

# ITU-T Focus Group Deliverable

(09/2023)

Focus Group on Artificial Intelligence for Health  
(FG-AI4H)

---

## FG-AI4H DEL10.4

**Topic Description Document for the Topic  
Group on falls among the elderly (TG-Falls)**



### DEL10.4 – Topic Description Document for the Topic Group on falls among the elderly (TG-Falls)

#### Summary

This topic description document (TDD) specifies a standardized benchmarking for AI-based systems for fall prevention and management for older people. It reports background, definitions, methods, and systems related to falls from the most consolidated scientific literature. It offers an overview of the state of the art of validation and benchmarking of existing systems for fall prediction. It proposes a methodology for benchmarking AI systems for falls based on systematic reviews of available datasets and individual participant data meta-analyses (IPD-MA) of the AI systems. It provides the protocol and preliminary results of such a systematic review and IPD-MA for the specific subtopic of fall prediction with wearable inertial sensors.

#### Keywords

Artificial intelligence, benchmarking, clinical relevance, data audit, data quality, elderly, ethics, falls, health, overview, regulations, topic description, topic groups.

#### Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

#### Change Log

This document contains Version 1 of the Deliverable DEL10.4 on *FG-AI4H Topic Description Document for the Topic Group on falls among the elderly (TG-Falls)* approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

#### Editor:

Pierpaolo Palumbo  
TG-Falls Topic Driver  
University of Bologna  
Italy

Tel: +39 3402378412  
E-mail: [pierpaolo.palumbo@unibo.it](mailto:pierpaolo.palumbo@unibo.it)

Inês Sousa  
Associação Fraunhofer Portugal Research  
– Fraunhofer AICOS  
Portugal

Tel: +351 220 430 326  
E-mail: [ines.sousa@fraunhofer.pt](mailto:ines.sousa@fraunhofer.pt)

#### Contributors:

Jose Albites Sanabria  
University of Bologna  
Italy

E-mail: [jose.albitessanabri2@unibo.it](mailto:jose.albitessanabri2@unibo.it)

Barry Greene  
Kinesis Health Technologies Ltd.  
Ireland

E-mail: [barry.greene@kinesis.ie](mailto:barry.greene@kinesis.ie)

Killian McManus  
Kinesis Health Technologies Ltd.  
Ireland

E-mail: [killian.mcmanus@kinesis.ie](mailto:killian.mcmanus@kinesis.ie)

Luca Palmerini  
University of Bologna  
Italy

E-mail: [luca.palmerini@unibo.it](mailto:luca.palmerini@unibo.it)

Pierpaolo Palumbo  
University of Bologna  
Italy

Tel: +39 3402378412

E-mail: [pierpaolo.palumbo@unibo.it](mailto:pierpaolo.palumbo@unibo.it)

Inês Sousa  
Associação Fraunhofer Portugal Research  
– Fraunhofer AICOS  
Portugal

Tel: +351 220 430 326

E-mail: [ines.sousa@fraunhofer.pt](mailto:ines.sousa@fraunhofer.pt)

Kimberley S. van Schooten  
University of New South Wales, Sydney,  
NSW  
Australia.

E-mail: [k.vanschooten@neura.edu.au](mailto:k.vanschooten@neura.edu.au)

Eugenio Zuccarelli  
CVS Health  
USA

E-mail: [eugenio.zuccarelli@gmail.com](mailto:eugenio.zuccarelli@gmail.com),  
[ez@alum.mit.edu](mailto:ez@alum.mit.edu)

© ITU 2025

Some rights reserved. This publication is available under the Creative Commons Attribution-Non Commercial-Share Alike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; <https://creativecommons.org/licenses/by-nc-sa/3.0/igo>). For any uses of this publication that are not included in this licence, please seek permission from ITU by contacting [TSBmail@itu.int](mailto:TSBmail@itu.int).

If you wish to reuse material from this publication that is attributed to a third party, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder.

## Table of Contents

	Page
1 Introduction.....	1
2 About the FG-AI4H topic group on Falls among the elderly .....	1
2.1 Documentation .....	1
2.2 Status of this topic group.....	3
2.3 Topic Group participation .....	6
3 Topic description .....	7
3.1 Definition of the AI task.....	7
3.2 Current gold standard .....	7
4 Ethical considerations .....	9
5 Existing work on benchmarking.....	10
5.1 Publications on benchmarking systems.....	10
5.2 Benchmarking by AI developers .....	11
5.3 Relevant existing benchmarking frameworks .....	12
6 Benchmarking by the topic group.....	12
6.1 Subtopic Fall prediction .....	12
7 Overall discussion of the benchmarking.....	20
8 Regulatory considerations .....	20
8.1 Existing applicable regulatory frameworks.....	20
8.2 Regulatory features to be reported by benchmarking participants.....	20
8.3 Regulatory requirements for the benchmarking systems .....	21
8.4 Regulatory approach for the topic group.....	21
References.....	22
Annex A – Glossary .....	26
Annex B – Declaration of conflict of interests .....	28



### DEL10.4 – Topic Description Document for the Topic Group on falls among the elderly (TG-Falls)

#### 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 (World Health Organization 2007), and these events represent more than 50 per cent 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 per cent of cases falls result in fractures, thus contributing to a significant increase 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 [7].

Current guidelines for fall prevention and management in older adults, indicate a need to implement fall risk screening to identify those at increased risk [23], [26]. These high-risk individuals should be offered fall prevention programs, which reduce fall risk by improving strength and balance and modifying behaviours [17]. In case of a fall, a prolonged time on the floor leads to worse physical and clinical outcomes. Thus, it is essential to promptly provide assistance to those who are not able to get up independently [11].

According to these recommendations, diverse tools have been proposed for fall prevention and management, including systems for fall risk prediction [12], fall preventive intervention recommendations [9], [22], and fall detection and classification [45]. In the last decade, some of these tools have been developed using Artificial Intelligence (AI) techniques. Although AI has the potential to increase the accuracy and efficacy of these tools, most of them have not been sufficiently validated. A platform for standardized benchmarking of these systems would allow to consistently evaluate their predictive accuracy and their efficacy in preventing and managing falls.

This topic description document specifies such a standardized benchmarking. It serves as deliverable No. 10.4 of the ITU/WHO Focus Group on AI for Health (FG-AI4H). It focuses on the subtopic of fall prediction with wearable inertial sensors, as specified in the following.

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

To develop this benchmarking framework, FG-AI4H decided to create the topic group TG-Falls (Falls among the elderly) at 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 topic description document (TDD). During the FG-AI4H meeting A in Geneva, Switzerland, on 25-27 September 2018, Inês Sousa from Associação Fraunhofer Portugal Research – Fraunhofer AICOS was nominated as the topic driver for TG-Falls. In December 2020, this role was taken by Pierpaolo Palumbo, from the University of Bologna, Bologna, Italy. In June 2023, Pierpaolo Palumbo expressed his intention to leave this role, while continuing to serve as a member of TG-Falls. He indicated that Kimberley van Schooten would be available to take over. Dr Kimberley van Schooten is from the Falls, Balance and Injury Research Centre, Neuroscience Research Australia, Sydney, Australia and from the School of Population Health, University of New South Wales, Sydney, Australia.

##### 2.1 Documentation

This document is the TDD for TG-Falls. It introduces the health topic including the AI task, outlines its relevance and 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 one subtopic at a level of detail that includes technological and operational implementation. There are individual subclauses for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative. In addition, the TDD addresses ethical and regulatory aspects.

This TDD has been developed cooperatively by all members of the TG-Falls over time and updated TDD iterations have been presented at each FG-AI4H meeting.

This final version of this TDD is released as deliverable "DEL 10.4 Falls among the elderly (TG-Falls)." The TG-Falls has submitted output documents reflecting updates to the work on this deliverable (Table 1) to each FG-AI4H meeting.

**Table 1 – Topic group output documents**

<b>Number</b>	<b>Title</b>
FGAI4H-R-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting R
FGAI4H-R-012-A01	Latest update of the Topic Description Document of the TF-Falls for meeting R
FGAI4H-Q-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting Q
FGAI4H-Q-012-A01	Latest update of the Topic Description Document of the TF-Falls for meeting Q
FGAI4H-P-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting P
FGAI4H-P-012-A01	Latest update of the Topic Description Document of the TF-Falls for meeting P
FGAI4H-O-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting O
FGAI4H-N-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting N
FGAI4H-N-012-A01	Latest update of the Topic Description Document of the TF-Falls for meeting N
FGAI4H-M-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting M
FGAI4H-M-012-A01	Latest update of the Topic Description Document of the TF-Falls for meeting M
FGAI4H-L-012-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting L
FGAI4H-L-012-A01	Latest update of the Topic Description Document of the TF-Falls for meeting L
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-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting H
FGAI4H-H-012-A02	Latest update of the Call for Topic Group Participation (CfTGP) for meeting H
FGAI4H-H-012-A01	Latest update of the Topic Description Document of the TG-Falls for meeting J
FGAI4H-G-010	Latest update of the Topic Description Document of the TG-Falls for meeting G
FGAI4H-F-010-A01	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting F



**Table 1 – Topic group output documents**

Number	Title
FGAI4H-F-010	Latest update of the Topic Description Document of the TG-Falls for meeting F
FGAI4H-E-012-A01	The presentation summarizing the latest update of the Topic Description Document of the TG-Falls for meeting E
FGAI4H-E-012	Latest update of the Topic Description Document of the TG-Falls for meeting E
FGAI4H-E-005-A05	Call for Topic Group participation: Standardizing benchmarking of AI to prevent falls among the elderly
FGAI4H-D-037	Meeting notes: Standardizing benchmarking of AI to prevent falls among the elderly
FGAI4H-C-014	Status report of: Reducing risk of falling among elderly
FGAI4H-B-026-R1	Proposal: Multifactorial screening of fall risk in community-dwelling adults

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 subclauses describe the update of the collaboration within TG-Falls for the official focus group meetings.

### **2.2.1 Status update for meeting B (Lausanne)**

The first submission was provided in response to the ITU-WHO 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 TG-Falls following the manifestation of interest sent by Pierpaolo Palumbo. The Personal Health Systems Laboratory is headed by Prof. Lorenzo Chiari. Pierpaolo Palumbo is a biomedical engineer and post-doctoral fellow, working on algorithms for health risk assessment, with a focus on fall risk in community-dwelling older adults and lower-limb amputees.

### **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 TG-Falls for benchmarking AI-based systems for fall prediction.

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

Demonstrations of interest to join the focus group were 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], [37].

### **2.2.7 Status update for meeting H (Brasilia)**

TG-Falls received demonstrations of potential interest from two groups that carry out epidemiological studies about ageing, with data on falls: InCHIANTI<sup>1</sup> and TILDA<sup>2</sup>. TG-Falls discussed the dataset eligibility criteria and methods for harmonizing heterogeneous datasets. Main concerns were about datasets which are not undisclosed and datasets which are not representative of the general population of older adults.

The TG-Falls held a conference call. The participants were:

- 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

The TG-Falls discussed on the following points:

- 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 heterogeneous and consider different variables and functional tests. These datasets may or may not include data from sensors.

The TG-Falls agreed upon the following action points:

- To systematize information regarding fall risk assessment;
- To continue the discussion of the variables to be considered, and methods/best practices for algorithm evaluation;
- To discuss with the working group how the benchmarking framework should deal with heterogeneous datasets.

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

The TG-Falls participants have met and discussed guidelines for standardization and evaluation of AI models to estimate the risk of falling. Conference call participants:

- Inês Sousa, Fraunhofer AICOS
- Pierpaolo Palumbo, University of Bologna
- Stefania Bandinelli, SOC Geriatria -USLToscana Centro, Firenze

---

<sup>1</sup> <http://inchiantistudy.net>

<sup>2</sup> <https://tilda.tcd.ie/>

- 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)**

The TG-Falls started preparing the workshop "Artificial Intelligence and fall prediction" within the EU Falls Festival 2021<sup>3</sup>. This EU Falls Festival 2021 was later postponed to 2022.

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

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

The TDD was re-formatted according to the new template (FG-AI4H-J-105).

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

The TDD has been updated with additions regarding the state-of-the-art and relevant information from the Prevention of Falls Network Europe (ProFaNE) consensus on definitions and measures for fall injury prevention trials. Kimberley van Schooten contributed a book chapter to the discussion of the TG-Falls [44]. Further updates were made on ethical considerations.

The TG-Falls asked permission to access the harmonized version of the datasets belonging to the Health and Retirement Study (HRS) family [38].

The TG-Falls started an internal discussion on whether it is appropriate to open a sub-topic on fall detection and whether some research groups could provide their data on falls recorded with wearable inertial sensors [8].

### **2.2.11 Status update for meeting M (E-meeting)**

TG-Falls discussed the intention to make a literature review and an expert consensus process to define different aspects of 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.

TDD was updated with descriptions of the basic features that a first version of the benchmarking should have. TG-Falls 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)**

TG-Falls started a literature review on datasets for AI systems for falls, agreed on the aim and started refining the search queries and other methodological details. TG-Falls joined the AI Trial Audit Project 2.0<sup>4</sup>. Within this project, the TG-Falls customized a questionnaire/checklist for qualitative assessment and started working on the code for quantitative assessment.

### **2.2.13 Status update for meeting P (Helsinki)**

In order to benchmark AI systems for falls, the TG-Falls decided to combine the systematic review with an individual-participant data meta-analysis (IPD-MA). TG-Falls drafted the protocol to submit to PROSPERO<sup>5</sup>. The systematic review and IPD-MA will be entitled "Fall risk assessment with wearable inertial sensors. A systematic review of datasets and individual-participant data meta-

---

<sup>3</sup> <https://geriatrie.be/event/eu-falls-festival-2021/>

<sup>4</sup> <https://aiaudit.org/>

<sup>5</sup> <https://www.crd.york.ac.uk/prospero/>

analysis". The main aim of the systematic review is to identify datasets available for the training and validating of models for wearable inertial sensor-based fall risk assessment. The main aim of the IPD-MA is to evaluate the prognostic value for falls of single features derived from wearable inertial sensors and validate sensor-based models for fall prediction. The TG-Falls also prepared a draft email and a draft form to be sent to the authors of the retrieved articles requesting datasets. The email includes an introduction to the FG-AI4H, to the TG-Falls, and to the systematic review, as well as an annex describes the rationale and the aims of the systematic review and the IPD-MA, and provides details on data protection and authorship policies.

#### **2.2.14 Status update for meeting R (Douala)**

TG-Falls registered the protocol for the systematic review in PROSPERO (2022 CRD420223673946). It also completed the title-and-abstract screening stage of the systematic review.

The TG-Falls drafted a commentary on the fall risk stratification algorithm recommended in the world guidelines for fall prevention and management [23], and also collected feedback from some FG-AI4H representatives (Eva Weicken and Markus Wenzel).

#### **2.2.15 Status update for meeting R (MIT Media Lab & Harvard Kennedy School, Cambridge, US)**

TG-Falls submitted the commentary on the world guidelines [23] to age and ageing. The commentary discussed the model design, validation, usability, potential effect, and future research.

The TG-Falls completed the full-text screening stage of the systematic review and defined a template for data extraction and quality assessment.

#### **2.2.16 Status update for meeting S (ITU HQ, Geneva)**

The commentary on the world guidelines in age and ageing has been accepted for publication.

TG-Falls started the data extraction phase of the systematic review.

Eugenio Zuccarelli from CVS Health, USA, joined the TG-Falls.

Pierpaolo Palumbo announced his intention to step aside as topic driver, while continuing to serve as TG-Falls member. Kimberly van Schooten expressed her availability to drive the TG-Falls.

### **2.3 Topic Group participation**

The participation in both, the FG-AI4H 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](mailto: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>

---

<sup>6</sup> [https://www.crd.york.ac.uk/prospero/display\\_record.php?RecordID=367394](https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=367394)

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

- [fgai4htgfalls@lists.itu.int](mailto: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**

Diverse AI-based systems have been proposed for fall prevention and management, including systems for fall risk prediction [12], fall preventive intervention recommendations [9], [22], and fall detection and classification [45]. The TG-Falls has decided to address the sub-topic of fall prediction in older adults. This clause contains a detailed description and background information of this specific health topic and how this can help to solve a relevant real-world problem.

The concrete benchmarking initiative implemented through the systematic review and IPD-MA is further focused on fall prediction with wearable inertial sensors.

#### **3.1 Definition of the AI task**

AI-based systems for fall prediction aim to provide an individual-specific risk score of falling in the future within a given time window (prediction window) [21], given information about the subject's risk factors for falls and/or their balance or motor ability. The individuals 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" (World Health Organization 2007). 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" [20]. Both definitions could be accepted for the purposes of this topic group.

The AI systems for fall prediction should run on records of a single individual or on a dataset containing multiple records of different individuals. The AI systems 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 levels expressing higher fall risk. In this case the AI predictions is called 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 the prediction window. In this case the AI prediction is called probabilistic on a dichotomous outcome. The values of the output variable should lay in the range between 0 and 1.
- Express the expected number of falls in the prediction window. In this case the AI prediction is called probabilistic on a count outcome. The values of the output variable should be non-negative.

The length of the prediction window is generally set at 12 months, but different time windows could be accepted, ranging from 6 to 24 months. Additionally, the AI systems could further provide suggestions on possible preventive actions to take.

#### **3.2 Current gold standard**

At present, a variety of tools for fall risk prediction have been proposed [3], [12]. The World Guidelines for falls prevention and management for older adults recommends following an algorithm for fall risk stratification which is a decision tree with three output risk categories: low risk,

intermediate risk, and high risk. Each risk category drives a distinct fall prevention intervention or treatment [23]. This algorithm is a combination of evidence-based and expert-based recommendations. It still needs to be validated with respect to its accuracy, impact, and clinical effectiveness (topic group on "Falls among the elderly" 2023).

The fall risk stratification algorithm proposed by the Word Guidelines is similar to the algorithm proposed by the American Geriatrics Society (AGS) and British Geriatrics Society (BGS) [31]. 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) [28].

Another tool for fall prediction is the Timed Up and Go Test (TUG) [32]. It is one of the most widespread functional tests in clinical practice. The output score is the time for completing a motor task consisting in getting up from a chair, walking 3 meters, going back to the chair, and sitting down. Its prognostic ability for falls has been evaluated several times over the years in different studies and populations. Two systematic reviews report much heterogeneity in its performance across studies and a relatively low average predictive accuracy. 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) [2], [35].

The predictive accuracy of classical or AI algorithms for fall prediction should be evaluated against prospective falls [20]. In other words, validation datasets should contain information on falls occurred during the prediction window i.e., after the observation window and the prediction time [21]. Moreover, falls should preferably be ascertained using prospective daily recording and a notification system with a minimum of monthly reporting [20]. These requirements have become *de facto* standards in the literature.

### **3.3 Relevance and impact of an AI solution**

Falls are a prevalent health issue among the elderly. Approximately one-third of community-dwelling adults aged 65 or older experience a fall annually (World Health Organization 2007), accounting for over 50% of hospitalizations due to injuries in this age group. Falls are also a leading cause of loss of independence and institutionalization. In 10% of cases, falls result in fractures, significantly increasing morbidity and mortality rates. The direct healthcare costs associated with falls are substantial, amounting to 25 billion euros per year in the European Union and 31 billion dollars annually in the United States [7].

While some preventive interventions have proven effective, their implementation across the entire population is often unfeasible or clinically inappropriate. Therefore, AI-based systems for fall prevention aim to identify individuals who should be prioritized for fall prevention interventions and determine the most suitable types of interventions for them.

A modelling approach estimated that using the AGS/BGS algorithm for fall risk assessment within a preventive intervention could reduce the number needed to treat (NNT) by about 17% (95% CI 4.1%-34.0%) compared to another preventive strategy not based on a risk assessment tool [28]. However, impact assessment studies for other tools with better predictive accuracy are lacking, and no study has evaluated the impact of a fall prediction tool using an experimental design. Two pragmatic, cluster-randomized controlled trials on fall injury prevention applied risk screening algorithms to select patients to target [5], [19]. Although neither trial proved the efficacy of the tested intervention, their specific experimental design does not allow for conclusions about the impact of the employed risk screening algorithms.

Creating a standardized platform for benchmarking fall prediction systems would enable rigorous and comparable assessments of these tools, highlighting their strengths and limitations and addressing concerns about over-fitting and over-optimism raised by some authors [12], [36]. Ultimately, we believe that benchmarking will drive progress in this field.



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 [28]. 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 pragmatic, cluster-randomized controlled trials on fall injury prevention applied risk screening algorithms for selecting the patients to target [5], [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 systems, 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 [12], [36]. In the end, it is believed that benchmarking will drive progress in this field.

### 3.4 Existing AI solutions

Existing AI solutions for fall prediction are multivariate prognostic models developed with supervised learning algorithms (e.g., logistic regression or other machine learning algorithms). A recent review found 72 prognostic models developed or validated in prospective cohorts [12]. Of those, only three were validated and had an AUC between 0.62 and 0.69.

Howcroft *et al.* [16] reviewed previous studies focusing on fall risk assessment with wearable 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; however, this criterion was only used in 15 per cent 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 [16].

Greene *et al.* [15] carried out a study involving 8521 participants ( $72.7 \pm 12.0$  years, 5392 female) from six countries, assessed using a digital falls risk assessment protocol. Data consisted of wearable sensor data captured during the TUG test along with data on falls risk factors from self-reported questionnaires, applied to previously trained and validated classifier models. They 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.

## 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 [DEL1](#) "*Ethics and governance of artificial intelligence for health*" which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This clause refers to DEL1 and should reflect the ethical considerations of TG-Falls.

Specific ethical considerations should include the fact that fear of falling is in 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 [34]. 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. Seeking ethical approval for reusing these data within the ITU/WHO benchmarking platform was not deemed necessary.

It is foreseen 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.

## **5 Existing work on benchmarking**

This clause focuses on the existing benchmarking processes in the context of AI and fall prediction 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 clause summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable [DEL7](#) *"AI for health evaluation considerations,"* [DEL7.1](#) *"AI4H evaluation process description,"* [DEL7.2](#) *"AI technical test specification,"* [DEL7.3](#) *"ML4H trial audits–Iteration 2.0 Playbook,"* and [DEL7.4](#) *"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 ProFaNE [20]. 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 [18].



Recently, the Mobilise-D consortium<sup>7</sup> has proposed a standardization protocol for storing and organizing data from wearable inertial sensors and related gold standards (reference systems e.g., stereophotogrammetric systems), for laboratory evaluation and for real-world monitoring. Inertial sensor data includes accelerations and angular velocities. Data organization encompasses format, structure and modalities ([27]. All data that are being 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.

Other methodological indications available in the literature for benchmarking predictive models regard metrics for evaluating the algorithmic performance. 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 [40]. 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 [14]. Calibration cannot be computed when the output of the prediction tool is not probabilistic [13].

Within the literature, no platform has been established to systematically evaluate multiple fall predictive systems 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.

Among the traditional tools for risk screening, TUG is one of the most widespread in clinical practice. 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. 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) [2], [35].

From this experience, it is believed 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.

## **5.2 Benchmarking by AI developers**

Some of the early studies on AI systems for fall prediction were affected by methodological flaws. For example, some studies were missing any form of validation [36]. Nowadays, all developers of AI solutions for fall prediction implement internal benchmarking systems for assessing the performance. However, external validation or side-by-side comparison with previous models are hardly ever performed.

Among multivariate tools for predicting falls in the elderly, FRAT-up was externally validated on four European datasets of longitudinal studies about ageing. It showed to be more accurate than simple traditional tools and exhibits much heterogeneity in its performance across different populations [29], [30]. 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 [33].

Among wearable sensor-based AI system for fall prediction, to the best of our knowledge none has been externally validated. Furthermore, none of these systems has ever been evaluated for their usability or clinical effectiveness, nor has any impact analysis been carried out.

---

<sup>7</sup> [www.mobilise-d.eu](http://www.mobilise-d.eu)

### 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. The TG has started trying the Trial Audit Project platform (<https://health.aiaudit.org/>; [DEL7.5](#) "FG-AI4H assessment platform"), developed by FG-AI4H on the bases of EvalAI [46]. TG-Falls is currently evaluating which software platform would be more suitable for the benchmarking of IPD-MA.

## 6 Benchmarking by the topic group

This clause describes all technical and operational details regarding the benchmarking process for the AI-based fall prediction task. It includes one subclause for the first version of the benchmarking, that will be iteratively improved over time.

It reflects the considerations of various deliverables: [DEL5](#) "Data specification" (introduction to deliverables 5.1-5.6), [DEL5.1](#) "Data requirements" (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL5.2](#) "Data acquisition", [DEL5.3](#) "Data annotation specification", [DEL5.4](#) "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), [DEL5.5](#) "Data handling" (which outlines how data will be handled once they are accepted), [DEL5.6](#) "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), [DEL6](#) "AI training best practices specification" (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL7](#) "AI for health evaluation considerations" (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), [DEL7.1](#) "AI4H evaluation process description" (which provides an overview of the state of the art of AI evaluation principles and methods for the evaluation process of AI for health), [DEL7.2](#) "AI technical test specification" (which specifies how an AI can and should be tested *in silico*), [DEL7.3](#) "ML4H trial audits–Iteration 2.0 Playbook", [DEL7.4](#) "Clinical Evaluation of AI for health" (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), [DEL7.5](#) "FG-AI4H assessment platform" (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL9](#) "AI for health applications and platforms" (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL9.1](#) "Mobile applications," and [DEL9.2](#) "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

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 clause outlines the first benchmarking version implemented thus far and the rationale behind it. It serves as an introduction to the subsequent clauses, where the actual benchmarking methodology is described.

The TG-Falls is implementing the first version of the benchmarking through a systematic review and IPD-MA of wearable sensor-based tools for fall prediction. The full-text screening of eligible articles has been completed. Currently, the TG-Falls is extracting data from the included articles and is evaluating their quality.

The implementation of the benchmarking is following a progressive and incremental approach: at the onset a simple version with a single dataset and basic functionalities will be implemented, 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 1

This clause includes all technological and operational details of the benchmarking process for the benchmarking version 1.

### 6.1.1.1 Overview

This clause provides an overview of the key aspects of this benchmarking iteration, version 1.

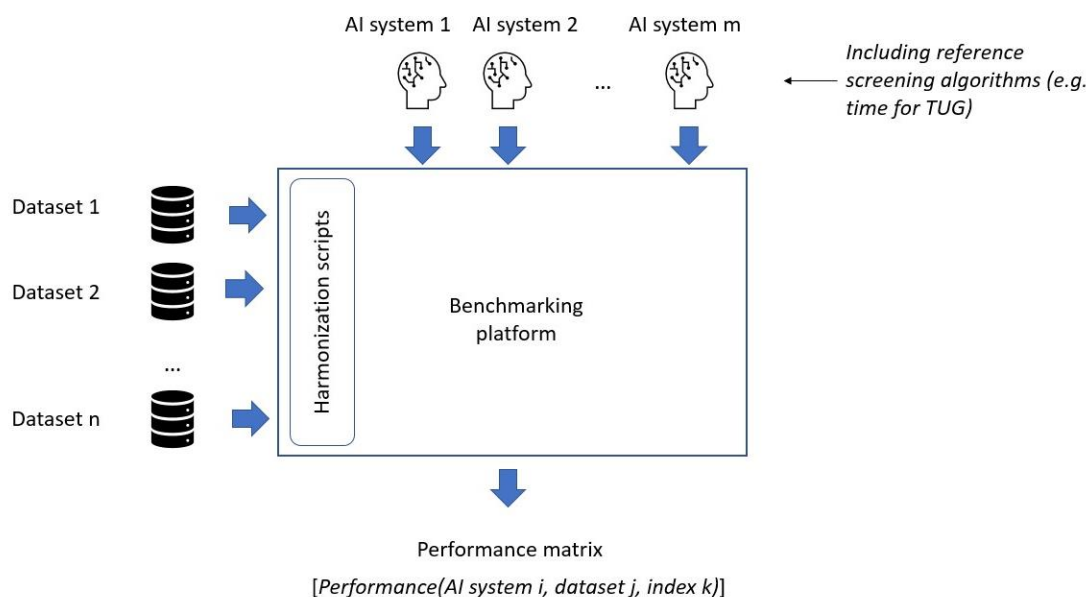
### 6.1.1.2 Benchmarking methods

This clause provides details about the methods of the benchmarking version 1. 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 clause 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 proposed benchmarking. 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 proposed benchmarking**

In Figure 1, each AI system for fall prediction is evaluated upon multiple datasets and multiple performance indices.

The performance matrix could be thought of 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 the performance scores. Table 2 provides an example of a performance matrix.

**Table 2 – Example of a performance matrix for an AI system evaluated on three datasets**

AI system			
	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%
...	...	–	...

Some datasets could be accessed promptly for the benchmarking version 1, including the FallSensing study, provided by the group of Inês Sousa, and the InCHIANTI datasets, currently analysed by Personal Health Systems laboratory of the University of Bologna. Procedures to have formal access to the dataset are underway.

#### 6.1.1.2.2 Benchmarking system dataflow

This clause describes the dataflow throughout the benchmarking architecture.

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

- Dataset identification. TG-Falls searches for datasets that are possibly suitable to be included in the benchmarking. This search is conducted by a systematic review of available datasets supporting wearable sensor-based tools for fall prediction, as per the protocol registered in PROSPERO<sup>8</sup>. TG-Falls is retrieving datasets from journal articles and data portals. Journal articles have been identified from systematic reviews on sensor-based fall risk assessment. Sources for retrieving systematic reviews have been: Pubmed, Web of Science and Scopus. 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. Eligible systematic reviews are in English and published not before 2017. Additional datasets will be searched for from data portals like Google Dataset Search, Mendeley Data, IEEE DataPort, Physionet, Figshare, Dataverse and Dryad.
- Eligibility check. The TG-Falls defines eligible all peer-reviewed articles/conference proceedings in English including datasets with the following characteristics:
  - Datasets including at least 20 individuals
  - Datasets where the predicting features comprised of at least one inertial sensor-based feature
  - Datasets from any community-dwelling population
  - Datasets with individual-level (not aggregated) information about falls
  - Falls collected after the predicting features (prospective design).

Information about falls should be available as occurrence of at least one fall in a given time period or a number of falls or the date of first fall occurrence. The datasets should also have an ethical waiver which does not impede their use for such an IPD-MA. Datasets not described in peer-reviewed articles will be excluded.

<sup>8</sup> [https://www.crd.york.ac.uk/prospero/display\\_record.php?RecordID=367394](https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=367394)

- Invitation of authors. TG-Falls will contact the authors of all the eligible datasets, inviting them to participate to the IPD-MA. Invitation emails and subsequent reminders will be sent according to current practice [41]. The authors willing to participate will have the possibility to choose among three data-sharing (DS) levels:
  - DS1: To share their dataset, including raw sensor data, into a secure centralised repository
  - DS2: To run the signal processing scripts prepared by the TG-Falls at their own premises and share data on digital biomarkers and falls at individual level
  - DS3: To run at their own premises the processing scripts prepared by the TG-Falls for calculating the digital biomarkers and their association with falls, and share the final association measures (e.g., odds ratios).

Depending on the chosen data-sharing level, the TG-Falls could perform a one-stage or two-stage IPD-MA. According to the guidelines for authorship of the International Committee of Medical Journal Editors (ICMJE), the data donors will be invited to contribute as authors to the publication of the results of the IPD-MA.

- Dataset entry. The available datasets will be included in the benchmarking using harmonization scripts, which create a harmonized dataset from each input dataset. Each input and harmonized dataset will be assigned a version number. Each dataset will be described in a description document which will be 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 shall be maintained according to their data management documents. Each harmonized dataset will be made of records, that are defined as the collection of AI input and label related to a single individual. 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 the AI input taken from only one wave.

#### **6.1.1.2.3 Safe and secure system operation and hosting**

This clause 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.

TG-Falls will store the data shared under DS1 or DS2 in secure private repositories of the Lepida datacenter<sup>9</sup>. The datasets will be maintained according to their data management documents, agreed upon with the contributing authors. Further details will be specified in future versions of this document.

#### **6.1.1.2.4 Benchmarking process**

TG-Falls will identify eligible datasets with a systematic review and retrieve the available ones sending invitation emails to the authors, as described above. The invitation emails will contain a

---

<sup>9</sup> <https://www.lepida.net/en/datacenter-cloud/home>



detailed protocol of the IPD-MA constituting the benchmarking process. The description of the process will include its scope and its time frame.

Each IPD-MA participant will have the possibility to propose one wearable sensor-based AI fall prediction system for benchmarking. They shall declare if by any means they have ever accessed any of the benchmarking datasets or parts thereof. Additionally, multivariate models will be developed and validated following published procedures for risk prediction models in IPD meta-analyses [1].

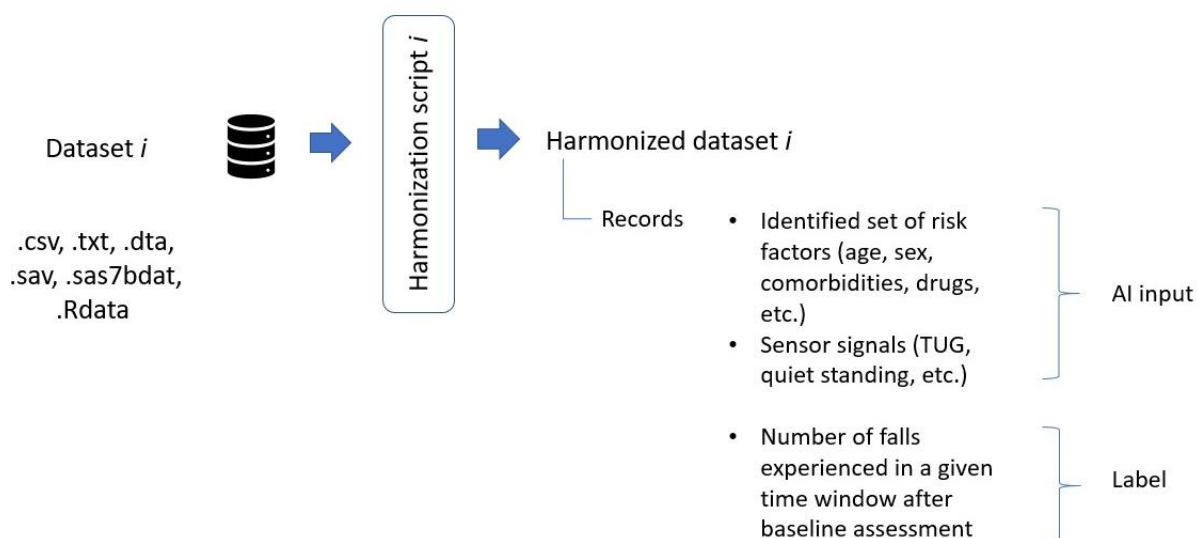
The algorithms could be coded as Python scripts or encrypted files. Further specifications on files will be defined in future versions of this document.

The participants shall agree that the results obtained by benchmarking their AI systems will be submitted for publication in scientific peer-reviewed journals.

### 6.1.1.3 AI input data structure for the benchmarking

This clause describes the input data provided to the AI solutions as part of the benchmarking of wearable sensor-based AI systems for fall prediction. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking.

The AI input data will encompass both clinical risk factors for falls and recordings from wearable inertial sensors during instrumented functional tests or real-world recordings (Figure 2). The AI input data will be arranged as much as possible using common data models (e.g., OMOP) [6] and standardized vocabularies (e.g., ICD-10, SNOMED, ICF, ATC, etc.). The wearable inertial sensor signals will be arranged according to the Mobilise-D convention [27]. More details will be specified in future versions of this document.



**Figure 2 – Dataset harmonization Schema representing the production of a harmonized dataset from an input dataset using a harmonization script**

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

### 6.1.1.4 AI output data structure

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

The 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 prediction window after the assessment. In this case the AI prediction is 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 the AI predictions are ordered non probabilistic and the values are not constrained in the range between 0 and 1.

In future releases of the benchmarking, it may be considered to accept AI systems that produce other subject-specific output variables, e.g., expressing the expected number of falls in a future time window. Additionally, the algorithms could further provide suggestions on possible preventive actions to take.

#### 6.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous clauses, 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 clause.

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 prediction window after the assessment. In future releases of the benchmarking, the label could be an integer expressing the number of times the subject fell down.

#### 6.1.1.6 Scores and metrics

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

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:  $\text{Youden index} = \text{sensitivity} + \text{specificity} - 1$ .

Table 3 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 AI system 3 is applicable and probabilistic.

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.

**Table 3 – 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%

Dataset 1 v1			
	AI system 1	AI system 2	AI system 3
Specificity	74%	–	81%
Brier score	–	–	0.018

#### 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.

The benchmarking will include the datasets identified by the systematic review and made available by the authors. The data will be stored in secure private repositories of the Lepida data center<sup>10</sup>. They will not be shared to third parties. Further details will be specified in future versions of this document.

#### 6.1.1.8 Data sharing policies

This clause 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 [DEL5.5](#) on *data handling* and [DEL5.6](#) on *data sharing practices*).

The authors willing to participate to the IPD-MA will have the possibility to choose among three data-sharing (DS) levels:

- DS1: To share their dataset, including raw sensor data, into a secure centralised repository
- DS2: To run the signal processing scripts prepared by the TG-Falls at their own premises and share data on digital biomarkers and falls at individual level
- DS3: To run at their own premises the processing scripts prepared by the TG-Falls for calculating the digital biomarkers and their association with falls, and share the final association measures (e.g., odds ratios).

The datasets should have an ethical waiver which does not impede their use for the IPD-MA. The datasets shall be maintained according to their data management documents, agreed upon with the contributing authors.

#### 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.

Where possible, the wearable sensor-based AI systems for fall prediction will be compared with the time to perform the TUG, gait speed, and risk level provided by the stratification algorithm of the World Guidelines for fall prevention and management (Montero-Odasso 2022).

#### 6.1.1.10 Reporting methodology

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

When sharing their datasets and submitting their AI systems for fall prediction for the benchmarking, the participants shall agree that the results of the IPD-MA will be submitted for publication to a peer-

<sup>10</sup> <https://www.lepida.net/en/datacenter-cloud/home>



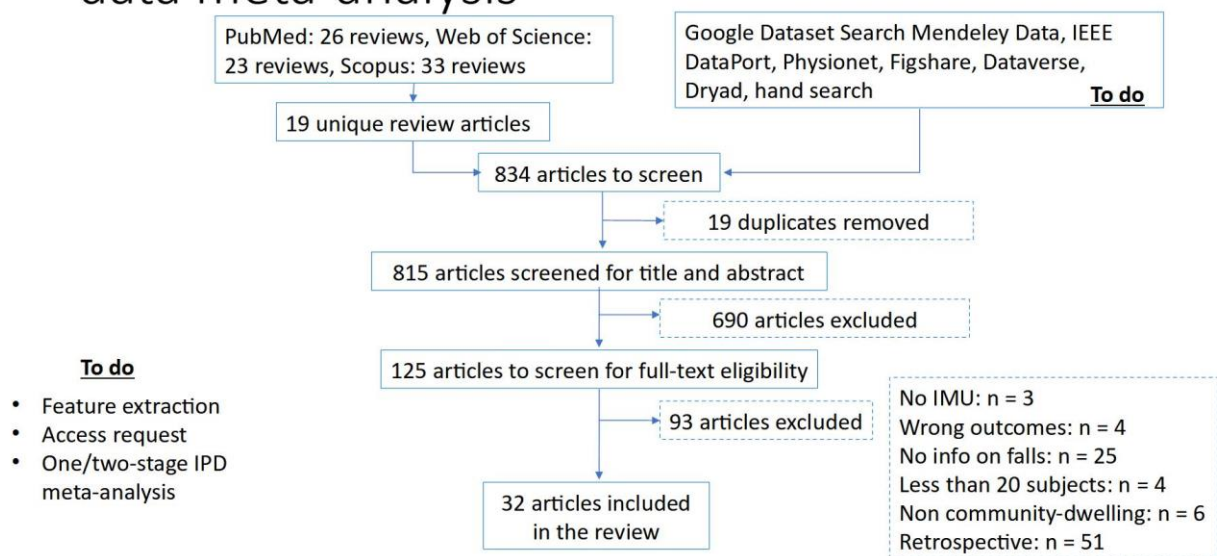
reviewed scientific journal. According to the guidelines for authorship of the International Committee of Medical Journal Editors (ICMJE), the data donors will be invited to contribute as authors to the publication of the results of the IPD-MA.

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 [10], the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) [24], and the Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data (PRISMA-IPD) checklist [39].

#### 6.1.1.11 Result

Within the systematic review of datasets for wearable sensor-based fall risk prediction, TG-Falls has retrieved 19 unique systematic reviews from Pubmed, Web of Science and Scopus. From this set of systematic reviews, 834 articles were identified for screening and 33 were identified as eligible and included in the systematic review (Figure 3). Further updates will be included in future versions of this document.

## Systematic review and individual participant data meta-analysis



**Figure 3 – PRISMA flowchart of the systematic review on datasets for wearable sensor-based fall prediction**

#### 6.1.1.12 Discussion of the benchmarking

This clause discusses insights of this benchmarking iteration and provides details about the outcome of the benchmarking process (e.g., giving an overview of the benchmark results and process). It will be completed in future versions of this document.

#### 6.1.1.13 Retirement

This clause 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 clause [DEL4](#) "AI software lifecycle specification" (identification of standards and best practices that are relevant for the AI of a health software life cycle).

Each dataset comprising the IPD-MA will be maintained according to their data management documents, agreed upon with the contributing authors. Whenever possible, the datasets will be made publicly available or securely stored for future use. Further details will be added in future versions of this document.

## **7 Overall discussion of the benchmarking**

This clause discusses the overall insights gained from benchmarking work in this topic group.

TG-Falls has identified a systematic review and IPD-MA as rigorous methods for benchmarking AI systems for fall prediction. This work is expected to produce a remarkable step forward in the literature of fall prediction.

## **8 Regulatory considerations**

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

The deliverables with relevance for regulatory considerations are [DEL2](#) *"Regulatory considerations on artificial intelligence for health"* (which provides an educational overview of some key regulatory considerations), [DEL2.1](#) *"Mapping of IMDRF essential principles to AI for health software"*, and [DEL2.2](#) *"Good practices for health applications of machine learning: Considerations for manufacturers and regulators"* (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). [DEL4](#) 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 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 are 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 clause summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for prevention of falls among the elderly. It will be completed in future versions of this document.

### **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 clause 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. It will be completed in future versions of this document.

### **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 clause outlines the regulatory requirements for software used for benchmarking in this topic group. It will be completed in future versions of this document.

#### **8.4 Regulatory approach for the topic group**

Building on the outlined regulatory requirements, this clause 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 [DEL2](#) *"Regulatory considerations on artificial intelligence for health"*. This clause will be completed in future versions of this document.

## References

- [1] Ahmed I, Debray TPA, Moons KGM, Riley RD (2014), *Developing and validating risk prediction models in an individual participant data meta-analysis*, BMC Med Res Methodol, 14(3).
- [2] Barry E, Galvin R, Keogh C, Horgan F, Fahey T (2014), *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 [Internet]. [cited 2014 Feb 21];14(1):14. Available from: <http://www.biomedcentral.com/1471-2318/14/14>
- [3] Beck Jepsen D, Robinson K, Ogliari G, Montero-Odasso M, Kamkar N, Ryg J, *et al.* (2022), *Predicting falls in older adults: an umbrella review of instruments assessing gait, balance, and functional mobility*, BMC Geriatr [Internet]. [cited 2022 Nov 9];22(1):1–27. Available from: <https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-022-03271-5>
- [4] Berg K, Wood-Dauphinée S, Williams JI, Gayton D (1989), *Measuring Balance in the Elderly Preliminary development of an Instrument*, Physiother Canada, 41(6):304–11.
- [5] Bhasin S, Gill TM, Reuben DB, Latham NK, Ganz DA, Greene EJ, *et al.* (2020), *A Randomized Trial of a Multifactorial Strategy to Prevent Serious Fall Injuries*, N Engl J Med, 383(2):129–40.
- [6] Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, *et al.* (2021), *Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases*. BMC Med Res Methodol [Internet];21(1):1–16. Available from: <https://doi.org/10.1186/s12874-021-01434-3>
- [7] Burns ER, Stevens JA, Lee R (2016), *The direct costs of fatal and non-fatal falls among older adults – United States*. J Safety Res [Internet]. [cited 2021 Jan 20]; 58:99–103. Available from: <https://pubmed.ncbi.nlm.nih.gov/27620939/> World Health Organization (2007), *WHO global report on falls prevention in older age*. Geneva, Switzerland: World Health Organization.
- [8] Casilari E, Santoyo-Ramón JA, Cano-García JM (2020), *On the Heterogeneity of Existing Repositories of Movements Intended for the Evaluation of Fall Detection Systems*, J Healthc Eng [Internet]. [cited 2021 May 8]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7738812/>
- [9] Chaieb S, Mrad A Ben, Hnich B (2021), *Interventions Recommendation System for Preventing future Falls in Older Adults*. Procedia Comput Sci.;192:192–201.
- [10] Collins GS, Reitsma JB, Altman DG, Moons KGM (2015), *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement*, Ann Intern Med [Internet]. [cited 2015 Sep 28];162(1):55–63. Available from: <http://annals.org/article.aspx?articleid=2088549>
- [11] Fleming J, Brayne C (2008), *Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90, Cambridge City over-75s Cohort (CC75C) study collaboration*. BMJ [Internet]. [cited 2019 Mar 17];337:a2227. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19015185>
- [12] Gade GV, Jørgensen MG, Ryg J, Riis J, Thomsen K, Masud T, *et al.* (2021), *Predicting falls in community-dwelling older adults: a systematic review of prognostic models*, BMJ Open [Internet]. [cited 2021 May 11];11(5):e044170. Available from: <https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2020-044170>
- [13] Gneiting T, Katzfuss M (2014), *Probabilistic Forecasting*, Annu Rev Stat Its Appl, Vol. 1, Issue 1, pp. 125–151.

- [14] Gneiting T, Raftery AE (2007), *Strictly Proper Scoring Rules, Prediction, and Estimation*. J Am Stat Assoc [Internet]. [cited 2013 Aug 15];102(477):359–78. Available from: <http://www.tandfonline.com/doi/abs/10.1198/016214506000001437>
- [15] Greene BR, McManus K, Redmond SJ, Caulfield B, Quinn CC (2019), *Digital assessment of falls risk, frailty, and mobility impairment using wearable sensors*, npj Digit Med [Internet]. [cited 2020 Jan 13];2(1):125. Available from: <http://www.nature.com/articles/s41746-019-0204-z>
- [16] Howcroft J, Kofman J, Lemaire ED (2013), *Review of fall risk assessment in geriatric populations using inertial sensors*, J Neuroeng Rehabil [Internet]. [cited 2013 Aug 16];10(1):91. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23927446>
- [17] Hopewell S, Copsey B, Nicolson P, Adedire B, Boniface G, Lamb S (2019). *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 [Internet]. [cited 2021 May 11];54(22):1340–50. Available from: <http://bjsm.bmj.com/>
- [18] Klenk J, Chiari L, Helbostad JL, Zijlstra W, Aminian K, Todd C, *et al.* (2013), *Development of a standard fall data format for signals from body-worn sensors: the FARSEEING consensus*. Zeitschrift für Gerontol und Geriatr [Internet]. [cited 2016 Feb 1];46(8):720–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24271252>
- [19] Lamb SE, Bruce J, Hossain A, Ji C, Longo R, Lall R, *et al.* (2020), *Screening and Intervention to Prevent Falls and Fractures in Older People*, N Engl J Med [Internet]. [cited 2021 Jun 17];383(19):1848–59. Available from: <https://www.nejm.org/doi/10.1056/NEJMoa2001500>
- [20] Lamb SE, Jørstad-Stein EC, Hauer K, Becker C (2005), *Development of a common outcome data set for fall injury prevention trials: the Prevention of Falls Network Europe consensus*, J Am Geriatr Soc [Internet]. [cited 2012 Aug 22];53(9):1618–22. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16137297>
- [21] Lauritsen SM, Thiesson B, Jørgensen MJ, Riis AH, Espelund US, Weile JB, *et al.* (2021), *The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards*. npj Digit Med;4(1):1–12.
- [22] Mebrahtu TF, Skyrme S, Randell R, Keenan A-M, Bloor K, Yang H, *et al.* (2021), *Effects of computerised clinical decision support systems (CDSS) on nursing and allied health professional performance and patient outcomes: a systematic review of experimental and observational studies*, BMJ Open [Internet]. [cited 2022 Jan 10];11(12):e053886. Available from: <https://bmjopen.bmj.com/content/11/12/e053886>
- [23] Montero-Odasso M, van der Velde N, Martin FC, Petrovic M, Tan MP, Ryg J, *et al.* (2022), *World guidelines for falls prevention and management for older adults: a global initiative*, Age Ageing [Internet]. [cited 2022 Oct 4];51(9). Available from: <https://pubmed.ncbi.nlm.nih.gov/36178003/>
- [24] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, *et al.* (2014). *Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist*. PLoS Med [Internet]. [cited 2015 Oct 26];11(10):e1001744. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4196729&tool=pmcentrez&rendertype=abstract>
- [25] Morse JM, Morse RM, Tylko SJ (1989), *Development of a Scale to Identify the Fall-Prone Patient*, Can J Aging / La Rev Can du Vieil [Internet]. [cited 2021 Jan 20];8(4):366–77. Available from: <https://www.cambridge.org/core/journals/canadian-journal-on-aging-la-revue-canadienne-du-vieillessement/article/abs/development-of-a-scale-to-identify-the-fallprone-patient/A0CDA5381DEC8DA4D7E7A1B1A74692E>



- [26] National Institute for Health and Care Excellence. *Falls in older people overview*, NICE Pathways [Internet]. [cited 2015 Mar 3]. Available from: <http://pathways.nice.org.uk/pathways/falls-in-older-people>
- [27] Palmerini L, Reggi L, Bonci T, Del Din S, Micó-Amigo ME, Salis F, *et al.* (2023), *Mobility recorded by wearable devices and gold standards: the Mobilise-D procedure for data standardization*, *Sci Data*;10(1):1–13.
- [28] Palumbo P, Becker C, Bandinelli S, Chiari L (2019), *Simulating the effects of a clinical guidelines screening algorithm for fall risk in community dwelling older adults*. *Aging Clin Exp Res* [Internet]. [cited 2018 Oct 19];31(8):1069–76. Available from: <http://link.springer.com/10.1007/s40520-018-1051-5>
- [29] Palumbo P, Klenk J, Cattelan L, Bandinelli S, Ferrucci L, Rapp K, *et al.* (2016), *Predictive Performance of a Fall Risk Assessment Tool for Community-Dwelling Older People (FRAT-up) in 4 European Cohorts*, *J Am Med Dir Assoc* [Internet]. [cited 2016 Sep 26];17(12):1106–13. Available from: <http://www.sciencedirect.com/science/article/pii/S1525861016302936>
- [30] Palumbo P, Palmerini L, Bandinelli S, Chiari L (2015), *Fall Risk Assessment Tools for Elderly Living in the Community: Can We Do Better?* Wang Y, editor. *PLoS One* [Internet]. [cited 2015 Dec 30];10(12):e0146247. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146247>
- [31] 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* [Internet]. 2011. Available from: <http://geriatricscareonline.org/toc/updated-american-geriatrics-societybritish-geriatrics-society-clinical-practice-guideline-for-prevention-of-falls-in-older-persons-and-recommendations/CL014>
- [32] Podsiadlo D, Richardson S (1991), *The timed "Up & Go": a test of basic functional mobility for frail elderly persons*, *J Am Geriatr Soc* [Internet]. [cited 2014 Mar 27];39(2):142–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1991946>
- [33] Rapp K, Freiberger E, Todd C, Klenk J, Becker C, Denking M, *et al.* (2014), *Fall incidence in Germany: results of two population-based studies, and comparison of retrospective and prospective falls data collection methods*, *BMC Geriatr* [Internet]. [cited 2014 Nov 18];14:105. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4179843&tool=pmcentrez&rendertype=abstract>
- [34] Scheffer AC, Schuurmans MJ, Van dijk N, Van der hooft T, De rooij SE (2008), *Fear of falling: Measurement strategy, prevalence, risk factors and consequences among older persons*, *Age Ageing* [Internet]. [cited 2021 May 11];37(1):19–24. Available from: <https://pubmed.ncbi.nlm.nih.gov/18194967/>
- [35] Schoene D, Wu SM-S, Mikolaizak AS, Menant JC, Smith ST, Delbaere K, *et al.* (2013), *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* [Internet]. [cited 2013 May 23];61(2):202–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23350947>
- [36] Shany T, Wang K, Liu Y, Lovell NH, Redmond SJ (2015), *Review: Are we stumbling in our quest to find the best predictor? Over-optimism in sensor-based models for predicting falls in older adults*, [Internet]. Vol. 2, *Healthcare Technology Letters*. IET Digital Library; [cited 2015 Aug 13]. p. 79–88. Available from: <http://digital-library.theiet.org/content/journals/10.1049/htl.2015.0019>
- [37] Silva JR, Sousa I, Cardoso JS (2020), *Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls*, *IEEE J Biomed Heal Informatics* [Internet]. [cited 2019 Nov 12];24(1):1–1. Available from: <https://ieeexplore.ieee.org/document/8894463/>
- [38] Sonnega A, Weir DR (2014), *The Health and Retirement Study: A Public Data Resource for Research on Aging*, *Open Heal Data* [Internet]. [cited 2015 May 29];2(1):e7. Available from: [openhealthdata.metajnl.com/articles/10.5334/ohd.am/](http://openhealthdata.metajnl.com/articles/10.5334/ohd.am/)

- [39] Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, Tierney JF; PRISMA-IPD Development Group (2015), *Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement*, JAMA.;313(16):1657-65. doi: 10.1001/jama.2015.3656. PMID: 25919529.
- [40] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* (2010), *Assessing the performance of prediction models: a framework for traditional and novel measures*, Epidemiology [Internet]. [cited 2012 Nov 8];21(1):128–38. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20010215>
- [41] Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, *et al.* (2021), *Data sharing practices and data availability upon request differ across scientific disciplines*, Sci Data;8(1):1–11.
- [42] Tinetti ME. Performance-oriented assessment of mobility problems in elderly patients. J Am Geriatr Soc [Internet]. 1986 Feb [cited 2014 Mar 25];34(2):119–26. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3944402>
- [43] Jose Luis Albites Sanabria, Barry R. Greene, Killian McManus, Luca Palmerini, Pierpaolo Palumbo, Inês Sousa, Kimberley S. van Schooten, Eva Weicken, Markus Wenzel (2023), *Fall risk stratification of community-living older people. Commentary on the world guidelines for fall prevention and management*, Topic Group on "Falls among the elderly" of the ITU/WHO Focus Group "Artificial Intelligence for Health", Age Ageing;52(10):afad162.
- [44] van Schooten KS, Brodie M (2021), *Fall detection and risk assessment with new technologies*. In: Lord SR, Sherrington C, Naganathan V, editors. Falls in Older People. Cambridge: Cambridge University Press.
- [45] Wang X, Ellul J, Azzopardi G (2020), *Elderly Fall Detection Systems: A Literature Survey*, Front Robot AI;7:71.
- [46] Yadav D, Jain R, Agrawal H, Chattopadhyay P, Singh T, Jain A, *et al.* (2019), *EvalAI: Towards Better Evaluation of AI Agents*, [cited 2022 Feb 8]; Available from: <https://www.aicrowd.com/>

## Annex A

### Glossary

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

Acronym/Term	Expansion	Comment
AI	Artificial Intelligence	
AI-MD	AI-based Medical Device	
AI4H	Artificial Intelligence for Health	
API	Application Programming Interface	
AUC	Area Under the ROC Curve	
CfTGP	Call for Topic Group Participation	
CHARMS	CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies	
CI	Confidence Interval	
DEL	Deliverable	
DS	Data Sharing	
FDA	Food and Drug Administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross Domestic Product	
GDPR	General Data Protection Regulation	
HR	Hazard Ratio	
HRS	Health and Retirement Study	
HRQoL	Health related Quality of Life	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual Property	
IPD-MA	Individual-Participant Data Meta-Analysis	
ISO	International Organization for Standardization	
ITU	International Telecommunication Union	
LMIC	Low-and Middle-Income Countries	
MDR	Medical Device Regulation	
NNT	Number Needed to Treat	
OR	Odds Ratio	
PII	Personal Identifiable Information	
PRISMA	Preferred Reporting Items for Systematic Review and Meta-Analyses	
ProFaNE	Prevention of Falls Network Europe	
RaR	Rate Ratio	
ROC	Receiver Operating Characteristic curve	
SaMD	Software as a Medical Device	



<b>Acronym/Term</b>	<b>Expansion</b>	<b>Comment</b>
TBD	To be defined	
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	
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
TUG	Timed Up and Go Test	
WHO	World Health Organization	
WG	Working Group	

## **Annex B**

### **Declaration of conflict of interests**

Barry Greene and Killian McManus are employees of Linus Health and declare the ownership of share or share options in Linus Health Inc. Luca Palmerini is co-founder and own shares of mHealth Technologies srl. Pierpaolo Palumbo holds copyright on FRAT-up software code, for fall risk assessment in older adults. Kimberly van Schooten holds copyright on software codes for fall risk assessment in older adults.

Linus Health is a digital health company focused on brain health. Mhealth Technologies srl is a mobile health company offering technological solutions for gait assessment. They further offer customized versions of FRAT-up, a free-to-use tool for fall risk assessment.

---