|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATION****STANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-E-024 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plen | Geneva, 30 May-1 June 2019 |
| **DOCUMENT** |
| **Source:** | University of Bologna |
| **Title:** | TG-Falls: Benchmarking fall prediction AI algorithms: general thoughts and experience |
| **Purpose:** | Discussion |
| **Contact:** | Pierpaolo PalumboUniversity of Bologna, Italy | Tel: +39 051 20 93188Email: pierpaolo.palumbo@unibo.it  |
| **Contact:** | Lorenzo ChiariUniversity of Bologna, Italy | Tel: +39 051 20 93095Email: lorenzo.chiari@unibo.it  |

|  |  |
| --- | --- |
| **Abstract:** | This document briefly discusses some points that the authors think are important for the definition of a procedure for benchmarking fall prediction AI algorithms and present some of the author's research experience. The discussed points are: target of prediction, predictors, populations, semantics of prediction, and performance metrics. |

Here we briefly discuss some topics that we think are important for the definition of a procedure for benchmarking fall prediction AI algorithms. Alongside this discussion, we present some of our research experience.

# Target of prediction

In 2005 the Prevention of Falls Network Europe (ProFaNE) issued a set of recommendations on outcome measures for fall injury prevention trials [1]. These recommendations were established after an international expert consensus process and still represent a cornerstone about standardization of fall measures also in the field of fall risk assessment.

Choosing different outcome measures (e.g. single or multiple falls, or duration of the follow-up period) as targets of prediction, clearly leads to different values of performance metrics for the same predictive algorithm. Simple probabilistic models can explain these differences and are potentially useful for adjustments [2].

# Predictors

Besides outcome measures, we think that datasets for training and validating AI algorithms for fall prediction should ideally contain: i) the principal acknowledged risk factors for falls (see e.g. [3]); ii) new/putative biomarkers for falls (e.g. digital biomarkers from wearable motion sensors); iii) measures from established clinical tools for fall risk assessment (e.g. the Timed Up and Go test or results of screening algorithms recommended by scientific associations and public health institutes).

From acknowledged fall risk factors, we have developed FRAT-up, a web-based fall risk assessment tool for community-dwelling older adults (<http://ffrat.farseeingresearch.eu/>) [4]. Regarding the search for new biomarkers, we focused on quantity, quality, and variability of turn measures, recorded with wearable inertial sensors during ambulatory monitoring. We have identified turn duration (DUR), mean and peak turn velocity (MV, PV), number of steps to turn (NSTEP), number of turns per hour (TPH), and day-to-day variability of turn angle (CV-ANG) as digital biomarkers having prognostic information on future falls [4]. Regarding established screening algorithms, we validated the algorithm indicated in the guidelines for fall prevention issued by the American Geriatrics Society and British Geriatrics Society [5].

# Populations

Fall incidence and fall risk factors prevalence rates vary greatly in different datasets, the reason being yet to fully uncover [6]. As expected, while validating FRAT-up we found much heterogeneity in its performance across different populations [7]. Benchmarking fall prediction algorithms on different datasets/populations may provide more robust estimates of their performance.

# Semantics of prediction

Algorithms for fall prediction can deliver their prediction according to different semantics. Namely, the prediction can be deterministic (e.g. no fall vs at least one fall), probabilistic (probability distribution over the possible outcome events, e.g. probability of falling at least once in 12 months, probability of experiencing 3 falls in 12 months, etc.), or fuzzy (dichotomic or with multiple levels, e.g. high vs intermediate vs low risk). In our works we have chosen probabilistic predictions, over dichotomic events (probability of no fall vs at least one fall in 12 months) or counts (probability distribution over the number of falls in 12 months) [8], [9].

# Performance metrics

Biostatistical literature on prognostic models indicates classical performance metrics to use for continuous or dichotomic risk indicators (e.g. sensitivity, specificity, positive/negative predictive value, area under the ROC curve (AUC), etc.). Performance is generally thought of as made of two components: discriminative ability (i.e. ability to discriminate between classes, e.g. fallers and non-fallers) and calibration (i.e. accordance between predictions and observed frequencies). For dichotomic events, discriminative ability is generally assessed with AUC, while calibration with calibration plots and the Hosmer-Lemeshow test.

We also think that algorithms should be assessed for their robustness with respect to missing data, since they often occur in clinical research and real-world scenarios. We employed a very simple procedure to assess performance degradation with increasing percentage of missing values on a model for prediction of depression in older adults [10].

Finally, besides the statistical predictive accuracy, fall predictive algorithms should be evaluated on the basis of the impact they would have once coupled with appropriate preventive interventions, as measured in terms of prevented falls and/or improved allocation of resources available for prevention (e.g. [5]).

References

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

[2] P. Palumbo, L. Palmerini, and L. Chiari, “A probabilistic model to investigate the properties of prognostic tools for falls,” *Methods Inf. Med.*, vol. 54, no. 2, pp. 189–197, 2015.

[3] S. Deandrea, E. Lucenteforte, F. Bravi, R. Foschi, C. La Vecchia, and E. Negri, “Risk Factors for Falls in Community-dwelling Older People: A Systematic Review and Meta-analysis,” *Epidemiology*, vol. 21, no. 5, pp. 658–668, 2010.

[4] J. M. Leach, S. Mellone, P. Palumbo, S. Bandinelli, and L. Chiari, “Natural turn measures predict recurrent falls in community-dwelling older adults: a longitudinal cohort study,” *Sci. Rep.*, vol. 8, no. 1, p. 4316, Dec. 2018.

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

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

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

[8] L. Cattelani *et al.*, “FRAT-up, a web-based fall risk assessment tool for elderly people living in the community,” *J. Med. Internet Res.*, vol. 17, no. 2, p. e41, 2015.

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

[10] L. Cattelani, M. Belvederi Murri, F. Chesani, L. Chiari, S. Bandinelli, and P. Palumbo, “Risk Prediction Model for Late Life Depression: Development and Validation on Three Large European Datasets,” *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, 2018.

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