International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.4

FG-AI4H 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; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; falls; elderly

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).

Editors:

Pierpaolo Palumbo
TG-Falls Topic Driver
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

Contributors:

Jose Albites Sanabria
University of Bologna
Italy
E-mail: jose.albitessanabri2@unibo.it

Barry Greene
Kinesis Health Technologies Ltd.
Ireland
E-mail: barry.greene@kinesis.ie

Killian McManus
Kinesis Health Technologies Ltd.
Ireland
E-mail: killian.mcmanus@kinesis.ie

Luca Palmerini
University of Bologna
Italy
E-mail: luca.palmerini@unibo.it

Pierpaolo Palumbo
University of Bologna
Italy
Tel: +39 3402378412
E-mail: pierpaolo.palumbo@unibo.it

Inês Sousa Associação Fraunhofer Portugal Research – Fraunhofer AICOS Portugal	Tel: +351 220 430 326 E-mail: ines.sousa@fraunhofer.pt
Kimberley S. van Schooten University of New South Wales, Sydney, NSW Australia.	E-mail: k.vanschooten@neura.edu.au
Eugenio Zuccarelli CVS Health USA	E-mail: eugenio.zuccarelli@gmail.com, ez@alum.mit.edu

CONTENTS

Page

Page

List of Tables

Page

List of Figures

Page

ITU-T FG-AI4H Deliverable 10.4

FG-AI4H 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% 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 (Burns et al. 2016).

Current guidelines for fall prevention and management in older adults, indicate to implement fall risk screening to identify those at increased risk (Montero-Odasso et al. 2022, National Institute for Health and Care Excellence). These high risk individuals should be offered fall prevention programs, which reduce fall risk by improving strength and balance and modifying behaviours (Hopewell et al. 2019). In case of a fall, prolonged time on 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 (Fleming et al. 2008).

According to these recommendations, diverse tools have been proposed for fall prevention and management, including systems for fall risk prediction (Gade et al. 2021), fall preventive intervention recommendations (Chaieb et al 2021, Mebrahtu et al. 2021), and fall detection and classification (Wang et al. 2020). 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 TG-Falls (Falls among the elderly) at the meeting A in Geneva, Switzerland, on 25-27 September 2018.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. During FG-AI4H meeting A in Geneva, Switzerland, on 25-27 September 2018, Inês Sousa from Associação Fraunhofer Portugal Research – Fraunhofer AICOS was nominated as topic driver for the TG-Falls. In December 2020, 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 the TG-Falls. He indicated that Kimberley van Schooten would be available to take over. Dr Kimberly 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 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 one subtopic 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. 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

Number	Title
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

Number	Title
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
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 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-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 the 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 the 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 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] (Silva et al. 2019).

2.2.7 Status update for meeting H (Brasilia)

The TG-Falls received demonstrations of potential interest from two groups that carry out epidemiological studies about ageing, with data on falls: InCHIANTI[1] and TILDA[2]. The TG-Falls discussed about 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.

[1] http://inchiantistudy.net/wp/

[2] https://tilda.tcd.ie/

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
- 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[3]. 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 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 (van Schooten and Brodie). 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 (Sonnega et al 2014).

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 (Casilari et al. 2020).

2.2.11 Status update for meeting M (E-meeting)

The TG-Falls 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.

TDD was updated with descriptions of the basic features that a first version of the benchmarking should have. The 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)

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

[3] https://eufallsfest2021.eu/index.php/workshops

[4] https://aiaudit.org/, https://health.aiaudit.org/

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). The TG-Falls drafted the protocol to submit to PROSPERO[5]. 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-analysis". The main aim of the systematic review is to identify the datasets available for training and validating 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 for requesting the datasets. The email includes an introduction to the FG-AI4H, to the TG-Falls, and to the systematic review. An annex describes the rational and the aims of the systematic and IPD-MA and provides details on data protection and authorship policy.

2.2.14 Status update for meeting R (Douala)

The TG-Falls registered the protocol for the systematic review in Prospero (PROSPERO 2022 CRD42022367394[6]). The TG-Falls 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 (Montero-Odasso 2022), also collecting 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)

The TG-Falls submitted the commentary on the world guidelines (Montero-Odasso 2022) to Age and Ageing. The commentary discusses about 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 has been accepted for publication in Age and Ageing.

The 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 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:

[5] https://www.crd.york.ac.uk/prospero/

[6] https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=367394

- 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

Diverse AI-based systems have been proposed for fall prevention and management, including systems for fall risk prediction (Gade et al. 2021), fall preventive intervention recommendations (Chaieb et al 2021, Mebrahtu et al. 2021), and fall detection and classification (Wang et al. 2020). The TG-Falls has decided to address the sub-topic of fall prediction in older adults. This section 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.

3.1.1 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) (Lauritsen et al. 2021), 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" (Lamb et al. 2005). 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 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 the prediction window. In this case we call the AI prediction to be 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 we call the AI prediction to be 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.

3.1.2 Current gold standard

At present, a variety of tools for fall risk prediction have been proposed (Gade at al. 2021, Beck Jepsen et al. 2022). The World Guidelines for falls prevention and management for older adults recommends to follow 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 (Montero-Odasso et al. 2022). 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) (Panel on Prevention of Falls in Older Persons 2011). 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) (Palumbo et al. 2018).

Another tool for fall prediction is the Timed Up and Go Test (TUG) (Podsiadlo et al. 1991). 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) (Barry et al. 2014, Schoene et al. 2013).

The predictive accuracy of classical or AI algorithms for fall prediction should be evaluated against prospective falls (Lamb et al. 2005). 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 (Lauritsen et al. 2021). Moreover, falls should preferably be ascertained using prospective daily recording and a notification system with a minimum of monthly reporting (Lamb et al. 2005). These requirements have become *de facto* standards in the literature.

3.1.3 Relevance and impact of an AI solution

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% 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 (Burns et al. 2016).

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 (Palumbo et al. 2018). 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 (Bhasin et al. 2020, Lamb et al. 2020). 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 (Gade et al. 2021, Shany et al. 2015). In the end, we believe that benchmarking will drive progress in this field.

3.1.4 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 (Gade et al. 2021). Of those, only three were validated and had an AUC between 0.62 and 0.69.

Howcroft et al. (Howcroft et al. 2013) 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% 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 (Howcroft et al. 2013).

Greene et al. (Greene et al. 2019) carried out a study involving 8521 participants (72.7 ± 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 "*AI4H ethics considerations,*" which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL1 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 (Scheffer et al. 2008). 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.

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.

5 Existing work on benchmarking

This section 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 section 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 *"Data and artificial intelligence assessment methods (DAISAM),"* 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 group (Lamb et al. 2005). 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 (Klenk et al. 2013).

Recently, the Mobilise-D consortium[7] 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 sensors data include accelerations and angular velocities. Data organization encompasses format, structure, and modalities (Palmerini et al. 2023). 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 (Steyerberg et al. 2010). 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 (Gneiting and Raftery 2007). Calibration cannot be computed when the output of the prediction tool is not probabilistic (Gneiting and Katzfuss 2014).

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) (Barry et al. 2014, Schoene et al. 2013).

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.

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 (Shany et al. 2015). 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 that

[7] www.mobilise-d.eu

simple traditional tools and exhibits much heterogeneity in its performance across different populations (Palumbo et al. 2015, Palumbo et al. 2016). 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 (Rapp et al. 2014).

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 any impact analysis has ever been done.

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 (Yadav et al). The TG-Falls is currently evaluating which software platform would be more suitable for the benchmarking IPD-MA.

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the AI-based fall prediction task. It includes one subsection 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 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), 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 based AI 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 section outlines the first benchmarking version implemented thus far and the rationale behind it. It serves as an introduction to the subsequent sections, 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: 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 1

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

6.1.1.1 Overview

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

6.1.1.2 Benchmarking methods

This section 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 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.

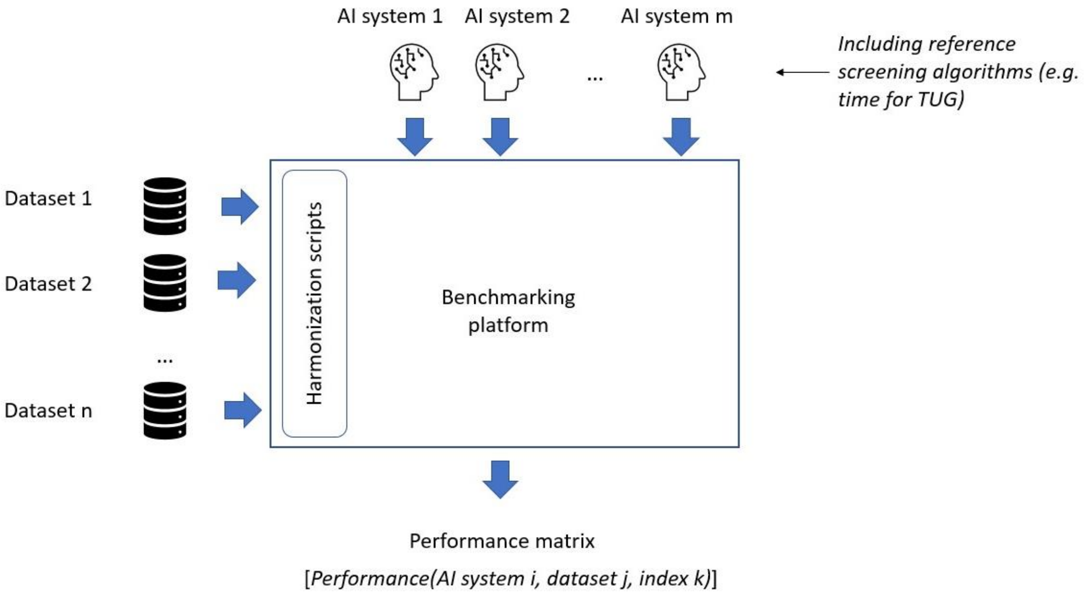

Figure 1: Overview of the benchmarking. Each AI system for fall prediction is evaluated upon multiple datasets and multiple performance indices.

The performance matrix could be thought 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 2 provides an example of 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 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 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 for the protocol registered in PROSPERO[8]. Briefly, the 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 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 number of falls OR 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.
- Invitation of authors. The 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 (Tedersoo et al. 2021). 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.

[8] https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=367394

- Data management. The datasets shall be maintained according to their data management documents. Each harmonized dataset will be made of records, that here we define 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 AI input taken from only one wave.

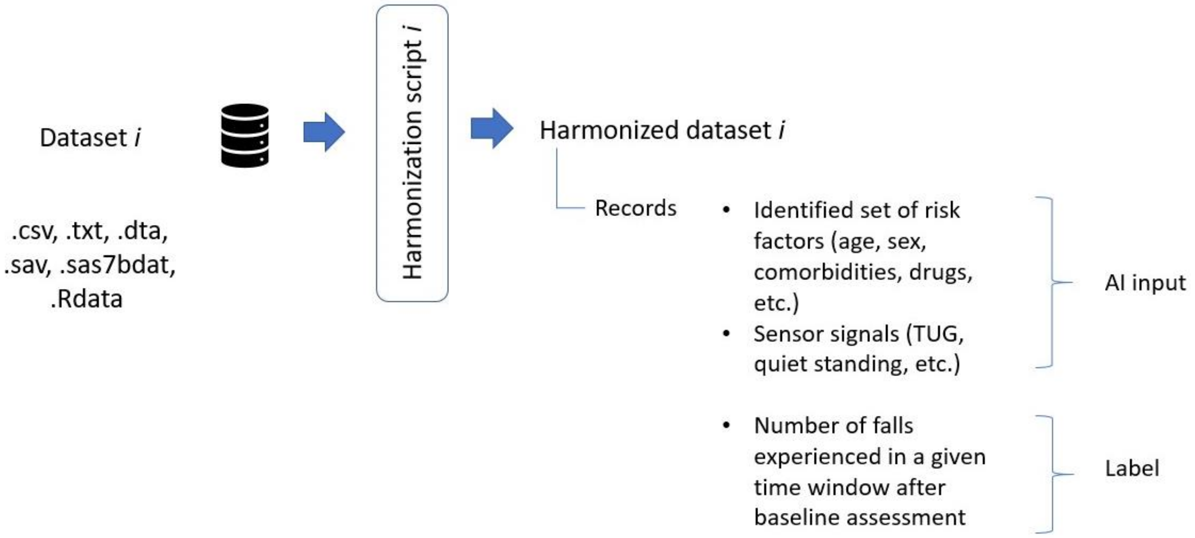
Dataset *i*
.csv, .txt, .dta, .sav, .sas7bdat, .Rdata
Harmonization script *i*
Harmonized dataset *i*
Records
Identified set of risk factors (age, sex, comorbidities, drugs, etc.)
Sensor signals (TUG, quiet standing, etc.)
AI input
Number of falls experienced in a given time window after baseline assessment
Label

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

6.1.1.2.3 Safe and secure system operation and hosting

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

The TG-Falls will store the data shared under DS1 or DS2 in secure private repositories of the Lepida datacenter[9]. 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

The 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 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 (Ahmed et al. 2014).

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

[9] https://www.lepida.net/en/datacenter-cloud/home

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 section 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, “Dataset harmonization”). The AI input data will be arranged as much as possible using common data models (e.g. OMOP) (Biedermann et al. 2021) 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 (Palmerini et al. 2023). More details will be specified in future versions of this document.

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

The 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 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. 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 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.

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 section 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: 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.

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 provides an example of evaluation grid for a dataset.

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%
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[10]. 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 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 DEL5.5 on *data handling* and DEL5.6 on *data sharing practices*).

[10] https://www.lepida.net/en/datacenter-cloud/home

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 section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and 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-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 (Collins et al. 2015), the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (Moons et al. 2014), and the Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data (PRISMA-IPD) checklist (Stewart et al. 2015).

6.1.1.11 Result

Within the systematic review of datasets for wearable sensor-based fall risk prediction, the 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.

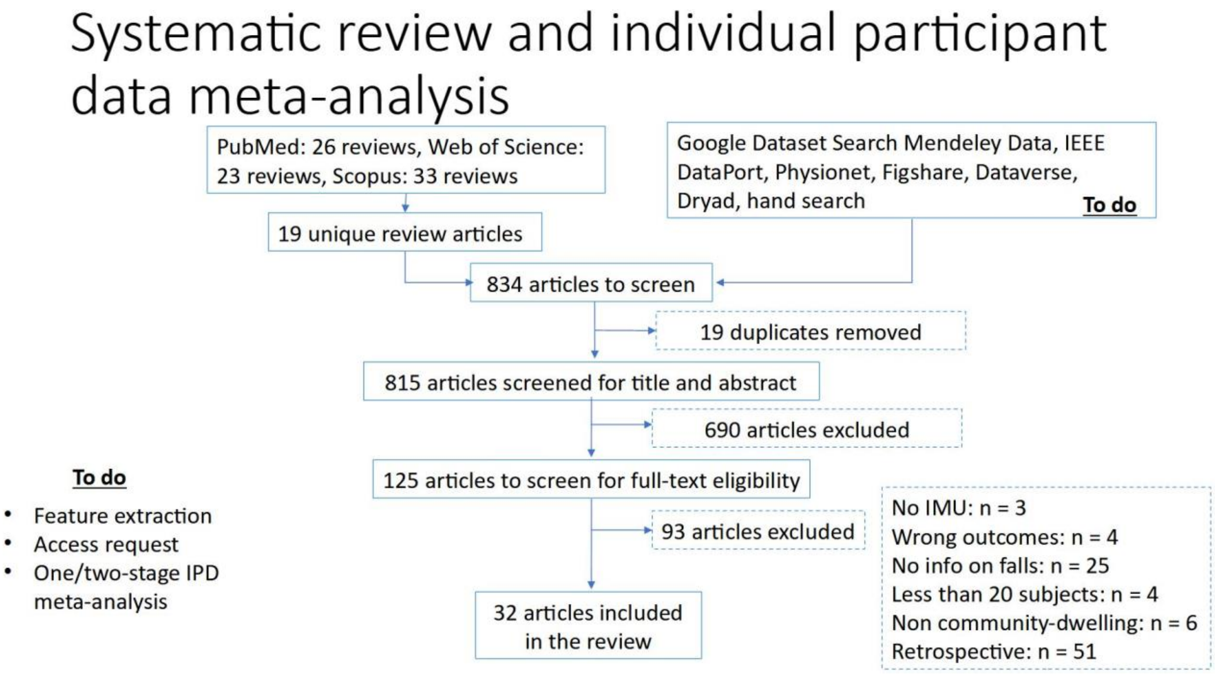
Systematic review and individual participant data meta-analysis
PubMed: 26 reviews, Web of Science: 23 reviews, Scopus: 33 reviews
Google Dataset Search Mendeley Data, IEEE DataPort, Physionet, Figshare, Dataverse, Dryad, hand search
To do
19 unique review articles
834 articles to screen
19 duplicates removed
815 articles screened for title and abstract
690 articles excluded
125 articles to screen for full-text eligibility
93 articles excluded
No IMU: n = 3
Wrong outcomes: n = 4
No info on falls: n = 25
Less than 20 subjects: n = 4
Non community-dwelling: n = 6
Retrospective: n = 51
32 articles included in the review
To do
• Feature extraction
• Access request
• One/two-stage IPD 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 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). It will be completed in future versions of this document.

6.1.1.13 Retirement

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section DEL4 "*AI software lifecycle specification"* (identification of standards and best practices that are relevant for the AI for health software life cycle).

Each dataset comprising the IPD-MA will be maintained according to the their data management documents, agreed upon with the contributing authors. Whenever possible, the datasets will 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 section discusses the overall insights gained from benchmarking work in this topic group.

The 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 *"AI4H regulatory considerations"* (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 *"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). 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 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. 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 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. 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 section 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 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 DEL2 *"AI4H regulatory considerations."* This section will be completed in future versions of this document.

References

1 Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14(3).

2 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 [Internet]. 2014 Jan [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. Predicting falls in older adults: an umbrella review of instruments assessing gait, balance, and functional mobility. BMC Geriatr [Internet]. 2022 Dec 1 [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. Measuring Balance in the Elderly Preliminary development of an Instrument. Physiother Canada. 1989;41(6):304–11.

5 Bhasin S, Gill TM, Reuben DB, Latham NK, Ganz DA, Greene EJ, et al. A Randomized Trial of a Multifactorial Strategy to Prevent Serious Fall Injuries. N Engl J Med. 2020;383(2):129–40.

6 Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol [Internet]. 2021;21(1):1–16. Available from: https://doi.org/10.1186/s12874-021-01434-3

7 Burns ER, Stevens JA, Lee R. The direct costs of fatal and non-fatal falls among older adults — United States. J Safety Res [Internet]. 2016 Sep 1 [cited 2021 Jan 20];58:99–103. Available from: https://pubmed.ncbi.nlm.nih.gov/27620939/World Health Organization. WHO global report on falls prevention in older age. Geneva, Switzerland: World Health Organization; 2007.

8 Casilari E, Santoyo-Ramón JA, Cano-García JM. On the Heterogeneity of Existing Repositories of Movements Intended for the Evaluation of Fall Detection Systems. J Healthc Eng [Internet]. 2020 [cited 2021 May 8];2020. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7738812/

9 Chaieb S, Mrad A Ben, Hnich B. Interventions Recommendation System for Preventing future Falls in Older Adults. Procedia Comput Sci. 2021 Jan 1;192:192–201.

10 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med [Internet]. 2015 Jan 6 [cited 2015 Sep 28];162(1):55–63. Available from: http://annals.org/article.aspx?articleid=2088549

11 Fleming J, Brayne C, Cambridge City over-75s Cohort (CC75C) study collaboration. Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. BMJ [Internet]. 2008 Nov 17 [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. Predicting falls in community-dwelling older adults: a systematic review of prognostic models. BMJ Open [Internet]. 2021 May 4 [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. Probabilistic Forecasting. Annu Rev Stat Its Appl. 2014;

14 Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. J Am Stat Assoc [Internet]. 2007 Mar [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. Digital assessment of falls risk, frailty, and mobility impairment using wearable sensors. npj Digit Med [Internet]. 2019 Dec 11 [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. Review of fall risk assessment in geriatric populations using inertial sensors. J Neuroeng Rehabil [Internet]. 2013 Aug 8 [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. 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]. 2019 Nov 1 [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. Development of a standard fall data format for signals from body-worn sensors : the FARSEEING consensus. Zeitschrift für Gerontol und Geriatr [Internet]. 2013 Dec [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. Screening and Intervention to Prevent Falls and Fractures in Older People. N Engl J Med [Internet]. 2020 Nov 5 [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. Development of a common outcome data set for fall injury prevention trials: the Prevention of Falls Network Europe consensus. J Am Geriatr Soc [Internet]. 2005 Sep [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. The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. npj Digit Med. 2021;4(1):1–12.

22 Mebrahtu TF, Skyrme S, Randell R, Keenan A-M, Bloor K, Yang H, et al. 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]. 2021 Dec 1 [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. World guidelines for falls prevention and management for older adults: a global initiative. Age Ageing [Internet]. 2022 Sep 2 [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. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med [Internet]. 2014 Oct [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. Development of a Scale to Identify the Fall-Prone Patient. Can J Aging / La Rev Can du Vieil [Internet]. 1989 [cited 2021 Jan 20];8(4):366–77. Available from: https://www.cambridge.org/core/journals/canadian-journal-on-aging-la-

revue-canadienne-du-vieillissement/article/abs/development-of-a-scale-to-identify-the-fallprone-patient/A0CDFA5381DEC8DA4D7E7A1B1A74692E

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. Mobility recorded by wearable devices and gold standards: the Mobilise-D procedure for data standardization. Sci Data. 2023;10(1):1–13.

28 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 [Internet]. 2019 Aug 19 [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, Cattelani L, Bandinelli S, Ferrucci L, Rapp K, 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 [Internet]. 2016 Dec 1 [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. Fall Risk Assessment Tools for Elderly Living in the Community: Can We Do Better? Wang Y, editor. PLoS One [Internet]. 2015 Dec 30 [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. The timed “Up & Go”: a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc [Internet]. 1991 Feb [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, Denkinger M, et al. Fall incidence in Germany: results of two population-based studies, and comparison of retrospective and prospective falls data collection methods. BMC Geriatr [Internet]. 2014 Jan [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. Fear of falling: Measurement strategy, prevalence, risk factors and consequences among older persons. Age Ageing [Internet]. 2008 Jan [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. 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]. 2013 Feb [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. 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; 2015 [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. Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls. IEEE J Biomed Heal Informatics [Internet]. 2020 [cited 2019 Nov 12];24(1):1–1. Available from: https://ieeexplore.ieee.org/document/8894463/

38 Sonnega A, Weir DR. The Health and Retirement Study: A Public Data Resource for Research on Aging. Open Heal Data [Internet]. 2014 Oct 16 [cited 2015 May 29];2(1):e7. Available from: 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. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA. 2015 Apr 28;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. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology [Internet]. 2010 Jan [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. Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data. 2021;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 Topic Group on “Falls among the elderly” of the ITU/WHO Focus Group “Artificial Intelligence for Health”: Jose Luis Albites Sanabria, Barry R. Greene, Killian McManus, Luca Palmerini, Pierpaolo Palumbo, Inês Sousa, Kimberley S. van Schooten, Eva Weicken, Markus Wenzel. Fall risk stratification of community-living older people. Commentary on the world guidelines for fall prevention and management. Age and Ageing. 2023 (*Accepted for publication*)

44 van Schooten KS, Brodie M. Fall detection and risk assessment with new technologies. In: Lord SR, Sherrington C, Naganathan V, editors. Falls in Older People. Cambridge: Cambridge University Press; 2021.

45 Wang X, Ellul J, Azzopardi G. Elderly Fall Detection Systems: A Literature Survey. Front Robot AI. 2020 Jun 23;7:71.

46 Yadav D, Jain R, Agrawal H, Chattopadhyay P, Singh T, Jain A, et al. EvalAI: Towards Better Evaluation of AI Agents. [cited 2022 Feb 8]; Available from: 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
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	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
IPD-MA	Individual-participant data meta-analysis	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MDR	Medical Device Regulation	
NNT	Number needed to treat	
PII	Personal identifiable information	
PRISMA	Preferred Reporting Items for Systematic Review and Meta-Analyses	
ProFaNE	Prevention of Falls Network Europe	
ROC	Receiver Operating Characteristic curve	
SaMD	Software as a medical device	
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.