

ForesightCares

AVAL: A VALIDATION PIPELINE FOR AVA MOBILITY

TECHNICAL REPORT

ForesightCares, Inc.
Charlotte, NC
info@foresightcares.com

ABSTRACT

Mobility tests such as the 30-second chair stand and the Timed Up and Go (TUG) are widely used to identify older adults at elevated risk of falling. Both tests are simple, but they depend on a trained observer to count repetitions or run a stopwatch, which limits how often and how widely they can be given. AVA Mobility is a vision-based assessment system that automates these tests using a single iPad camera. It performs pose estimation and movement-event detection entirely on the device, with no wearable sensors, external hardware, or cloud processing. AVA performs well in controlled testing, but its accuracy in the community settings where it is meant to be used had not been formally measured.

This report introduces **AVAL**, a validation pipeline that compares AVA's results against measurements made by physical therapists (PTs) and against a commercial wearable motion-capture system (Noraxon MR4). Data were collected from 66 older adults across four community sites between June and October 2025. AVAL links each participant's video to their reference measurements, runs AVA on the video, reconstructs the test outcome, and computes error metrics. TUG timing is evaluated with MAPE-based accuracy, mean absolute error, root mean squared error, and 95th-percentile error. Chair-stand counts are evaluated by the share of predictions that match the clinical count exactly or within one repetition. We also compare the low, fair, and high mobility categories each system assigns using CDC STEADI thresholds.

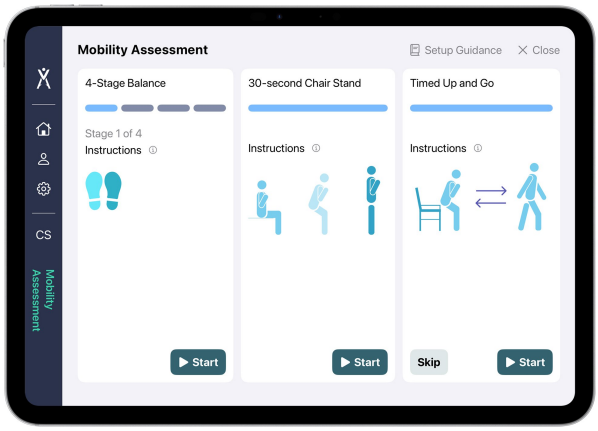
On the TUG, AVA's timings had a mean absolute percentage error (MAPE) of 6.0% relative to PT timing, with a mean absolute error of 0.71 s; the MR4 overestimated TUG time by 1.59 s on average. On the chair stand, AVA matched the clinical count exactly for all but one participant, and every prediction was within one repetition. AVA also agreed more closely than the MR4 with PT-assigned mobility categories on both tests (TUG: 88.1% vs. 71.2%; chair stand: 100% vs. 87.5%, on the same participants). These results indicate that AVA can deliver accurate, scalable, and privacy-preserving mobility assessment in real-world community settings.

1 Introduction

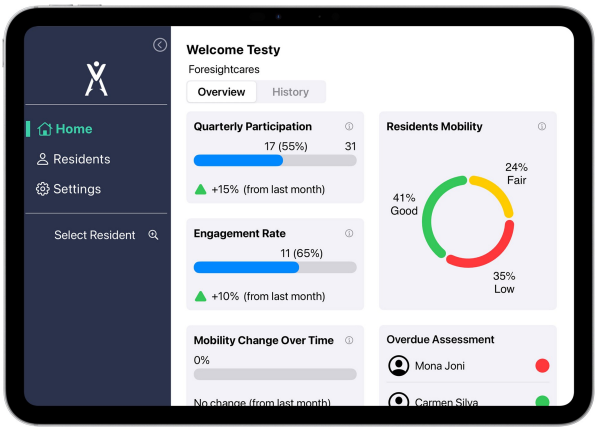
Falls are the leading cause of injury, hospitalization, and loss of independence among older adults [1]. To catch mobility problems early, the Centers for Disease Control and Prevention (CDC) recommends routine functional assessments, most notably the 30-second chair stand and the Timed Up and Go (TUG) test [2]. These tests are inexpensive and quick, but each one still needs a trained observer to count repetitions, time the task, and judge whether it was performed correctly. This makes it hard to assess large populations regularly, and results can vary from one observer or location to another.

Until now, removing the human observer has generally meant using a multi-sensor motion-capture system. The Noraxon MR4, for example, uses sixteen wearable inertial measurement units (IMUs) that are strapped or adhered to the body and must be calibrated before each assessment. Setup takes roughly 30 minutes per participant and requires a trained operator, which makes such systems costly and impractical for routine screening in community settings.

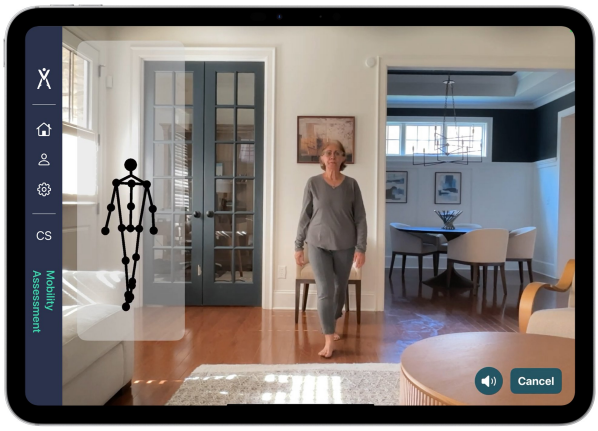
We developed **AVA Mobility** to remove these barriers. AVA is a lightweight pose-estimation and event-detection system that runs on an iPad and administers standardized mobility tests using only the device's camera. It requires no wearable sensors, no calibration, and no technical expertise from the person running the test. Figure 1 shows the AVA interface and the two tests evaluated in this report.



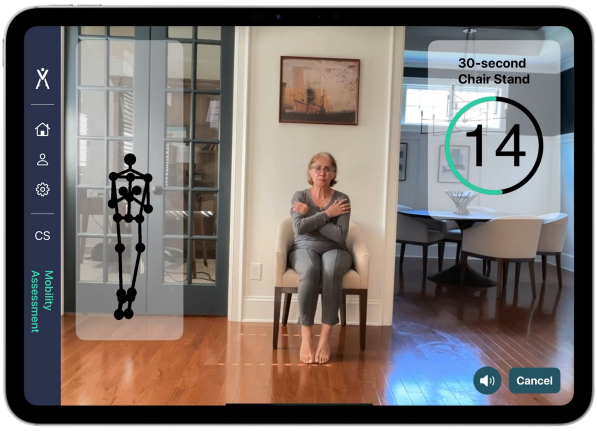
(a) Available AVA assessments



(b) AVA home screen



(c) Participant performing the Timed Up and Go (TUG) test



(d) Participant performing the 30-second chair stand test

Figure 1: The AVA Mobility interface and the two assessments evaluated in this report.

AVA performs reliably in controlled testing, but it had not yet been formally validated in the conditions it is designed for. Community settings vary in lighting, background, room layout, and participant ability. Showing that AVA stays accurate across this everyday variation is a necessary step before broad clinical and community deployment.

This report introduces **AVAL** (AVA Validation), a pipeline that benchmarks AVA against physical-therapist (PT) measurements collected at four senior-living and community sites. The same assessments were also recorded with the MR4 motion-capture system, which gives a second automated system to compare against. AVAL manages the video and metadata database, runs AVA on each recording, and reconstructs the chair-stand and TUG outcomes for comparison.

Our main contributions are:

- A labeled dataset of 66 older adults from four community sites, with each assessment captured on iPad video, by the MR4 motion-capture system, and by manual clinical measurement.

- **AVAL**, a repeatable pipeline for validating AVA Mobility against reference measurements.
- An evaluation of AVA on clinically administered TUG and chair-stand tests, covering continuous error metrics, tolerance-based accuracy, agreement on mobility categories, and a site-by-site comparison.
- An analysis of how errors are distributed and of the conditions that produce occasional outliers.

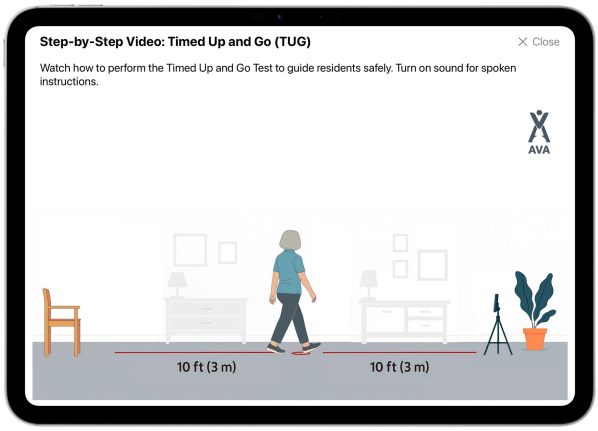
Together, these results provide the first formal validation of AVA and show that vision-based mobility assessment can work reliably outside the clinic.

2 AVA Mobility

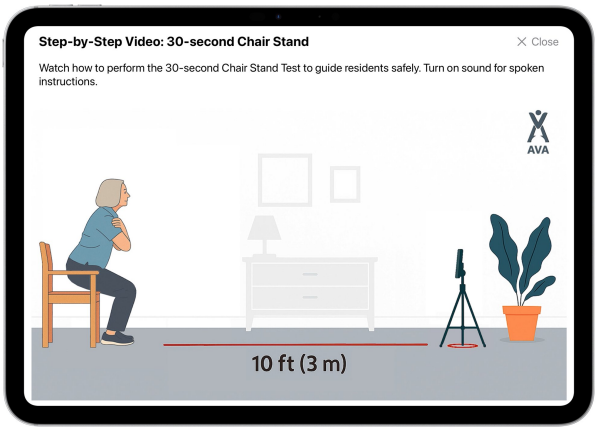
AVA Mobility turns camera video into clinical mobility measurements entirely on an iOS device. A GPU-accelerated pose detector tracks 33 body landmarks at 30 frames per second, with no cloud connection and no motion-capture hardware. AVA then applies custom biomechanical algorithms to these landmarks to detect the movement events that clinical tests depend on and to convert them into standard scores.

Tracking body landmarks is only the first step. The harder problem, and AVA's core contribution, is detecting clinical events reliably from a single frontal camera: identifying exactly when a person rises from a chair, sits back down, turns, takes a step, or loses balance. Single-camera pose estimates are noisy, especially under variable lighting, cluttered backgrounds, and inconsistent subject positioning. AVA therefore emphasizes stability over time, so that a detected event reflects a real movement rather than a momentary tracking error. This event-detection method is protected under U.S. Patent No. 12,343,138.

AVA is built to work within the limits of a mobile device, including restricted processing power, the need for real-time feedback, and unpredictable surroundings. Unlike motion-analysis systems that need controlled settings, multiple cameras, or specialized equipment, AVA works from one live video feed. It also guides the participant through each test with on-screen, step-by-step instructions (Figure 2), so the test can be run without a trained administrator.



(a) TUG instructions



(b) 30-second chair stand instructions

Figure 2: Step-by-step instructions that AVA Mobility shows participants before each assessment.

3 Data Collection

3.1 Study Design, Period, and Ethics

This was a prospective, cross-sectional method-comparison study. Each participant was assessed once, and each assessment was measured simultaneously by AVA, the MR4 motion-capture system, and physical therapists. Data were collected between **June 2025 and October 2025**. The study was approved by the Advarra Institutional Review Board (protocol Pro00080873) and conducted in accordance with the Declaration of Helsinki [3]. All participants gave informed consent before taking part, including consent to video recording and to wearing the MR4 sensors. Reporting follows the STARD 2015 [4] and GRRAS [5] guidelines where applicable.

3.2 Setting and Procedure

Data were collected at four community sites. Individuals with a pacemaker were excluded for safety, because they could not wear the MR4 sensors. Participants varied widely in mobility, balance, gait, and familiarity with standardized mobility tests. This range was intentional: it allowed us to evaluate AVA under the conditions it will meet in practice.

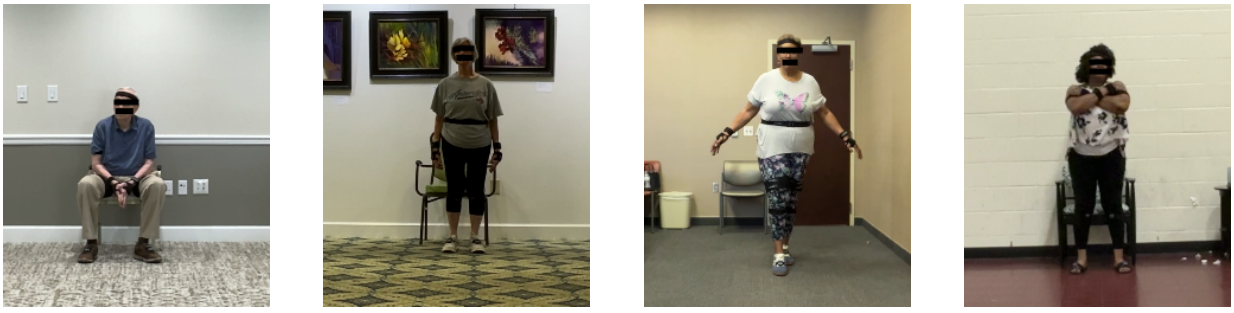
Each participant completed the following assessments, in the same order, while wearing the Noraxon MR4 system:

- a 30-second chair stand test;
- a Timed Up and Go (TUG) test;
- a 4-stage balance test (feet together, semi-tandem, tandem, and single-leg stance).

The balance test is outside the scope of this report and is not analyzed here.

The MR4 system consists of sixteen wireless IMUs placed on the major body segments and joints. Its accompanying software computes assessment metrics such as TUG duration and chair-stand repetitions, and these outputs are what we compare against. Operators ran the full MR4 calibration sequence before each assessment.

A single iPad was placed about 20 feet directly in front of the participant, far enough that the full movement stayed in frame. This gave AVA the same frontal view at every site. Physical therapists (PTs) manually recorded chair-stand repetitions and TUG times, and these PT measurements serve as the reference standard throughout the report.



(a) Community 1 (b) Community 2 (c) Community 3 (d) Community 4

Figure 3: Assessment spaces at the four data-collection sites. Site names are withheld for privacy.

Figure 3 shows the assessment spaces at each site. Tests took place in community gyms, multipurpose rooms, and small conference rooms, and the sites differed in lighting, room size, background activity, and furniture layout. No special lighting, markers, or dedicated testing space was provided, so these recordings reflect the conditions AVA would face in a typical independent-living community.

4 Dataset

The dataset combines, for each assessment, the iPad video, the PT measurement, and the MR4 output. It is the basis for every result reported below.

4.1 Demographics

Participants were recruited from independent-living communities and partner church organizations. A total of 66 older adults took part (48 female, 18 male), aged 55 to 98. Table 1 breaks the sample down by site. Some participants used a walking aid during the walking portion of the TUG. No walking aids were used while standing up from the chair or during the chair stand test.

Age differed noticeably between sites. Communities 1 and 3 are independent-living communities and enrolled older participants, while Communities 2 and 4 enrolled younger volunteers. All sites followed the same data-collection protocol.

Table 1: Participant demographics by community.

Community	Participants	Sex (F / M)	Mean Age (years)	SD (years)
1	13	7 / 6	87.6	5.1
2	7	5 / 2	72.3	11.1
3	27	19 / 8	83.3	5.8
4	19	17 / 2	69.9	7.9
Total	**66**	**48 / 18**	**78.3**	**8.3**

4.2 Excluded Recordings

Not every enrolled participant contributes to every analysis. A small number of recordings were excluded because of problems during collection, such as incomplete MR4 calibration, interrupted recordings, or unusable video. These were removed before analysis so that errors in the reference data would not be counted against AVA. Figure 4 shows the participant flow, and Section 6.1 gives the final sample sizes.

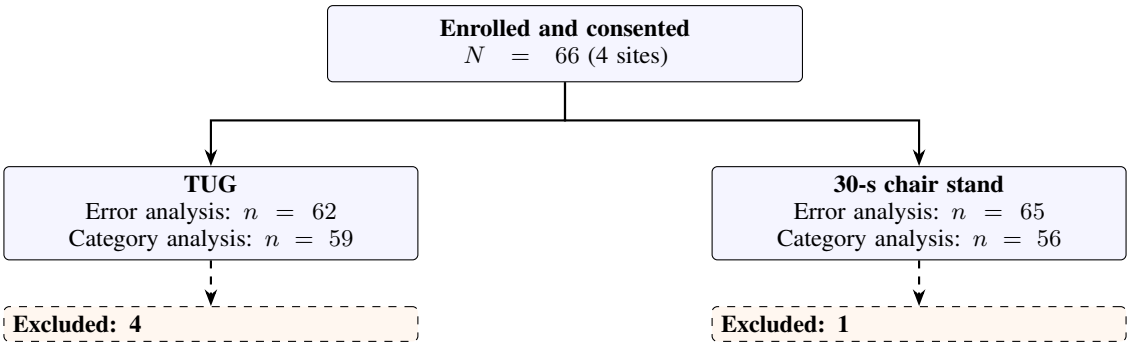


Figure 4: Participant flow from enrollment to analysis. Category analyses include fewer participants than error analyses (see Section 7).

5 Methods

5.1 Outcome Definitions

Both tests are scored from the same two events, which AVA detects in the video: the *sit-to-stand* transition (rising from the chair) and the *stand-to-sit* transition (sitting back down).

30-second chair stand. The participant stands up fully and sits back down as many times as possible in 30 seconds. The score is the number of complete stands. The PTs counted a stand when the participant reached a full standing position, and, following standard practice, they also counted a final stand if the participant was at least halfway up when time expired. AVA counts a repetition only when a complete cycle has been detected, that is, a sit-to-stand transition followed by a stand-to-sit transition, so each repetition is counted when the participant returns to sitting, not at full standing. AVA does not apply a halfway rule: a final stand that is still in progress when the 30-second window closes is not counted. AVA and the PTs therefore used different rules for the last repetition, and this difference can affect the count by at most one stand.

Timed Up and Go (TUG). The participant rises from a chair, walks 3 meters, turns, walks back, and sits down. The score is the total time taken. AVA measures this as the time from the start of the first sit-to-stand transition to the end of the final stand-to-sit transition. Each TUG was timed independently by two PT observers. The reference value for each assessment is the mean of their two timings, and the range between the two PTs shows how much manual timing itself varies. Both AVA and the MR4 are compared against this reference. The start and stop cues given to the two PT observers, and whether they could see AVA or MR4 outputs while timing, were not recorded in a form that can be reported here. Because AVA's timing starts at the onset of the first sit-to-stand transition and ends at the completion of the final stand-to-sit transition, any difference between these events and the PTs' start and stop cues would appear as a systematic offset between AVA and the PT reference.

5.2 AVAL Pipeline

Figure 5 shows how AVAL evaluates AVA Mobility. Each assessment produces three independent results: AVA's result, computed automatically from the iPad video; the reference result recorded by a physical therapist; and the MR4 motion-capture result. AVAL brings these three results together for each participant and compares them.

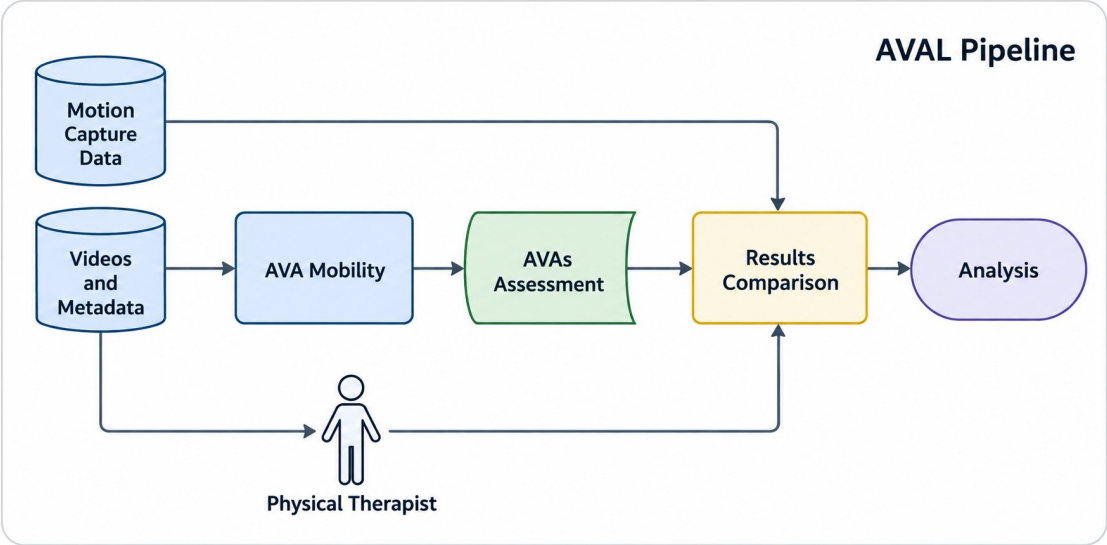


Figure 5: The AVAL pipeline. Each recorded video and its metadata are processed by AVA Mobility to produce AVA's assessment. The same assessment is scored by a physical therapist and recorded by the MR4 motion-capture system. AVAL compares the three results for each participant and summarizes the errors in the analysis stage.

Data sources. AVAL draws on two data stores. The first holds every iPad video together with its metadata, such as participant ID, site, and assessment type. The second holds the MR4 motion-capture results for the same assessments. At the start of each run, AVAL builds a table that links each participant to their video, their PT measurements, and their MR4 results, so records stay correctly matched across all four sites.

AVA Mobility. AVAL runs each video through AVA Mobility frame by frame, exactly as a live iPad camera feed would be delivered. AVA estimates the participant's body landmarks, detects the sit-to-stand and stand-to-sit transitions, and filters the resulting signals over time so that brief tracking errors are not counted as movements.

AVA's assessment. From the detected transitions, AVA computes the test scores:

- the number of completed chair stands within the 30-second window; and
- the TUG duration, from the start of the first sit-to-stand transition to the end of the final stand-to-sit transition.

Reference measurements. In parallel, a physical therapist scores each assessment, giving the reference TUG times and chair-stand counts described in Section 6.1, and the MR4 system provides its own measurements of the same assessment.

Results comparison and analysis. Finally, AVAL pairs each AVA result with the matching manual measurement and MR4 result, computes the errors, and summarizes them by site and overall as tables and plots. These summaries are reported in Section 7.

5.3 Error Metrics

Let y_i be the reference (manual) measurement for participant i, $\hat{y}_i$ the value predicted by AVA or the MR4, and N the number of participants. We evaluate TUG timing with four complementary metrics.

Mean absolute percentage error (MAPE). The absolute percentage error expresses each error relative to the participant's own TUG time, so a 1-second error counts more for a fast 8-second walk than for a slow 20-second one:

$$e_i = \left| \frac{\hat{y}_i - y_i}{y_i} \right|, \qquad \text{MAPE} = \frac{100\%}{N} \sum_{i=1}^{N} e_i. \tag{1}$$

A MAPE of 6% means that predictions differ from the PT reference by 6% of the participant's TUG time on average. MAPE describes timing error only; agreement on mobility categories is reported separately.

Mean absolute error (MAE). The average size of the error in seconds, which is the most direct measure of the typical error:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{y}_i - y_i \right|. \tag{2}$$

Root mean squared error (RMSE). Squaring the errors before averaging penalizes large errors more heavily than MAE does, so RMSE is sensitive to the occasional large miss that could affect a clinical decision:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i \right)^2}. \tag{3}$$

Mean signed error (bias). The average of $\hat{y}_i - y_i$, which shows whether a system systematically over- or underestimates TUG time. A positive value means the system reports longer times than the PTs.

T5 error. The 95th percentile of the absolute errors. Ninety-five percent of predictions have an error at or below this value, so it describes near-worst-case performance without being driven by a single outlier.

Chair stand. Because the chair stand produces a small whole-number count, we report the number and percentage of participants whose predicted count matches the reference exactly, and the number and percentage within ± 1 repetition, pooled across all participants.

Mobility categories. Agreement between each system's mobility category and the PT category is reported as the percentage of participants placed in the same category and as Cohen's κ [6], which corrects for agreement expected by chance.

6 Validation Protocol

AVAL applied the same evaluation procedure at every site, so participant inclusion, alignment with the reference measurements, and outcome reconstruction were handled identically throughout.

6.1 Evaluation Scope

The analysis included assessments with complete MR4 calibration, usable video, a reliable PT measurement, and a successful AVA run. This left 62 of 66 TUG assessments and 65 of 66 chair-stand assessments; all exclusions were due to calibration problems or missing data, as described in Section 4.2. Because every included assessment has valid AVA, MR4, and PT results, AVA and the MR4 are compared on exactly the same participants.

TUG reference times are the mean of the two PT observers' timings, which reduces the effect of either PT's timing alone. Chair-stand reference counts were checked manually.

6.2 Site Comparison

Each site was first evaluated on its own, and the results were then pooled. AVA was not tuned, retrained, or given different thresholds for any site. The results therefore reflect how well a single fixed model generalizes to new environments.

6.3 Tolerance Bands

The TUG scatter plots include $\pm 5\%$ and $\pm 10\%$ bands around the reference value as visual reference bands. These are not established clinical thresholds: they show the size of each error relative to the participant's own TUG time, and readers should not interpret them as limits of clinically acceptable error. For the chair stand, the ± 1 repetition tolerance is included only to show how close the single non-exact prediction was.

6.4 Reproducibility

All assessments were processed with the AVAL pipeline, which links each participant to their video, MR4 results, and manual measurements, then computes errors in the same way for every site.

7 Results

This section reports AVA's accuracy on the TUG and chair stand, pooled across all sites and site by site, and compares it with the MR4 motion-capture system. Both AVA and the MR4 are measured against the same manual reference.

7.1 TUG Error and Accuracy

Table 2 summarizes AVA's TUG results. Pooled across all sites, AVA's MAPE relative to PT timing was 6.0%, meaning its timings differed from the PT reference by about 6% of the TUG time on average. The pooled mean absolute error was 0.71 s, so a typical AVA timing was well under one second from the reference. The RMSE of 0.85 s is only slightly higher than the MAE, which indicates that errors were consistently small and that few large misses were present. The T5 error of 1.60 s shows that 95% of AVA timings fell within about 1.6 s of the reference.

Table 2: AVA Timed Up and Go timing error relative to PT timing, pooled across all sites.

Participants	MAPE (%)	MAE (s)	RMSE (s)	T5 error (s)	Mean signed error (s)
62	6.0	0.709	0.854	1.595	$+0.17$

Figure 6 plots each system's TUG time against the manual reference, with the range between the two PTs shown for each participant. Most AVA timings fall inside the $\pm 10\%$ band. The MR4, by contrast, consistently reports longer times than the PTs, typically by about 1 to 3 seconds.

Table 3 compares the two systems directly on the same participants. AVA's mean signed error was $+0.17$ s, close to zero, whereas the MR4 overestimated TUG time by $+1.59$ s on average. On mobility categories, AVA agreed with the PTs for 88.1% of TUG results ($\kappa = 0.79$) and 100% of chair-stand results ($\kappa = 1.00$), compared with 71.2% ($\kappa = 0.55$) and 87.5% ($\kappa = 0.78$) for the MR4.

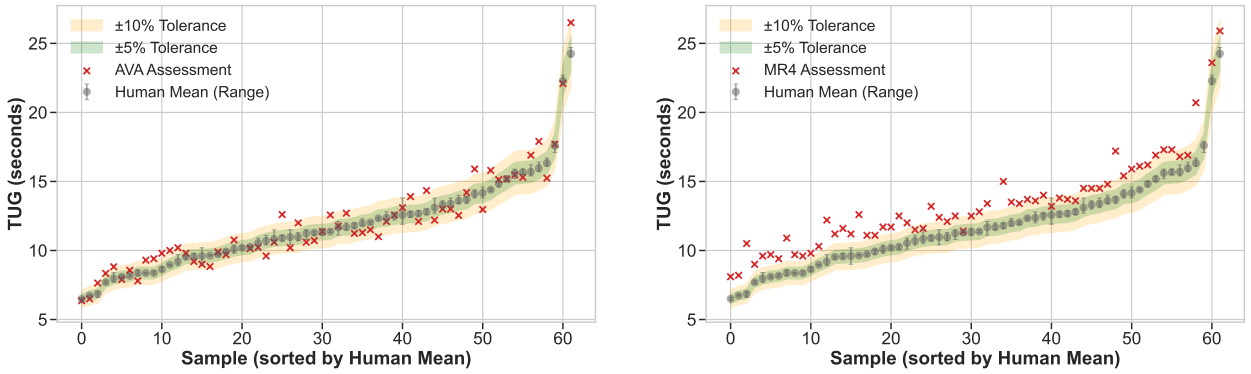
Table 3: Agreement with PT measurements for AVA and the MR4 on the same participants.

Test	Metric	AVA	MR4
TUG	Mean signed error (s)	$+0.17$	$+1.59$
	Category agreement, $n = 59$	52/59 (88.1%)	42/59 (71.2%)
	Cohen's κ	0.79	0.55
Chair stand	Category agreement, $n = 56$	56/56 (100%)	49/56 (87.5%)
	Cohen's κ	1.00	0.78

Figure 7 shows the distribution of signed TUG errors (system time minus PT mean time) for each system. AVA's errors are centered close to zero (mean $+0.17$ s), with both over- and underestimates. Nearly all of the MR4's errors are positive (mean $+1.59$ s), which reflects the systematic overestimate seen in the scatter plots.

Figure 8 breaks the errors down by site. AVA's errors are centered near zero at every site, and their spread is similar across sites. The MR4's errors are positive at every site, with medians between about 1.2 and 1.6 seconds.

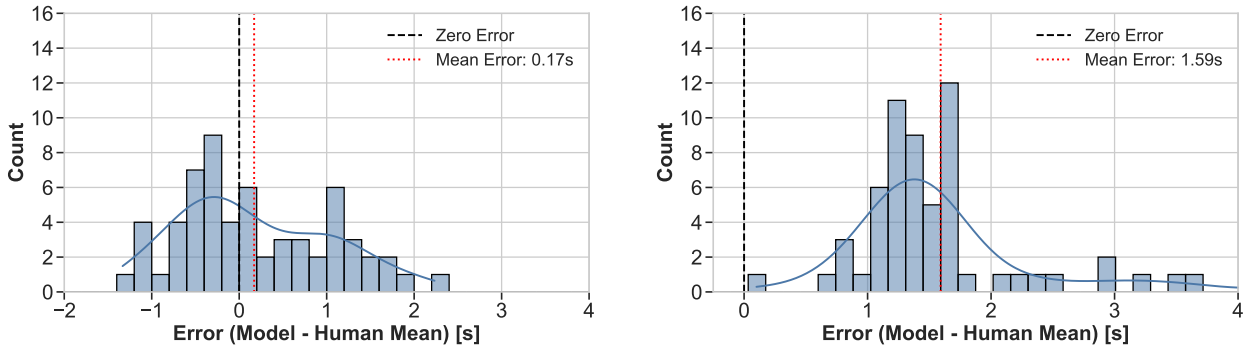
Taken together, these results show that AVA produces accurate and consistent TUG timings across a wide range of participant abilities and testing environments.



(a) AVA vs. manual timing

(b) MR4 vs. manual timing

Figure 6: Predicted versus manually measured TUG times for (a) AVA and (b) the MR4 motion-capture system. Tolerance bands mark $\pm 5\%$ and $\pm 10\%$ of the reference value, and error bars show the range across observers.



(a) AVA

(b) MR4

Figure 7: Distribution of signed TUG errors (system minus PT mean). Negative values are underestimates and positive values overestimates; the dashed line marks zero and the dotted line the mean.

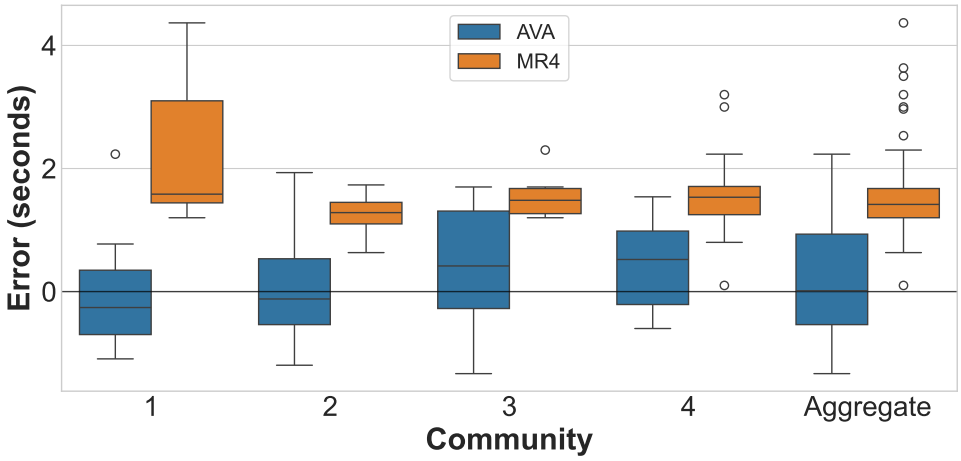


Figure 8: Signed TUG errors (system minus PT mean) by community for AVA and the MR4.

7.2 Chair Stand Error and Accuracy

AVA performed equally well on the 30-second chair stand. Pooled across all 65 analyzed participants, AVA's count matched the PT count exactly for 64 (98.5%), and all 65 (100%) were within one repetition. Every site therefore had 100% agreement within one repetition, and the single exact-count mismatch occurred at one site.

The one mismatched prediction was off by a single repetition. AVA therefore identifies sit-to-stand cycles reliably despite differences in environment, recording conditions, and participant mobility. These findings complement the TUG results and support AVA's use for mobility screening in community settings.

7.3 Agreement on Mobility Categories

In practice, a clinician usually acts on a mobility *category* rather than a raw time or count. We therefore also tested whether AVA assigns participants to the same category as the manual measurements. We use the CDC's STEADI (Stopping Elderly Accidents, Deaths and Injuries) framework [2], which provides age- and sex-specific normative values for the chair stand and standard timing cut-offs for the TUG. From these we define three categories: low, fair, and high mobility.

For the chair stand, the expected number of repetitions falls with age and differs between men and women. Figure 9 lists the STEADI cut-offs, and Figure 10 shows the same cut-offs graphically so the shift across age groups is easier to see.

Age	Men	Women
64 or less	14	12
65 - 69	12	11
70 - 74	12	10
75 - 79	11	10
80 - 84	10	9
85 - 89	8	8
90+	7	4

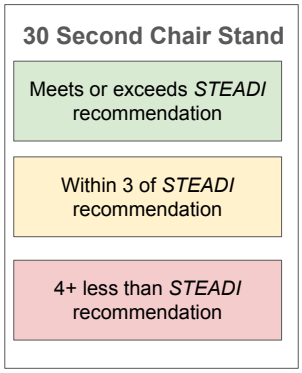
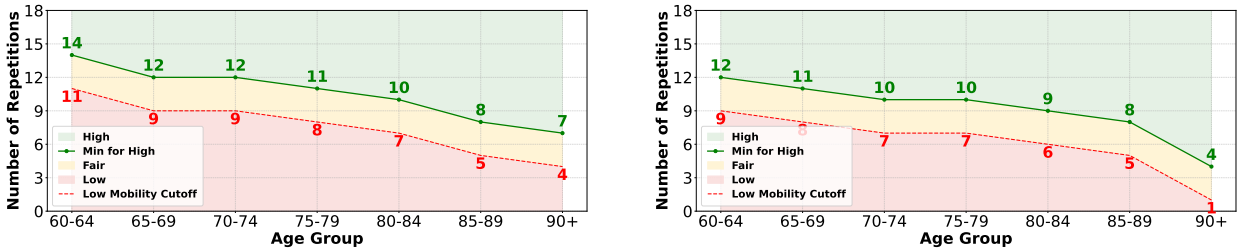


Figure 9: CDC STEADI normative values for the 30-second chair stand.



(a) Men (b) Women

Figure 10: STEADI chair stand mobility categories by age group.

For the TUG, the categories are based on completion time alone (Table 4). A time of 11 seconds or less indicates high mobility, a time above 11 but below 15 seconds indicates fair mobility, and a time of 15 seconds or more indicates low mobility. These categories were defined for this study; the 11-second boundary is lower than the 12-second fall-risk cut-off commonly cited from STEADI [2].

Figure 11 shows every participant's TUG result as measured by the PTs, AVA, and the MR4. For display, each TUG time is converted to a *task-speed index*: the 6-meter course distance divided by the total TUG time. This index is not gait speed, because the total TUG time also includes rising from the chair, turning, and sitting down; it is a monotonic transformation of TUG time, so it preserves the ordering of participants and the category boundaries (6 m divided by 11 s and 15 s gives 0.55 and 0.40 m/s). The values are overlaid on the three category bands. AVA's values follow the

Table 4: Timed Up and Go mobility categories used in this report. The intervals are continuous, so every TUG time, including fractional values such as 11.5 s or 14.5 s, falls in exactly one category. The same boundaries define the shaded bands in Figure 11.

TUG time	Mobility category
≤ 11 s	High
> 11 s to < 15 s	Fair
≥ 15 s	Low

manual values closely. The MR4's values sit consistently lower, because its longer times convert to slower speeds, and this pushes some participants into a lower category.

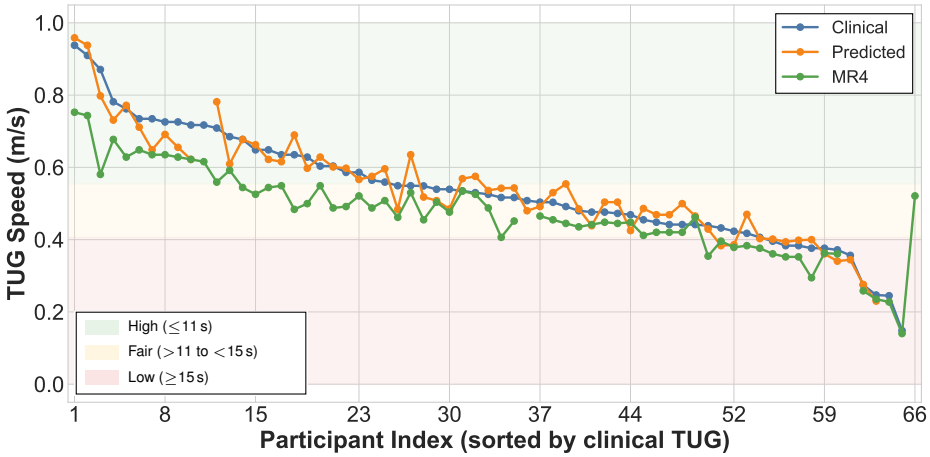


Figure 11: TUG results for each participant, shown as a task-speed index (6 m divided by total TUG time, which includes rising, turning, and sitting; not gait speed) and overlaid on the mobility-category bands. Participants are sorted by PT-measured TUG time.

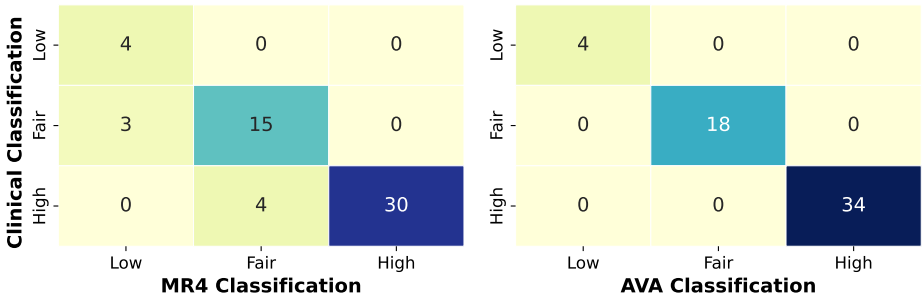
Figure 12 shows confusion matrices comparing each system's category against the category from the manual measurement. Counts on the diagonal are correct classifications, and counts off the diagonal are misclassifications.

On the chair stand (Figure 12a), AVA placed every participant in the correct category, including the one participant whose count was off by a repetition. The MR4 misclassified seven participants. On the TUG (Figure 12b), AVA misclassified 7 participants and the MR4 misclassified 17. Most of the MR4's errors placed participants one category lower than the manual result, which is consistent with its tendency to overestimate TUG time. The "fair" category, which spans only a few seconds, was where both systems made most of their errors, since a small timing difference near a boundary can change the category. Overall, AVA agreed with manual categorization more closely than the MR4 on both tests.

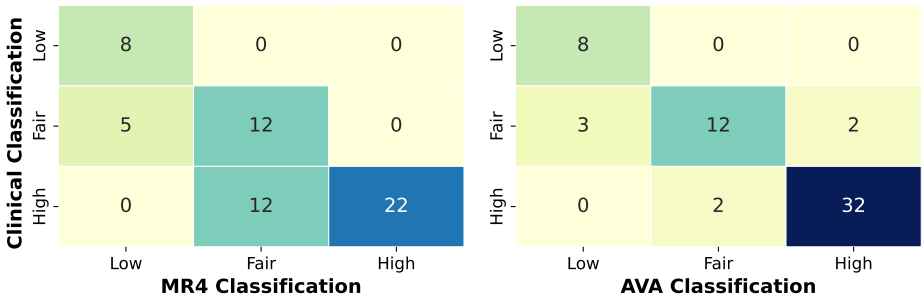
8 Limitations

AVA performed reliably across all four sites, but several practical factors can affect measurement quality. Very dim or very bright lighting can reduce pose-estimation accuracy, and people or objects passing between the camera and the participant can briefly block the view. AVA does not require technical expertise, but the iPad still needs to be positioned correctly, because a camera that is too low or too close can cut off parts of the movement. Clothing that blends into the background, such as dark trousers against a dark floor, can also make body landmarks harder to detect.

The study itself also has limitations. Some recordings were excluded because of calibration failures or missing data in the reference measurements. The start and stop cues of the two PT observers and their blinding to AVA and MR4 outputs were not documented, and AVA and the PTs applied different rules to the final chair-stand repetition. The sample of 66 participants is moderate in size, weighted toward women (48 of 66), and drawn from four sites in one region. The balance test was recorded but not analyzed. Future work will extend AVAL to additional mobility tasks,



(a) 30-second chair stand



(b) Timed Up and Go

Figure 12: Mobility-category agreement between each automated system (MR4 left, AVA right) and the manual measurements. Rows show the category from manual measurement, and columns show the category each system assigned.

including the balance test, add automated quality-control checks, and validate AVA on participants with a wider range of functional impairments and in a wider range of environments.

9 Conclusion

This report introduced **AVAL**, a standardized pipeline for validating AVA Mobility in real community settings. We used it to compare AVA against physical-therapist measurement and the MR4 motion-capture system on the TUG and 30-second chair stand at four community sites. On the TUG, AVA's timings had a mean absolute percentage error of 6.0% relative to PT timing across 62 valid assessments, with a mean absolute error of 0.71 s and a mean signed error of only $+0.17$ s. Its errors were concentrated near zero, and performance was consistent across sites despite differences in lighting, room layout, and background activity. On the chair stand, AVA's count matched the manual count exactly for all but one of 65 participants, and that one prediction was off by a single repetition. AVA also agreed more closely with manual measurement than the MR4 did. The MR4 consistently overestimated TUG time by about 1 to 3 seconds, which pushed many participants into a lower mobility category. AVA showed no such offset: it placed every participant in the correct chair-stand category and misclassified far fewer TUG results. Because the mobility category drives the clinical decision in fall-risk screening, this difference matters more than small differences in average error.

These results carry practical weight. The MR4 requires sixteen body-worn sensors, a calibration routine, and a trained operator, amounting to roughly 30 to 40 minutes of setup per participant. AVA runs the same assessments with a single iPad in under five minutes, with no wearable hardware and no data leaving the device. Community staff can therefore run standardized fall-risk tests themselves. This makes it realistic to screen residents regularly rather than occasionally, to detect declines in mobility earlier, and to remove the observer-to-observer variation that makes results hard to compare across sites or over time. Future work will extend validation to additional STEADI assessments, particularly the 4-stage balance test, and to larger and more diverse populations. It will also include a detailed analysis of how AVA's

errors are distributed and when they change a clinical decision. Longer-term studies should examine whether repeated AVA screening predicts falls and whether acting on its results reduces them. Overall, AVAL shows that AVA Mobility delivers clinically aligned accuracy while greatly reducing assessment time, equipment, and staffing requirements. These findings support AVA as a practical, privacy-conscious, and scalable tool for routine, community-wide mobility screening.

References

[1] CDC. Important facts about falls. `https://www.cdc.gov/falls/data-research/index.html`, 2023. Accessed: 2025-11-14.

[2] Centers for Disease Control and Prevention. Steadi: Stopping elderly accidents, deaths and injuries – chair stand test normative data, 2023. URL `https://www.cdc.gov/steadi/index.html`. Accessed: 2025-01-15.

[3] World Medical Association. World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–2194, 2013.

[4] Patrick M. Bossuyt, Johannes B. Reitsma, David E. Bruns, Constantine A. Gatsonis, Paul P. Glasziou, Les Irwig, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, 351:h5527, 2015.

[5] Jan Kottner, Laurent Audigé, Stig Brorson, Allan Donner, Byron J. Gajewski, Asbjørn Hróbjartsson, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1): 96–106, 2011.

[6] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46, 1960.