTI can be considered rigorously representative of the Portuguese or Italian older population. How do we take this into account?
* What is the approximate time schedule of our activities?

A conference call to which the subjects expressing interest in the activities of the Topic Group have been invited, was held.

Participants:

Inês Sousa, Fraunhofer AICOS

Pierpaolo Palumbo, University of Bologna

Stefania Bandinelli, SOC Geriatria -USLToscana Centro, Firenze

Barry Greene, Chief Technology Officer, Kinesis Health Technologies, Ireland

Salman Khan, Assistant Professor in the department of electrical engineering, University of Engineering and Technology Peshawar, Pakistan

Brief summary of the points discussed:

* A systematic assessment of all solutions and studies regarding fall risk assessment is missing;
* Quality levels and standards for algorithm evaluation should be defined;
* Most datasets available are heterogenous and consider different variables and functional tests, may include data from sensors or not.

Action Points:

* Systematize information regarding fall risk assessment;
* Continue the discussion of the variables to be considered, and methods/best practices for algorithm evaluation;
* Discuss with the Working Group how should the Benchmarking Framework deal with heterogenous datasets.

### Status update for meeting J (E-meeting)

The Topic Group participants have met and discussed guidelines for standardization and evaluation of AI models to estimate the risk of falling.

Participants:

Inês Sousa, Fraunhofer AICOS

Pierpaolo Palumbo, University of Bologna

Stefania Bandinelli, SOC Geriatria -USLToscana Centro, Firenze

Barry Greene, Chief Technology Officer, Kinesis Health Technologies, Ireland

Arnab Paul, CEO Patient Planet, WHO Roster of Expert – DigitalHealth, India

### Status update for meeting K (E-meeting)

Preparation of the workshop Artificial Intelligence and fall prediction within the EU Falls Festival 2021 (<https://eufallsfest2021.eu/index.php/workshops>).

Notification that this EU Falls Festival 2021 has been postponed to 2022.

Notification by Barry Greene (Kinesis Health Technologies Ltd.) about the development of a new app for self-assessment of balance and fall risk.

Pierpaolo Palumbo (University of Bologna) nominated interim topic driver for the period December 2020-September 2021.

Re-formatting of the TDD according the new template (FG-AI4H-J-105).

### Status update for meeting L (E-meeting)

In this TDD, we have updated the state of the art and have included relevant information from the consensus reported by the Prevention of Falls Network Europe (ProFaNE) group on definitions and measures for fall injury prevention trials [9]. Kimberley van Schooten has contributed a book chapter to the discussion of the Topic Group [10].

Updates on Ethical considerations.

We have asked permission to access the harmonized version of the datasets belonging to the Health and Retirement Study (HRS) family [11].

We are starting an internal discussion on whether we should open a sub-topic on fall detection and whether some research groups could provide their data on falls recorded with wearable inertial sensors [12].

## Topic group participation

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

* <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Falls.pdf>

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

* <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Falls.aspx>

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

* <https://itu.zoom.us/my/fgai4h>

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

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

* <https://itu.int/go/fgai4h/join>

In addition to the general FG-AI4H mailing list, the topic group TG-Falls has a dedicated mailing list*:*

* fgai4htgfalls@lists.itu.int

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

* <https://itu.int/go/fgai4h>

# Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in Falls among the elderly 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-Falls currently has no subtopics.

### 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 in more detail in chapter 4). This section corresponds to [DEL03](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7997F2C1-5A1D-4409-B2A0-CBC4E9CE8CDA%7D&file=DEL03.docx&action=default) *“AI requirements specifications*,” which describes the functional, behavioural, and operational aspects of an AI system.

AI-based systems for fall prediction aim to provide a subject-specific risk score of falling in the future within a given time window, given information about the subject’s risk factors for falls and/or their balance or motor ability. The subjects under assessment are older adults.

According to the World Health Organization a fall is defined as “an event which results in a person coming to rest inadvertently on the ground or floor or other lower level” [13]. Similarly, according to the Prevention of Falls Network Europe Consensus (ProFaNE), a fall is defined as “an unexpected event in which the participants come to rest on the ground, floor, or lower level” [9]. Both definitions could be accepted for the purposes of this Topic Group.

The algorithms should run on records of a single subject or on a dataset containing multiple records of different subjects.

The algorithms should be able to run on records with missing values or on datasets with missing variables.

The algorithms should produce one or more of the following subject-specific outputs:

1. the probability of falling at least once in a future time window,
2. the expected number of falls in a future time window,
3. an ordered, fuzzy label for the risk of falling (e.g., low, medium, high).

The length of the time window for fall prediction is generally set at 12 months, but different time windows could be accepted, ranging from 6 to 24 months.

Additionally, the algorithms could further provide suggestions on possible preventive actions to take.

### Current gold standard

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

The current gold standard for establishing the predictive ability of AI algorithms for fall prediction is the occurrence of falls, recorded prospectively with respect to the time of risk factor assessment. Preferably, falls should be ascertained using prospective daily recording and a notification system with a minimum of monthly reporting [9].

At present, a variety of tools for fall risk assessment have been proposed. The Timed Up and Go Test (TUG) is one of the most widespread. Its performance has been evaluated many times over the years in different studies and population. Two systematic reviews report much heterogeneity in its performances across studies and a relatively low average predictive accuracy [14], [15]. In particular, the sensitivity was estimated to be 0.31 (95% confidence interval (CI) 0.13-0.57), the specificity 0.74 (95% CI 0.52-0.88), and the area under the Receiver Operating Characteristic (ROC) curve (AUC) = 0.57 (95% CI 0.54-0.59) [14].

Another tool for discriminating people at risk of falling is the algorithm proposed by the American Geriatrics Society (AGS) and British Geriatrics Society (BGS) within their guidelines for fall prevention [16]. Its sensitivity was estimated to be 0.36 (95% CI 0.23-0.53) and its specificity 0.84 (95% CI 0.79-0.88) [17].

* 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?

### 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.

Falls are one of the most common health problems in the elderly population, about a third of community-dwelling adults aged 65 years or older fall each year [1], and these events represent more than 50% of the hospitalizations due to lesions in this age group. Falls are also considered one of the main causes for loss of independence and institutionalization. In 10% of cases falls result in fractures, thus contributing to significant increases in morbidity and mortality. Direct health care costs associated with this phenomenon are high, reaching yearly costs of 25 billion euros in the European Union and 31 billion dollars in the United States of America [2].

Some preventive interventions have been shown to be effective, but their implementation on the whole population is unfeasible or not clinically appropriate. Thus, AI-based systems for fall prevention are aimed to identify those to prioritize for fall prevention interventions, and the most appropriate type of interventions for them.

It was estimated that deploying the AGS/BGS algorithm for fall risk assessment within a preventive intervention could decrease the number needed to treat (NNT) of about 17% (95% CI 4.1%–34.0%) with respect to another preventive strategy not based on a risk assessment tool [17]. Impact assessment studies for other tools with better predictive accuracy are lacking.

The creation of a standardized platform for benchmarking fall prediction tools, would allow to assess these tools in a rigorous and comparable manner, informing about strengths and limitations of each tool, overcoming concerns about over-fitting and over-optimism raised by some authors [7], [18].

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

### Existing AI solutions

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

A recent review found 72 prognostic models developed or validated in prospective cohorts [7]. Of those, only three were validated and had an AUC between 0.62 and 0.69.

[to revise]

Howcroft et al. [21] reviewed previous studies focusing on the fall risk assessment with inertial sensors. The authors concluded that future research should i) consider investigating the relationship between the models’ predictive variables and specific fall risk factors and ii) focus on groups with an increased fall risk due to some diseases. A weak point of most studies is not having used separate datasets for model training and validation, which could have impacted the models’ applicability beyond the training set population. Another aspect to be considered is that clinical assessment thresholds were not used consistently across the research studies included in the review. The prospective fall occurrence rate is considered to be the most reliable criterion for dividing subjects into non-fallers and fallers [21]; however, this criterion was only used in 15% of the studies. Regarding the retrospective fall assessment, the most relevant limitations are the inaccurate recording of fall histories most commonly assessed by self-reported questionnaires and the fact that balance, strength, and gait parameters can change due to past falls.

Greene et al. [22] reported in 2019 that 8521 participants (72.7 ± 12.0 years, 5392 female) from six countries were assessed using a digital falls risk assessment protocol. Data consisted of wearable sensor data captured during the Timed Up and Go (TUG) test along with data on falls risk factors from self-reported questionnaires, applied to previously trained and validated classifier models. We found that 25.8% of patients reported a fall in the previous 12 months. Of the 74.6% of participants that had not reported a fall, 21.5% were found to have a high predicted risk of falls. Overall 26.2% of patients were predicted to be at high risk of falls. 29.8% of participants were found to have slow walking speed, while 19.8% had high gait variability and 17.5% had problems with transfers.

* 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.

# Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable [DEL01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*AI4H ethics considerations,”* which was developed by the working group on “Ethical considerations on AI4H” (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-Falls.

Specific ethical considerations should include the fact that fear of falling is itself a risk factor for falls and a disabling condition leading to a decline in physical and mental performance and loss of quality of life [23]. Therefore, fall risk communication should be made with care, possibly by a health professional. Furthermore, an indication of the presence of high fall risk should be accompanied by a plan for risk mitigation and a comprehensive explanation of preventive measures.

The data that will be used for the benchmarking will come exclusively from studies (‘parent studies’) already approved by competent Ethical Committees. We do not think that it is needed to seek for ethical approval for reusing these data within the ITU/WHO benchmarking platform. However, for most parent studies it is needed to submit a proposal for getting access to the data. Within these proposals, it is usually required to indicate the time period when the data will be used and other details regarding data handling and processing. The period that the benchmarking platform will be operative and some other details are currently unknown.

We foresee that most parent studies will be population-based, thus being representative of the target population. Furthermore, they will come from different geographical areas and countries with different income levels. Other datasets may be based just on convenience samples. In this case, either unbiasedness should be sought with statistical techniques (e.g., using inverse probability weights) or a disclaimer about the nature of the data should be written next to the performance results.

* 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?

# Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and fall prevention 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.

## Publications on benchmarking systems

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

Some methodological elements regarding data specification, data requirements, and data acquisition come from the consensus on definitions and measures for fall injury prevention trials, published in 2005 by the Prevention of Falls Network Europe (ProFaNE) [9]. Among the outputs of the consensus:

* They identified physical activity, psychological consequences, and generic health related quality of life (HRQoL) as domains of interest for fall injury prevention.
* They proposed a formal definition of falls and the way to phrase it in questionnaires for fall ascertainment considering the lay perspective.
* They indicated methods for fall data acquisition. They recommended prospective daily recording, a notification system with a minimum of monthly reporting, and telephone or face-to-face interviews to rectify missing data and to acquire further details on falls and injuries.
* They set specifications for fall data summary. In particular, they recommended reporting the number of falls, the number of fallers/non-fallers/frequent fallers, the fall rate per person year, and the time to first fall.

The predictive ability of tools for fall prediction (as it is for other prognostic tools) is usually evaluated on two aspects: discriminative ability and calibration [24]. The AUC or the *c* statistics are generally used for evaluating the discriminative ability. Calibration can be evaluated visually from calibration curves, or with the Brier Score, which also involves aspects related to the discriminative ability [25]. Calibration cannot be computed when the output of the prediction tool is not probabilistic [26].

The results of the validation procedure should be reported following the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist [27].

Although no formal platform has been established for benchmarking multiple

Within the literature, no platform has been established to systematically evaluate multiple predictive tools for falls on a common set of data. Instead, there are examples of tools tested on multiple populations, either in original studies or in systematic reviews collecting the results from different original studies.

The Timed Up and Go Test (TUG) is one of the most widespread among the traditional tools for risk screening. Although it is not based on AI, it is worth discussing because its performance has been evaluated many times over the years in different studies and population. Two systematic reviews report much heterogeneity in its performances across studies and a relatively low average predictive accuracy [14], [15]. In particular, the sensitivity was estimated to be 0.31 (95% confidence interval (CI) 0.13-0.57), the specificity 0.74 (95% CI 0.52-0.88), and the area under the Receiver Operating Characteristic (ROC) curve (AUC) = 0.57 (95% CI 0.54-0.59) [14].

Among the new tools for predicting falls in the elderly that consider multiple risk factors, FRAT-up was validated on four European datasets of longitudinal studies about ageing. It showed to be more accurate that simple traditional tools and exhibits much heterogeneity in its performance across different populations [28], [29]. Its discriminative ability was quantified with an AUC between 0.562 to 0.699 (mean 0.646, 95% CI 0.584–0.708). Calibration was also poor and heterogeneous across populations.

Heterogeneity across datasets and populations was also found on fall incidence and fall risk factors prevalence rates, the reason being yet to be fully uncovered [30]. From this experience we believe that benchmarking fall prediction algorithms on different datasets/populations is necessary to obtain robust estimates of their performance. Furthermore, these datasets should be as much as possible representative of their target populations.

.

* 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)?

## Benchmarking by AI developers

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

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

## Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

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

# Benchmarking by the topic group

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

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

## Subtopic Fall prediction

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

The benchmarking of tools for fall prediction 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?

Question: shall all the Topic Groups of the Focus Group share the same platform?

### Benchmarking version 0

This section includes all technological and operational details of the benchmarking process for the benchmarking version 0.

#### Overview

This section provides an overview of the key aspects of this benchmarking iteration, version 0.

This version 0 of the benchmarking platform is a prototype. Its overall scope is to trigger discussion among the members of the Topic Group and other members of the Focus Group about the technical specifications and functional requirements of the platform.

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

#### Benchmarking methods

This section provides details about the methods of the benchmarking version 0. 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).

##### 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.

This version 0 of the benchmarking platform runs locally on a personal computer.

It is made of datafiles (.csv, .RData, .dta, .sav, .sas7bdat) and scripts (R, Python, Stata).

It receives input data files and AI algorithms to be tested. Each set of input datafiles represents a study population on which the AI algorithm should be applied.

It outputs scores representing the performance of the AI algorithms on the populations represented by the input data files.

The input data files are pre-processed by harmonization scripts. The harmonization scripts create harmonized datafiles from the input data files. The harmonized data files have the same format and same semantics.

The AI algorithms interact with the harmonized data files through interface scripts. The interface scripts apply the AI algorithms on the harmonized data files and produce performance metrics which are the output of the platform.

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

##### Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

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

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

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

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

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

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

##### Benchmarking process

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

Questions to discuss:

shall all the Topic Groups of the Focus Group have the same benchmarking process (technological platform, rules, time schedule, result publication…)?

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

#### AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of AI-based prevention of falls among the elderly. 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.

* The data shall be a collection of diverse datasets.
* The datasets should contain information on falls and acknowledged and/or putative fall risk factors (either clinical or sensor-based).
* The subjects represented in the datasets should be older adults.
* The variables of the datasets should be named as much as possible with harmonized conventions. Standard nomenclatures to consider could be: ICD-10, SNOMED, ICF, ATC for drugs, etc.
* Each dataset could be openly accessible, undisclosed, or partly accessible and partly undisclosed.
* Each dataset should be accompanied by documentation with a protocol on data collection methods.
* Others…
* 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

#### AI output data structure

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

The algorithms should run on records of a single subject or on a dataset containing multiple records of different subjects.

The algorithms should produce one or more of the following subject-specific outputs:

1. the probability of falling at least once in a future time window,
2. the expected number of falls in a future time window,
3. an ordered, fuzzy label for the risk of falling (e.g., low, medium, high).

The length of the time window for fall prediction is generally set at 12 months, but different time windows could be accepted, ranging from 6 to 24 months.

Additionally, the algorithms could further provide suggestions on possible preventive actions to take.

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

#### Test data label/annotation structure

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

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

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

#### Scores and metrics

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

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

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

#### 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.

Table 2: Datasets available/considered so far

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Country | Type of risk factors | Falls [to check…] | Public/undisclosed |
| FallSensing | Portugal | Clinical and sensor-based | Prospective and retrospective | Possibility to keep part of the dataset undisclosed |
| InCHIANTI | Italy | Clinical and sensor-based | Prospective and retrospective |  |
| HRS | USA | Clinical | Retrospective and prospective on the next wave | Public |
| MHAS | Mexico | Clinical | Retrospective and prospective on the next wave | Public |
| ELSA | UK | Clinical | Retrospective and prospective on the next wave | Public |
| SHARE | Different European countries | Clinical | Retrospective and prospective on the next wave | Public |
| KLoSA | Korea | Clinical | Retrospective and prospective on the next wave | Public |
| IFLS | Indonesia | Clinical | Retrospective and prospective on the next wave | Public |
| TILDA | Ireland | Clinical | Retrospective and prospective on the next wave | (Possibility to discuss to keep a wave of TILDA undisclosed for a short period of time) |
| CHARLS | China | Clinical | Retrospective and prospective on the next wave | Public |
| LASI | India | Clinical | Retrospective | Public |
| CRELES | Costa Rica | Clinical | Retrospective and prospective on the next wave | Public |
| ELSI | Brazil | Clinical | Retrospective and prospective on the next wave | Public |
| Datasets from Kinesis (?) |  | Clinical and sensor-based |  | (??) |

FallSensing is made of 403 annotated data samples. The data have been kept undisclosed. Only a small part of it can be made publicly available (1 or 2%) for model training, while the rest can be used for model testing. However, since the data acquisition protocol is published in an open access journal, it can be easily replicated by peers.

The protocol of the InCHIANTI is much similar to the one adopted in FallSensing. Data on prospective falls and instrumented functional tests are available for the last two waves (FU4 and FU5). The data are generally shared with other research groups on the basis of formal agreements with a non-disclosure clause.

TILDA dataset could become available in 2021.

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

#### Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) on *data handling* and [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) on *data sharing practices*).

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

#### Baseline acquisition

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

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

#### Reporting methodology

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

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

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

#### Result

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

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

#### Discussion of the benchmarking

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

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

#### Retirement

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

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

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

### Benchmarking version [X]

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

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

# Overall discussion of the benchmarking

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

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

# Regulatory considerations

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

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *“*[*Regulatory considerations on AI for health”* *(WG-RC)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“Guidelines for AI based medical device (AI-MD): Regulatory requirements”* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) identifies standards and best practices that are relevant for the “*AI software lifecycle specification*.*”* The following sections discuss how the different regulatory aspects relate to the TG-Falls.

## 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 prevention of falls among the elderly.

* 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?

## Regulatory features to be reported by benchmarking participants

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

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

## Regulatory requirements for the benchmarking systems

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

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

## Regulatory approach for the topic group

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

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

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

# References

[1] World Health Organization, *WHO global report on falls prevention in older age*. Geneva, Switzerland: World Health Organization, 2007.

[2] E. R. Burns, J. A. Stevens, and R. Lee, “The direct costs of fatal and non-fatal falls among older adults — United States,” *J. Safety Res.*, vol. 58, pp. 99–103, Sep. 2016, doi: 10.1016/j.jsr.2016.05.001.

[3] S. Hopewell, B. Copsey, P. Nicolson, B. Adedire, G. Boniface, and S. Lamb, “Multifactorial interventions for preventing falls in older people living in the community: A systematic review and meta-analysis of 41 trials and almost 20 000 participants,” *Br. J. Sports Med.*, vol. 54, no. 22, pp. 1340–1350, Nov. 2019, doi: 10.1136/bjsports-2019-100732.

[4] J. M. Morse, R. M. Morse, and S. J. Tylko, “Development of a Scale to Identify the Fall-Prone Patient,” *Can. J. Aging / La Rev. Can. du Vieil.*, vol. 8, no. 4, pp. 366–377, 1989, doi: 10.1017/S0714980800008576.

[5] K. Berg, S. Wood-Dauphinée, J. I. Williams, and D. Gayton, “Measuring Balance in the Elderly Preliminary development of an Instrument,” *Physiother. Canada*, vol. 41, no. 6, pp. 304–311, 1989.

[6] M. E. Tinetti, “Performance-oriented assessment of mobility problems in elderly patients.,” *J. Am. Geriatr. Soc.*, vol. 34, no. 2, pp. 119–26, Feb. 1986, Accessed: Mar. 25, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3944402.

[7] G. V. Gade *et al.*, “Predicting falls in community-dwelling older adults: a systematic review of prognostic models,” *BMJ Open*, vol. 11, no. 5, p. e044170, May 2021, doi: 10.1136/bmjopen-2020-044170.

[8] J. R. Silva, I. Sousa, and J. S. Cardoso, “Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls,” *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, 2019, doi: 10.1109/JBHI.2019.2951230.

[9] S. E. Lamb, E. C. Jørstad-Stein, K. Hauer, and C. Becker, “Development of a common outcome data set for fall injury prevention trials: the Prevention of Falls Network Europe consensus.,” *J. Am. Geriatr. Soc.*, vol. 53, no. 9, pp. 1618–22, Sep. 2005, doi: 10.1111/j.1532-5415.2005.53455.x.

[10] K. S. van Schooten and M. Brodie, “Fall detection and risk assessment with new technologies,” in *Falls in Older People*, .

[11] “Gateway to Global Aging Data.” https://g2aging.org/ (accessed May 11, 2021).

[12] E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, “On the Heterogeneity of Existing Repositories of Movements Intended for the Evaluation of Fall Detection Systems,” *J. Healthc. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/6622285.

[13] World Health Organization Department of Ageing and Life Course, *WHO global report on falls prevention in older age*. Geneva, Switzerland: World Health Organization, 2008.

[14] E. Barry, R. Galvin, C. Keogh, F. Horgan, and T. Fahey, “Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta- analysis.,” *BMC Geriatr.*, vol. 14, no. 1, p. 14, Jan. 2014, doi: 10.1186/1471-2318-14-14.

[15] D. Schoene *et al.*, “Discriminative ability and predictive validity of the timed up and go test in identifying older people who fall: systematic review and meta-analysis.,” *J. Am. Geriatr. Soc.*, vol. 61, no. 2, pp. 202–8, Feb. 2013, doi: 10.1111/jgs.12106.

[16] Panel on Prevention of Falls in Older Persons and American Geriatrics Society and British Geriatrics Society, “Prevention of Falls in Older Persons: AGS/BGS Clinical Practice Guideline,” 2011. [Online]. Available: http://geriatricscareonline.org/toc/updated-american-geriatrics-societybritish-geriatrics-society-clinical-practice-guideline-for-prevention-of-falls-in-older-persons-and-recommendations/CL014.

[17] P. Palumbo, C. Becker, S. Bandinelli, and L. Chiari, “Simulating the effects of a clinical guidelines screening algorithm for fall risk in community dwelling older adults,” *Aging Clin. Exp. Res.*, pp. 1–8, Oct. 2018, doi: 10.1007/s40520-018-1051-5.

[18] T. Shany, K. Wang, Y. Liu, N. H. Lovell, and S. J. Redmond, “Review: Are we stumbling in our quest to find the best predictor? Over-optimism in sensor-based models for predicting falls in older adults,” *Healthcare Technology Letters*, vol. 2, no. 4. IET Digital Library, pp. 79–88, Aug. 03, 2015, Accessed: Aug. 13, 2015. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/htl.2015.0019.

[19] K. L. Perell, A. Nelson, R. L. Goldman, S. L. Luther, N. Prieto-Lewis, and L. Z. Rubenstein, “Fall risk assessment measures: an analytic review.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 56, no. 12, pp. M761-6, Dec. 2001, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11723150.

[20] M. Vassallo, L. Poynter, J. C. Sharma, J. Kwan, and S. C. Allen, “Fall risk-assessment tools compared with clinical judgment: an evaluation in a rehabilitation ward.,” *Age Ageing*, vol. 37, no. 3, pp. 277–81, May 2008, doi: 10.1093/ageing/afn062.

[21] J. Howcroft, J. Kofman, and E. D. Lemaire, “Review of fall risk assessment in geriatric populations using inertial sensors.,” *J. Neuroeng. Rehabil.*, vol. 10, no. 1, p. 91, Aug. 2013, doi: 10.1186/1743-0003-10-91.

[22] B. R. Greene, K. McManus, S. J. Redmond, B. Caulfield, and C. C. Quinn, “Digital assessment of falls risk, frailty, and mobility impairment using wearable sensors,” *npj Digit. Med.*, vol. 2, no. 1, p. 125, Dec. 2019, doi: 10.1038/s41746-019-0204-z.

[23] A. C. Scheffer, M. J. Schuurmans, N. Van dijk, T. Van der hooft, and S. E. De rooij, “Fear of falling: Measurement strategy, prevalence, risk factors and consequences among older persons,” *Age Ageing*, vol. 37, no. 1, pp. 19–24, Jan. 2008, doi: 10.1093/ageing/afm169.

[24] E. W. Steyerberg *et al.*, “Assessing the performance of prediction models: a framework for traditional and novel measures.,” *Epidemiology*, vol. 21, no. 1, pp. 128–38, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.

[25] T. Gneiting and A. E. Raftery, “Strictly Proper Scoring Rules, Prediction, and Estimation,” *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[26] T. Gneiting and M. Katzfuss, “Probabilistic Forecasting,” *Annu. Rev. Stat. Its Appl.*, 2014, doi: 10.1146/annurev-statistics-062713-085831.

[27] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement.,” *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55–63, Jan. 2015, doi: 10.7326/M14-0697.

[28] P. Palumbo, L. Palmerini, S. Bandinelli, and L. Chiari, “Fall Risk Assessment Tools for Elderly Living in the Community: Can We Do Better?,” *PLoS One*, vol. 10, no. 12, p. e0146247, Dec. 2015, doi: 10.1371/journal.pone.0146247.

[29] P. Palumbo *et al.*, “Predictive Performance of a Fall Risk Assessment Tool for Community-Dwelling Older People (FRAT-up) in 4 European Cohorts,” *J. Am. Med. Dir. Assoc.*, vol. 17, pp. 1106–1113, 2016, doi: 10.1016/J.JAMDA.2016.07.015.

[30] K. Rapp *et al.*, “Fall incidence in Germany: results of two population-based studies, and comparison of retrospective and prospective falls data collection methods.,” *BMC Geriatr.*, vol. 14, p. 105, Jan. 2014, doi: 10.1186/1471-2318-14-105.

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

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

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

Figure 1: Example of a figure

Annex A:
Glossary

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

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

Annex B:
Declaration of conflict of interests

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

Pierpaolo Palumbo holds copyrights on a tool for fall risk assessment for older adults.

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).

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

1. https://www.nice.org.uk/guidance/cg161/chapter/recommendations#multifactorial-assessment-or-multifactorial-falls-risk-assessment [↑](#footnote-ref-1)
2. http://inchiantistudy.net/wp/ [↑](#footnote-ref-2)
3. https://tilda.tcd.ie/ [↑](#footnote-ref-3)