The Optimal Sequence and Selection of Screening Test Items to Predict Fall Risk in Older Disabled Women: The Women’s Health and Aging Study

Sarah E. Lamb,1 Chris McCabe,2 Clemens Becker,3 Linda P. Fried,4 and Jack M. Guralnik5

1Warwick Clinical Trials Unit and 2Health Sciences Research Institute, Warwick Medical School, University of Warwick, Coventry, United Kingdom.
3Klinik für Geriatrische Rehabilitation, Robert Bosch-Krankenhaus, Stuttgart, Germany.
4The Johns Hopkins University Medical Institutions, Baltimore, Maryland.
5Epidemiology, Demography and Biometry Program, National Institute on Aging, Bethesda, Maryland.

Background. Falls are a major cause of disability, dependence, and death in older people. Brief screening algorithms may be helpful in identifying risk and leading to more detailed assessment. Our aim was to determine the most effective sequence of falls screening test items from a wide selection of recommended items including self-report and performance tests, and to compare performance with other published guidelines.

Methods. Data were from a prospective, age-stratified, cohort study. Participants were 1002 community-dwelling women aged 65 years old or older, experiencing at least some mild disability. Assessments of fall risk factors were conducted in participants’ homes. Fall outcomes were collected at 6 monthly intervals. Algorithms were built for prediction of any fall over a 12-month period using tree classification with cross-set validation.

Results. Algorithms using performance tests provided the best prediction of fall events, and achieved moderate to strong performance when compared to commonly accepted benchmarks. The items selected by the best performing algorithm were the number of falls in the last year and, in selected subpopulations, frequency of difficulty balancing while walking, a 4 m walking speed test, body mass index, and a test of knee extensor strength. The algorithm performed better than that from the American Geriatric Society/British Geriatric Society/American Academy of Orthopaedic Surgeons and other guidance, although these findings should be treated with caution.

Conclusions. Suggestions are made on the type, number, and sequence of tests that could be used to maximize estimation of the probability of falling in older disabled women.

Key Words: Accidental falls—Multiphasic screening—Aged.

Falls are a major cause of disability, dependence, and death in older people (1). Women are at particular risk of falling (1). Current guidelines suggest that all older people should be given a short screen for fall risk and, if at risk, referred on for further detailed assessment and targeted intervention (2). Effective interventions are exercise and multifactorial programs (3,4). Screening recommendations from the American Geriatric Society/British Geriatric Society/American Academy of Orthopaedic Surgeons (AGS/BGS/AAOS) Panel on Falls Prevention are that older people should be asked about falls in the last year and, in selected subpopulations, frequency of difficulty balancing while walking, a 4 m walking speed test, body mass index, and a test of knee extensor strength. The algorithm performed better than that from the American Geriatric Society/British Geriatric Society/American Academy of Orthopaedic Surgeons and other guidance, although these findings should be treated with caution.

The aim of this analysis was to suggest improvements in test selection, sequence, and cut points for screening tools by using a data-driven approach to develop algorithms to estimate fall risk in community-dwelling older women with a range of physical abilities, and to investigate the performance of recommended methods of screening for falls in community-dwelling populations (the AGS/BGS/AAOS and the Ganz guideline).

METHODS

Study Design and Setting
Data were from the Women’s Health and Aging Study (WHAS). The Institutional Review Board of The Johns
Hopkins University approved the study, and all women gave informed, written consent. Study methods have been described in detail elsewhere (10).

Sample
The Health Care Financing Administration enrollment file was used to obtain an age-stratified random sample of all female Medicare beneficiaries living in Baltimore, Maryland (n = 6521). Women living in nursing homes or who had moved out of the area were excluded. The remainder (5316) were invited for eligibility checks; 4137 women agreed. Women were eligible if they reported at least mild disability in 2 or more of the following domains: mobility (including walking, climbing steps, bed/chair transfers, or doing heavy housework); upper extremity activities (including raising arm above the head, grasping/handling objects, or lifting/carrying 10 lbs); basic self-care (including bathing or showering, dressing, eating, or toileting); and higher functioning (including using the telephone, doing light housework, preparing meals, or shopping for personal items). Women with substantial cognitive impairment (<18 on the Mini-Mental State Examination (MMSE) (11)) were excluded. Of 1409 women meeting the eligibility criteria, 1002 agreed to participate, with no significant disparities between eligible women who participated and those who refused (10).

Women were examined in their home by trained interviewers and nurses (10). Follow-up evaluations were conducted every 6 months for 3 years. The analysis presented here used the baseline data (predictor variables), and the first two follow-up time points (response variables). The baseline examination included a range of variables relevant to falls screening, but excluded tests that required specialist nonportable equipment. We selected potential items on the basis of existing recommendations that the item should be used to screen for risk of falling, injurious falls or if it was a commonly occurring and frequently reported risk factor for falls (7,9).

Predictor Variables

Activities of Daily Living.—Participants self-rated the amount of difficulty with bathing, transferring from bed to chair, toileting, dressing, and eating, using response categories of no difficulty, a little difficulty, some difficulty, a lot of difficulty, or inability to perform the task. An overall score was derived as described by Volpato and colleagues (12).

Age.—Age was entered as a continuous variable.

Body mass index.—Height (m) was measured barefoot, in standing and with a stadiometer. Weight (kg) was measured with a calibrated bathroom type digital scale, and with light indoor clothes but no shoes, jewelry, or heavy clothing. Body mass index (weight/height$^2$) was entered as a continuous variable and as a dichotomous variable (<30 kg/m$^2$).

Depressive symptoms.—Depressive symptoms were assessed using the Geriatric Depression Scale (13). Scores were entered as a continuous variable, and were categorized as 0–9 for no depression, 10–13 for mild to moderate depression, and ≥14 for moderate to severe depression (14).

Chronic diseases.—Women were asked whether a physician had ever told them that they had diabetes, osteoporosis, stroke, Parkinson’s disease, arthritis, myocardial infarction, angina, chronic heart failure, high blood pressure, any other heart problems, cancer, lung diseases, and/or previous hip fracture. Women were asked if they experienced any difficulties with control of urine or if they experienced any wetting. The responses were coded as yes or no.

Multisite pain.—Pain severity in the back, knees, hips, and/or feet was classed as (i) no pain or mild pain in only one site, (ii) pain in two sites or moderate to severe pain in only one site, or (iii) pain in three or four sites regardless of severity (15).

Cardiovascular risk factors.—We asked the frequency with which women experienced fainting, dizziness, or spinning in the last year.

Medications.—We included polypharmacy (≥4 prescribed medications) and use of sedative/hypnotic medications. Other classes of medication have less consistent or weaker associations with falls and were not considered (16,17).

Vision and hearing.—Visual acuity was evaluated using a Snellen chart (18). If customary, women wore glasses. In addition we asked “Do you have difficulty recognizing a face across a small room?” (19). Hearing deficit was elicited by asking “Do you have any difficulty in hearing to converse in a small room?”

Gait and balance.—The literature suggests an array of tests for detecting gait and balance impairment associated with falling. The most frequently suggested are timed performance tests (2). We considered usual and fastest walk speed over 4 m, chair stand time, and elements of the Tinetti balance protocol (20). Details of the test protocols are published elsewhere (10). We also summarized these tests into the Short Physical Performance Battery (21).

Self-reported gait and balance.—Self-report may be more predictive of falls than are performance tests (22). Women were asked about the frequency with which they had difficulty balancing while walking and dressing. Responses were recorded as never, sometimes, often, very often, or always. Also, women were asked (i) if they considered their balance to be poor, and (ii) how much difficulty they had with walking 2–3 blocks.

Cognition.—The MMSE (11,23) was entered continuously or using a cut point of ≤ 24 to indicate mild to moderate impairment.

Physical activity.—Physical activity was summarized as the number of city blocks walked in the last week, with >8 blocks used to denote higher levels of activity.
Alcohol consumption.—Participants were asked the number of alcoholic drinks they consumed per week.

Maximum knee extensor strength.—Women sat in a chair with the hip flexed to 90° and knee flexed to 85°. A handheld muscle dynamometer (Nicholas Manual Muscle Tester, model BK-7454; Fred Sammons Inc.) was placed a few inches above the ankle joint. The women were encouraged to extend with maximum effort. Maximum force (kg) was recorded for two 5-second contractions from each leg. Sufficient time was allowed between contractions for recovery. The average from both legs was used (24).

Fear of falling.—Women were asked if they had any fear of falling in the last year and, if so, how frequently (none, a little, often, a lot).

History of falling prior to the baseline assessment.—Participants were asked if they had fallen during the last 12 months and the number of falls in the last year.

Response Variable
Falling was defined as falling on the ground or at some other level such as chair level. At each of the 6 monthly follow-ups, women were asked if they had fallen in the preceding 6 months and how many falls they had experienced. The response variable was any reported falls in the 12-month follow-up period.

Analysis
We built two algorithms using tree-based classification (TC), which is a recursive partitioning technique that optimizes the selection and sequence of questions and cut points. We used an exhaustive chi-square automatic interaction detection method (25) because of the facility for multiway variable splits (26). The algorithm is built by first selecting the variable most closely associated to the outcome; it then identifies cut points that distinguish risk groups maximally (using chi-square statistics). Then, within each of the risk groups identified, the algorithm is reapplied separately to each subgroup (or partition), and so on until the tree comes to a node (or end point). The algorithm calculates the probability of falling for individuals within each node, with a possible range of 0–1. To ensure simple algorithms, node sizes were set to a minimum of 100 for parent nodes and 50 for final nodes.

The tree calculates a separate probability for each node. A probability threshold can then be selected for the entire tree and is used to assign the node as “predicted to fall” or “predicted not to fall.” This threshold can be varied depending on the desired measurement properties of the algorithm. We calculated performance across a variety of thresholds. For each probability threshold, we calculated the sensitivity, specificity, positive and negative likelihood ratios (PLR and NLR, respectively), and diagnostic odds ratio (DOR) for prediction of fall events during the follow-up period. The PLR calculates how many more times a person who experiences...
a fall in the follow-up year is likely to have a positive test result, and the NLR calculates the number of times a person who has no fall in the follow-up year is likely to have a negative test result. The DOR is the ratio of PLR and NLR, and should ideally be \( >4 \) (27). A strategy with strong evidence would have a PLR \( >5 \) and an NLR \( <0.2 \) (28).

We built two algorithms, one based on self-report items alone and a second incorporating a selection of performance tests. Predictor variables were entered with no prespecified cut points. For variables that also have commonly accepted clinical cut points, these were included as additional candidates. Cross-set (n – 1) methods were used to maximize validity in preference to split set validation (29). Trees were grown in 20 randomly sampled sets of the data, and final trees are an average of all 20 models. Missing predictor values were minimal (\(<1\%\)), and were incorporated.

The WHAS contained all the variables necessary to construct the Ganz and AGS/BGS/AAOS guideline algorithms, but neither guideline details the exact choice of tests (9) or cut points (2,9). Therefore, we estimated screening performance using timed walking speed and, separately, self-report balance. We used definitions of slow speed as \(<0.42 \text{ m/s} \) based on previously published cohorts (30), as well as the cut points identified by the tree methods.

Statistical analysis was carried out using SPSS (version 15; SPSS Inc., Chicago, IL) and S-PLUS (version 7; Insightful, Seattle, WA).

**RESULTS**

**Sample**

A total of 885 women were available for analysis at 12 months. Sixty-two (6%) women died by 12 months, and 107 (10.7%) were lost to follow-up (unable to be located or refused to participate). Women who died in the follow-up period were older and more frail (Table 1). With the exception of a small difference in depression, there were no statistically or clinically significant differences between women lost to follow-up and those retained in the analysis (Table 1). Thirty-nine percent of women experienced at least one fall in the follow-up period.

**Self-Report Algorithms**

The self-report algorithm is shown in Figure 1. The algorithm diagram is read from top downward. The dark boxes indicate the succession of items selected by the algorithm, and the light boxes are the partitions/nodes identified. Each light box specifies the optimal cut point for the item selected by the tree, the probability of falling detected using the item, and the numbers of women falling within that part of the tree at that point. The first question selected was “Have you experienced any falls in the last 12 months?” The algorithm identified three splits: no falls, one fall only, or greater than one fall. For recurrent falls, no further variables were defined as discriminatory. For women with no prior falls, the next most discriminatory question was “How often do you have problems balancing while walking?” with the maximum point of discrimination being between rare/no problems versus sometimes/often and always. For women reporting one prior fall, activities of daily living difficulty was the next item selected, and the maximal discriminatory cut point was between none, a little, or some problems and a lot of problems or inability to perform the tasks. The range of fall probabilities generated for the whole algorithm was 0.30–0.59.

![Figure 1. Tree diagram showing algorithm for prediction of any fall (self-report items only). *ADLs = activities of daily living.](https://academic.oup.com/biomedgerontology/article-abstract/63/10/1082/559232/188)
Performance and Self-Report Algorithm

The algorithm is shown in Figure 2. The first question selected was “Have you experienced any falls in the last 12 months?” with splits identified as no falls, one fall only, or greater than one fall. For no prior falls, the next question was self-reported problems balancing while walking, and, within the partition of women with rare or no problems, a test of knee extensor strength identified women with an elevated risk (probability of falling 0.41 vs 0.24). For women who reported one prior fall, the algorithm selected a 4 m usual speed walk test as the next item. This was a highly discriminatory test. Finally, in women who reported recurrent falls, body mass index ≤ 29.2 distinguished a group of women at very high risk of falling. The range of probabilities generated by the model was from 0.24 to 0.76.

Predicting Fall Events From the Algorithms

The sensitivity, specificity, likelihood, and DORs of the probability thresholds generated are shown in Table 2. Predictive performance was maximal in the combined performance and self-report algorithm, followed by the self-report algorithm, AGS/BGS/AAOS and Ganz guidance, respectively. Better predictive performance was achieved if higher thresholds of probability were used to assign fall event status.

DISCUSSION

We have investigated the performance of the AGS/BGS/AAOS screening and the Ganz algorithm in a cohort of older women living in the community, and modeled an alternative sequence of questions and cut points from a wide selection of candidate variables (2,9). The data suggest that strategies developed through more sophisticated modeling methods might yield improved screening. However, performance of the items, sequence, and cut points must be confirmed in independent samples before recommendation on implementation can be made. In particular, it is important to recognize that the WHAS is representative of older women with disability, and it is not clear how the cutoffs would work in other populations.

Despite refinement and only when using high levels of fall probability to assign fall event status, the properties of all algorithms and guidance were moderate to strong. The superior predictive performance was reflected in higher PLRs, lower NLRs, and higher DORs (28,29). The highest levels of accuracy that are observed in other clinical fields were not achieved, nor was performance consistently strong across all indicators. There are a number of potential reasons. First, most studies in other fields ascertain case status within a relatively short time period, which enhances screening accuracy. Second, our method of ascertaining falls during follow-up is likely to
have been biased as we used recall over a 6-month period. Fall events may have been underestimated as a consequence (31). The third potential source of variability is measurement error in the individual components of the algorithm.

The structure of the algorithm confirms the importance of falls history as the first line of questioning. However, there were notable departures in structure from both the AGS/BGS/AAOS and Ganz guidance. There has been substantial debate about whether performance tests should be used in screening algorithms, which tests to select, and how best to use them (e.g., in selected persons or in all people). Screening performance was improved substantially by inclusion of a 4 m walk test in women who reported a previous fall, and the algorithm consistently selected self-reported balance problems in preference to a balance performance test. The most likely explanation is that the self-report question contained temporal reference to the frequency of problems over a prolonged time period, which may not be captured in a single measurement. As in previous systematic reviews, a test of knee extensor force was helpful in identifying women at risk of falling (24). The algorithm suggests that, for screening purposes, knee strength measures are relevant for persons with no history of falls or gait or balance problems, and would not need to be implemented in all strands of the screening process.

The algorithm demonstrated sensitivity in excess of many other published fall risk measures, including many that also report from their test sample only (6). The most likely explanation is the use of multiple tests, sequenced within each risk stratum. The TC technique used was a flexible method of analysis, offering rapid detection of optimized cut points, and a recursive partitioning method that was simple to implement. An important and unresolved question is what benchmark of screening performance should be met by a test designed for population screening, and the level of risk at which various types of intervention should be implemented. These questions can only be decided by modeling, empirical studies, and using expert groups that consider the range of clinical effectiveness and costs generated by various screening and linked treatment strategies. Such activities were beyond the scope of this analysis, but the data presented in this article provide empirical estimation of screening performance and guidance on test selection and sequence. One argument is that, in the WHAS, the lowest level of fall probability that could be detected (24%) would merit intervention for all women. Alternatively, targeting of highly resource-intensive interventions may be desirable.

Fall incidence was similar to that in other population-based cohorts. The candidate variables were predominantly intrinsic risk factors. It is well accepted that interaction between extrinsic and intrinsic factors is important, but that the assessment of intrinsic factors is the most practical approach for first-line screening (2,9). We included all of the most commonly recommended items for fall screening (2,9). The only notable exception was the Timed Up and Go Test, which is a gait and balance test in which the time to complete a sit to stand, short walk and turn is measured (32). It is a composite test, integrating three functions into one timed evaluation. To minimize the possibility that it is the combination of items that is important in screening, we summed the scores of the performance tests in accordance with Guralnik and colleagues (21). Another limitation of the algorithms is that they have been generated in a cohort of women. Women have a greater propensity than men do to fall (1). However, gender has not emerged as an important split in previous algorithms (33). Men tend to fall outside or
as the result of engaging in vigorous activities more often, but otherwise the risk profile is similar (15,34). Formal validation of the algorithm should be undertaken in mixed gender and/or male-only cohorts as well as in cohorts that are representative of the population more generally. The recall period for fall events was quite long (6 months), but this is reflective of the intervals recommended in various clinical guidelines (2).

Conclusion

We have undertaken quantitative modeling of variables that have been recommended as screening items previously, and have generated a series of algorithms that vary in the degree to which they use sophisticated techniques and/or additional equipment. We have presented comparative statistics that inform judgments about the trade-offs in performance between screening options. We recommend further testing of an algorithm incorporating a sequenced use of performance tests as well as self-reported items for establishing risk of fall events and estimating fall probability.

Acknowledgments

This work was supported by the National Institute on Aging (NIA; contract N01-AI21122 and the NIA Intramural Research Program. We acknowledge with thanks the assistance of Angeliki Chorti and Christelle Evaert in the preparation of the manuscript.

Correspondence

Address correspondence to Sarah E. Lamb, DPhil, MSc, Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK. E-mail: S.Lamb@warwick.ac.uk

References