ForesightCares

AVAL.v2: MULTI-ASSESSMENT VALIDATION AND ERROR ANALYSIS OF ON-DEVICE MOBILITY AI FOR FALL-RISK SCREENING

TECHNICAL REPORT

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

ABSTRACT

Vision-based Artificial Intelligence (AI) can automate standardized fall-risk assessments. Before such systems are used for clinical decisions, however, they must be validated with the rigor expected of any measurement instrument. That validation needs to cover more than average error: it must show how errors are distributed, where they change clinical decisions, and which parts of the model cause them. We present **AVAL.v2**, an extended validation framework for the AVA Mobility on-device AI pipeline. It evaluates three assessments from the Centers for Disease Control and Prevention (CDC) Stopping Elderly Accidents, Deaths, and Injuries (STEADI) toolkit: the Timed Up and Go (TUG), the 30-second chair stand, and, new in this version, the 4-stage balance test.

AVAL.v2 evaluates model output at three levels. The first is continuous measurement agreement, using error metrics, Bland–Altman limits of agreement, concordance and intraclass correlation, and bootstrap confidence intervals. The second is tolerance coverage, the share of participants whose error falls below each threshold. The third is agreement on clinical risk zones, with each error classified by its direction (over-severity or under-severity). The framework also supports component-level ablation of pipeline settings.

We evaluated 80 older adults (72.5% female, aged 54–98), enrolled between February and September 2026, against physical-therapist (PT) scoring. On the TUG ($n = 71$), AVA achieved a mean absolute error of 0.386 s (95% CI 0.314–0.462), a mean relative error of 3.5%, a concordance correlation of 0.989, and 95% limits of agreement of -1.11 to $+0.67$ s, with no evidence of proportional bias (slope -0.001, 95% CI -0.033 to 0.031). On the chair stand ($n = 79$), AVA's count matched the PT's exactly for 76 participants and was one stand lower for the other three, and 78 of 79 participants were placed in the correct risk zone ($\kappa = 0.98$). On the 4-stage balance test, pass/fail agreement was 100% for side-by-side stance, 98.6% for semi-tandem, 90.9% for full tandem, and 92.0% for single-leg stance. Most balance errors were early terminations of the hold, which make the result more severe rather than less. An ablation of the single-leg stance-check configuration showed a trade-off between exact-duration accuracy and pass/fail agreement. All seven TUG zone disagreements occurred within 0.6 s of a zone boundary.

These results extend our earlier validation to a larger cohort and a new assessment. They also identify specific, measurable targets for improving the model.

Keywords AI validation · Error analysis · Pose estimation · Fall risk · Timed Up and Go · Chair stand · 4-stage balance · Method agreement

1 Introduction

Falls are the leading cause of injury-related hospitalization and loss of independence among older adults [1]. The U.S. Centers for Disease Control and Prevention (CDC) Stopping Elderly Accidents, Deaths, and Injuries (STEADI) initiative recommends three brief functional tests to identify people at elevated risk [2]. The Timed Up and Go (TUG) test measures functional mobility [3]. The 30-second chair stand measures the lower-body strength [4]. The 4-stage balance test measures static balance [5]. Each test is quick and inexpensive, but each still depends on a trained observer, which limits how often and how widely it can be given.

Camera-based movement analysis has advanced rapidly. Deep-learning pose estimators can track body landmarks from ordinary video [6], including in real time on mobile devices [7]. Single-camera video has been used to estimate gait parameters and clinical movement measures [8, 9], smartphone video can now support full musculoskeletal analysis [10], and earlier work established depth cameras for assessing physical function [11]. Most of these systems, however, were evaluated in laboratory settings, with trained operators, or in healthy adults, and fewer have been validated against clinician scoring of standardized fall-risk tests in older adults in everyday settings.

AVA Mobility automates these tests using a single iPad camera and on-device Artificial Intelligence (AI). In our first technical report (AVAL [12]), we introduced a validation pipeline and showed that AVA's TUG timings and chair-stand counts were closely in agreement with manual measurement at four community sites. That report answered the question *"Is AVA accurate?"* This report addresses a harder set of questions that must be answered before an AI system is trusted with clinical decisions:

- **How are errors distributed?** Is the error unbiased, constant across the measurement range, and free of heavy tails?
- **When do errors matter?** How often does a measurement error change a participant's clinical risk zone, and in which direction?
- **Where do errors come from?** Which components or settings of the pipeline produce the remaining errors, and how does changing them affect accuracy?
- **Does performance transfer to a new task?** Does the approach generalize to static balance, where the model must detect a subtle loss of posture rather than a large, obvious movement?

Figure 1 illustrates why the second question cannot be answered from average error alone. The three examples are hypothetical TUG results. Participant A has a 0.8 s error but remains well inside the green zone, so the error has no clinical consequence. Participant C has an even larger error, 1.0 s, but both values lie deep in the red zone, so the risk assessment is unchanged. Participant B has the *smallest* error of the three, 0.5 s, yet it is the only one that changes the result: the physical-therapist (PT) time is just above the green/yellow cut-off and the AVA time is just below it, so a participant who should be flagged for intermediate risk is reported as lower risk. The clinical impact of an error therefore depends on its magnitude, its direction, and how close the true value lies to a decision boundary. Two systems with identical mean absolute error (MAE) can have very different clinical consequences. This observation motivates the decision-level analysis and error taxonomy in Section 4, and the real cohort shows the same pattern. The largest TUG error in this study (1.64 s, participant P058) caused no zone change, whereas a 0.57 s error for participant P008, whose PT time was only 0.04 s above the cut-off, did (Section 5.2).

Contributions. This report makes the following contributions.

1. **AVAL.v2 framework.** We restructure AVAL into a modular system with a video feed service, an AI pipeline, a denoising stage, and assessment-specific engines, so each component can be evaluated and ablated independently (Section 2).
2. **A three-level evaluation methodology.** The methodology combines measurement agreement, tolerance coverage, and clinical-decision agreement with an error taxonomy based on direction. We specify it in enough detail to be reproduced (Section 4).
3. **A larger cohort and a new assessment.** We evaluate 80 older adults and add the 4-stage balance test, with separate evaluation of each stage (Sections 3 and 5.4).
4. **Extensive results.** We report complete results, including confidence intervals, stratified analyses, per-case failure analysis, a configuration ablation, and a cross-assessment analysis of how risk zones relate across tests (Section 5).

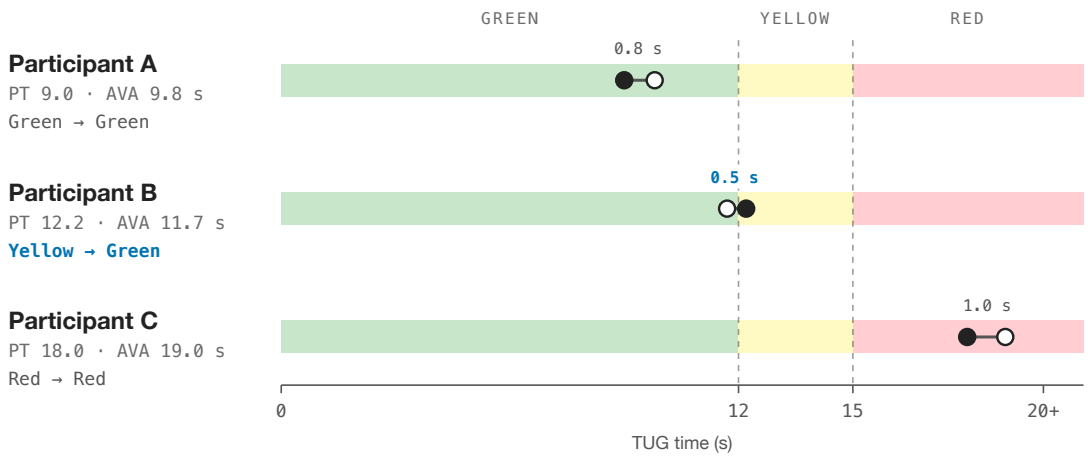


Figure 1: Error magnitude is not the same as clinical impact. Three hypothetical TUG results show that a small error near a zone boundary (Participant B) can change the risk zone, while larger errors far from a boundary (Participants A and C) do not. Each bar shows the green, yellow, and red TUG zones; filled and open markers show the PT and AVA times. Dashed lines mark the 12 s and 15 s cut-offs, and each label gives the error.

5. **Improvement targets.** We translate the error analysis into concrete targets for improving the model (Section 7).

Table 1 summarizes how this report differs from the first.

Table 1: Scope of the first AVAL report compared with AVAL.v2. MR4: wearable motion-capture suit; MAPE: mean absolute percentage error; RMSE: root mean squared error; T5: 95th-percentile absolute error; CCC: Lin's concordance correlation coefficient; ICC: intraclass correlation coefficient; CI: confidence interval.

	AVAL (v1) [12]	**AVAL.v2 (this report)**
Primary question	Is AVA accurate in the community?	How do AVA's errors behave, when do they affect decisions, and how can the model improve?
Participants	66	80
Assessments	TUG, chair stand	TUG, chair stand, 4-stage balance
Reference	Manual timing; MR4 motion capture	PT scoring
Agreement statistics	MAPE, MAE, RMSE, T5	Adds bias, limits of agreement, CCC, ICC, bootstrap CIs, coverage curves, κ
Decision analysis	Category confusion matrices	Adds over- and under-severity taxonomy, boundary analysis, and per-stage pass/fail
Model analysis	None	Configuration ablation, phase segmentation, failure-case review

2 The AVAL.v2 Framework

2.1 System Architecture

Figure 2 shows the AVAL.v2 system flow. The architecture is designed so that the offline validation path uses the same code as the deployed application. Every component that runs on the iPad in production also runs, unchanged, during validation.

Data source and video feed service. All assessment videos are held in a central dataset. The video feed service replays each recording frame by frame at its native frame rate, exactly as a live iPad camera would deliver it. This lets the pipeline process stored video under the same timing conditions as on the device.

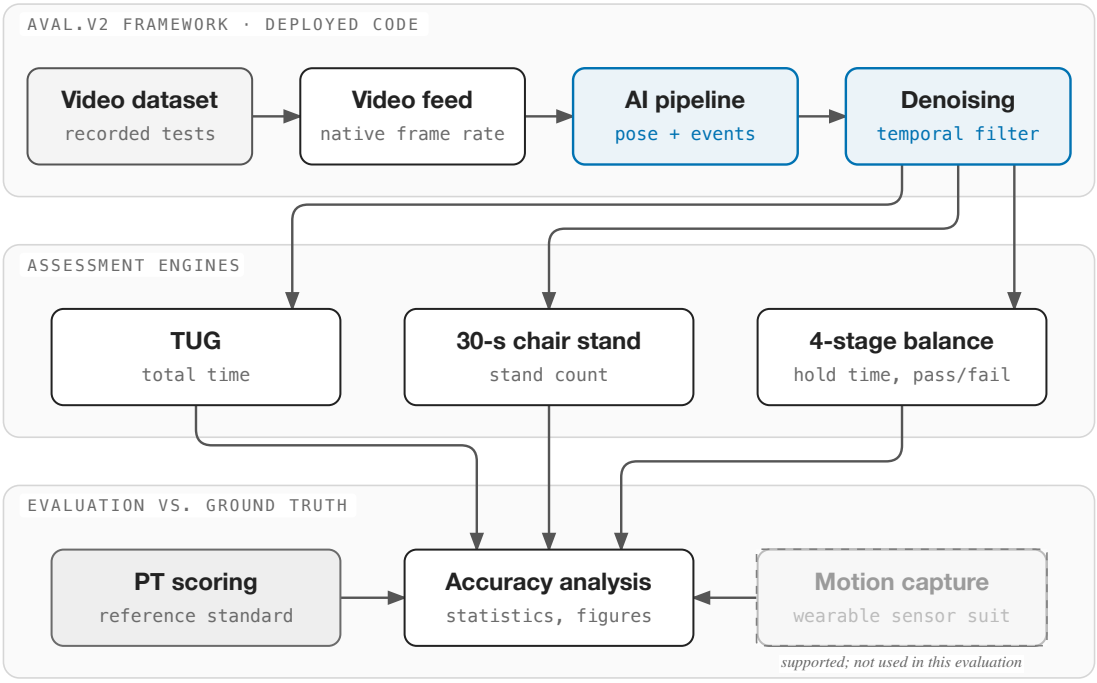


Figure 2: AVAL.v2 system flow. Recorded assessment videos are streamed through the same AI pipeline used in the deployed app, then denoised and scored by one engine per assessment. The outputs are compared against ground-truth sources in the evaluation stage. Blue marks the AI pipeline and denoising stages, which this report evaluates and ablates. Motion capture (dashed) is a supported reference source that was not used in this evaluation.

AI pipeline. For each frame, a *person detector* locates the participant and a *feature extractor* estimates 33 body landmarks. From these landmarks AVA derives biomechanical features such as joint angles, trunk orientation, and foot position. AVA's event model then converts the feature time series into movement events and stance states.

Denoising and filtering. Frame-level landmark estimates from a single camera are noisy. Before any event is accepted, a temporal filtering stage smooths the signals and suppresses brief spurious changes. The configuration of this stage, including the temporal window used to confirm a stance, directly affects accuracy. Section 5.5 evaluates it through ablation.

Assessment engines. Each test has its own engine, which turns the filtered signals into clinical outputs:

- **TUG engine:** total test time plus phase timing (rise, walk out, turn, walk back, turn, sit) and step count.
- **Chair stand engine:** number of completed stands in 30 seconds and duration of each cycle.
- **4-stage balance engine:** hold duration for each stance (capped at 10 s), a pass/fail result for each stance, and postural sway.

This report evaluates the primary clinical outputs: TUG total time, chair-stand count, and balance hold duration with its pass/fail result. Secondary outputs such as step count, cycle duration, and sway area are produced by the engines but not yet validated. They are left for future work.

Evaluation. Engine outputs are compared with ground truth in the accuracy-analysis module, which produces all the statistics and figures in this report. The framework supports several ground-truth sources, including PT scoring and wearable motion capture. This report uses PT scoring as the reference standard for every assessment.

2.2 Event Detection Example: TUG Phase Segmentation

Figure 3 shows how the TUG engine works for one participant (P021). The engine tracks knee angle over time and uses two thresholds: 132° to decide the participant is standing and 112° to decide they are sitting. Using separate thresholds

for the two directions (hysteresis) stops small fluctuations near a single cut-off from registering as repeated transitions. The test starts at the sit-to-stand transition ($t = 0$). It then passes through walk-out, first-turn, walk-back, second-turn, and sitting phases. The stand-to-sit transition is detected at 9.6 s, and the participant is judged fully settled at 10.4 s. For this participant AVA reported 10.37 s, compared with a PT time of 10.70 s.

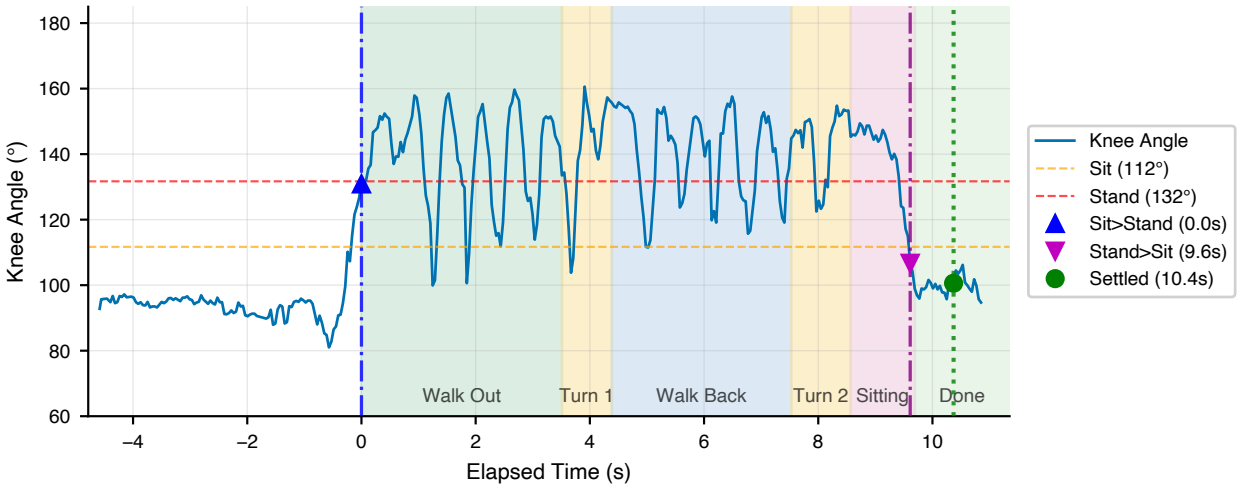


Figure 3: TUG phase segmentation for participant P021. Knee angle (blue) is compared with the stand (132°) and sit (112°) thresholds. Shaded regions mark the automatically detected phases.

3 Study Design and Dataset

3.1 Study Design, Period, and Ethics

AVAL.v2 was a prospective, cross-sectional method-comparison study in which each participant was measured once by AVA and by the reference standard. Data were collected between **February 2026 and September 2026**. The study was approved by the Advarra Institutional Review Board (protocol Pro00085739) and conducted in accordance with the Declaration of Helsinki [13]. All participants gave informed consent before taking part, including consent to video recording. Reporting follows the Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015 guidelines [14] and the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [15], where applicable.

3.2 Setting

Participants were recruited from independent-living communities, senior centers, and community activity groups in the Piedmont region of North Carolina.

3.3 Participants

Eighty older adults took part (58 female, 72.5%; 22 male, 27.5%). Their mean age was 78.5 years (standard deviation [SD] 9.9; median 78.5, interquartile range [IQR] 72–87; range 54–98). Table 2 reports age at enrollment by sex, and Figure 4 shows the age and sex distributions.

3.4 Protocol and Reference Standard

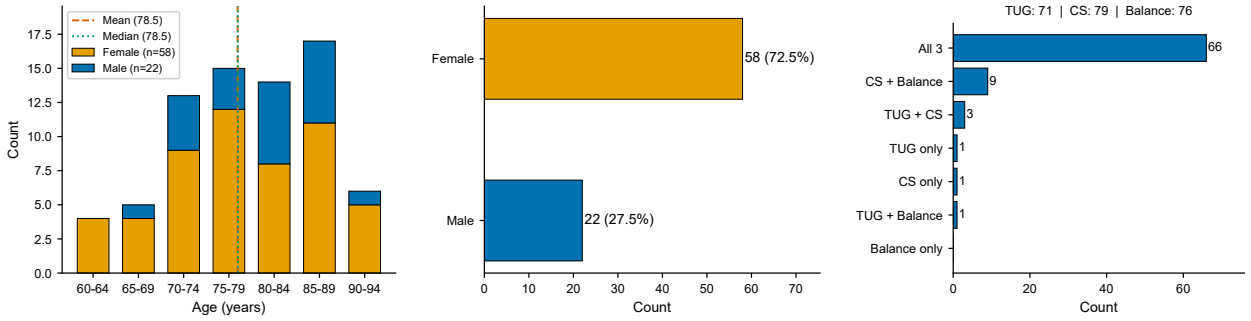
Each participant was asked to complete the TUG, the 30-second chair stand, and the 4-stage balance test, following STEADI administration guidelines [2]. Each assessment was recorded by a single iPad positioned in front of the participant. A physical therapist (PT) scored every assessment: TUG time in seconds, number of chair stands, and hold duration for each balance stance in whole seconds, capped at 10 s. The PT scores are the reference standard for all analyses. PT scoring was performed live with a stopwatch.

Not every participant completed or had usable data for every test. Table 3 gives the final sample sizes, and Figure 4c shows how they overlap: 66 participants have all three assessments.

Figure 5 shows the flow of participants from enrollment to each analysis, following STARD [14].

Table 2: Age at enrollment by sex. Counts for 60–94 years are taken from the cohort age histogram; the 6 participants outside those bins are shown together: four were aged 54–59 and two were aged 95 or older.

Age group (years)	**Female,** n	**Male,** n	**Total,** n **(%)**
<60 or ≥ 95	5	1	6 (7.5)
60–64	4	0	4 (5.0)
65–69	4	1	5 (6.3)
70–74	9	4	13 (16.3)
75–79	12	3	15 (18.8)
80–84	8	6	14 (17.5)
85–89	11	6	17 (21.3)
90–94	5	1	6 (7.5)
Total	**58**	**22**	**80 (100)**



(a) Age distribution by sex (b) Sex (c) Assessments completed per participant

Figure 4: Cohort characteristics. The age histogram shows the 74 participants aged 60–94; the mean and median in its legend are for all 80 participants.

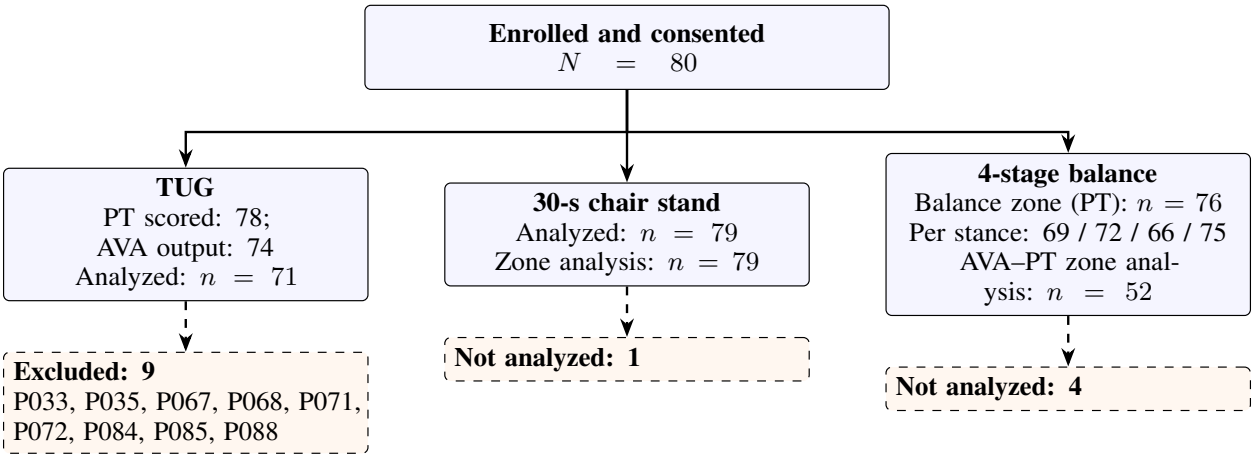


Figure 5: Participant flow from enrollment to analysis for each assessment. Per-stance balance counts are side-by-side / semi-tandem / full tandem / single-leg.

3.5 Sample Size and Statistical Precision

With the achieved sample, the 95% confidence interval (CI) for TUG MAE has a half-width of about 0.07 s. The Wilson 95% confidence interval [16] for TUG zone agreement is 81.0–95.1%, and for chair-stand zone agreement 93.2–99.8%.

3.6 Clinical Risk Zones

Each assessment result is mapped to one of three study-defined categories, informed by the CDC STEADI materials [2]: green (lower risk), yellow (intermediate), and red (higher risk). STEADI provides cut-offs and normative values for the individual tests but does not define this three-color scheme, and we are not aware of a validation of these exact categories against fall outcomes. The categories are therefore used here only to compare AVA's and the PT's classifications on a common scale, not to predict falls. The zone definitions are:

- **TUG:** green < 12 s; yellow 12 to < 15 s; red ≥ 15 s.
- **Chair stand:** the count is first normalized by subtracting the CDC age- and sex-specific norm. Green: at or above the norm (≥ 0); yellow: 1 to 3 stands below it; red: more than 3 stands below it.
- **4-stage balance:** a stance is passed if it is held for the full 10 s. Green: all four stances passed; yellow: full tandem passed but single-leg failed; red: full tandem not passed.

PT chair-stand scoring followed the STEADI administration rules: the participant sat with arms crossed over the chest; a stand was counted only when the participant reached a full standing position; if the participant was more than halfway up when 30 s elapsed, the final stand was counted; and a participant who needed their arms to stand was stopped and scored as zero. The CDC normative table covers ages 60 to 94. Six participants fell outside this range (four aged 54–59 and two aged 95 or older), and the norm assignment for these participants was not documented for this report. The two analysis outputs used here did not apply the norms identically (Section 5.3), so chair-stand categories for participants outside 60–94 years should be interpreted with caution.

Table 3: Cohort characteristics and PT-scored functional performance.

Characteristic	Value
Demographics ($N = 80$)	
Female, n (%)	58 (72.5)
Male, n (%)	22 (27.5)
Age, mean (SD), years	78.5 (9.9)
Age, median (IQR), years	78.5 (72–87)
Age range, years	54–98
TUG ($n = 71$)	
Time, mean (SD), s	11.3 (3.4)
Time, median (IQR), s	11.3 (9.2–12.7)
Green / Yellow / Red, n (%)	46 (64.8) / 16 (22.5) / 9 (12.7)
30-second chair stand ($n = 79$)	
Count, mean (SD)	9.6 (3.3)
Count, median (IQR)	9 (8–12)
Count relative to CDC norm, mean (SD)	0.2 (3.1)
Green / Yellow / Red, n (%)	42 (53.2) / 30 (38.0) / 7 (8.9)
4-stage balance ($n = 76$)	
Stages passed, mean (SD)	3.2 (0.7)
Stages passed, median (IQR)	3 (3–4)
Green / Yellow / Red, n (%)	28 (36.8) / 34 (44.7) / 14 (18.4)

4 Evaluation Methodology

AVAL.v2 evaluates the model at three levels, each answering a different question. Using all three together avoids a common weakness of AI validation, where a good average error hides clinically important failures.

4.1 Level 1: Measurement Agreement

Let y_i be the PT reference value for participant i, $\hat{y}_i$ the AVA prediction, $e_i = \hat{y}_i - y_i$ the signed error, and n the number of participants.

Error magnitude. We report the mean absolute error, MAE $= \frac{1}{n}\sum |e_i|$, which gives the typical size of an error in the original units. We also report the median absolute error, which is not affected by outliers, and the root mean squared error, RMSE $= \sqrt{\frac{1}{n}\sum e_i^2}$, which weights large errors more heavily. The mean relative error, MRE $= \frac{100\%}{n}\sum |e_i|/y_i$, expresses error as a percentage of the participant's own result.

Bias and limits of agreement. The mean signed error $\bar{e}$ measures systematic bias: a positive value means AVA overestimates. Following Bland and Altman [17], the 95% limits of agreement (LoA) are $\bar{e} \pm 1.96\, s_e$, where s_e is the sample standard deviation of the errors. The LoA give the range expected to contain 95% of individual differences between AVA and the PT. We test for *proportional bias*, meaning error that grows or shrinks with the size of the measurement, by regressing e_i on the mean $(\hat{y}_i + y_i)/2$.

Association and concordance. Pearson's r measures linear association, but it can be high even when the two methods disagree by a constant offset. We therefore also report Lin's concordance correlation coefficient (CCC) [18], which penalizes departure from the line of perfect agreement. We also report the two-way random-effects, absolute-agreement, single-measure intraclass correlation, ICC(2,1) [19].

Uncertainty. We report 95% confidence intervals for MAE, median absolute error, RMSE, MRE, and bias. These use a nonparametric percentile bootstrap with 10,000 resamples of participants [20]. For proportions (agreement, sensitivity, specificity) we use Wilson score intervals [16].

4.2 Level 2: Tolerance Coverage

A single summary statistic hides the shape of the error distribution. The *coverage curve* $C(\tau)$ is the percentage of participants whose absolute error is at most τ, plotted for increasing tolerances τ. We report the tolerance needed to reach 50%, 75%, 90%, 95%, and 100% coverage. The 95% value plays the role of the T5 error in the first report. The 100% value is the worst case in the sample.

Balance-specific outcomes. Balance hold durations are recorded in whole seconds and capped at 10 s. This produces many exact ties at 10 s, so continuous metrics alone are misleading. For each stance we therefore report five outcomes:

- **Exact match:** AVA duration equals PT duration.
- **Full hold:** both AVA and PT record the full 10 s.
- **Partial match:** both record the same duration below 10 s.
- **Both partial:** both record some duration below 10 s (both call it a fail), whether or not the durations are equal.
- **Pass/fail agreement:** AVA and PT agree on pass or fail. This equals *full hold* plus *both partial*.

For the cases where the durations do not match, we report the MAE, bias, standard deviation, median, and range of the error.

4.3 Level 3: Clinical Decision Agreement

A clinician usually acts on the risk zone, not on the raw number. We compare AVA and PT zones using a confusion matrix, overall percentage agreement, Cohen's κ [21], and linearly weighted κ. Weighted κ treats a green-to-red disagreement as worse than a green-to-yellow one.

Error taxonomy by direction. Every zone or pass/fail disagreement is classified by its clinical direction:

- **Over-severity (OS):** AVA assigns a *higher*-risk zone than the PT, or fails a stance the PT passed. This is a false alarm. It is conservative, but it can lead to unnecessary follow-up.
- **Under-severity (US):** AVA assigns a *lower*-risk zone than the PT, or passes a stance the PT failed. This is a missed risk and the more important error to minimize in a screening setting.

For balance pass/fail, where "fail" is treated as the positive finding, we also report sensitivity (the share of PT failures that AVA detected) and specificity (the share of PT passes that AVA confirmed).

Boundary analysis. For each zone disagreement we record how far the PT reference value was from the nearest zone boundary. This separates disagreements caused by small measurement errors near a cut-off, which are unavoidable for any measurement with nonzero error, from disagreements that point to a real failure of the model.

4.4 Ablation and Stratified Analyses

The modular design lets us change a single pipeline setting while keeping everything else fixed, then re-run the whole evaluation. In this report we ablate the temporal window used by the stance-check filter in the balance engine. We compare a 100 ms and a 250 ms configuration on the single-leg stance, which is the most challenging stance (Section 5.5). We also stratify TUG errors by PT-assigned zone, and all three assessments' zone distributions by age group.

4.5 Exclusions

A participant was excluded from an analysis if either the AVA output or the PT reference was missing, or if the recording was flagged during quality review. For the TUG, 74 AVA results and 78 PT results were available. Nine participants were excluded (P033, P035, P067, P068, P071, P072, P084, P085, P088), leaving 71 matched pairs. For the chair stand, one participant was not analyzed. For the balance test, per-stance sample sizes were 69, 72, 66, and 75.

5 Results

Throughout this section, "AVAL" in figure labels refers to the AVA Mobility predictions produced by the AVAL.v2 framework, and "PT" refers to the physical-therapist reference.

5.1 Cohort Functional Profile

Figure 6 shows the PT-scored results. TUG times ranged from about 5 s to over 24 s. Chair-stand counts ranged from 10 stands below the CDC norm to 7 above it. The cohort-level balance pass rates could not be reconciled with the paired per-stance samples in Table 9, which include 17 PT full-tandem failures among 66 participants, so they are not reported; per-stance PT pass and fail counts are given in Table 9. The cohort therefore covers the full range of each test, including a meaningful number of participants in the red zone.



(a) TUG time

(b) Chair stands relative to CDC norm

Figure 6: Distribution of PT-scored results, with the study-defined risk zones (Section 3.6) shaded.

Risk zones by age. Figure 7 shows how risk zones change with age. TUG and balance results worsen sharply with age. For the TUG, 83% of participants younger than 70 were in the green zone, compared with 14% of those aged 90 and over; 71% of the oldest group were in the red zone. The share of balance results in the green zone dropped from 46–58% below age 80 to 14–25% at age 80 and above. The chair stand shows the *opposite* pattern: 23% green below age 70 and 75% green at 90 and over. This is because chair-stand zones are defined relative to age- and sex-specific norms, while TUG and balance zones use the same cut-offs at every age. For an AI system this matters because the same measurement error has different clinical consequences depending on how each test's zones are defined.

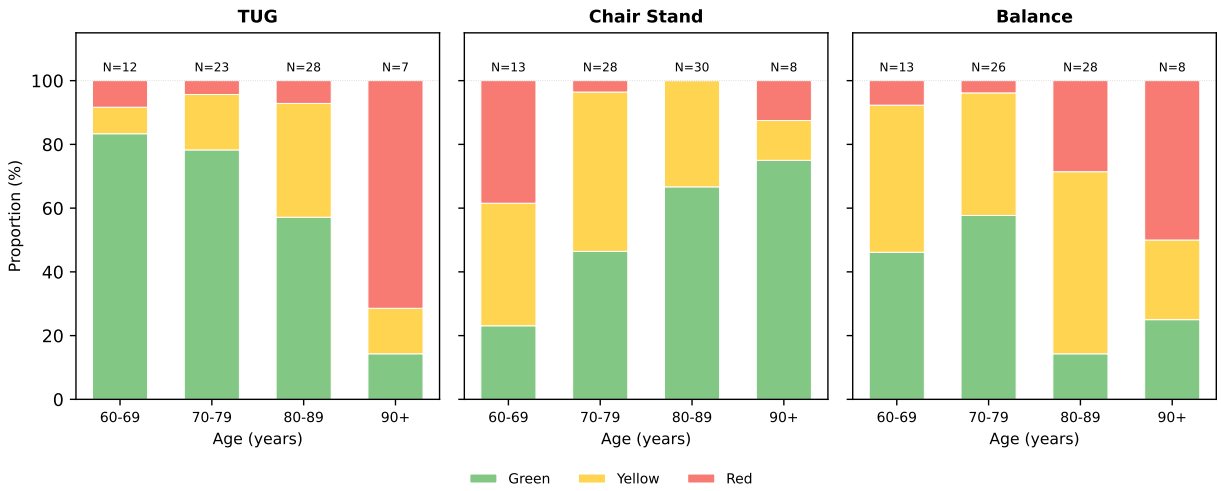
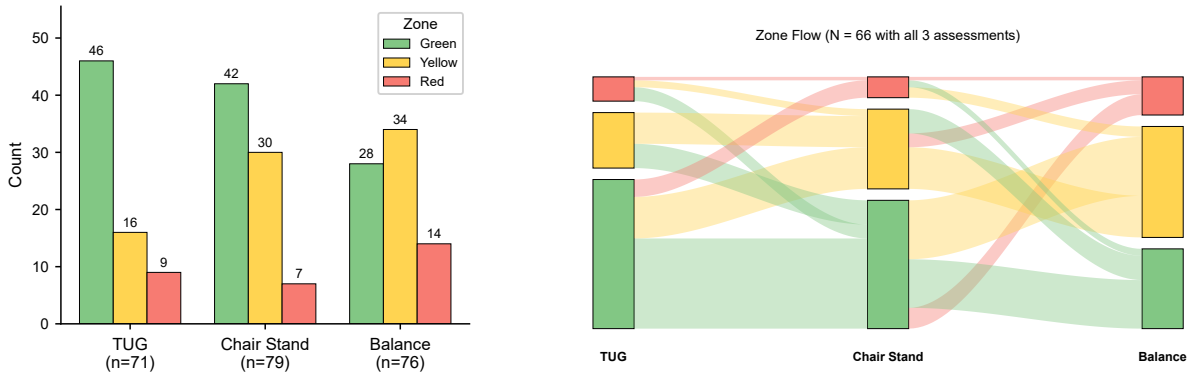


Figure 7: PT-assigned risk-zone proportions by age group for each assessment. Group sizes are shown above each bar. The youngest group, labelled 60–69, also includes the four participants aged 54–59, and the oldest group (90+) includes the two aged 95 or older, so these group sizes exceed the 60–69 and 90–94 counts in Table 2.

How zones relate across assessments. Among the 66 participants with all three tests, being in the green zone on one test did not reliably predict the zone on the others (Figures 8a and 8b). Overall, 64.8% of TUG results, 53.2% of chair-stand results, and 36.8% of balance results were in the green zone. Many participants in the TUG or chair-stand green zone fell into the balance yellow or red zone. This supports running all three tests: each one picks up impairments the others miss, so the AI must be accurate on all of them, not just one.



(a) Zone counts per assessment

(b) Zone flow across assessments ($n = 66$)

Figure 8: PT-assigned risk zones across the three assessments.

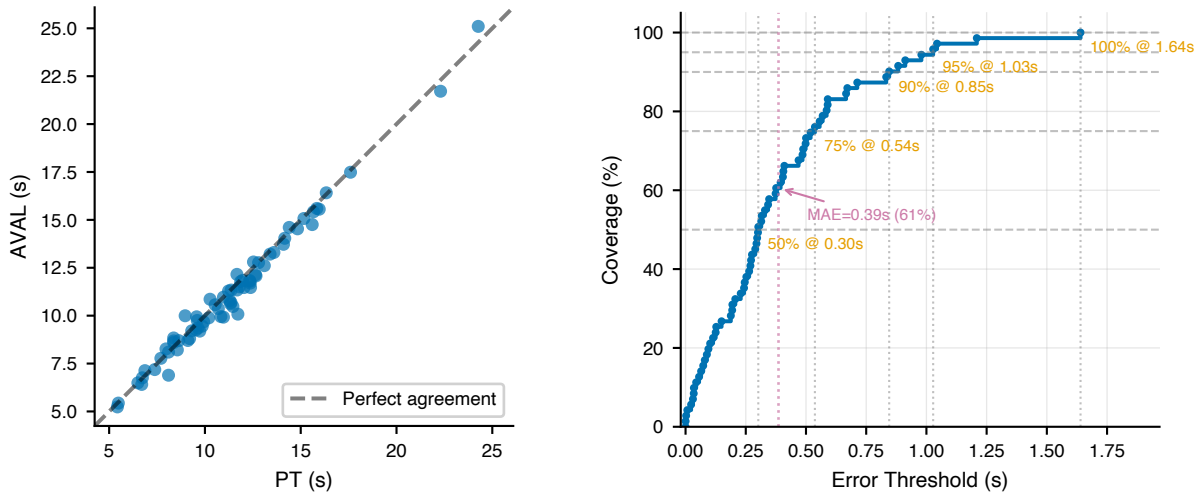
5.2 Timed Up and Go

Measurement agreement. Table 4 summarizes TUG agreement for the 71 matched participants. AVA's MAE was 0.386 s (95% CI 0.314–0.462) and its median absolute error was 0.302 s, so a typical error was about a third of a second. The mean relative error was 3.5%. RMSE (0.501 s) was about 30% larger than MAE; four participants had absolute errors above 1 s, and the largest was 1.64 s. Agreement was near perfect by every correlation measure ($r = 0.991$, CCC $= 0.989$, ICC(2,1) $= 0.989$). Figure 9a shows that the points lie close to the line of perfect agreement across the whole range from 5 to 25 s.

Bias. AVA timed the TUG slightly shorter than the PT on average (bias -0.221 s). Its time was shorter in 50 participants, longer in 19, and identical in 2. The bias is statistically distinguishable from zero, since the confidence interval excludes zero, but it is small: about 2% of the median TUG time. In the Bland–Altman analysis (Figure 10a), the regression slope of error on mean was -0.001 s per second (95% CI -0.033 to 0.031; $p = 0.94$). The data therefore

Table 4: TUG agreement between AVA and PT ($n = 71$). Confidence intervals are 95% percentile bootstrap intervals (10,000 resamples).

Metric	Value	95% CI
Mean absolute error (s)	0.386	0.314 – 0.462
Median absolute error (s)	0.302	0.260 – 0.390
Root mean squared error (s)	0.501	0.404 – 0.598
Mean relative error (%)	3.5	2.8 – 4.2
Bias, mean signed error (s)	−0.221	−0.328 – −0.115
SD of error (s)	0.45	–
95% limits of agreement (s)	−1.11 to +0.67	–
Pearson r	0.991	–
Lin's CCC	0.989	–
ICC(2,1), absolute agreement	0.989	–



(a) AVA vs. PT TUG time

(b) Coverage curve

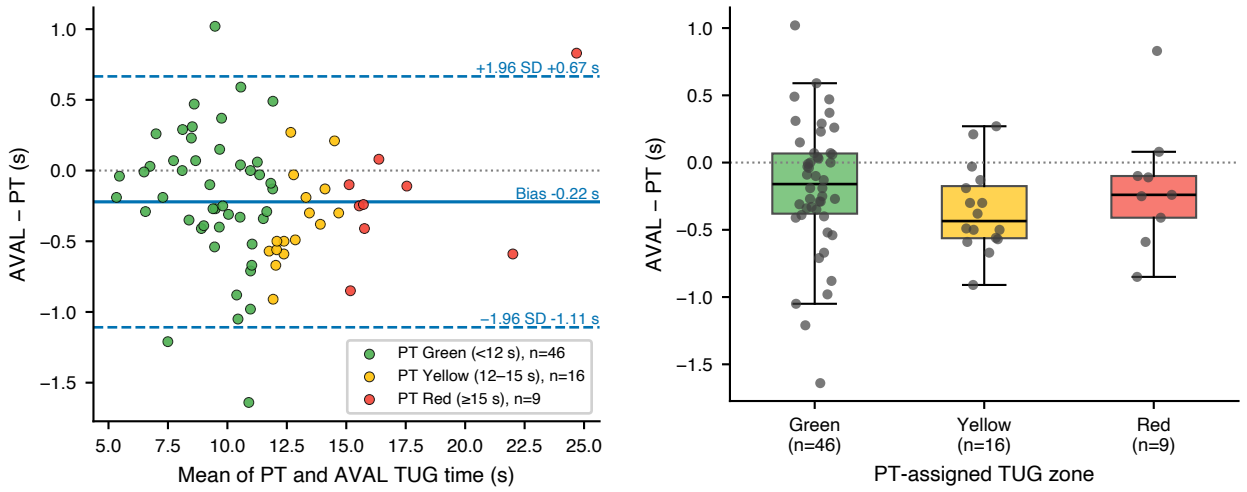
Figure 9: TUG measurement agreement ($n = 71$).

show no evidence that AVA's absolute error changes with TUG time, although the interval is compatible with a small trend in either direction, up to about 0.3 s over a 10-second range. Consistent with roughly constant absolute error, relative error was lower among slower participants, from 3.9% in the green zone to 2.1% in the red zone (Table 6). The 95% limits of agreement (−1.11 to +0.67 s) mean that an individual AVA measurement can be expected to lie within about one second of the PT's.

Coverage. Table 5 lists the coverage milestones. Half of all TUG timings were within 0.30 s of the PT's, 90% were within 0.85 s, and 95% were within 1.03 s; 94.4% (67 of 71) were within 1 s. The largest error in the sample was 1.64 s. Measured in relative terms, 80.3% of predictions were within 5% of the PT's value and 95.8% were within 10%.

Table 5: TUG coverage: absolute-error tolerance needed to reach each share of participants, and the share of participants within fixed tolerances.

Coverage	Error ≤	Tolerance	Participants within
50%	0.30 s	±0.25 s	38.0%
75%	0.54 s	±0.50 s	73.2%
90%	0.85 s	±1.00 s	94.4%
95%	1.03 s	±5% relative	80.3%
100%	1.64 s	±10% relative	95.8%



(a) Bland–Altman plot

(b) Signed error by PT zone

Figure 10: TUG error structure. (a) Difference against mean, with bias and 95% limits of agreement; points are colored by PT-assigned zone. (b) Signed error in each PT zone; boxes show the interquartile range and dots show individual participants.

Stratified analysis. Table 6 breaks TUG error down by PT zone. MAE was very similar across zones (0.377–0.412 s), which shows that accuracy is maintained for the slowest, highest-risk participants. The yellow zone had the largest negative bias (−0.352 s) and the smallest spread (SD 0.317 s), so AVA's errors in this range were small but consistently in the same direction. This matters because the yellow zone lies between the two decision boundaries (Figure 10b).

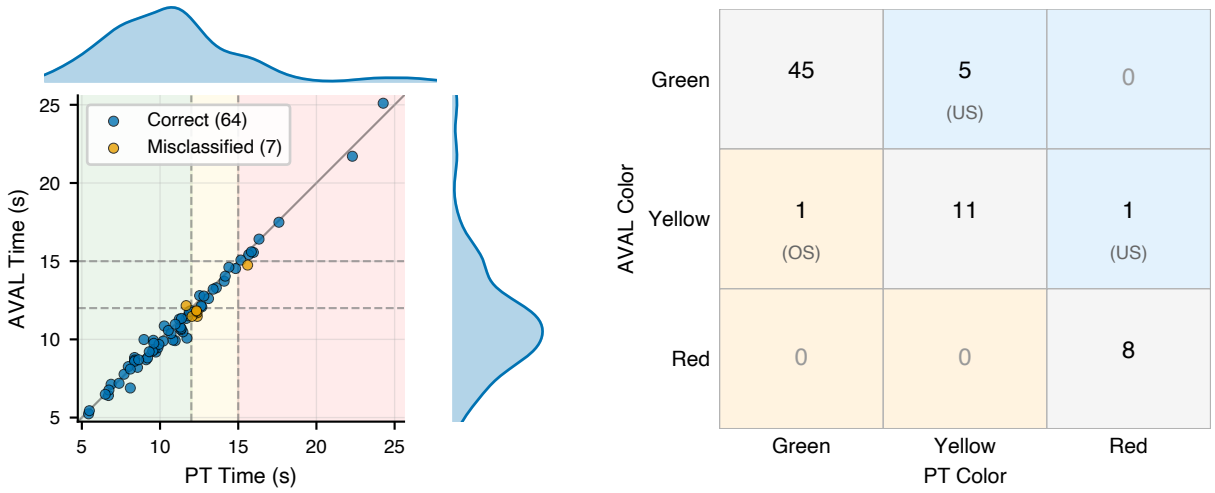
Table 6: TUG error stratified by PT-assigned zone. AE: absolute error; MRE: mean relative error.

PT zone	n	MAE (s)	Median AE (s)	RMSE (s)	Bias (s)	SD (s)	MRE (%)
Green (<12 s)	46	0.377	0.290	0.516	−0.183	0.488	3.9
Yellow (12–15 s)	16	0.412	0.435	0.468	−0.352	0.317	3.2
Red (≥ 15 s)	9	0.384	0.250	0.480	−0.182	0.471	2.1
All	71	0.386	0.302	0.501	−0.221	0.45	3.5

Clinical decision agreement. AVA assigned the same TUG zone as the PT for 64 of 71 participants (90.1%, 95% CI 81.0–95.1; $\kappa = 0.80$, weighted $\kappa = 0.85$; Figure 11). No participant was moved by two zones. Six disagreements were under-severity: five PT-yellow participants were placed in green, and one PT-red participant was placed in yellow. One was over-severity, a PT-green participant placed in yellow. The under-severity errors match the direction of AVA's negative bias.

Boundary analysis. Table 7 lists all seven disagreements. In every case the PT time was within 0.6 s of a zone boundary; six of the seven PT times fell between 11.67 and 12.38 s, and each AVA error was below 1 s. Inter-rater variability between PTs was not measured in this study, so we cannot say how often two PTs would disagree on these participants. As an exploratory check, we removed the mean bias by adding 0.221 s to every AVA time. This reduced MAE to 0.339 s and left the total zone agreement at 64 of 71, but the disagreements were not the same. Two participants moved into agreement (P013 and P048), two who had agreed now disagreed (P002 and P036, both PT-green, moved to yellow), and five disagreements were unchanged. Under-severity errors fell from six to four and over-severity errors rose from one to three. The unchanged total therefore conceals a shift in the direction of the errors, and it does not show by itself whether the boundary disagreements are random or systematic.

Largest errors. Four participants had absolute errors above 1 s: P058 (−1.64 s, 14.0%), P050 (−1.21 s, 14.9%), P062 (−1.04 s, 9.5%), and P055 (+1.03 s, 11.5%). None of them changed zone, because their PT times (8.1–11.7 s) were well inside the green zone. They are, however, the highest-priority cases to review in video.



(a) AVA vs. PT time with zone boundaries

(b) Zone confusion matrix

Figure 11: TUG clinical-zone agreement. In (b), OS marks over-severity and US marks under-severity.

Table 7: All TUG zone disagreements, with the distance of the PT time from the nearest zone boundary.

Participant	PT (s)	AVA (s)	Error (s)	PT → AVA zone	Type	Distance to boundary (s)
P075	12.38	11.47	−0.91	Yellow → Green	US	0.38
P029	12.37	11.70	−0.67	Yellow → Green	US	0.37
P013	12.35	11.79	−0.56	Yellow → Green	US	0.35
P048	12.33	11.83	−0.50	Yellow → Green	US	0.33
P008	12.04	11.47	−0.57	Yellow → Green	US	0.04
P019	15.60	14.75	−0.85	Red → Yellow	US	0.60
P016	11.67	12.16	+0.49	Green → Yellow	OS	0.33

5.3 30-Second Chair Stand

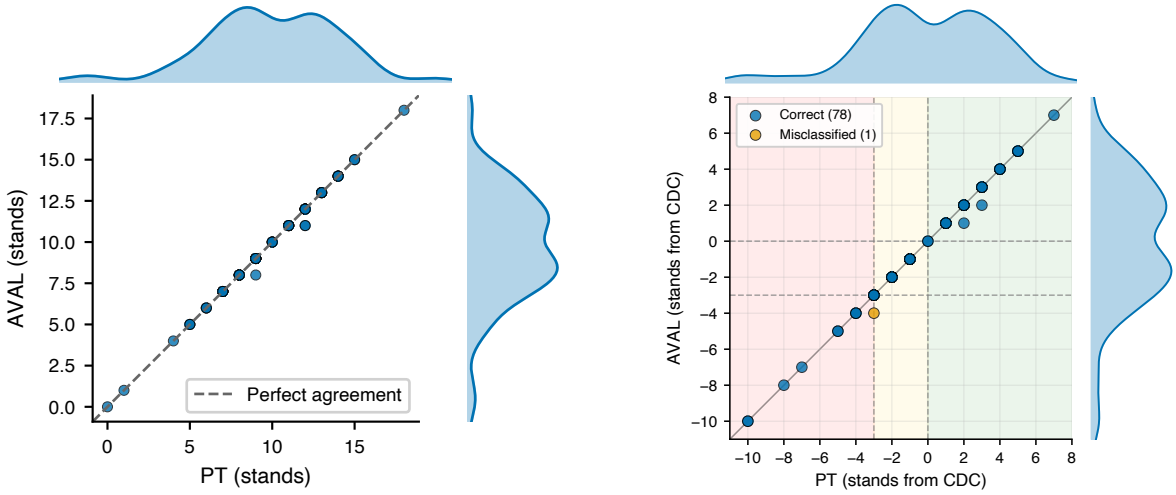
Measurement agreement. Figure 12a compares AVA and PT chair-stand counts for 79 participants, from 0 to 18 stands. AVA's count matched the PT's exactly for 76 of 79 participants (96.2%, 95% CI 89.4–98.7). The other three were each undercounted by exactly one stand, so no prediction was off by more than one repetition. The raw-count MAE was 0.04 stands (3 of 79), and the mean signed error was −0.04 stands.

Clinical decision agreement. Because the errors were so small, zone agreement was nearly perfect: 78 of 79 participants (98.7%, 95% CI 93.2–99.8; $\kappa = 0.98$, weighted $\kappa = 0.98$; Figure 13). The one disagreement was an over-severity error. The PT count was exactly 3 stands below the norm, the lowest value still in the yellow zone. AVA's one-stand undercount therefore moved the participant into red. There were no under-severity errors. The PT zone totals in Figure 13 (40 green, 30 yellow, 9 red) differ slightly from those in Table 3 (42, 30, 7) because the two analyses applied the CDC age- and sex-specific norms differently. Agreement in Figure 13 is not affected, because AVA's and the PT's counts there were categorized with the same norms.

5.4 4-Stage Balance

The 4-stage balance test is new in AVAL.v2 and technically the most demanding. Instead of detecting a large movement such as standing up, the model must recognize the correct foot position and then detect the moment that position is lost. The first loss can be a small step, a touch-down of the lifted foot, or a hand reaching for support.

Stopping protocol and unattempted stances. The four stances were administered in order of difficulty, and each was timed for up to 10 s; a stance counted as passed only if it was held for the full 10 s. The STEADI administration guidance, which the protocol followed, stops the test once a participant cannot hold a stance for 10 s. The PT records available for this analysis do not document, participant by participant, whether testing stopped after a failed stance or



(a) Raw counts

(b) Counts relative to CDC norm, with zones

Figure 12: Chair-stand agreement between AVA and PT ($n = 79$).

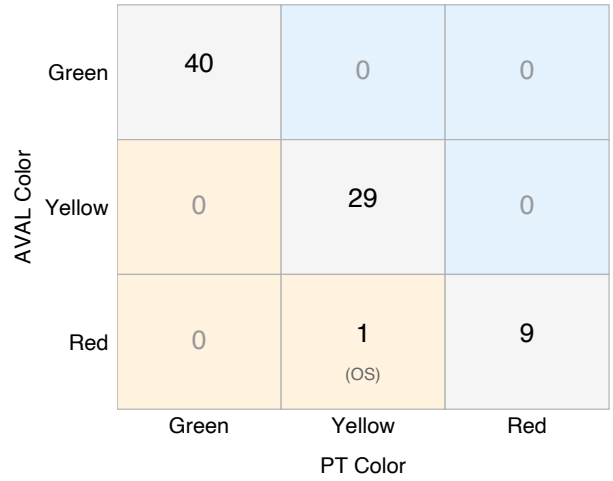


Figure 13: Chair-stand zone confusion matrix (rows: AVA; columns: PT).

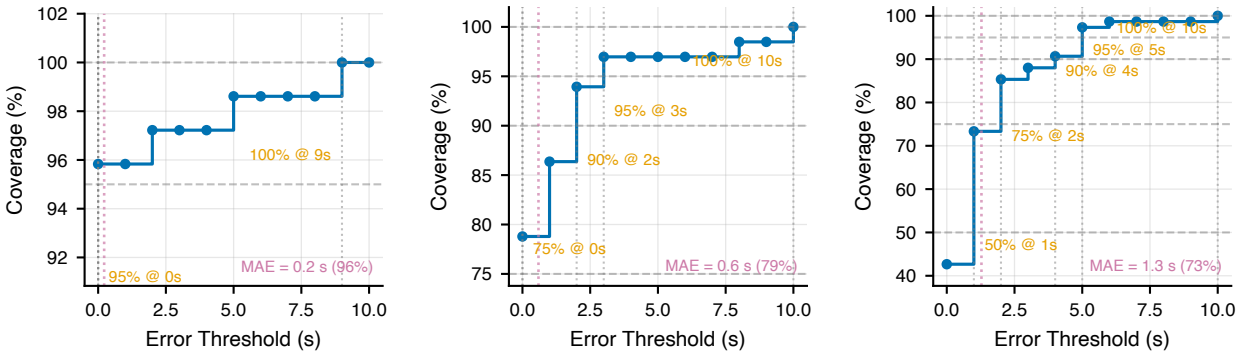
how a stance that was not attempted was recorded. Each per-stance analysis therefore includes every participant who had both an AVA and a PT result for that stance, which is why the per-stance sample sizes differ (69, 72, 66, and 75), and a stance recorded as not held may include stances that were not attempted.

Per-stance duration agreement. Table 8 reports duration agreement for each stance. Accuracy falls as the stance gets harder. Side-by-side stance was perfect (69 of 69 exact matches). Semi-tandem was near perfect: 95.8% exact matches and an overall MAE of 0.2 s. Full tandem reached 78.8% exact matches with an overall MAE of 0.6 s. Single-leg stance was the hardest, with 42.7% exact matches and an overall MAE of 1.3 s. In every stance with errors, the mean signed error was negative: when AVA and the PT disagreed, AVA usually recorded a *shorter* hold. This means the model tends to end a hold early.

Coverage. Figure 14 shows the coverage curves. For semi-tandem, 95.8% of durations matched exactly and every case was within 9 s. For full tandem, 86.4% were within 1 s, 93.9% within 2 s, and 97.0% within 3 s. For single-leg stance, a 1-second tolerance raised coverage from 42.7% (exact) to 73.3%. Coverage reached 90.7% within 4 s and 97.3% within 5 s. Much of the single-leg disagreement is therefore off by only one second, which is the resolution at which durations are recorded.

Table 8: Balance hold-duration agreement by stance (single-leg: 100 ms configuration). Error statistics are computed over the mismatched cases only; the overall MAE is computed over all participants.

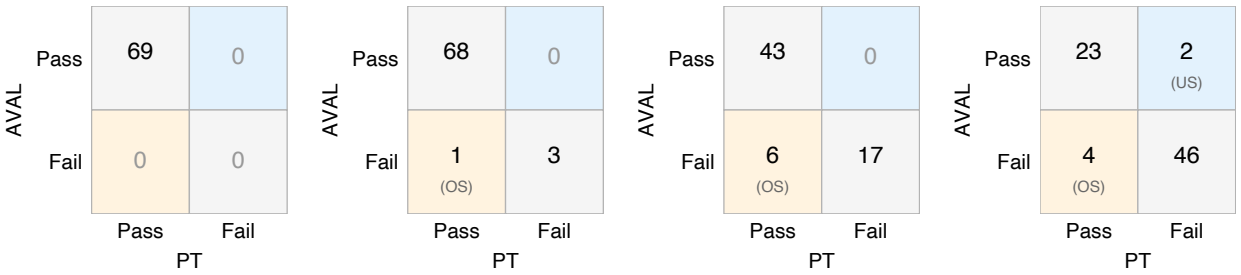
	Side-by-side	**Semi-tandem**	**Full tandem**	**Single-leg**
n	69	72	66	75
Exact match	100.0%	95.8%	78.8%	42.7%
Full hold (both 10 s)	100.0%	94.4%	65.2%	30.7%
Partial match (equal, < 10 s)	0.0%	1.4%	13.6%	12.0%
Both partial (both < 10 s)	0.0%	4.2%	25.8%	61.3%
Pass/fail agreement	**100.0%**	**98.6%**	**90.9%**	**92.0%**
Mismatched cases, n	0	3	14	43
MAE (s)	–	5.33	2.79	2.23
Bias (s)	–	−5.33	−2.07	−1.63
SD (s)	–	3.51	3.36	2.47
Median error (s)	–	−5.0	−1.5	−1.0
Range (s)	–	−9 to −2	−10 to +2	−10 to +5
Overall MAE, all cases (s)	0.0	0.2	0.6	1.3



(a) Semi-tandem (b) Full tandem (c) Single-leg (100 ms)

Figure 14: Balance hold-duration coverage curves: the share of participants whose absolute duration error is at most each threshold.

Pass/fail agreement. Figure 15 and Table 9 give the pass/fail results for each stance. AVA detected nearly every PT-recorded failure. Sensitivity was 100% for semi-tandem and full tandem and 95.8% for single-leg stance. Most disagreements were over-severity errors: AVA failed a hold the PT had passed (1 in semi-tandem, 6 in full tandem, 4 in single-leg). Under-severity errors occurred only in single-leg stance, where AVA passed 2 holds the PT had failed. The main area for improvement in the balance engine is therefore specificity on the harder stances, not sensitivity.



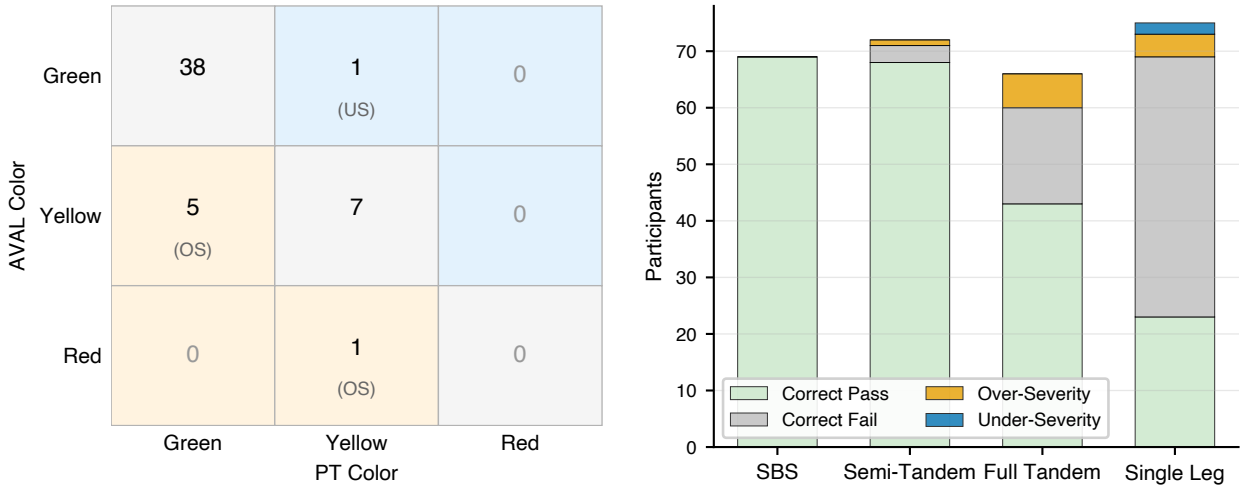
(a) Side-by-side (b) Semi-tandem (c) Full tandem (d) Single-leg

Figure 15: Per-stance pass/fail confusion matrices (rows: AVA; columns: PT; single-leg: 100 ms configuration).

Table 9: Per-stance pass/fail agreement, with failure treated as the positive finding (single-leg: 100 ms configuration). Agreement, sensitivity, and specificity are given with 95% Wilson score intervals.

Stance	n	PT fail	Agreement (95% CI)	κ	Sensitivity	Specificity	OS / US
Side-by-side	69	0	100.0% (94.7–100)	–	–	100.0% (69/69; 94.7–100)	0 / 0
Semi-tandem	72	3	98.6% (92.5–99.8)	0.85	100.0% (3/3; 43.8–100)	98.6% (68/69; 92.2–99.7)	1 / 0
Full tandem	66	17	90.9% (81.6–95.8)	0.79	100.0% (17/17; 81.6–100)	87.8% (43/49; 75.8–94.3)	6 / 0
Single-leg	75	48	92.0% (83.6–96.3)	0.82	95.8% (46/48; 86.0–98.8)	85.2% (23/27; 67.5–94.1)	4 / 2

Balance risk zones. We assigned balance zones to the participants who had AVA results for all four stances ($n = 52$). This analysis used a different balance rule from the one defined in Section 3.6: green, three or four stances passed; yellow, two passed; red, zero or one passed. Because most participants passed at least three stances, this rule places most of them in the green zone, which is why the subset contains 43 PT-green participants, whereas Table 3, which uses the Section 3.6 rule, has 28 PT-green participants among 76. The agreement results below are therefore for this three-or-more rule. AVA and the PT agreed for 45 of them (86.5%, 95% CI 74.7–93.3; $\kappa = 0.60$, weighted $\kappa = 0.62$; Figure 16a). Six of the seven disagreements were over-severity, mostly PT-green participants placed in yellow, consistent with the early single-leg terminations seen in Table 9. One was under-severity (PT yellow, AVA green). The κ value is lower than for the other tests partly because this subset had few participants outside the green zone: no one in it was PT-red. With so few non-green participants, κ drops sharply for each disagreement.



(a) Balance zone confusion matrix

(b) Outcome breakdown by stance

Figure 16: Balance clinical agreement (single-leg: 100 ms configuration). (a) Zone agreement for participants with AVA results on all four stances ($n = 52$), using the three-or-more-stances rule described in Section 5.4. (b) Correct passes, correct fails, over-severity, and under-severity for each stance. SBS: side-by-side.

5.5 Ablation: Stance-Check Window for Single-Leg Stance

Single-leg stance had the most duration errors, so we used it to test how sensitive the balance engine is to the temporal window of its stance-check filter. We compared the 100 ms configuration with a 250 ms configuration and kept every other part of the pipeline the same. Table 10 and Figure 17 give the results.

Neither configuration is better on every measure. The 250 ms window produced more exact duration matches (+6.6 percentage points), all of them among holds below 10 s. Its errors, when they occurred, were larger and more variable, including one overestimate of 10 s. The 100 ms window produced fewer exact matches but smaller, tighter errors. It also gave better pass/fail agreement and one fewer under-severity error, and it reached 90% and 95% coverage one second earlier. For screening, the pass/fail decision and the avoidance of missed risk matter most, so the 100 ms configuration is preferred. The trade-off suggests that a hybrid approach could keep the advantages of both: a longer window to confirm a stance and a shorter one to detect its loss.

Table 10: Single-leg stance ablation: 100 ms vs. 250 ms stance-check window ($n = 75$). The better value in each row is in bold.

Metric	100 ms	250 ms
Exact duration match	42.7% (32)	**49.3% (37)**
Full hold	30.7% (23)	30.7% (23)
Partial match	12.0% (9)	**18.7% (14)**
Pass/fail agreement	**92.0% (69)**	90.7% (68)
Pass/fail κ	**0.82**	0.80
Under-severity errors	**2**	3
Over-severity errors	4	4
Mismatched cases	43	**38**
MAE, mismatched (s)	**2.23**	2.76
Bias, mismatched (s)	−1.63	**−1.13**
SD, mismatched (s)	**2.47**	3.47
Error range (s)	−10 **to** +5	−10 to +10
Overall MAE, all cases (s)	**1.3**	1.4
Coverage: 90% within	**4 s**	5 s
Coverage: 95% within	**5 s**	6 s

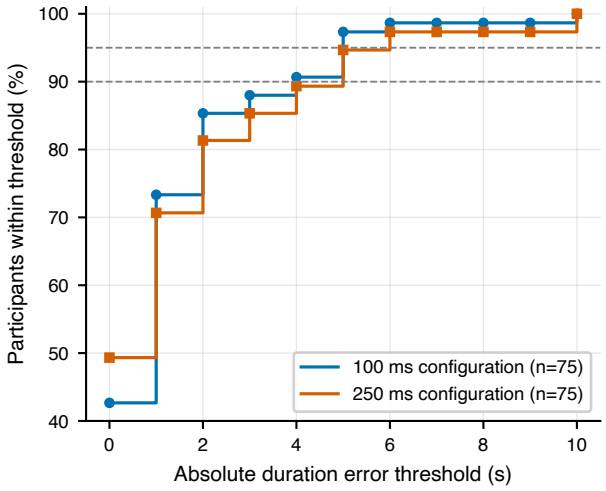


Figure 17: Single-leg duration coverage curves for the two stance-check configurations.

6 Comparison with AVAL v1

Table 11 lists headline TUG and chair-stand results from the two reports side by side. The two studies are not directly comparable. They differ in cohort and sites, in the reference procedure (in v1 the TUG reference was the mean of two PT observers' timings; in v2 it is PT scoring of each assessment), and in the TUG category boundaries (v1 used 11 s as the upper limit of the high-mobility category, whereas v2 uses 12 s). The model version may also differ. Error values were lower in v2 than in v1, but because the cohorts, reference procedures, and category boundaries all changed, these differences cannot be attributed to changes in the model, and the table should not be read as evidence of model improvement.

7 Discussion

7.1 What the Error Analysis Shows

The three levels of evaluation lead to one consistent picture. For the **TUG and chair stand**, which rely on large movements, AVA's measurement error is small and centered on a small bias, with no evidence that it changes across the measurement range. All of the remaining clinical disagreements involved sub-second or single-repetition errors

Table 11: Headline results in AVAL v1 and AVAL.v2. The studies used different cohorts, reference procedures, and TUG category boundaries, so the values are not directly comparable.

Metric	**v1** [12]	**v2 (this report)**
TUG n	62	71
TUG MAE (s)	0.709	0.386
TUG RMSE (s)	0.854	0.501
TUG mean relative error	6.0%	3.5%
TUG 95th-percentile error (s)	1.595	1.03
TUG zone agreement[a]	88.1% ($\kappa = 0.79$)	90.1% ($\kappa = 0.80$)
Chair stand n	65	79
Chair stand within ± 1	100%	100%
Chair stand zone agreement	100%	98.7% ($\kappa = 0.98$)
4-stage balance	not evaluated	90.9–100% pass/fail agreement per stance

[a]Not directly comparable: v1 placed the high/fair boundary at 11 s and v2 at 12 s, and the two studies used different reference procedures.

at participants whose reference values lay close to a zone cut-off; how often two PTs would disagree on the same participants was not measured. At this point, further gains in zone agreement for these tests will come less from improving the model and more from how results near a cut-off are handled.

For the **4-stage balance test**, the error analysis finds a specific, fixable problem. Errors tended toward shorter hold durations, but occurred in both directions. This produces a small number of over-severity errors on the harder stances, while almost every true failure is still detected. From a safety standpoint this is the preferable way to fail. It still lowers specificity, which in turn lowers balance zone agreement.

7.2 Improvement Targets

The results point to four specific targets for the next model iteration.

1. **Early termination of balance holds.** Separate confirming a stance from detecting its loss, using the trade-off seen in the ablation. Also review the 10 full-tandem and single-leg over-severity cases to find which signal (foot displacement, arm movement, or trunk sway) triggers the early stop.
2. **Ambiguity near zone boundaries.** All TUG zone disagreements occurred within 0.6 s of a cut-off, which is inside the limits of agreement. We propose reporting a "borderline" flag whenever a result falls within the measured limits of agreement of a boundary. This would direct those results to clinician review rather than forcing a hard zone assignment.
3. **Residual TUG bias.** The -0.22 s bias is consistent and could be corrected by calibration. Our exploratory analysis shows that this would lower MAE and leave total zone agreement unchanged, but it would change which participants disagree and shift errors from under-severity toward over-severity. The correction must be estimated and validated on held-out data.
4. **Outlier review.** The four TUG cases with errors above 1 s and the balance cases with 10 s errors should be reviewed in video and added to a regression test set, so that each future model version is checked against them.

7.3 Limitations

This evaluation has several limitations. Every result depends on a single reference standard, PT scoring. Balance durations were recorded in whole seconds, which limits how finely duration agreement can be measured, and the PT's own measurement error is not estimated. Two or more independent PT raters would allow AVA's agreement to be compared directly with inter-rater agreement. The cohort is predominantly female (72.5%) which may limit generalizability. The cohort and samples in some subgroups are small: nine TUG participants were in the red zone, and eight participants were aged 90 and over. Sample sizes also differ across assessments and stances because of missing data. The ablation was run on the same data used for evaluation, so the preferred configuration should be confirmed on an independent set. Finally, secondary engine outputs (TUG phase times, step count, chair-stand cycle duration, and sway area) are not yet validated. The sample was not stratified by site, so we cannot yet report how performance varies across environments.

8 Conclusion

AVAL.v2 extends the validation of AVA Mobility from "Is it accurate?" to "How and where does it fail, and what should be improved?" We evaluated 80 older adults from independent-living communities, senior centers, and community activity groups, using a three-level methodology that examines measurement agreement, tolerance coverage, and clinical-decision agreement together. AVA agreed closely with PT scoring on all three STEADI assessments. On the TUG, its mean absolute error was 0.386 s (95% CI 0.314–0.462), its mean relative error was 3.5%, and its concordance with the PT was 0.989. The error showed a small, consistent bias (-0.22 s) with no evidence of proportional bias, and 94% of timings were within 1 s of the PT. AVA's chair-stand count matched the PT's exactly for 76 of 79 participants and was one stand lower for the other three. On the new 4-stage balance test, pass/fail agreement ranged from 90.9% to 100% across the four stances, and AVA detected nearly every PT-recorded failure. For clinical decisions, AVA assigned the same risk zone as the PT for 90.1% of TUG results ($\kappa = 0.80$) and 98.7% of chair-stand results ($\kappa = 0.98$). Every TUG zone disagreement involved a PT time within 0.6 s of a cut-off; removing AVA's bias left total zone agreement unchanged but changed which participants disagreed. The largest TUG error in the study caused no zone change at all, which confirms that error magnitude alone does not determine clinical impact (Figure 1).

The error analysis also shows where AVA should improve and how it should be deployed. On the harder balance stances, AVA tended to end holds early. This produced a small number of conservative false alarms while preserving high sensitivity, and the ablation showed that the behavior can be tuned, with a trade-off between exact-duration accuracy and pass/fail agreement. Four improvement targets follow directly: separating stance confirmation from loss-of-stance detection, flagging results within the limits of agreement of a zone boundary for clinician review, calibrating the residual TUG bias on held-out data, and adding the largest-error cases to a permanent regression test set. With such a flag, AVA could handle the clear majority of TUG and chair-stand screenings automatically and direct the few uncertain results to a clinician. More broadly, the approach used here offers a general template for validating AI systems used in functional assessment. It combines a modular framework that runs the production pipeline unchanged on recorded data, a three-level evaluation that links measurement error to clinical decisions, an error taxonomy that separates false alarms from missed risk, boundary analysis, and component-level ablation. Future studies should validate the improved model on an independent cohort and compare AVA's agreement with the agreement between independent PT raters. They should also validate the secondary outputs, such as TUG phase timing, step count, and postural sway, and include larger and more diverse cohorts. Finally, they should examine whether repeated AVA screening predicts falls and supports earlier intervention.

References

[1] Centers for Disease Control and Prevention. 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. `https://www.cdc.gov/steadi/index.html`, 2023. Accessed: 2025-01-15.

[3] Diane Podsiadlo and Sandra Richardson. The timed "Up & Go": A test of basic functional mobility for frail elderly persons. *Journal of the American Geriatrics Society*, 39(2):142–148, 1991.

[4] C. Jessie Jones, Roberta E. Rikli, and William C. Beam. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Research Quarterly for Exercise and Sport*, 70(2):113–119, 1999.

[5] Jane E. Rossiter-Fornoff, Steven L. Wolf, Leslie I. Wolfson, and David M. Buchner. A cross-sectional validation study of the FICSIT common data base static balance measures. *The Journals of Gerontology: Series A*, 50(6): M291–M297, 1995.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 172–186, 2021.

[7] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. BlazePose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[8] Łukasz Kidziński, Bryan Yang, Jennifer L. Hicks, Apoorva Rajagopal, Scott L. Delp, and Michael H. Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature Communications*, 11:4054, 2020.

[9] Jan Stenum, Cristina Rossi, and Ryan T. Roemmich. Two-dimensional video-based analysis of human gait using pose estimation. *PLOS Computational Biology*, 17(4):e1008935, 2021.

[10] Scott D. Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp. OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology*, 19(10):e1011462, 2023.

[11] Ross A. Clark, Benjamin F. Mentiplay, Emma Hough, and Yong Hao Pua. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. *Gait & Posture*, 68:193–200, 2019.

[12] ForesightCares, Inc. AVAL: A validation pipeline for AVA mobility. Technical report, ForesightCares, Charlotte, NC, 2026.

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

[14] 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.

[15] 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.

[16] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[17] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.

[18] Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1): 255–268, 1989.

[19] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[20] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

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