INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2022-2024

FG-AI4H-P-012-A01

ITU-T Focus Group on AI for Health

Original: English

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

DOCUMENT

Source:	TG-Falls Topic Driver
Title:	Att.1 – TDD update (TG-Falls)
Purpose:	Discussion

Contact:	Pierpaolo Palumbo TG-Falls Topic Driver University of Bologna Italy	Tel: +39 3402378412 E-Mail: pierpaolo.palumbo@unibo.it
Contact:	Inês Sousa Associação Fraunhofer Portugal Research – Fraunhofer AICOS Portugal	Tel: +351 220 430 326 Email: 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 6 (to submit as FGAI4H-P-012-A01 to meeting P)

- Updates on the systematic review in section 2.2.13 "Status update for meeting P(Helsinki)"

Version 5 (submitted as FGAI4H-N-012-A01 to meeting N)

- Updates on the challenge on fall prediction within the Trial Audit Project 2.0
- Updates on the systematic review of datasets for training and testing AI systems for falls

Version 4 (submitted as FGAI4H-M-012-A01 to meeting M)

- Ideation of a literature review and expert consensus process
- Initiation of the participation in the ML4H Trial Audits 2.0 project

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 Palumbo University of Bologna Italy	Tel: +39 3402378412 Email: pierpaolo.palumbo@unibo.it
Inês Sousa Associação Fraunhofer Portugal Research – Fraunhofer AICOS Portugal	Tel: +351 220 430 326 Email: ines.sousa@fraunhofer.pt
Barry Greene Kinesis Health Technologies Ltd. Ireland	Email: barry.greene@kinesis.ie
Kimberley S. van Schooten University of New South Wales, Sydney, NSW Australia.	Email: k.vanschooten@neura.edu.au
Luca Palmerini University of Bologna Italy	Email: luca.palmerini@unibo.it

Jose Albites Sanabria
University of Bologna
Italy
Email: jose.albitessanabri2@unibo.it

CONTENTS

Page

Page

List of Tables

List of Figures

FG-AI4H Topic Description Document

Topic group - Falls among the elderly

1 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], 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).

2 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.

[1] https://www.nice.org.uk/guidance/cg161/chapter/recommendations#multifactorial-assessment-or-multifactorial-falls-risk-assessment

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.

2.1 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

2.2 Status of this topic group

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

2.2.1 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 A102. The document was presented remotely.

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

2.2.2 Status update for meeting C (New York)

The topic group description was refined and presented remotely.

2.2.3 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.

2.2.4 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.

2.2.5 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.

2.2.6 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].

2.2.7 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] and TILDA[3]. The following questions were raised in the internal discussion of the Topic Group:

First, 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)?

Second, 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?

[2] http://inchiantistudy.net/wp/
[3] https://tilda.tcd.ie/

Third, 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?

Finally, 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.

2.2.8 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

2.2.9 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).

2.2.10 Status update for meeting L (E-meeting)

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 have been made 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].

2.2.11 Status update for meeting M (E-meeting)

We have discussed about the intention to make a literature review and an expert consensus process to define different aspects of the benchmarking, including available datasets, eligibility requirements for datasets and AI algorithms, criteria for performance evaluations, and populations of interest. The literature review and expert consensus process will be conducted following a priority list on the different possible lines of research.

We have updated the TDD, describing the basic features that a first version of the benchmarking should have, and started a discussion on its implementation with Marc Lecoultre and Pradeep Balachandran from the AI4H Open Code Project.

2.2.12 Status update for meeting N (E-meeting)

We have started a literature review on datasets for AI systems for falls. We have agreed on the aim and are now refining the search queries and other methodological details. We have joined the AI Trial Audit Project 2.0 (https://aiaudit.org/, https://health.aiaudit.org/). Within this project, we have customized a questionnaire/checklist for qualitative assessment and are working on the code for the quantitative assessment.

2.2.13 Status update for meeting P (Helsinki)

We have continued our work on the systematic review, which is currently entitled “Fall risk assessment with wearable inertial sensors. A systematic review of datasets and individual-participant data meta-analysis”.

Main question of the systematic review is: which are the available datasets for training and validating models for wearable inertial sensor-based fall risk assessment? The main question of the individual participant data (IPD) meta-analysis is: which is the prognostic value for falls of features derived from wearable inertial sensors? The prognostic value of single sensor-based features will be determined with crude and adjusted odds ratios (ORs), rate ratios (RaRs), and hazard ratios (HRs). The prognostic value of a multivariate model will be assessed with calibration and discrimination measures.

We will retrieve datasets from journal articles and data portals. Journal articles will be identified from systematic reviews on sensor-based fall risk assessment. Sources for retrieving systematic reviews will be: Pubmed, Web of Science, and Scopus. Data portals will include Google Dataset Search, Mendeley Data, IEEE DataPort, Physionet, Figshare, Dataverse, and Dryad. A search for systematic reviews on sensor-based fall risk assessment has been made in March 2022 and will be re-run just before the final analyses. Systematic reviews shall be in English and published not

before 2017. Searches on data portals still have to be made.to be performed. Datasets not described in peer-reviewed articles will be excluded.

We have retrieved 19 unique systematic reviews on sensor-based fall risk assessment from Pubmed, Web of Science, and Scopus. From this set of systematic reviews, we have identified 956 articles that we are currently screening.

We have drafted the protocol to submit to PROSPERO (https://www.crd.york.ac.uk/prospero/). We have also prepared a draft email and a draft form to be sent the authors of the retrieved articles for requesting the datasets. The email includes an introduction to the FG AI4H, to the TG Falls, and to the systematic review. The additional request form further includes the rational and aims of the systematic and individual-participant data (IPD) meta-analysis, details on data protection, and the authorship policy.

2.3 Topic group participation

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

- 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

3 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.

3.1.1 Definition of the AI task

This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is *not* about the benchmarking process (this will be discussed in more detail in chapter 4). This section corresponds to DEL03 *"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 AI systems should generate an output variable with one entry for each record. This output variable should either:

- Be an ordered variable, with higher numbers expressing higher fall risk. In this case we call the AI predictions to be ordered non probabilistic and the values are not constrained in the range between 0 and 1
- Express the probability to fall at least once during a time window after the assessment. In this case we call the AI prediction to be probabilistic on a dichotomic outcome and the values of the output variable should lay in the range between 0 and 1
- Express the expected number of falls in a time window after the assessment. In this case we call the AI prediction to be probabilistic on a count outcome and the values of the output variable should be non-negative.

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.

3.1.2 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?

3.1.3 Relevance and impact of an AI solution

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

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.

Taking a modelling approach, 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. Furthermore, no study has ever evaluated the impact of a fall prediction tool using an experimental design. Two recent pragmatic, cluster-randomized controlled trials on fall injury prevention applied risk screening algorithms for selecting patients to target [18,19]. Although neither was able to prove the efficacy of the tested intervention, their specific experimental design does not allow to draw conclusions on the impact of the employed risk screening algorithms.

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 [7,20]. In the end, we hope that benchmarking will drive progress in this field.

- 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)?

3.1.4 Existing AI solutions

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

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.

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.

4 Ethical considerations

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

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?

5 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.

5.1 Publications on benchmarking systems

While a representative comparable benchmarking for fall prediction tools does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL07 *"AI for health evaluation considerations,"* DEL07_1 *"AI4H evaluation process description,"* DEL07_2 *"AI technical test specification,"* DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL07_4 *"Clinical Evaluation of AI for health"*.

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.

Other important decisions on fall data were taken in 2013 with the FARSEEING consensus. They include an endorsement of the ProFANE fall definition, methods and variables for reporting falls, clinical variables for describing subjects' characteristics, requirements on sensors, and information to describe signal characteristics [24].

Finally, a standardization protocol for data storage and organization (format, structure, modalities) is being finalized by the Mobilise-D consortium (www.mobilise-d.eu). The protocol will provide guidelines to store data (e.g. accelerations, angular velocities, etc.) from wearable sensors (e.g. inertial measurement units) and related gold standards (reference systems, e.g., stereophotogrammetric systems), both for laboratory evaluation and for real-world monitoring. At the end of the project, all data that will be collected during the Mobilise-D project will be available in such format, enabling their sharing and re-use. Such standardization protocol could also be used to format similar data thus ensuring the increase of the available amount of directly comparable data. The paper describing such standardization protocol is currently under submission. The first representative subjects recorded with such a protocol can be expected to be available by the end of 2021, together with the published article.

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 [25]. The AUC or the *c* statistics are generally used for evaluating the discriminative ability. Calibration can be evaluated: i) visually from calibration curves, ii) with the calibration intercept and slope, or iii) with the Brier Score, which also involves aspects related to the discriminative ability [26]. Calibration cannot be computed when the output of the prediction tool is not probabilistic [27].

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 showed much heterogeneity in its performance 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 and 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)?

5.2 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.

The Topic Group is planning to make a literature review and collect expert knowledge through a Delphi consensus process on AI tools for fall prediction, including information on target populations, sensors, data, algorithms, and benchmarking methods.

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

5.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable DEL07_5 *"FG-AI4H assessment platform"* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

The TG has started using the Trial Audit Project platform (https://health.aiaudit.org/), which is based on EvalAI [Yadav et al].

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

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the 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 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL05_1*"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL05_2 *"Data acquisition"*, DEL05_3 *"Data annotation specification"*, DEL05_4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL05_5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL05_6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL07*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL07_1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods andtiator for the evaluation process of AI for health), DEL07_2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL07_4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL07_5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL09 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL09_1 *"Mobile based AI applications,"* and DEL09_2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 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?

At the moment we are working to implement the first version of the benchmarking. The Topic Group intends to define all details of the benchmarking methodology throughout a process based on a literature review, a Delphi consensus process among domain experts, and interactions with the Working Groups of the FG-AI4H and members of the Open Code Initiative (OCI).

Domain experts will also be contacted for making an overview of datasets about falls that fulfil some technical requirements and those that could be available for the benchmarking.

The implementation of the benchmarking will follow a progressive and incremental approach: we will start implementing a simple version with a single dataset and basic functionalities, and later proceed towards richer versions, with multiple datasets from different populations, multiple endpoints (e.g., injurious falls in addition to falls), and advanced functionalities.

6.1.1 Benchmarking version 0

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

6.1.1.1 Overview

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

This version 0 of the benchmarking is a prototype. Its overall scope is to create a first working version of the benchmarking to enrich with further functionalities in later versions.

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

6.1.1.2 Benchmarking methods

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

6.1.1.2.1 Benchmarking system architecture

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

Figure 1 provides an overview of the benchmarking we envision. The platform receives input data files and AI systems to be tested. Each input datafile represents a study population on which the AI systems should be applied. The input data files are pre-processed by harmonization scripts, which create harmonized datafiles from the input data files. The harmonized data files have the same format and same semantics. The AI systems interact with the harmonized data files through interface scripts. The interface scripts apply the AI algorithms on the harmonized data files and produce a performance matrix representing the output of the platform.

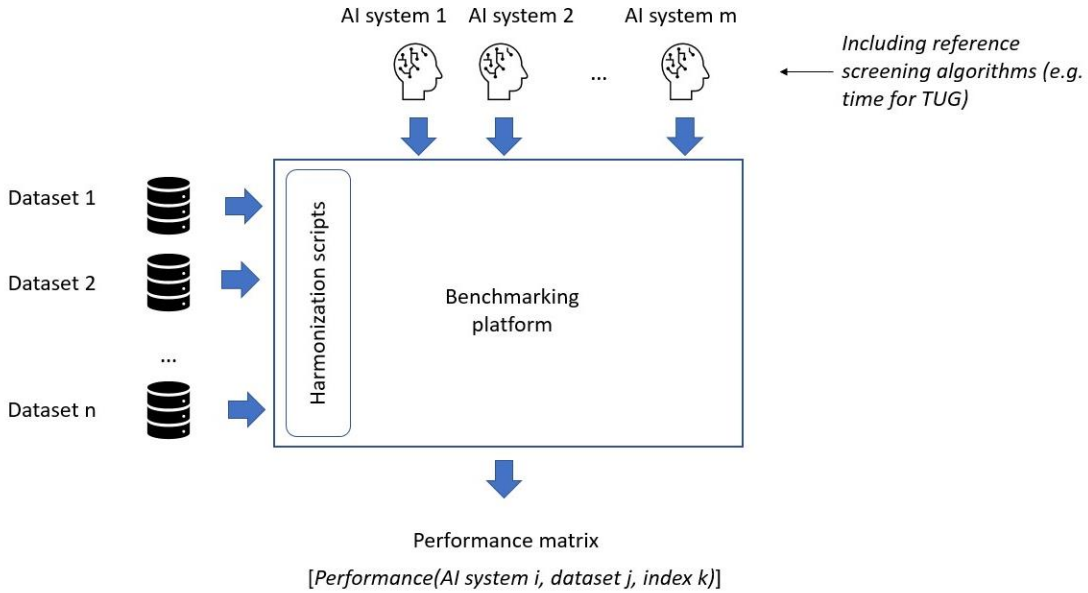

Figure 1: Overview of the benchmarking. Each AI system for fall prediction is evaluated upon multiple datasets and multiple performance indices.

The performance matrix could be thoughts as a three-dimensional matrix whose dimensions are the AI systems, the datasets, and the performance scores. For each AI system and each dataset, it contains: the indication on whether the AI system could be applied on the dataset and performance scores. Table “Performance matrix” provides an example.

Table

Table 2: “Performance matrix”. Example of a performance matrix for an AI system evaluated on three datasets.

AI system$_i$			
	Dataset 1 v2.0	Dataset 2 v1.4	Dataset 3 v2.1
Applicable	Yes	No	Yes
AUC	0.65	--	0.71
Brier score	0.021	--	0.018
Sensitivity – threshold 20%	62%	--	58%
Specificity – threshold 20%	74%	--	81%
...	...	--	...

For the benchmarking version 0 we plan to use only one input dataset, from the FallSensing study, provided by the group of Inês Sousa. Procedures to have formal access to the dataset are underway.

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

6.1.1.2.2 Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

The datasets employed within the benchmarking are searched, screened, included, and treated as follows:

- Dataset identification. The Topic Group searches for datasets that are possibly suitable to be included in the benchmarking. This search is conducted by screening the literature, sending emails to authors, advertising the Topic Group activities at meetings and conferences, receiving spontaneous reports from data owners, and any convenient formal or informal means. For the benchmarking version 0 we plan to use only one input dataset, already available to the Topic Group.
- Eligibility check. The Topic Group checks whether each of the identified datasets is eligible to be included in the benchmarking. The eligibility criteria will be fully specified after a literature review and an expert consensus process. They will concern data content (on AI input and label, including aspects related to validity, accuracy, and potential for harmonization), ethical waiver, and constraints set by data owners on data management. For the benchmarking version 0, we consider eligible the dataset that we have already identified.
- Dataset entry.
    - Eligible datasets are included in the benchmarking using harmonization scripts, which create a harmonized dataset from each input dataset.
    - Each input and harmonized dataset is assigned a version number.
    - Each dataset is described in a description document which is made available to all benchmarking participants. This description document shall contain information regarding the dataset population, the variables and the signals available in the dataset, the protocol used for data collection, the format in which these data are stored.
    - A data management document shall specify the data management rules for each dataset, including those for data maintenance, update, and deletion. Defining the lifecycle of the single input data files populating the benchmarking platform is critical, as each of them could be made available from the data owner under different conditions.
    - Each dataset could be openly accessible, undisclosed, or partly accessible and partly undisclosed.
- Data management. The datasets should be maintained according to their data management documents.

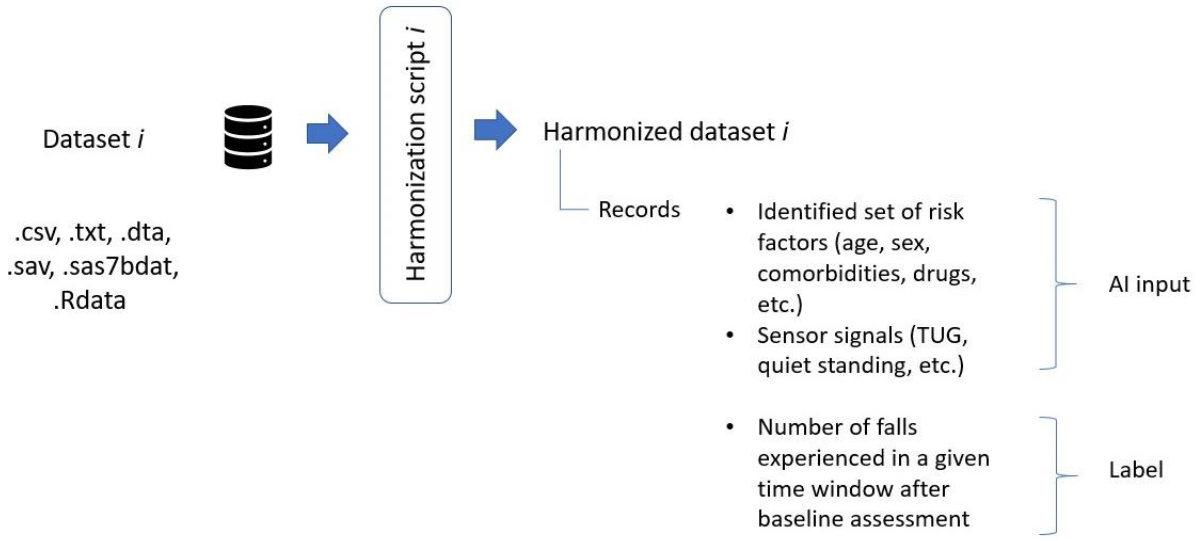

Figure 2: "Dataset harmonization" Schema representing the production of a harmonized dataset from an input dataset using a harmonization script.

Each harmonized dataset will be made of records, that here we define as the collection of AI input and label related to a single subject. Sometimes fall datasets come from longitudinal studies where the subjects are assessed on multiple waves. In this case, it may happen that information about falls may play the role of the label for prediction on one wave and the role of risk factor (thus AI input) for the subsequent wave. In order to keep inputs and labels separated and prevent AI systems from dishonest behaviour, each record will contain AI input taken from only one wave.

The AI systems will be submitted to the benchmarking in a way yet to define.

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

6.1.1.2.3 Safe and secure system operation and hosting

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.

[TBD]

- 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?

6.1.1.2.4 Benchmarking process

A description of the benchmarking process and its rules will be visible to everyone visiting the website of the benchmarking. The description of the process will include its scope, its timeframe, and a description of the benchmarking datasets. We may envision either periodic calls that open and close on specific deadlines, e.g., twice a year, or to leave always open the possibility to submit AI systems [TBD].

Individuals or groups willing to participate in the benchmarking with their AI system for fall prediction, shall register in the benchmarking. They shall declare if by any means they have ever accessed any of the benchmarking datasets or parts thereof. In this case, the submitted AI system will not be evaluated on these datasets. The participants shall also agree that the results obtained by benchmarking their AI systems will be published. The Topic Group will check the validity of the registration. Once registered, the participants will receive some sample records from the benchmarking datasets.

The registered participants shall proceed by submitting the algorithms of their AI systems. The algorithms could be coded as Python scripts or encrypted files [TBD]. Further specifications on files will be defined after interactions with the OCI members.

Each participant will be given the results of their AI system. The results of multiple AI systems could also be published in the benchmarking website or in scientific peer-reviewed journals. The publication strategy will be discussed together with the other members of the Focus Group.

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

6.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of 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 AI input data will be made of a table and additional text files. The table will contain risk factors for falls, while the text files will contain recordings from wearable inertial sensors during instrumented functional test (Figure 2, “Dataset harmonization”).

For this version 0 of the benchmarking, the input data structure will be as much as possible similar to the data structure of the input dataset. In other words, the harmonization script of the included dataset will be kept to a minimum. In future versions of the benchmarking, the variables of the harmonized 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.

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

- 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

6.1.1.4 AI output data structure

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

For this version 0 of the benchmarking, the AI systems should generate an output variable with one entry for each record. This output variable should either:

- Express the probability to fall at least once during the 12 months after the assessment. In this case we call the AI prediction to be probabilistic and the values of the output variable should lay in the range between 0 and 1
- Be an ordered variable, with higher numbers expressing higher fall risk. In this case we call the AI predictions to be ordered non probabilistic and the values are not constrained in the range between 0 and 1.

In future releases of the benchmarking, we may consider to accept also AI systems that produce other subject-specific output variables, e.g., expressing the expected number of falls in a future time window. The length of the time window for fall prediction could also be different, e.g., 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?

6.1.1.5 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.

In this version 0 of the benchmarking, the label of each record is represented with a number that could be either 1 or 0, depending on whether the subject represented by the record fell down or not in the 12 months after the assessment. In future releases of the benchmarking, the label could be an integer expressing the number of times the subject fell down. In future releases, we also may accept time windows different from 12 months.

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

6.1.1.6 Scores and metrics

*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.

In this version 0, the benchmarking should output a dichotomic variable expressing whether each submitted AI system is applicable on the test dataset. In addition, each applicable AI system shall be evaluated on the test dataset according to the following scores:

- The AUC
- The sensitivity and specificity at a cut-off maximising the Youden index
- The Brier score (only for probabilistic AI systems)

The Youden index is given by: Youden index = sensitivity + specificity –1.

Table “Evaluation grid” provides an example of the evaluation grid for three AI systems on a test dataset, where: AI system 1 is applicable and non-probabilistic, AI system 2 is not applicable, and the AI system 3 is applicable and probabilistic.

In future versions of the benchmarking, more scores will be computed.

Table 3: Evaluation grid. Example of a performance matrix for three AI systems evaluated on the test dataset.

Dataset 1 v1			
	AI system 1	AI system 2	AI system 3
Applicable	Yes	No	Yes
Probabilistic 0-1	No	Yes	Yes
AUC	0.65	--	0.71
Sensitivity	62%	--	58%
Specificity	74%	--	81%
Brier score	--	--	0.018

- Who are the stakeholders and what decisions should be supported by the scores and metrics of the benchmarking?
- What general criteria have been applied for selecting scores and metrics?
- What scores and metrics have been chosen/defined for robustness?
- What scores and metrics have been chosen/defined for medical performance?
- What scores and metrics have been chosen/defined for non-medical performance?
    - Metrics for technical performance tracking (e.g., monitoring and reporting when the performance accuracy of the model drops below a predefined threshold level as a function of time; computational efficiency rating, response times, memory consumption)
- What scores and metrics have been chosen/defined for model explainability?
- Describe for each aspect
    - The exact definition/formula of the score based on the labels and the AI output data structures defined in the previous sections and how they are aggregated/accumulated over the whole dataset (e.g., for a single test set entry, the result might be the probability of the expected correct class which is then aggregated to the average probability of the correct class)
    - Does it use some kind of approach for correcting dataset bias (e.g., the test dataset usually has a different distribution compared to the distribution of a condition in a real-world scenario. For estimating the real-world performance, metrics need to compensate this difference.)
    - What are the origins of these scores and metrics?
    - Why were they chosen?
    - What are the known advantages and disadvantages?
    - How easily can the results be compared between or among AI solutions?

- Can the results from benchmarking iterations be easily compared or does it depend too much on the dataset (e.g., how reproducible are the results)?

- How does this consider the general guidance of WG-DAISAM in DEL07_3 "Data and artificial intelligence assessment methods (DAISAM)"?
- Have there been any relevant changes compared to previous benchmarking iterations? If so, why?

6.1.1.7 Test dataset acquisition

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

In this version 0 of the benchmarking, we will include only the dataset FallSensing, which 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.

For future versions of the benchmarking, we will include multiple datasets. The datasets will be identified by different means, including a literature review. The eligibility criteria will be defined after a literature review and expert consensus process.

Table “Examples of datasets on ageing with information about falls” provides a list of datasets about ageing with information on falls. 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.

A systematic literature review to identify datasets to be used for training and testing AI systems for falls is underway.

Table 4: Examples of datasets on ageing with information about falls

• 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

• Name	• Country	• Type of risk factors	• Falls [to check…]	• Public/undisclosed
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		(??)

- How does the overall dataset acquisition and annotation process look?
- How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
- Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
- How was the dataset documented and which metadata were collected?
    - Where were the data acquired?
    - Were they collected in an ethical-conform way?
    - Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
    - Do the data contain 'sensitive information' (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
    - What kind of data anonymization or deidentification has been applied?
    - Are the data self-contained (i.e., independent from externally linked datasets)?
    - How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?

    - What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?

- Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
- Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
- How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
- Specific data governance derived by the general data governance document (currently F-103 and the deliverables beginning with DEL05)
- How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
- Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
- How was the annotation organized?
    - How many annotators/peer reviewers were engaged?
    - Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
    - How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
    - If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
    - How was the annotation of each case documented?
    - Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
- Were data/label update/amendment policies and/or criteria in place?
- How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
- How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
- Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
- What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)

6.1.1.8 Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data

sharing period, patient consent, and update procedure (see also DEL05_5 on *data handling* and DEL05_6 on *data sharing practices*).

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

6.1.1.9 Baseline acquisition

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

The AI systems will be compared with the time to perform the Timed Up and Go test, which was assessed in the test dataset.

- Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
- Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group, and ethnicity)?
- Was a baseline assessed as part of the benchmarking?

- How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous section.
- What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

6.1.1.10 Reporting methodology

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.

When registering for the benchmarking, the participants shall agree that the results obtained by benchmarking their AI systems will be published. Each participant will be given the results of their AI system. The results of multiple AI systems could also be published in the benchmarking website or in scientific peer-reviewed journals. The publication strategy will be discussed together with the other members of the Focus Group.

The results of the validation procedure should be reported as much as possible following the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist [31] and other applicable guidelines.

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

6.1.1.11 Result

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

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

6.1.1.12 Discussion of the benchmarking

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

- What was the general outcome of this benchmarking iteration?
- How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
- Are there real benchmarking results and interesting insights from this data?
    - How was the performance of the AI system compared to the baseline?
    - How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
    - How did the results change in comparison to the last benchmarking iteration?
- Are there any technical lessons?
    - Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
    - How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
- Are there any lessons concerning data acquisition?
    - Was it possible to collect enough data?
    - Were the data as representative as needed and expected?
    - How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
    - Was it possible to find annotators?
    - Was there any relevant feedback from the annotators?
    - How long did it take to create the dataset?
- Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
    - Are metrics missing?

    - Do the stakeholders need different reports or additional metadata (e.g., do they need the "offline capability" included in the AI metadata so that they can have a report on the best offline system for a certain task)?
- Are there insights on the benchmarking process?
    - How was the interest in participation?
    - Are there reasons that someone could not join the benchmarking?
    - What was the feedback of participants on the benchmarking processes?
    - How did the participants learn about the benchmarking?

6.1.1.13 Retirement

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 "*AI software lifecycle specification"* (identification of standards and best practices that are relevant for the AI for health software life cycle).

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

6.1.2 Benchmarking version [X]

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

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

7 Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 6.1.1.12).

- What is the overall outcome of the benchmarking thus far?
- Have there been important lessons?
- Are there any field implementation success stories?
- Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
- Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
- Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
    - Did the AI system perform as predicted relative to the baselines?

    - Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust, and quality of service)?

- Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?
- What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

8 Regulatory considerations

*Topic Driver: This section reflects the requirements of the working group on **Regulatory considerations on AI for health (WG-RC)** 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)* compiled the requirements that consider these challenges.

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

8.1 Existing applicable regulatory frameworks

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

8.2 Regulatory features to be reported by benchmarking participants

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

prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

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

8.3 Regulatory requirements for the benchmarking systems

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

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

8.4 Regulatory approach for the topic group

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 *"AI4H regulatory considerations."*

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

- How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

9 References

1. World Health Organization *WHO global report on falls prevention in older age*; World Health Organization: Geneva, Switzerland, 2007; ISBN 978 92 4 156353 6.
2. Burns, E.R.; Stevens, J.A.; Lee, R. The direct costs of fatal and non-fatal falls among older adults — United States. *J. Safety Res.* **2016**, *58*, 99–103, doi:10.1016/j.jsr.2016.05.001.
3. Hopewell, S.; Copsey, B.; Nicolson, P.; Adedire, B.; Boniface, G.; Lamb, S. 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.* **2019**, *54*, 1340–1350, doi:10.1136/bjsports-2019-100732.
4. Morse, J.M.; Morse, R.M.; Tylko, S.J. Development of a Scale to Identify the Fall-Prone Patient. *Can. J. Aging / La Rev. Can. du Vieil.* **1989**, *8*, 366–377, doi:10.1017/S0714980800008576.
5. Berg, K.; Wood-Dauphinée, S.; Williams, J.I.; Gayton, D. Measuring Balance in the Elderly Preliminary development of an Instrument. *Physiother. Canada* **1989**, *41*, 304–311.
6. Tinetti, M.E. Performance-oriented assessment of mobility problems in elderly patients. *J. Am. Geriatr. Soc.* **1986**, *34*, 119–26.
7. Gade, G.V.; Jørgensen, M.G.; Ryg, J.; Riis, J.; Thomsen, K.; Masud, T.; Andersen, S. Predicting falls in community-dwelling older adults: a systematic review of prognostic models. *BMJ Open* **2021**, *11*, e044170, doi:10.1136/bmjopen-2020-044170.
8. Silva, J.R.; Sousa, I.; Cardoso, J.S. Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls. *IEEE J. Biomed. Heal. Informatics* **2019**, 1–1, doi:10.1109/JBHI.2019.2951230.
9. Lamb, S.E.; Jørstad-Stein, E.C.; Hauer, K.; Becker, C. Development of a common outcome data set for fall injury prevention trials: The Prevention of Falls Network Europe consensus. *J. Am. Geriatr. Soc.* **2005**, *53*, 1618–1622, doi:10.1111/j.1532-5415.2005.53455.x.
10. van Schooten, K.S.; Brodie, M. Fall detection and risk assessment with new technologies. In *Falls in Older People*.
11. Gateway to Global Aging Data.
12. Casilari, E.; Santoyo-Ramón, J.A.; Cano-García, J.M. On the Heterogeneity of Existing Repositories of Movements Intended for the Evaluation of Fall Detection Systems. *J. Healthc. Eng.* **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*; World Health Organization: Geneva, Switzerland, 2008; ISBN 9789241563536.
14. Barry, E.; Galvin, R.; Keogh, C.; Horgan, F.; Fahey, T. 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.* **2014**, *14*, 14, doi:10.1186/1471-2318-14-14.
15. Schoene, D.; Wu, S.M.-S.; Mikolaizak, A.S.; Menant, J.C.; Smith, S.T.; Delbaere, K.; Lord, S.R. 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.* **2013**, *61*, 202–8, doi:10.1111/jgs.12106.

16. Panel on Prevention of Falls in Older Persons; American Geriatrics Society and British Geriatrics Society *Prevention of Falls in Older Persons: AGS/BGS Clinical Practice Guideline*; 2011;

17. Palumbo, P.; Becker, C.; Bandinelli, S.; Chiari, L. Simulating the effects of a clinical guidelines screening algorithm for fall risk in community dwelling older adults. *Aging Clin. Exp. Res.* **2018**, doi:10.1007/s40520-018-1051-5.

18. Bhasin, S.; Gill, T.M.; Reuben, D.B.; Latham, N.K.; Ganz, D.A.; Greene, E.J.; Dziura, J.; Basaria, S.; Gurwitz, J.H.; Dykes, P.C.; et al. A Randomized Trial of a Multifactorial Strategy to Prevent Serious Fall Injuries. *https://doi.org/10.1056/NEJMoa2002183* **2020**, *383*, 129–140, doi:10.1056/NEJMOA2002183.

19. Lamb, S.E.; Bruce, J.; Hossain, A.; Ji, C.; Longo, R.; Lall, R.; Bojke, C.; Hulme, C.; Withers, E.; Finnegan, S.; et al. Screening and Intervention to Prevent Falls and Fractures in Older People. *https://doi.org/10.1056/NEJMoa2001500* **2020**, *383*, 1848–1859, doi:10.1056/NEJMOA2001500.

20. Shany, T.; Wang, K.; Liu, Y.; Lovell, N.H.; Redmond, S.J. Review: Are we stumbling in our quest to find the best predictor? Over-optimism in sensor-based models for predicting falls in older adults. *Healthc. Technol. Lett.* **2015**, *2*, 79–88, doi:10.1049/htl.2015.0019.

21. Howcroft, J.; Kofman, J.; Lemaire, E.D. Review of fall risk assessment in geriatric populations using inertial sensors. *J. Neuroeng. Rehabil.* **2013**, *10*, 91, doi:10.1186/1743-0003-10-91.

22. Greene, B.R.; McManus, K.; Redmond, S.J.; Caulfield, B.; Quinn, C.C. Digital assessment of falls risk, frailty, and mobility impairment using wearable sensors. *npj Digit. Med.* **2019**, *2*, 125, doi:10.1038/s41746-019-0204-z.

23. Scheffer, A.C.; Schuurmans, M.J.; Van dijk, N.; Van der hooft, T.; De rooij, S.E. Fear of falling: Measurement strategy, prevalence, risk factors and consequences among older persons. *Age Ageing* **2008**, *37*, 19–24, doi:10.1093/ageing/afm169.

24. Klenk, J.; Chiari, L.; Helbostad, J.L.; Zijlstra, W.; Aminian, K.; Todd, C.; Bandinelli, S.; Kerse, N.; Schwickert, L.; Mellone, S.; et al. Development of a standard fall data format for signals from body-worn sensors. *Z. Gerontol. Geriatr.* **2013**, *46*, 720–726, doi:10.1007/s00391-013-0554-0.

25. Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **2010**, *21*, 128–38, doi:10.1097/EDE.0b013e3181c30fb2.

26. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378, doi:10.1198/016214506000001437.

27. Gneiting, T.; Katzfuss, M. Probabilistic Forecasting. *Annu. Rev. Stat. Its Appl.* **2014**, doi:10.1146/annurev-statistics-062713-085831.

28. Palumbo, P.; Palmerini, L.; Bandinelli, S.; Chiari, L. Fall Risk Assessment Tools for Elderly Living in the Community: Can We Do Better? *PLoS One* **2015**, *10*, e0146247, doi:10.1371/journal.pone.0146247.

29. Palumbo, P.; Klenk, J.; Cattelani, L.; Bandinelli, S.; Ferrucci, L.; Rapp, K.; Chiari, L.; Rothenbacher, D.; Berger, J.S.; Jordan, C.O.; 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.* **2016**, *17*, 1106–1113, doi:10.1016/J.JAMDA.2016.07.015.

30. Rapp, K.; Freiberger, E.; Todd, C.; Klenk, J.; Becker, C.; Denkinger, M.; Scheidt-Nave, C.;

Fuchs, J. Fall incidence in Germany: results of two population-based studies, and comparison of retrospective and prospective falls data collection methods. *BMC Geriatr.* **2014**, *14*, 105, doi:10.1186/1471-2318-14-105.

31. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **2015**, *162*, 55–63, doi:10.7326/M14-0697.

D. Yadav et al., "EvalAI: Towards Better Evaluation of AI Agents," Accessed: Jan. 05, 2022. [Online]. Available: https://www.aicrowd.com/.

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 on falls amongst the elderly.
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	
TBD	To be defined	
ROC	Receiver Operating Characteristic curve	
AUC	Area under the ROC curve	
CI	Confidence interval	
NNT	Number needed to treat	
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis	Checklist for reporting the results from the development and/or the validation of prediction models for health
[…]		

Annex B:
Declaration of conflict of interests

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

Company/Institution/Individual XYZ

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

_____