

Markerless Motion Capture in Routine Clinical Upper Limb Assessments: Validity and Insights Beyond Ordinal Scoring

Tim Unger

tim.unger@lilui.org

Lake Lucerne Institute

Olivier Lambercy

ETH Zurich

Roger Gassert

ETH Zurich

Andreas R. Luft

Cereneo

R. James Cotton

Shirley Ryan AbilityLab

Chris Easthope Awai

Lake Lucerne Institute

Research Article

Keywords: ARAT, Markerless Motion Capture, Precision Neurorehabilitation, AI

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10559678/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Markerless Motion Capture in Routine Clinical Upper Limb Assessments: Validity and Insights Beyond Ordinal Scoring

Tim Unger^{1,2*}, Olivier Lambercy², Roger Gassert²,
Andreas R. Luft^{3,4}, R. James Cotton⁵, Chris Easthope Awai¹

¹*Data Analytics & Rehabilitation Technology (DART), Lake Lucerne
Institute AG, Rubistrasse 9, Vitznau, 6354, Switzerland.

²Rehabilitation Engineering Laboratory, ETH Zurich, Gloriastrasse 37,
Zurich, 8006, Switzerland.

³Lake Lucerne Institute AG, Rubistrasse 9, Vitznau, 6354, Switzerland.

⁴cereneo, Seestrasse 18, Vitznau, 6354, Switzerland.

⁵Shirley Ryan AbilityLab, Department of Physical Medicine and
Rehabilitation, Northwestern University, 355 East Erie Street, Chicago,
60611, IL, USA.

*Corresponding author(s). E-mail(s): tim.unger@llui.org;
Contributing authors: olivier.lambercy@hest.ethz.ch;
roger.gassert@hest.ethz.ch; andreas.luft@cereneo.ch;
rcotton@sralab.org; chris.awai@llui.org;

Abstract

The Action Research Arm Test (ARAT) is a widely-used upper limb outcome measure in neurorehabilitation, but its ordinal scoring is subjective and suffers from limited sensitivity and specificity. We evaluated whether artificial-intelligence (AI)-based markerless motion capture (MMC), embedded into ARAT assessments during clinical routine, accurately reconstructs upper limb movement and yields valid, objective kinematic metrics carrying clinically meaningful information beyond the ordinal score. Across 47 sessions from 20 mixed-neurological patients (1,174 ARAT tasks), biomechanical reconstruction was accurate and

robust across impairment levels, and kinematic metrics showed the discrimination pattern expected of a construct-valid measure. In longitudinal case studies, the metrics added the specificity and sensitivity the ordinal score lacks: a domain decomposition exposed patient-specific recovery profiles underlying equal ARAT gains (specificity), and kinematic improvement continued to be detected after the ARAT had saturated (sensitivity). MMC in clinical routine can thus provide valid, objective, sensitive, and specific kinematic measurement complementing ordinal scoring.

Keywords: ARAT, Markerless Motion Capture, Precision Neurorehabilitation, AI

1 Introduction

Neurological conditions such as stroke and Parkinson’s disease affect hundreds of millions of people worldwide, making them a leading cause of motor impairment [1]. Every year, more than 12 million people suffer a stroke [2], and approximately half of the survivors still experience upper limb impairment after rehabilitation [3], often limiting their independence in the long term [1]. With aging societies, the number of affected individuals is expected to rise, increasing pressure on healthcare systems [2] and demanding more effective and efficient rehabilitation strategies [4, 5]. Precision neurorehabilitation has been proposed as the framework to meet this demand: a closed loop in which interventions are continually adapted to a patient’s measured response and outcomes re-assessed to steer the next decision. Realising it depends on the quality of the measurements that close the loop, yet the outcome measures available in routine practice remain largely subjective, human-scored assessments that characterise functional state only coarsely. Outcome measures that capture functional state objectively, sensitively, and specifically are therefore a prerequisite for precision neurorehabilitation and for the development of improved surrogate biomarkers of recovery against which novel interventions can be evaluated.

54 When it comes to motor recovery, it is currently not feasible to measure movement
55 objectively in clinical routine, and hence outcome measures rely on therapists observ-
56 ing the patient performing standardised tasks and then scoring them on an ordinal
57 scale based on their visual perception. The Action Research Arm Test (ARAT) is one
58 of the most frequently applied measures in neurorehabilitation to assess upper limb
59 function in patients [6]. It is a 19-item measure assessing upper limb functional perfor-
60 mance on pick-and-place tasks (grasp, grip, pinch) and gross motor tasks, each rated
61 on a four-point ordinal scale: 0 indicating inability to perform any part of the task, 1
62 indicating partial performance, 2 indicating task completion but with increased time
63 or difficulty, and 3 indicating normal motor-performance [6]. The ARAT’s ordinal
64 scale introduces several inter-related limitations [4, 7, 8]. First, subjectivity [6, 9]: dis-
65 tinguishing between Score 2 and Score 3 relies on a clinician’s judgement of whether
66 movement is “normal”, a determination that is inherently subjective and susceptible
67 to inter-rater variability, particularly at the functional boundary where movement
68 quality is marginal. Despite continued efforts to harmonise scoring criteria across neu-
69 rorehabilitation clinics [10–13], residual between-rater and between-site differences
70 persist, and reliable scoring itself requires training and accrues over a learning curve.
71 Second, sensitivity [14]: like many ordinal clinical scales, the ARAT suffers from floor
72 and ceiling effects, meaning patients at the extremes of performance cannot be fur-
73 ther differentiated by the scale, and meaningful improvements in movement quality or
74 speed may occur without any change in ordinal score. Third, specificity [7]: an ordinal
75 score conveys no information about why a particular rating was assigned. Two patients
76 receiving Score 2 on a specific task may do so for entirely different reasons (one con-
77 strained by limited range of motion, another by slow and effortful movement), yet
78 the ARAT provides no means to distinguish these distinct movement quality profiles,
79 limiting its value for treatment planning and outcome monitoring.

80 Kinematic analysis of upper limb movement offers a principled route to address
81 these limitations, providing continuous, observer-independent measurements of move-
82 ment quality. Marker-based optical motion capture (OMC) has demonstrated this
83 potential clearly: kinematic metrics derived from an instrumented drinking task can
84 quantify upper limb movement quality in people with stroke with high sensitivity
85 and validity [15, 16]. However, OMC’s substantial time requirements, technical com-
86 plexity, and cost render it incompatible with clinical routine, limiting its role to
87 occasional laboratory-based assessments rather than integrated rehabilitation moni-
88 toring. A practical kinematic measurement solution, one that is both accurate and
89 minimally disruptive to clinical workflows, is therefore needed to bridge the gap.

90 Markerless motion capture (MMC), an emerging technology, derives full-body
91 kinematics from multi-camera red-green-blue (RGB) video using AI-based computer
92 vision, supported by a growing ecosystem of research and commercial tools [17–22].
93 Rather than tracking physical markers attached to the body, as in active or passive
94 marker-based systems, it detects anatomical landmarks directly in the image using
95 pose estimation, removing the need for any on-body instrumentation. This approach
96 has shown high accuracy in laboratory validation against marker-based optical motion
97 capture [23] and minimal disruption to clinical workflows [24]. While these prior
98 contributions establish that MMC can be accurately reconstructed under controlled
99 laboratory conditions and integrated into clinical workflow without disruption, the
100 present study takes the next step: evaluating whether the kinematic metrics collected
101 during routine clinical sessions are themselves valid and clinically informative.

102 Evaluating MMC-based kinematics in this context calls for an incremental strat-
103 egy, building from the underlying reconstruction up to the metrics derived from it
104 and, ultimately, to the clinical value those metrics offer. The starting point is the
105 kinematic reconstruction itself: before any metric can be trusted, the reconstructed
106 joint trajectories must be of adequate accuracy in the clinical environment, and this

107 accuracy must hold across the full range of patient impairment rather than degrading
108 where motor function is worst.

109 Once an accurate signal is established, the central question becomes whether
110 metrics derived from it validly capture the upper limb function construct. This is
111 methodologically non-trivial because no gold-standard continuous measure of upper
112 limb function exists in clinical routine: validation must proceed against the very ordi-
113 nal system whose limitations motivate the work. This apparent circularity can be
114 turned into an advantage by asking not simply whether kinematic metrics agree with
115 the ARAT, but whether they agree *in the way a valid measure should*. A metric that
116 genuinely captures upper limb function should discriminate strongly between score
117 groups where the clinical boundary is robust, yet discriminate less where the ordinal
118 scale is itself unreliable. Observing this differential pattern would provide evidence
119 of construct validity that simple agreement cannot, precisely because it tracks the
120 underlying functional construct rather than the ordinal, clinical scale.

121 A metric established as valid in this cross-sectional sense can then be examined
122 longitudinally, both to confirm that its validity holds as patients recover over time
123 and to expose the added clinical value it offers beyond the ordinal score: the ability
124 to track recovery continuously, to reveal which specific aspects of movement drive a
125 given change, and to remain sensitive to improvement even after the clinical scale has
126 saturated. Together, these strands of evidence would establish whether MMC-derived
127 kinematics can serve as a valid and clinically informative complement to ordinal
128 scoring.

129 The aim of the present study is to evaluate the validity and added clinical value
130 of AI-based markerless motion capture for upper limb assessment in routine neurore-
131 habilitation, by quantifying the accuracy of the underlying kinematic reconstruction,
132 the construct validity of metrics derived from it at the population and longitudinal
133 levels, and the clinical information those metrics carry beyond ordinal scoring.

We therefore test the following hypotheses (Table 1):

Table 1: Evaluation framework. Each hypothesis provides a distinct level of evidence, building from the accuracy of the underlying kinematic reconstruction from the AI-based markerless motion capture algorithm (H1), through construct validity of kinematic metrics at the population (H2) and longitudinal (H3a) levels, to the added clinical value of kinematic metrics beyond ordinal scoring (H3b, H3c). Threshold symbols: η_H^2 , Kruskal–Wallis effect size (small-effect cut-off 0.06); AUC, area under the ROC curve (“excellent” ≥ 0.85 , “acceptable” ≥ 0.7 [25]); MCID, minimal clinically important difference (15 pp on the compound scale), adopted as an approximate threshold translated from consensus recommendations [4]. Operational definitions are given in Methods.

	Claim	Evidence method	type / Level	Data	Threshold
H1	Reconstruction accuracy & robustness	Reprojection error vs. established range; stability across score groups	Trial (precondition)	Cohort (1,174 trials)	$\overline{\text{error}} \in [10, 20] \text{ px}$ AND $\eta_H^2 < 0.06$
H2	Construct validity (population)	Known-groups discrimination (Tier 1 strong, Tier 2 attenuated)	Population (cross-sectional)	Cohort ($n = 20$)	Tier 1 AUC ≥ 0.85 (majority); $0.5 < \text{Tier 2} < \text{Tier 1}$
H3a	Longitudinal construct validity	S_{overall} agrees with clinical change (MAD)	Longitudinal (within-patient)	Case studies ($n = 2$)	$\text{MAD}(S_{\text{overall}}) < \text{MCID}$, per analysis group
H3b	Specificity (added value)	Domain-level recovery profiles where present	Longitudinal	Case studies ($n = 2$)	Reported, not a pass criterion
H3c	Responsiveness beyond ceiling (added value)	Domain change after clinical ceiling reached	Longitudinal	Case study ($n=1$)	$\Delta S_{\text{domain}} > \text{MCID}$ for ≥ 1 domain (post-ceiling)

H1 (Reconstruction accuracy and robustness). Biomechanical reconstruction from an ad-hoc three-camera setup (see Supplementary Figure A1) yields kinematic time series of adequate accuracy for metric extraction (cohort-level mean reprojection error within the 10–20 pixel (px) range previously established as acceptable for this reconstruction algorithm in clinical use, modestly above the 5–10 px laboratory standard [23]), with reconstruction quality remaining stable across the full range of clinical impairment (effect of ARAT score group on reprojection error $\eta_H^2 < 0.06$, conventional small-effect threshold).

H2 (Construct validity at the population level). MMC-derived kinematic metrics capture the intended movement-quality construct, evidenced by Tier 1

(Score 0/1 vs. Score 2/3) combined-feature discrimination of area under the receiver operating characteristic curve ($AUC \geq 0.85$ (conventional “excellent discrimination” threshold [25]) in the majority of analysis groups, with Tier 2 (Score 2 vs. Score 3) discrimination remaining above chance ($AUC > 0.5$, exceeding the conventional “acceptable discrimination” threshold of 0.7 [25] where possible) but consistently lower than Tier 1 across analysis groups, the pattern expected of a valid construct measure that tracks the underlying function rather than the noisy scale.

H3a (Longitudinal construct validity). During sub-ceiling recovery, the overall kinematic compound score S_{overall} , which aggregates all kinematic metrics, tracks the clinical ARAT trajectory within the adopted minimal clinically important difference (mean absolute deviation, $MAD < 15$ percentage points (pp) on the normalised compound scale, one minimal clinically important difference (MCID) [4]), evaluated per analysis group, establishing longitudinal construct validity in each analysis group where the criterion holds. Calculation of the metrics is described in Figure 1.

H3b (Specificity: domain-level recovery profiles beyond the ordinal score). Given the longitudinal validity established in H3a, the domain decomposition into range-of-motion (ROM), velocity, and compensation scores (S_{ROM} , S_{Vel} , S_{Comp}) exposes patient-specific recovery profiles invisible to the ordinal score when they exist. The case studies illustrate this added clinical value through the observed contrast between two patients on the Grasp analysis group.

H3c (Responsiveness beyond the clinical ceiling). Given the validity established in H3a, after clinical scores plateau at the ARAT ceiling, at least one domain-level compound score continues to change by more than the adopted MCID (≥ 15 pp on the normalised compound scale, one MCID [4]) during the post-ceiling window, demonstrating responsiveness where ordinal scoring saturates.

H1 and H2 are tested at the cohort level and constitute the primary hypotheses. H3a–H3c are evaluated through two longitudinal case studies and are exploratory, intended to illustrate the added clinical value of kinematic metrics and to motivate

173 future cohort-level investigation. H3b and H3c each build on H3a: the longitudinal
174 validity established in H3a is what licenses the domain divergence (H3b) and the post-
175 ceiling change (H3c) to be read as genuine recovery rather than measurement noise;
176 H3b and H3c are otherwise independent properties of the validated measure.

177 Demonstrating these properties together would establish AI-based markerless
178 motion capture as a valid, clinically meaningful complement to ordinal scoring inside
179 an unmodified neurorehabilitation routine, opening a path toward continuous, multi-
180 dimensional, observer-independent characterisation of upper limb recovery in standard
181 clinical practice.

182 **2 Methods**

183 **Participants and Data Collection**

184 Data were collected from 20 participants (10 stroke, 6 Parkinson’s disease, 4 other
185 neurological conditions) during routinely scheduled ARAT sessions at the Swiss neu-
186 rorehabilitation clinic cereneo, yielding 47 sessions and 1,304 raw ARAT tasks (trials)
187 (1,174 retained after outlier exclusion) in total. Three webcams (Logitech Brio 4K:
188 1280×720 px, 60 Hz) were set up ad-hoc around the assessment table (see Supple-
189 mentary Figure A1), while therapists used a tablet application to mark start and stop
190 of each individual trial and assign clinical ARAT scores in real time. All participants
191 provided informed consent for research use of their data; ethics approval was granted
192 by the local ethics committee (EKNZ, 2024-00196).

193 **Kinematic Reconstruction and Preprocessing**

194 Upper limb kinematics were reconstructed using a differentiable biomechanics algo-
195 rithm [17], previously validated against gold-standard OMC for upper limb reconstruc-
196 tion in people with stroke [23]. The resulting joint angle, angular velocity, wrist and

trunk position time series were low-pass filtered using a zero-phase fourth-order Butterworth filter at 5 Hz to remove high-frequency noise while preserving physiologically relevant movement dynamics. Trials were excluded if any derived metric fell outside three times the interquartile range of its respective ARAT score group ($IQR \times 3.0$).

Data Quality of Kinematic Reconstruction

Reconstruction quality was quantified using the mean reprojection error, which is commonly used as an internal measure of reconstruction quality in markerless motion capture systems [17, 23]. It is defined as the Euclidean distance in pixels (px) between the detected 2D keypoints in each camera view and the reprojection of the corresponding reconstructed 3D joint positions back onto that view. For each trial, the reprojection error was averaged across all keypoints, camera views, and frames to yield a single trial-level quality measure, with lower values indicating closer agreement between the biomechanical reconstruction and the observed image data.

To evaluate H1, we defined two criteria for adequate and robust reconstruction. First, for *adequacy*, the cohort-level mean reprojection error had to fall within the 10–20 px range previously established as acceptable for this reconstruction algorithm in clinical use, modestly above the 5–10 px laboratory baseline [23]; values within this range indicate reconstruction quality comparable to controlled clinical-routine conditions, while substantially larger errors would indicate inadequate quality. Second, for *robustness across impairment*, reconstruction quality had to be stable across the range of clinical ability, such that reprojection error did not depend systematically on ARAT score; a large, systematic increase in error with greater impairment would indicate performance-dependent reconstruction bias and reject this criterion. Reprojection error was compared across the four ARAT score groups using a Kruskal–Wallis test, with η_H^2 as the effect size and a negligible-to-small effect ($\eta_H^2 < 0.06$) taken as consistent with robustness; pairwise Dunn’s post-hoc tests with Holm correction

and rank-biserial correlation r characterised any group differences. H1 was considered supported if both the adequacy and robustness criteria were met.

Kinematic Metrics

Table 2: Standard ARAT task grouping and analysis groups used in the present study. Tasks sharing the same gross motor movement pattern were pooled for analysis; tasks with a unique movement pattern were analysed individually.

ARAT Subtest	Task	Analysis Group
Grasp	Block 10 cm	Grasp (pooled)
	Block 7.5 cm	
	Block 5 cm	
	Block 2.5 cm	
	Stone	
	Ball 7.5 cm	
Grip	Water pouring	Water Pouring (individual)
	Tube 2.25 cm	Grip (pooled)
	Tube 1 cm	
	Washer	
Pinch	Ball 6 mm (index)	Pinch (pooled)
	Ball 6 mm (middle)	
	Ball 6 mm (ring)	
	Ball 1.5 cm (index)	
	Ball 1.5 cm (middle)	
	Ball 1.5 cm (ring)	
Gross Movement	Hand behind head	Hand Behind Head (individual)
	Hand to mouth	Hand to Mouth (individual)
	Hand on top of head	Hand on Top of Head (individual)

Metric Selection

Movement-quality metric sets with established psychometric evidence in neurological populations currently exist for the drinking task [15, 16], but no equivalent evidence-based metric set exists for the ARAT. We therefore drew on the broader upper-limb kinematic literature, which has converged on a shared time-series substrate of joint angles, peak joint and end-effector (hand) velocities, and trunk displacement as a compensation indicator [4, 26], to construct a general-purpose metric set designed to

233 represent overall movement quality across all seven ARAT analysis groups rather than
234 to be optimised for any one of them.

235 Ten kinematic metrics were extracted at the trial level. The guiding principle was
236 to retain metrics that are robustly computable from a single trial without requiring
237 movement phase segmentation, yielding stable estimates that characterise upper limb
238 movement directly from the kinematic time series. This trial-level approach avoids any
239 additional error introduced by automated phase classification, which would propagate
240 uncertainty into every downstream metric. Metrics were derived from the kinematic
241 time series most consistently implicated in upper limb movement characterisation in
242 the literature: end-effector velocity, elbow angle and angular velocity, shoulder flexion
243 angle and angular velocity, shoulder abduction, and trunk displacement [15, 23]. From
244 these underlying time series, stable trial-level summary metrics were computed as
245 peak values and ranges of motion, capturing the magnitude and extent of movement
246 during each trial. Smoothness metrics were excluded as they are more susceptible to
247 kinematic jitter introduced by the reduced camera count in the clinical environment
248 relative to the laboratory setup. All metrics are listed in Figure 1 (raw metrics).

249 Task grouping

250 The analysis covered all 19 ARAT tasks, performed bilaterally yielding 38 task
251 instances. As it is good clinical practice to not evaluate a movement based on a sin-
252 gle repetition [27], kinematic metrics and clinical scores of the individual tasks were
253 averaged per ARAT subgroup wherever they were sharing the same gross movement
254 pattern. This resulted in 7 analysis groups (see Table 2). A secondary benefit of this
255 averaging is that the per-analysis-group clinical score becomes a continuous quantity
256 on $[0, 3]$ rather than the discrete ordinal values 0–3 of the individual tasks, providing
257 finer resolution that aligns more naturally with the continuous nature of the kinematic
258 compound scores in subsequent analyses.

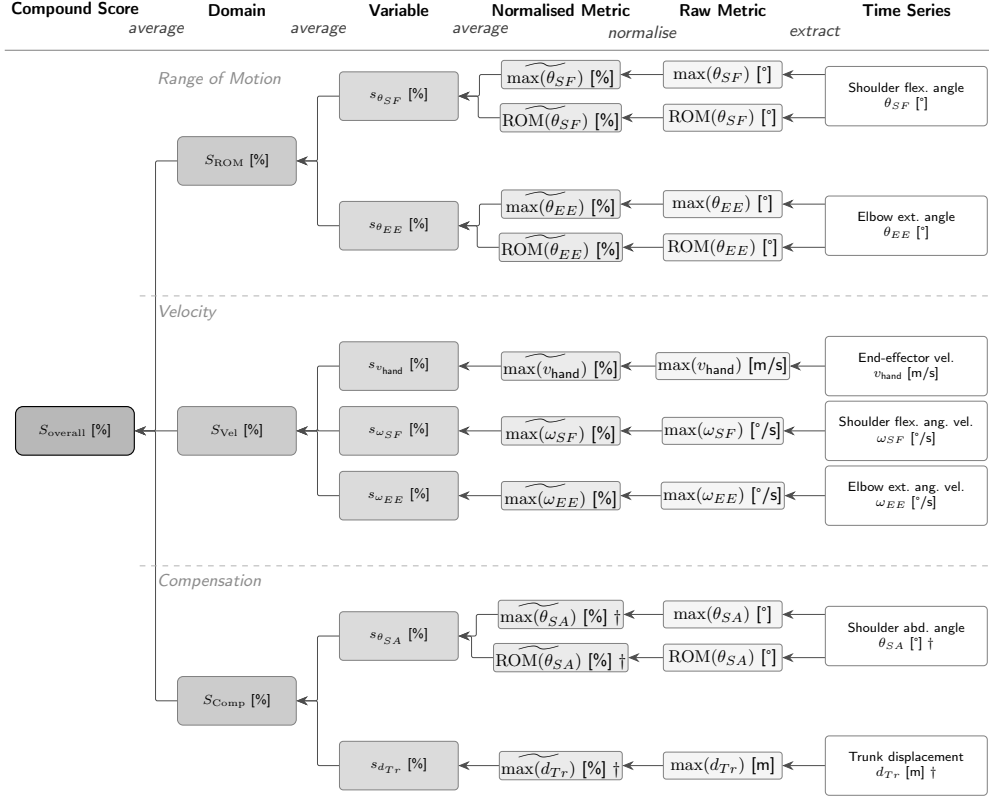


Fig. 1: Hierarchical construction of compound scores from kinematic metrics. Reading right to left: seven kinematic time series (right) are summarised by ten raw metrics (max or range of motion, ROM). Each raw metric is normalised to a Q75-anchored percentage scale on which 100 % corresponds to the per-metric 75th percentile (Q75) of Score 3 (clinically normal) cohort sessions and 0 % to the per-task cohort minimum, yielding a normalised metric $\tilde{m}$; values are not clipped and may exceed 100 % to indicate performance beyond the normal-performance threshold. Metrics marked with $\dagger$ are compensation metrics for which higher raw values indicate worse function; for these the normalisation is inverted (Q25 used as the 100 % anchor instead of Q75) so that higher normalised values consistently reflect better function across all metrics. Normalised metrics are averaged within each time series to yield seven variable scores; variable scores are averaged within each of three domains (Range of Motion, Velocity, Compensation); and the three domain scores are averaged to yield the overall compound score. All aggregated scores are expressed on the same Q75-anchored percentage scale. Background shading darkens with increasing aggregation, from raw inputs (white) to the overall compound score. Symbols: θ joint angle; ω joint angular velocity; v linear velocity; d linear displacement; subscripts denote joint (SF: shoulder flexion; EE: elbow extension; SA: shoulder abduction; Tr: trunk).

Compound Metrics

To support clinical interpretability, kinematic metrics were aggregated into compound scores summarizing movement across three domains (Range-of-Motion, Velocity, Compensation; see Figure 1). Referencing movement quality to a normal motor-performance benchmark is an established way to interpret upper-limb kinematic metrics [15]. Lacking a separate able-bodied cohort, we approximated this reference with Score 3 (clinically normal) sessions. Because Score 3 carries the scoring subjectivity this study documents, we anchored the normal threshold at the 75th percentile (Q75) of Score 3 rather than the median, a more conservative choice robust to that subjectivity.

Compound scores are expressed on this anchored scale: 100 % corresponds to the per-metric Q75 of Score 3 sessions and 0 % to the per-task cohort minimum. Values are not clipped and may exceed 100 %, indicating performance beyond the normal threshold. For metrics where higher values indicate worse performance (trunk displacement, shoulder abduction), Q25 replaces Q75 as the 100 % anchor, so higher scores always indicate better function. Where one kinematic variable is represented by multiple metrics (e.g. peak and ROM of shoulder flexion), those metrics were averaged into a single variable score so that each variable contributes equally within its domain (see Figure 1).

Domain scores were computed as the unweighted mean of their constituent variable scores: S_{ROM} (shoulder flexion, elbow extension), S_{Vel} (end-effector, shoulder flexion, elbow extension), and S_{Comp} (trunk displacement, shoulder abduction). The overall score (S_{overall}) is the unweighted mean of the three domain scores. Compound-score *levels* are reported in percent (%) while *changes* and the MCID threshold are reported in percentage points (pp).

284 Data Analysis

285 Construct Validity at the Population Level (H2)

286 Construct validity at the population level was evaluated through a two-tier known-
287 groups discrimination analysis [28] applied independently to each ARAT analysis
288 group (Table 2). *Tier 1* compared Score 0/1 (the task could not be completed) versus
289 Score 2/3 (the task was completed) trials, representing the clinically robust bound-
290 ary where strong discrimination was expected. *Tier 2* compared Score 2 (completed
291 with difficulty or abnormal time) versus Score 3 (completed normally) trials, probing
292 the more subjective upper boundary of the ARAT scale where inter-rater variability
293 is greatest and where attenuated discrimination would itself support construct valid-
294 ity by indicating that the metrics track the underlying functional construct rather
295 than the ordinal, clinical scale. Group differences for each metric were evaluated using
296 Mann–Whitney U tests with rank-biserial correlation r as the effect size measure,
297 and significance thresholds at $p < .05$, $p < .01$, and $p < .001$. Discriminative capacity
298 was further quantified using logistic regression with 5-fold stratified cross-validation,
299 with performance reported as AUC. To assess the relative contribution of different
300 movement domains, four feature sets were compared independently: ROM-only (four
301 metrics), Velocity-only (three metrics), Compensation-only (three metrics), and all
302 ten metrics combined. H2 was considered supported if Tier 1 combined-feature AUC
303 met the conventional “excellent discrimination” threshold of 0.85 [25] in the majority
304 of analysis groups, and Tier 2 combined-feature AUC remained above chance level
305 (AUC > 0.5) while consistently lower than Tier 1 across analysis groups.

306 Longitudinal Case Studies (H3)

307 Two participants with the most recorded sessions were selected to illustrate the longi-
308 tudinal behaviour of the kinematic metrics, examined in detail on the Grasp analysis
309 group: P10, followed over four sessions all below the clinical ARAT ceiling (sub-ceiling

window), and P20, followed over nine sessions across approximately four months,
 reaching the clinical ceiling at session 3 and remaining there afterwards (sub-ceiling
 window: sessions 1–3; post-ceiling window: sessions 3–9). For each participant, the
 per-analysis-group clinical ARAT score and the kinematic compound scores (S_{overall} ,
 S_{ROM} , S_{Vel} , S_{Comp}) were plotted across sessions alongside the individual metric
 trajectories. Given the few sessions per sub-ceiling window (3–4), inferential trend
 tests are not meaningful at this n ; we therefore characterised longitudinal agreement
 descriptively, benchmarking against an adopted MCID (15 pp on the compound scale)
 translated from consensus recommendations for kinematic measurement of upper limb
 quality after stroke [4]. As this threshold is adopted rather than estimated for the
 present measure and cohort, it is an approximate benchmark; no test–retest data were
 available to estimate a measurement-error-based threshold directly.

Sub-ceiling agreement and specificity (H3a, H3b). For each analysis group, the
 mean clinical ARAT score was rescaled to a common 0–100 axis across the sub-ceiling
 window, and the kinematic compound scores (already expressed on the normalised
 scale defined above) were used directly. Per-session change vectors for the clinical and
 each kinematic trajectory were then computed on this common percentage-point axis.
 The mean absolute deviation (MAD) between the clinical change vector and each
 kinematic change vector was reported as the index of longitudinal agreement, with
 smaller MAD indicating closer tracking of the clinical trajectory and larger MAD
 indicating divergence either in magnitude or direction. The S_{overall} MAD was used
 as the validity criterion, and domain compound MADs were additionally reported
 to illustrate patient-specific recovery profiles where present. H3a was evaluated per
 analysis group: the S_{overall} MAD was compared against the MCID (15 pp) separately
 for each analysis group, with analysis groups falling below MCID establishing longi-
 tudinal construct validity and exceedances expected to coincide with the metric-task
 mismatches identified cross-sectionally (H2). Where patient-specific recovery profiles

337 were present, domain MAD divergence then illustrated the added clinical value the
338 validated compound makes interpretable (H3b).

339 *Post-ceiling responsiveness (H3c).* For P20, the post-ceiling window (sessions in
340 which the clinical ARAT score had reached and remained at the maximum of 3) was
341 analysed separately. Changes were computed at two levels with distinct referents. For
342 raw individual kinematic metrics, $\Delta\%$ was expressed as a percentage of the cohort-
343 wide range of that metric ($\Delta\% = 100 \times (\text{last} - \text{first})/\text{cohort range}$). For compound
344 and domain scores, Δ was reported in percentage points (pp) on the normalised scale,
345 interpretable as the gap closed toward (or, when negative, opened from) the per-
346 metric normal motor-performance threshold. H3c was considered supported if at least
347 one domain compound score changed by ≥ 15 pp (one MCID) within the post-ceiling
348 window.

349 **3 Results**

350 **Data Set**

351 A total of 1,304 ARAT tasks (trials) were recorded across 47 assessment sessions from
352 20 participants. Participant characteristics are summarised in Supplementary Table
353 [A3](#); the cohort comprised 15 males and 5 females (mean age 71.2 years, standard
354 deviation (SD) 9.9, range 55–90) with diagnoses of stroke ($n=10$), Parkinson’s disease
355 ($n=6$), and other neurological conditions ($n=4$). Thirteen participants contributed
356 more than one session, with the two longest monitored participants followed over
357 approximately four months (9 sessions) and nine weeks (4 sessions), respectively (the
358 distribution of sessions per participant is shown in Supplementary Figure [A2](#)). Follow-
359 ing score-aware interquartile-range (IQR)-based outlier detection, 130 trials (10.0%)
360 were excluded, leaving 1,174 trials for analysis.

Data Quality of Kinematic Reconstruction (H1)

Across all analysed trials, the mean reprojection error of the biomechanical reconstruction was 12.5 ± 3.0 px (median 12.1 px, IQR 10.5–14.0 px). This falls within the acceptable 10–20 px range previously established for this reconstruction algorithm in clinical use, modestly above the 5–10 px laboratory standard [23] as expected given the reduced three-camera configuration and more frequent occlusions inherent to a clinical setting. The adequacy criterion for H1 is therefore met.

Reprojection error was largely stable across the four ARAT score groups (all trials pooled by their clinical ARAT score, 0–3, each trial contributing one mean reprojection-error value) (Table 3, Figure 2). Median reprojection error differed by at most 2.5 px across the four score groups (13.9 px for Score 0 vs. 11.4 px for Score 1, the largest pairwise gap), a small difference relative to the overall error distribution (IQR 3.5 px across all trials). A Kruskal–Wallis test indicated a statistically significant difference across score groups ($H(3) = 9.85$, $p = .020$), but the associated effect size was negligible ($\eta_H^2 = 0.0053$), indicating that score group explained essentially none of the variance in reconstruction quality and satisfying the robustness criterion ($\eta_H^2 < 0.06$). Pairwise Dunn’s tests with Holm correction revealed no significant differences between any score groups after correction (all $p_{\text{Holm}} > .05$), with effect sizes ranging from negligible to small ($|r| \leq 0.26$). The largest, albeit non-significant, difference involved Score 0 trials, which showed slightly higher median error (13.9 px) than Scores 1–3 (11.4–12.3 px), consistent with the greater movement variability and more frequent occlusions expected in the most impaired performances.

With both adequacy (12.5 px within the 10–20 px acceptable range) and robustness ($\eta_H^2 = 0.0053 < 0.06$) criteria met, H1 is supported: adequate and uniform kinematic reconstruction is achievable across the full range of clinical ability in a routine, reduced-camera clinical setting.

Construct Validity at the Population Level (H2)

We assessed construct validity through two-tier known-groups discrimination. Tier 1 (Score 0/1 vs. Score 2/3) probes the robust boundary between completed and uncompleted tasks, where strong discrimination is expected; Tier 2 (Score 2 vs. Score 3) probes the more subjective boundary between completions of differing quality, where discrimination lower than Tier 1 would itself support validity.

Tier 1: Score 0/1 vs. Score 2/3 Discrimination

Tier 1 discrimination between Score 0/1 and Score 2/3 trials was strong across six of seven analysis groups, with combined-feature AUC ranging from 0.91 to 1.00 (Table 4; Figure 3 illustrates the underlying metric distributions for all trials in the Grasp analysis group), demonstrating that the MMC-derived kinematic metrics reliably capture the fundamental distinction between task completion and failure to complete.

For pick-and-place tasks (Grasp, Grip, Pinch), combined-feature discrimination was near-perfect (AUC 0.96–>0.99, see Table 4). The most consistent individual

Table 3: Reprojection error of MMC by ARAT score group over all 1304 trials (pre outlier removal). Group-level descriptives (median, interquartile range), Kruskal–Wallis omnibus test, and pairwise Dunn’s post-hoc tests with Holm correction. r denotes the rank-biserial correlation as the pairwise effect size. Effect-size labels follow conventional thresholds: $\eta_H^2 < 0.01$ negligible, 0.01–0.06 small, 0.06–0.14 medium, > 0.14 large; $|r| < 0.10$ negligible, 0.10–0.30 small, 0.30–0.50 medium, ≥ 0.50 large. Significance: $*p < .05$, $**p < .01$, $***p < .001$, n.s. = not significant.

Score	n	Group descriptives				Pairwise (Dunn’s, Holm-corrected)				
		Median (px)	IQR (px)	Mean (px)	SD (px)	Pair	Δ_{med} (px)	r	$ r $	p_{Holm}
0	28	13.90	[10.73, 16.27]	14.04	3.97	0 vs 1	-2.55	-0.257	small	0.059
1	113	11.36	[10.05, 14.32]	12.16	3.32	0 vs 2	-2.07	-0.249	small	0.096
2	650	11.84	[10.47, 13.79]	12.47	3.00	0 vs 3	-1.58	-0.229	small	0.236
3	513	12.32	[10.74, 13.96]	12.61	2.82	1 vs 2	+0.48	+0.061	negl.	0.365
Kruskal–Wallis: $H(3)=9.85$, $p=0.020$						1 vs 3	+0.97	+0.107	small	0.236
$\eta_H^2=0.0053$ (negligible)						2 vs 3	+0.49	+0.062	negl.	0.236

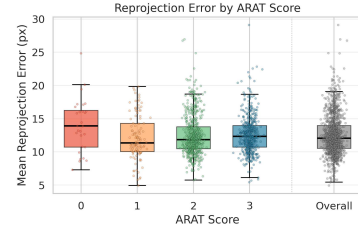


Fig. 2: Mean per-trial reprojection error (pixels) by clinical ARAT score (0–3) and pooled across all 1304 trials (right). Largely overlapping distributions confirm that reconstruction quality does not systematically depend on patient impairment level ($\eta_H^2=0.0053<0.06$; Table 3).

discriminators were elbow extension metrics (both ROM $r = 0.77$ – 0.93 and peak velocity $r = 0.82$ – 0.93), reflecting the requirement to achieve near-full arm extension to transport the object to the target. Score 2/3 trials also showed substantially higher end-effector and joint velocities than Score 0/1 trials (see Supplementary Table A1). ROM metrics discriminated strongly on their own (ROM-only AUC ≥ 0.95), with velocity-only somewhat lower (0.90 – 0.94); combining the domains yielded the highest discrimination in every case (AUC ≥ 0.96).

Gross motor tasks (Hand Behind Head, Hand on Top of Head, Hand to Mouth) similarly achieved high combined-feature discrimination (AUC 0.91 – 1.00 ; per-group metric distributions are shown in Supplementary Figure A3, A4), with the dominant single domain varying across tasks (ROM for Hand to Mouth, velocity for Hand Behind Head, compensation for Hand on Top of Head) and the combined feature set matching or exceeding the best single domain in every case. Hand to Mouth showed the strongest discrimination of all analysis groups (AUC = 1.00), with Score 0/1 trials characterised by higher shoulder flexion.

Trunk displacement was a significant discriminator across all pick-and-place and gross motor tasks, with Score 0/1 trials showing greater trunk compensation. As a standalone domain, compensation metrics discriminated Score 0/1 from Score 2/3 trials well above chance across all analysis groups (compensation-only AUC 0.72 – 0.97); for Grip (compensation-only AUC 0.96) and Hand to Mouth (0.97) compensation nearly matched ROM, and for Hand on Top of Head it was the strongest single domain (0.87 vs. ROM 0.84 , velocity 0.86), indicating that compensatory trunk and shoulder recruitment alone carries substantial information about whether a task can be completed.

The Water Pouring task was an exception, showing worse performance than the other tasks (combined-feature AUC 0.75) which is attributable to metric-task mismatch rather than reconstruction failure: the current metric set and metrics category

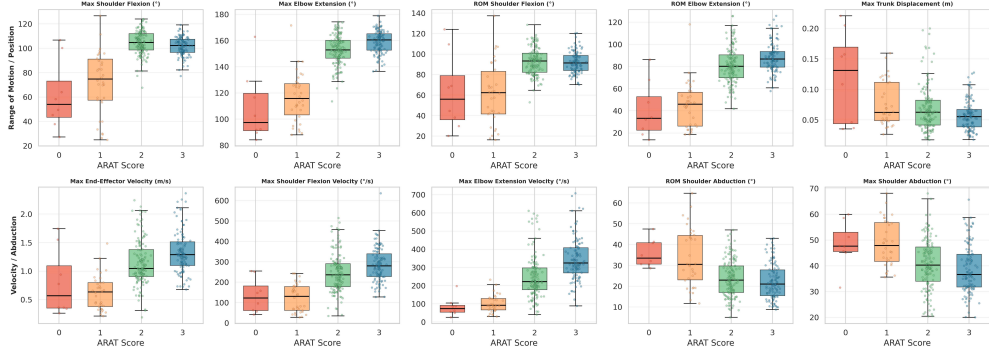


Fig. 3: Distribution of kinematic metrics across clinical ARAT scores (0–3) for all trials in the Grasp analysis group. Individual trials shown as jittered points. Outliers removed ($\text{IQR} \times 3.0$).

assignment (ROM, VEL, COMP) is mostly optimised for reach-and-transport movements. Such per-task construct-validity limitations bound the generalisation of H2 but do not affect the other analysis groups in which the metric set is well-matched to the task.

Together, Tier 1 discrimination exceeded the $\text{AUC} \geq 0.85$ “excellent discrimination” threshold [25] in six of seven analysis groups (AUC 0.91–1.00), supporting H2 at the Score 0/1 vs. Score 2/3 boundary.

Tier 2: Score 2 vs. Score 3 Discrimination

As anticipated, Tier 2 discrimination between Score 2 and Score 3 trials was substantially lower than Tier 1 across all analysis groups (combined-feature AUC 0.70–0.84; Table 4), reflecting reasonable overlap between the two groups. Compared to Tier 1, velocity metrics contributed more strongly to discrimination in Tier 2, sometimes achieving higher discriminative ability than ROM alone, for example, in the Grasp group (velocity-only: 0.75 vs. ROM-only: 0.70). This reflects the overall faster movement execution in Score 3 trials compared to Score 2 (see Supplementary Table A1).

444 For pick-and-place tasks (AUC 0.77–0.84), maximum elbow extension velocity
445 showed the largest individual discriminative ability. Trunk displacement remained a
446 significant discriminator for Grip and Pinch, but not for Grasp. Gross motor tasks
447 showed slightly lower Tier 2 discrimination (AUC 0.70–0.82) compared to pick-and-
448 place tasks, consistent with their lower performance in Tier 1; velocity metrics also
449 dominated for gross motor tasks. In Water Pouring, velocity metrics now drove
450 discrimination (AUC 0.83), reversing the pattern observed in Tier 1.

451 Velocity metrics were a more dominant domain in Tier 2 than in Tier 1, suggesting
452 that once the task is completed, movement speed strongly differentiates performance
453 quality. Compensation metrics were the weakest standalone domain at the Score 2 vs.
454 Score 3 boundary (compensation-only AUC 0.59–0.68 across most tasks). This likely
455 reflects the limited sensitivity of the present general-purpose compensation metrics
456 (trunk displacement, shoulder abduction) to the subtler compensatory differences that
457 distinguish good from normal completions, rather than an absence of compensatory
458 information at this boundary [29–31]; the per-task suitability of the metric set is
459 considered further in the Discussion. Hand to Mouth was the notable exception, where
460 compensation remained the strongest single domain (AUC 0.73), driven by persistent
461 shoulder abduction and trunk displacement in slower, more effortful completions, the
462 one analysis group where the compensatory metrics distinguish Score 2 from Score 3
463 completions.

464 Tier 2 discrimination (AUC 0.70–0.84) remained above chance in all seven analysis
465 groups and reached the conventional “acceptable discrimination” threshold of 0.7 [25]
466 throughout, while remaining consistently lower than Tier 1 across the full set, the
467 differential pattern expected of a valid construct measure tracking the underlying
468 function rather than the ordinal scale. Combined, the Tier 1 and Tier 2 results support
469 H2.

Table 4: Known-groups construct validity across all seven task groups. Top: pick-and-place tasks (Grasp, Grip, Pinch); bottom: gross-motor and pouring tasks (Water Pouring, Hand Behind Head, Hand on Top of Head, Hand to Mouth). Rank-biserial r from Mann–Whitney U tests. LR AUC = logistic regression area under the ROC curve using the feature subset indicated. * $p < .05$, ** $p < .01$, *** $p < .001$.

Pick-and-place tasks						
Metric	Grasp		Grip		Pinch	
	Tier 1 (42 vs 248)	Tier 2 (140 vs 108)	Tier 1 (12 vs 249)	Tier 2 (110 vs 139)	Tier 1 (44 vs 271)	Tier 2 (194 vs 77)
$\max(\theta_{SF})$ [°]	0.79***	-0.20**	0.62***	-0.31***	0.73***	-0.34***
$\max(\theta_{EE})$ [°]	0.89***	0.33***	0.95***	0.25***	0.93***	-0.01
$\text{ROM}(\theta_{SF})$ [°]	0.61***	-0.03	0.38*	-0.02	0.44***	0.11
$\text{ROM}(\theta_{EE})$ [°]	0.88***	0.28***	0.93***	0.15*	0.77***	0.08
$\max(d_{Tr})$ [m]	-0.23*	-0.17*	-0.88***	-0.23**	-0.62***	-0.31***
$\max(v_{\text{hand}})$ [m/s]	0.74***	0.30***	0.61***	0.37***	0.63***	0.26***
$\max(\omega_{SF})$ [°/s]	0.77***	0.28***	0.53**	0.37***	0.58***	0.25**
$\max(\omega_{EE})$ [°/s]	0.89***	0.52***	0.93***	0.45***	0.82***	0.33***
$\text{ROM}(\theta_{SA})$ [°]	-0.50***	-0.10	-0.52**	-0.14	-0.27**	-0.04
$\max(\theta_{SA})$ [°]	-0.50***	-0.14	-0.66***	-0.25***	-0.25**	-0.05
LR AUC (vel. only)	0.94	0.75	0.94	0.72	0.90	0.62
LR AUC (ROM only)	0.95	0.70	0.97	0.77	0.96	0.70
LR AUC (comp.)	0.76	0.61	0.96	0.63	0.82	0.62
LR AUC (all)	0.96	0.78	>0.99	0.84	0.99	0.77

Gross-motor and pouring tasks								
Metric	Water Pouring		Hand Behind Head		Hand on Top of Head		Hand to Mouth	
	Tier 1 (5 vs 70)	Tier 2 (44 vs 26)	Tier 1 (12 vs 58)	Tier 2 (43 vs 15)	Tier 1 (8 vs 71)	Tier 2 (35 vs 36)	Tier 1 (3 vs 81)	Tier 2 (20 vs 61)
$\max(\theta_{SF})$ [°]	0.65*	0.06	0.81***	-0.01	0.77***	0.30*	-1.00***	-0.04
$\max(\theta_{EE})$ [°]	-0.19	0.15	-0.25	-0.47**	-0.21	-0.21	0.04	-0.12
$\text{ROM}(\theta_{SF})$ [°]	0.34	0.13	0.71***	0.06	0.64**	0.23	-0.97***	-0.08
$\text{ROM}(\theta_{EE})$ [°]	-0.42	-0.18	-0.14	-0.30	-0.36	-0.26	0.04	-0.12
$\max(d_{Tr})$ [m]	-0.55*	0.14	-0.67***	-0.08	-0.81***	-0.29*	-0.96***	-0.47**
$\max(v_{\text{hand}})$ [m/s]	-0.31	0.55***	0.86***	0.40*	0.78***	0.61***	0.64	0.47**
$\max(\omega_{SF})$ [°/s]	0.03	0.61***	0.80***	0.39*	0.72***	0.46***	-0.18	0.30*
$\max(\omega_{EE})$ [°/s]	-0.40	0.24	0.55**	0.15	0.37	0.35*	0.69*	0.47**
$\text{ROM}(\theta_{SA})$ [°]	-0.66*	-0.34*	-0.16	-0.18	-0.38	-0.10	-0.64	0.12
$\max(\theta_{SA})$ [°]	-0.09	-0.10	-0.15	0.04	-0.41	-0.04	-0.98***	-0.39**
LR AUC (vel. only)	0.69	0.83	0.92	0.66	0.86	0.75	0.86	0.72
LR AUC (ROM only)	0.92	0.64	0.91	0.69	0.84	0.61	0.98	0.57
LR AUC (comp.)	0.75	0.60	0.72	0.68	0.87	0.59	0.97	0.73
LR AUC (all)	0.75	0.83	0.97	0.70	0.91	0.73	1.00	0.82

Exploratory Longitudinal Case Studies (H3)

We illustrate longitudinal behaviour in two participants: P10 (four sub-ceiling sessions) and P20 (nine sessions, reaching the clinical ceiling at session 3), with per-session trajectories shown in Figures 4 and 5. As detailed in Methods, longitudinal change is reported descriptively as the mean absolute deviation (MAD) between clinical and kinematic change vectors, benchmarked against the adopted MCID (15 pp).

476 We evaluate longitudinal agreement (MAD) across the six analysis groups retained
477 after excluding Water Pouring, and illustrate the recovery-profile contrast in detail on
478 the Grasp analysis group.

479 Water Pouring is excluded from the longitudinal case-study analyses below for the
480 metric-task mismatch discussed at Tier 1. The remaining six analysis groups (Grasp,
481 Grip, Pinch, Hand Behind Head, Hand on Top of Head, Hand to Mouth) are retained.

482 **Sub-Ceiling: Overall Concordance with Patient-Specific Domain** 483 **Divergence (H3a, H3b)**

484 The longitudinal case studies provide two related forms of evidence. First, the overall
485 kinematic compound score (S_{overall}) tracked the clinical ARAT score closely across
486 sub-ceiling sessions in both patients, indicating that the overall kinematic summary
487 captures the same functional change as the clinical score (longitudinal construct
488 validity). Second, beneath this consistent overall agreement, the domain compound
489 scores (range of motion S_{ROM} , velocity S_{Vel} , and compensation S_{Comp}) diverged from
490 the clinical trajectory in patient-specific ways, exposing which aspects of movement
491 actually drove each patient’s improvement, information the overall score and the
492 ordinal ARAT necessarily compress away. The two findings only acquire their mean-
493 ing together: the overall agreement validates the metric, and that validation is what
494 licenses the domain divergence to be read as genuine recovery profile rather than
495 measurement disagreement. A per-analysis-group S_{overall} MAD below the MCID was
496 taken as evidence of longitudinal construct validity for that analysis group.

497 We illustrate this pattern in detail on the Grasp analysis group and summarise
498 the remaining five analysis groups in Supplementary Table A2. For Grasp, S_{overall}
499 tracked the clinical ARAT trajectory closely for both patients (MAD $S_{\text{overall}} = 5.6$ pp
500 for P10, 5.9 pp for P20, both well within the 15 pp MCID), supporting longitudinal

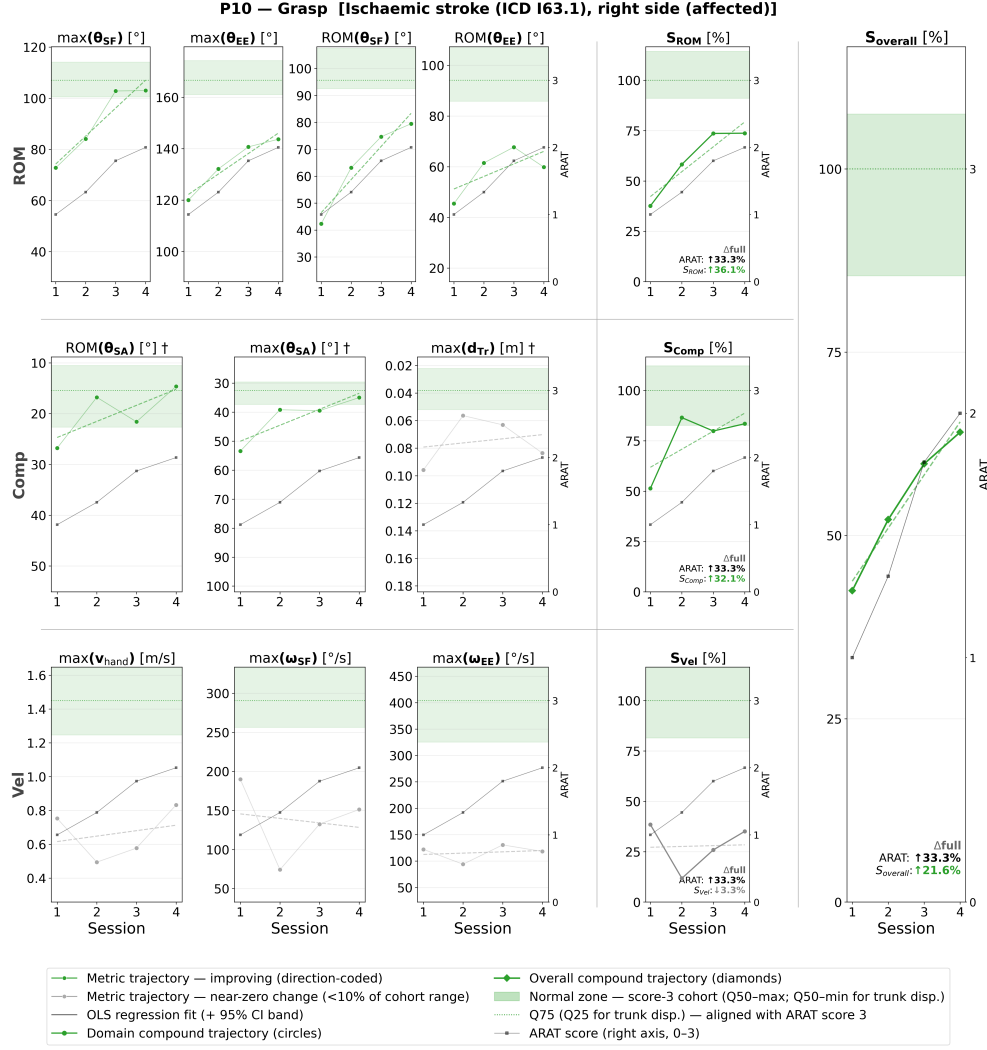


Fig. 4: Longitudinal kinematic monitoring of P10’s affected side (ischaemic stroke) across four ARAT sessions, Grasp analysis group. *Panel grid:* rows are the three kinematic domains (ROM, Compensation, Velocity), each with its raw metrics, a domain compound score (S_{ROM} , S_{Comp} , S_{Vel}), and at far right the overall compound ($S_{Overall}$). Compound panels are shown on the normalised 0–100 % scale (100 % = per-metric Q75 of Score 3 sessions, Q25 for inverted compensation metrics marked †; see Figure 1); values are unclipped and may exceed 100 %. Raw-metric panels show absolute units. The patient’s mean clinical ARAT score across the Grasp analysis group tasks (0–3) is overlaid on the secondary axis (black squares). *Reference bands:* green shading marks the normal motor-performance zone (Q50–max of Score 3 sessions; Q50–min for inverted metrics); the dotted line marks Q75, aligned with ARAT score 3. *Annotation:* each compound panel reports Δ_{full} , the first-to-last-session change (pp), for both ARAT and the compound; green ≥ 15 pp (MCID), red ≤ -15 pp, grey otherwise. P10 stayed below the clinical ceiling throughout, so the full window is entirely sub-ceiling.

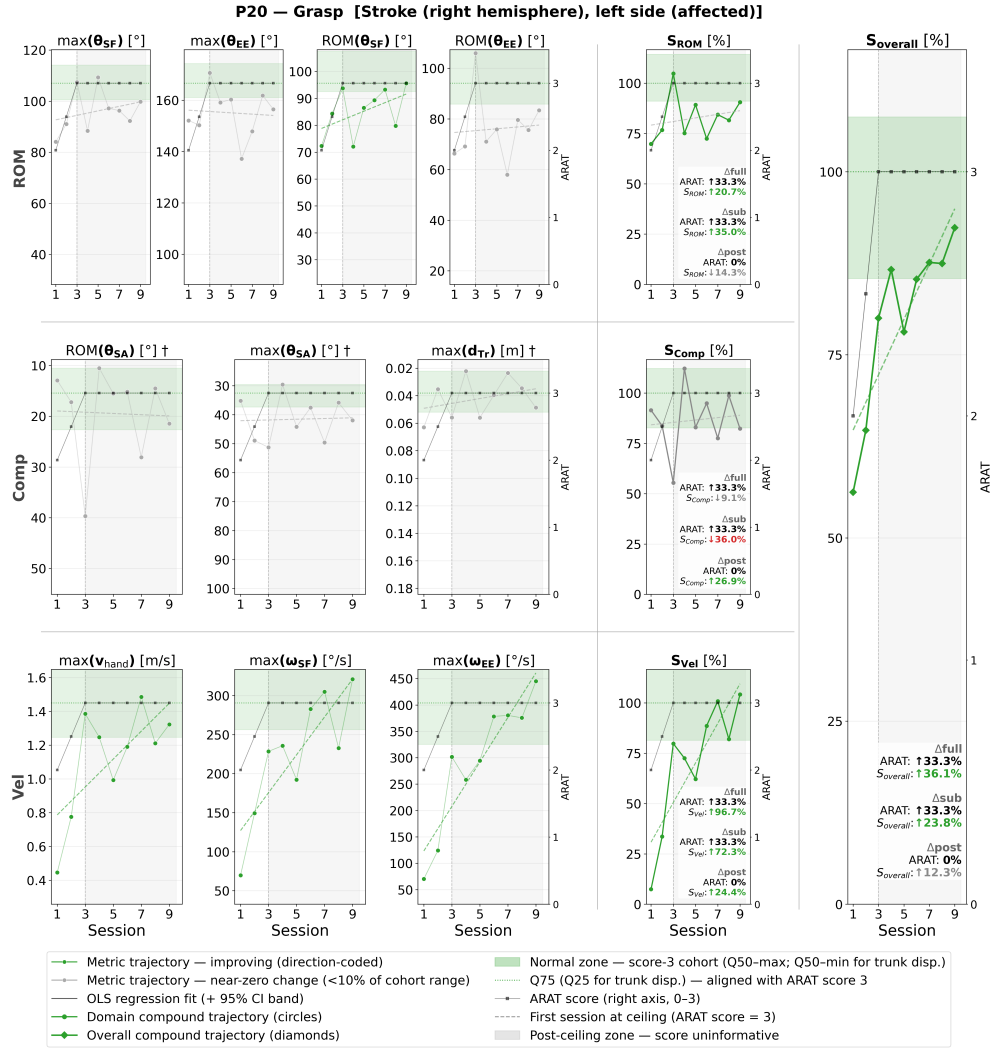


Fig. 5: Longitudinal kinematic monitoring of P20's affected side (stroke, right hemisphere) across nine ARAT sessions over approximately four months, Grasp analysis group. Panel grid, scale, reference bands, and the overlaid mean clinical ARAT score as in Figure 4. The vertical dashed line marks the first session at which the clinical ceiling was reached and held; the grey background marks the post-ceiling zone. *Annotation:* each compound panel reports three changes (pp) for both ARAT and the compound: Δ_{full} (first to last), Δ_{sub} (sub-ceiling, to first ceiling-attaining session), and Δ_{post} (post-ceiling, from that session to last); by construction $\Delta_{ARAT, post} = 0$. Colouring as in Figure 4.

501 construct validity. S_{overall} itself rose by $\Delta S_{\text{overall}} = +21.6$ pp for P10 over the four sub-
 502 ceiling sessions and by $+23.8$ pp for P20, against clinical ARAT changes of $+33.3$ pp
 503 for both patients on the rescaled 0–100 axis. Beneath this agreement, the two patients
 504 showed mechanistically opposite domain profiles. P10’s improvement was driven by
 505 gains in joint range of motion ($\Delta S_{\text{ROM}} = +36.1$ pp; MAD = 5.4 pp) and reduced com-
 506 pensatory movement ($\Delta S_{\text{Comp}} = +32.1$ pp; MAD = 6.8 pp, mostly through reduced
 507 shoulder abduction, $\Delta \geq 18^\circ$ from first to last session), while velocity remained essen-
 508 tially flat ($\Delta S_{\text{Vel}} = -3.3$ pp; MAD = 28.4 pp), a range-of-motion-driven recovery
 509 without a speed component. P20 showed the opposite profile on the same task: veloc-
 510 ity rose strongly ($\Delta S_{\text{Vel}} = +72.3$ pp; MAD = 16.2 pp) while compensation diverged
 511 ($\Delta S_{\text{Comp}} = -36.0$ pp; MAD = 31.3 pp), a recovery driven by faster, more vigorous
 512 execution with compensation worsening over the window.

513 The same pattern of overall concordance with patient-specific domain divergence
 514 held across the remaining five analysis groups (Supplementary Table A2). Evaluated
 515 per analysis group, the S_{overall} MAD fell below the MCID in all six analysis groups
 516 for P20 (range 5.9–13.6 pp) and in three of six for P10 (Grasp, Hand to Mouth,
 517 Pinch; range 5.6–8.5 pp). The three P10 exceedances (Grip 18.2 pp; Hand Behind
 518 Head 17.4 pp; Hand on Top of Head 18.1 pp) fell precisely in the analysis groups
 519 where the general-purpose metric set is known to be undersized: Grip, where finger
 520 kinematics are absent, and the two overhead tasks, where the set is optimised for reach-
 521 and-transport rather than overhead positioning, paralleling the metric-task mismatch
 522 flagged for Water Pouring at Tier 1. The exceedances thus reflect the same task-metric
 523 mismatch identified cross-sectionally rather than a breakdown of longitudinal validity,
 524 and do not affect the Grasp analysis group result on which the recovery contrast rests.

525 These case studies are consistent with H3a and H3b: the per-analysis-group valid-
 526 ity criterion (S_{overall} MAD below the MCID, H3a) was met in both patients in the
 527 analysis groups where the metric set matched the task (all six for P20; Grasp, Hand to

Mouth, and Pinch for P10). Within this validated frame, the patient-specific domain divergence (H3b), most cleanly illustrated by the opposite ROM-and-Compensation versus Velocity recovery profiles on Grasp, exposes recovery profiles that S_{overall} and the ordinal ARAT necessarily compress away: S_{overall} confirms what the clinical scale measures where the metric set matches the task, and the domain compounds explain it where patient-specific profiles are present. As exploratory evidence from two patients, these results motivate cohort-level investigation of the coupled finding.

Post-Ceiling: Continued Kinematic Recovery After the Clinical Scale Saturates (H3c)

After P20 first reached the maximum mean clinical ARAT score of 3 in the Grasp analysis group at session 3, the clinical score remained at ceiling for all subsequent sessions, providing no further information about recovery. Kinematic metrics, however, continued to change throughout the remaining six sessions (see Figure 5). On the normalised compound scale, S_{overall} rose by a further $\Delta S_{\text{overall}} = +12.3$ pp across the post-ceiling window. Velocity metrics showed the strongest post-ceiling change: S_{Vel} increased by $\Delta S_{\text{Vel}} = +24.4$ pp during the post-ceiling window, exceeding the 15 pp MCID. ROM metrics showed smaller positive changes (shoulder flexion ROM: $+27.8\%$ of cohort range, raw metric), and trunk displacement decreased modestly (-8.5% of cohort range, raw metric). P20 surpassed the normal motor-performance threshold of the velocity compound only several sessions after reaching the clinical score ceiling, indicating that kinematic normalisation toward the per-metric normal motor-performance was achieved only towards the end of monitoring, a distinction entirely invisible to the ordinal clinical scale. In this single post-ceiling case, kinematic measurement therefore detected clinically relevant change after the ordinal scale had saturated, consistent with H3c.

553 4 Discussion

554 Effective and efficient neurorehabilitation requires therapy to be steered by what is
555 actually measured, and precision neurorehabilitation has been proposed as the frame-
556 work to achieve this [32]. Its central bottleneck, however, is measurement [4]. This
557 study evaluated whether AI-based markerless motion capture, used inside routine
558 ARAT assessments in a neurorehabilitation clinic, accurately and robustly recon-
559 structs upper limb movement and yields kinematic metrics that are valid and clinically
560 informative beyond the ordinal score, and could therefore help close that measurement
561 loop at scale.

562 We addressed this through four hypotheses, spanning reconstruction accuracy
563 (H1), cross-sectional construct validity (H2), and longitudinal validity, specificity, and
564 beyond-ceiling responsiveness in two exploratory case studies (H3a–H3c). All were
565 supported within their pre-specified thresholds: H1 and H2 held at the cohort level,
566 and the longitudinal criterion (H3a) held in all six analysis groups for P20 and three
567 of six for P10. We discuss each in turn, then the limitations and future directions the
568 work opens.

569 Reconstruction accuracy is the precondition for every later claim, and it held where
570 it is most likely to fail: despite an ad-hoc three-camera setup in routine sessions,
571 reprojection error stayed within the acceptable clinical-use range and did not degrade
572 with impairment, making the underlying signal trustworthy enough to build metrics
573 on.

574 Validating kinematic measures against a clinical score is complicated by the obsta-
575 cle Schwarz et al. [26] identify: kinematic metrics provide complementary rather than
576 redundant information, so classical convergent correlation with the ordinal scale can-
577 not be the test. A known-groups approach [28] sidesteps this by asking not whether
578 the metrics agree with the ARAT but whether they agree *in the way a valid measure*
579 *should*. They did: discrimination was strong where the clinical boundary is robust

(Tier 1, completed vs. uncompleted) and attenuated where it is subjective (Tier 2, quality of completion). This differential pattern, rather than uniform agreement, is the signature of a continuous measure tracking the underlying construct and generalises to any setting where a continuous measure must be validated against a noisy ordinal reference.

The longitudinal case studies most directly motivate clinical deployment beyond the ordinal score. On the Grasp analysis group, the overall compound tracked the clinical ARAT closely in both patients, while their domain decompositions revealed mechanistically opposite recovery profiles: both gained the same ARAT points, yet one recovered through range of motion and reduced compensation and the other through faster execution. Overall agreement and domain decomposition are interpretable only together: the overall agreement legitimises the compound as tracking the same signal as the ARAT, which is what lets the domain decomposition be read as recovery profile rather than measurement noise. The same clinical improvement was thus driven by different movement profiles, directly addressing the third ARAT limitation. Crucially, kinematic measurement remained responsive beyond the clinical ceiling: in P20, the compound captured continued, velocity-driven recovery across the post-ceiling window that the saturated ARAT could not register, and resolved its profile, not merely that overall improvement continued but also in which domain.

Several considerations bound these results, each with a clear path forward now that the measurement infrastructure is demonstrated to be viable. The longitudinal evidence rests on two case studies, motivating but not establishing the coupled finding at the cohort level; a prospective cohort study with structured assessment schedules is the natural next step. The MCID adopted here is an approximation translated from consensus recommendations [4] rather than estimated for this measure and cohort; a larger cohort would allow movement-specific cut-offs and measure-specific reliability bounds to be established directly, for which criterion-anchored estimation against

607 patient-perceived change provides a template [33]. ARAT scores were also assigned by
608 treating therapists without parallel inter-rater reliability data, leaving the contribution
609 of clinical subjectivity to the Tier 2 attenuation unquantified. A final, cross-cutting
610 limitation is that the added insight depends on each metric’s domain assignment hold-
611 ing for the task at hand: because every feature is assigned to a single domain across
612 the full ARAT battery, the same feature can mean different things across tasks, and
613 this surfaces wherever the assignment breaks down. Cross-sectionally, Water Pouring
614 showed the lowest Tier 1 discrimination (combined-feature AUC 0.75), as shoulder
615 abduction, assigned as compensation for reach-and-transport, may instead be a legiti-
616 mate pouring strategy; longitudinally, the same mismatch excluded Water Pouring
617 and drove P10’s three MAD exceedances (Grip, lacking finger kinematics; the two
618 overhead tasks, assigned for reach-and-transport rather than overhead positioning).
619 Where a metric’s domain assignment is loosely coupled to a task’s actual recovery,
620 the compound tracks the clinical trajectory less faithfully; the principled response is
621 to surface such cases rather than report misleading agreement. This is not an accu-
622 racy limitation but a representational one: the present hand-crafted, domain-assigned
623 metrics are a deliberately general first representation rather than the last word.

624 Beyond extending the hand-crafted metric set, more fundamental measurement-
625 side directions follow naturally and are made viable by the routine acquisition
626 demonstrated here. Scoring the whole movement by its deviation from an able-bodied
627 reference, rather than reducing it to pre-chosen scalars, sidesteps the per-task assign-
628 ment problem, since the reference itself encodes what healthy execution of each task
629 looks like; this is established practice in gait analysis, where indices such as the Gait
630 Deviation Index quantify deviation from an able-bodied reference set [34], and is
631 beginning to be explored for the upper limb [35]. Further still, because the underlying
632 reconstruction yields full biomechanical trajectories rather than isolated metrics, the
633 data are directly consumable by emerging movement foundation models that learn

the shared structure of human movement and answer clinical questions, including outcome scoring, without task-specific feature engineering [36]. Realising these data-driven directions requires able-bodied and patient data at population scale, and would also replace the present Score 3 normal motor-performance proxy with measured able-bodied references; the routine, low-burden acquisition demonstrated here is one route to generating that volume of data. In parallel, the underlying tracking can be enriched along a hardware axis: returning to a five-camera configuration, as used in laboratory validation, would support reliable finger kinematics and reduce occlusion-driven artefacts [37, 38], while at the opposite extreme single-camera reconstruction [39–41] could lower the deployment barrier further toward ubiquitous, in-room measurement.

5 Conclusion

The present results represent a first exploration of a new measurement territory in which full kinematic reconstruction during routine clinical assessment becomes possible through passive camera instrumentation. Passive continuous kinematic measurement during routine assessment opens a new mode of motor evaluation: observer-independent, multi-dimensional, sensitive past the clinical ceiling, capable of revealing recovery profiles that ordinal scoring cannot, and able to generate the volume and continuity of structured data needed to bring modern data-driven methods into routine practice and close the loop for precision upper-limb motor neurorehabilitation.

Supplementary information. Supplementary Material is available in the Appendix A.

Acknowledgements. We thank all participants of the study.

6 Declarations

- Funding: This study was enabled through generous funding from the P & K Foundation.
- Conflict of interest/Competing interests: The authors declare no competing interests.
- Ethics approval and consent to participate: This study was approved by the cantonal ethics committee (EKNZ, 2024-00196). All participants provided informed consent for research use of their data.
- Consent for publication: Not applicable.
- Data availability: The datasets generated and analysed in this study contain potentially identifiable clinical and video-derived data and are therefore not publicly available owing to data-protection and participant-privacy constraints. De-identified data that can be shared in compliance with the applicable data-protection regulations and ethical approvals are available from the corresponding author on reasonable request.
- Materials availability: Not applicable
- Code availability: Analysis code available upon request.
- Author contribution: T.U. conceived and designed the study, developed the analysis pipeline, performed the data analysis, interpreted the results, and drafted the manuscript. C.E.A. supervised the project, contributed to the conception and design of the study, and revised the manuscript. R.J.C. provided technical guidance on the markerless motion capture and biomechanical reconstruction methodology and revised the manuscript. A.R.L. provided clinical oversight and access to the clinical site at which the data were collected, and revised the manuscript. O.L. and R.G. contributed to the scientific framing and methodological direction of the work and revised the manuscript. All authors reviewed and approved the submitted version.

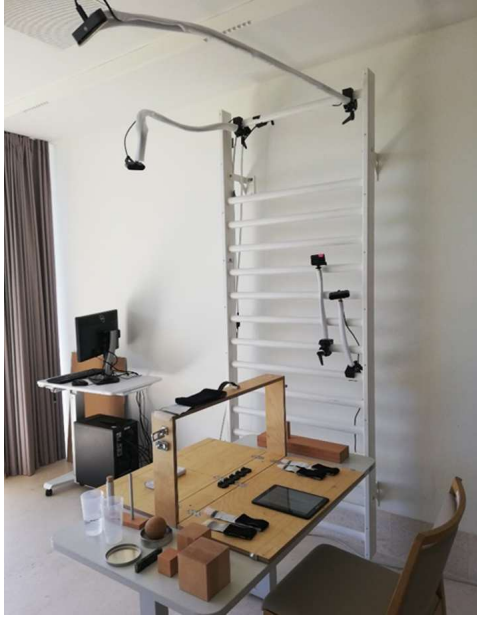


Fig. A1: Ad-hoc camera setup around the ARAT assessment table.

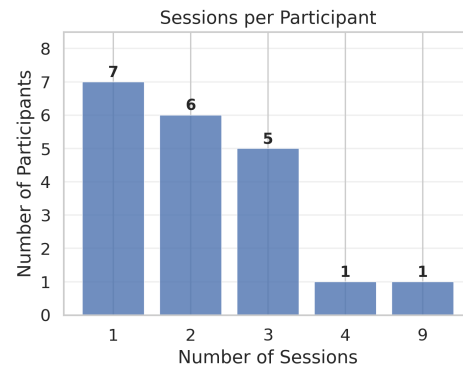
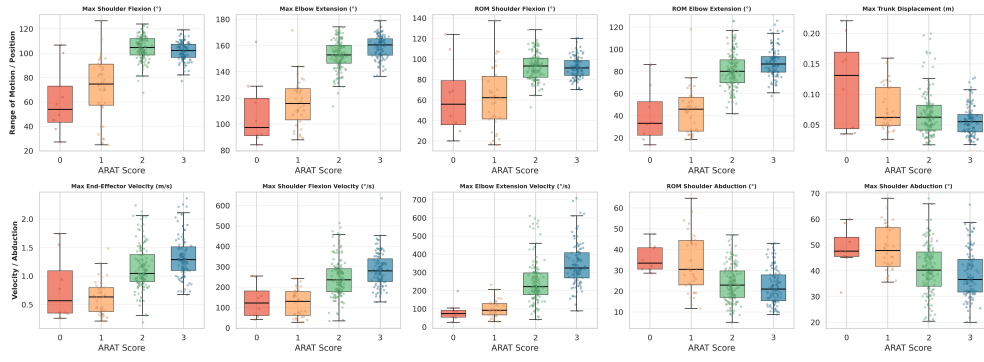


Fig. A2: Sessions per participant.

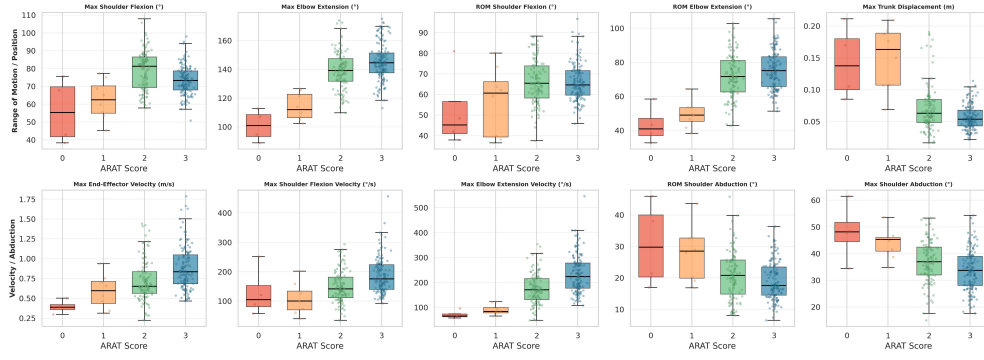
Table A1: Population descriptive statistics across all seven task groups. Top: pick-and-place tasks (Grasp, Grip, Pinch); bottom: gross-motor and pouring tasks (Water Pouring, Hand Behind Head, Hand on Top of Head, Hand to Mouth). Raw kinematic metric rows show the median across trials in raw units (degrees, m, m/s); compound score rows show the median with the interquartile range $[Q_{25}, Q_{75}]$ on the Q75-anchored 0–100 % scale, where 100 % corresponds to the per-metric 75th percentile of Score 3 cohort sessions and values are not clipped at the upper bound (i.e., compound scores may exceed 100 % when performance surpasses the normal-performance threshold). Inverted compensation metrics (trunk displacement and shoulder abduction) use Q25 as the 100 % anchor so that higher normalised values always indicate better function.

<i>Pick-and-place tasks</i>												
Metric	Grasp				Grip				Pinch			
	S0 (8)	S1 (34)	S2 (140)	S3 (108)	S0 (4)	S1 (8)	S2 (110)	S3 (139)	S0 (10)	S1 (34)	S2 (194)	S3 (77)
$\max(\theta_{SF})$ [°]	53.9	74.8	104.8	102.2	55.4	62.5	81.3	73.3	34.4	88.9	109.3	101.3
$\max(\theta_{EE})$ [°]	97.4	115.7	152.9	160.4	100.9	111.9	139.2	144.6	99.7	117.9	153.4	152.8
$\max(\theta_{SA})$ [°]	47.6	47.8	40.3	36.6	48.1	45.3	36.9	33.7	32.9	47.0	42.1	41.0
$\text{ROM}(\theta_{SF})$ [°]	56.0	62.4	93.3	91.3	45.2	60.7	65.4	64.6	34.9	56.9	92.1	93.8
$\text{ROM}(\theta_{EE})$ [°]	33.2	45.9	80.1	86.8	40.9	49.0	71.6	75.1	29.8	55.8	83.1	83.3
$\text{ROM}(\theta_{SA})$ [°]	33.5	30.5	22.9	21.1	29.7	28.5	20.8	17.6	21.3	29.0	24.1	21.9
$\max(d_{Tr})$ [m]	0.131	0.062	0.063	0.055	0.137	0.163	0.063	0.054	0.096	0.114	0.062	0.048
$\max(v_{hand})$ [m/s]	0.569	0.635	1.05	1.29	0.388	0.596	0.652	0.836	0.362	0.733	0.969	1.18
$\max(\omega_{SF})$ [°/s]	121.9	130.5	236.2	280.2	105.0	100.2	141.7	175.9	150.8	154.5	209.1	255.9
$\max(\omega_{EE})$ [°/s]	73.9	91.8	222.0	324.8	65.8	84.0	170.8	223.6	55.4	108.8	220.7	288.2
S_{ROM} [%]	34	50	89	92	32	45	86	85	24	59	92	93
IQR	[19, 48]	[27, 67]	[82, 97]	[87, 100]	[17, 47]	[39, 52]	[71, 100]	[75, 97]	[18, 69]	[30, 80]	[85, 100]	[85, 98]
S_{Vel} [%]	24	32	62	81	25	38	55	74	26	40	66	84
IQR	[10, 43]	[13, 38]	[50, 84]	[68, 97]	[19, 33]	[21, 46]	[43, 70]	[60, 102]	[14, 39]	[25, 56]	[49, 87]	[63, 100]
S_{Comp} [%]	63	64	81	87	52	52	80	89	84	63	81	85
IQR	[39, 71]	[53, 81]	[70, 91]	[78, 97]	[32, 63]	[40, 61]	[68, 92]	[79, 96]	[66, 96]	[53, 70]	[71, 90]	[74, 95]
S_{Overall} [%]	33	46	78	85	34	44	75	83	46	55	79	88
IQR	[32, 43]	[34, 56]	[72, 84]	[80, 94]	[30, 38]	[36, 48]	[67, 82]	[74, 94]	[40, 57]	[39, 63]	[74, 88]	[81, 95]

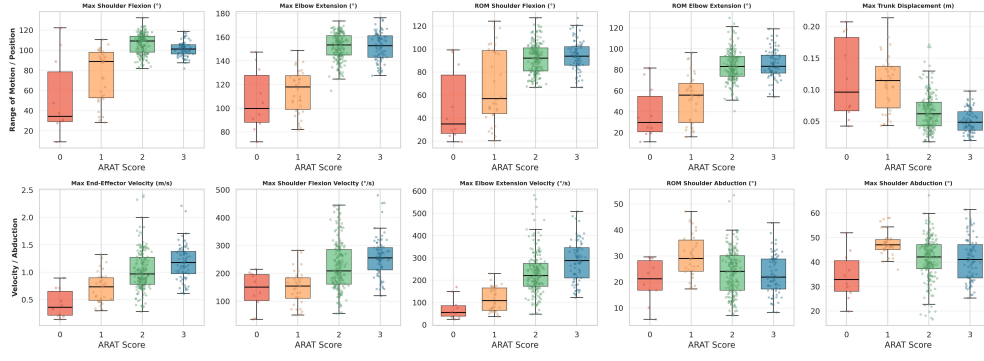
<i>Gross-motor and pouring tasks</i>												
Metric	Water				HBH				HOH			
	S0 (1)	S1 (4)	S2 (44)	S3 (26)	S0 (0)	S1 (12)	S2 (43)	S3 (15)	S0 (0)	S1 (8)	S2 (35)	S3 (36)
$\max(\theta_{SF})$ [°]	–	45.3	64.1	67.7	–	78.6	125.6	126.4	–	99.9	118.3	129.5
$\max(\theta_{EE})$ [°]	–	105.1	93.8	97.4	–	135.8	132.0	117.6	–	148.6	134.3	129.7
$\max(\theta_{SA})$ [°]	–	32.2	34.0	30.7	–	48.9	46.9	46.3	–	55.8	46.2	45.7
$\text{ROM}(\theta_{SF})$ [°]	–	44.6	52.7	58.3	–	72.0	120.2	124.1	–	95.1	114.3	120.5
$\text{ROM}(\theta_{EE})$ [°]	–	30.9	18.8	16.1	–	101.4	92.8	83.8	–	91.0	84.0	75.0
$\text{ROM}(\theta_{SA})$ [°]	–	22.6	15.0	11.0	–	41.0	34.2	30.9	–	48.2	33.3	30.5
$\max(d_{Tr})$ [m]	–	0.060	0.044	0.049	–	0.050	0.030	0.028	–	0.044	0.028	0.026
$\max(v_{hand})$ [m/s]	–	0.609	0.463	0.591	–	1.06	1.93	2.45	–	1.20	1.81	2.45
$\max(\omega_{SF})$ [°/s]	–	117.7	99.9	126.0	–	212.4	365.6	459.7	–	211.6	334.0	398.4
$\max(\omega_{EE})$ [°/s]	–	134.2	47.4	58.7	–	210.8	285.5	286.5	–	170.2	217.1	305.0
S_{ROM} [%]	–	86	74	80	–	89	103	84	–	73	85	77
IQR	–	[51, 119]	[59, 104]	[63, 91]	–	[42, 105]	[87, 118]	[77, 95]	–	[63, 102]	[69, 106]	[71, 93]
S_{Vel} [%]	–	125	53	85	–	37	73	81	–	36	58	81
IQR	–	[92, 135]	[42, 73]	[70, 98]	–	[27, 49]	[51, 87]	[77, 95]	–	[28, 46]	[31, 76]	[66, 93]
S_{Comp} [%]	–	77	87	94	–	76	97	97	–	60	83	87
IQR	–	[63, 85]	[71, 97]	[80, 100]	–	[63, 97]	[81, 105]	[91, 99]	–	[49, 64]	[71, 95]	[79, 98]
S_{Overall} [%]	–	92	74	84	–	70	88	87	–	52	73	84
IQR	–	[73, 105]	[64, 84]	[76, 94]	–	[59, 74]	[83, 93]	[83, 95]	–	[49, 71]	[62, 84]	[74, 89]



(a) Grasp

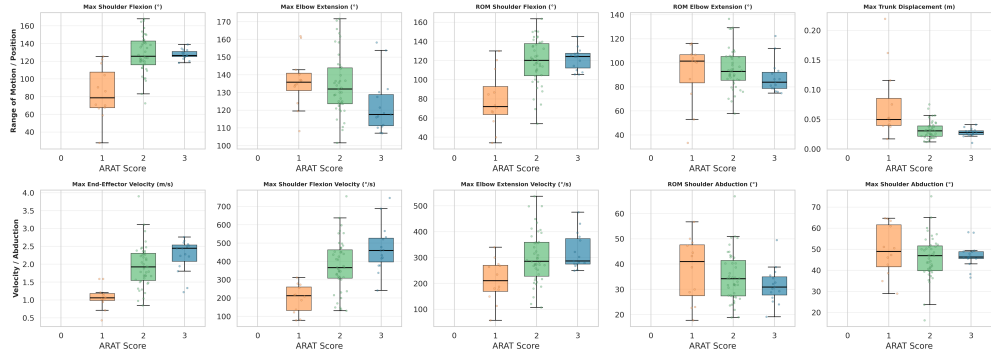


(b) Grip

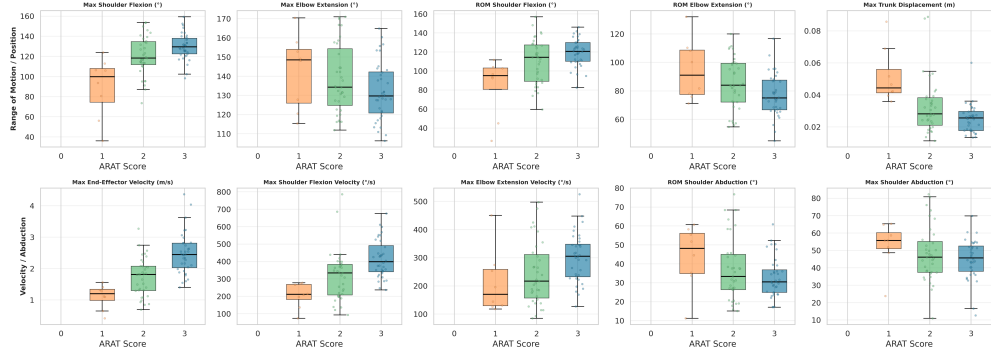


(c) Pinch

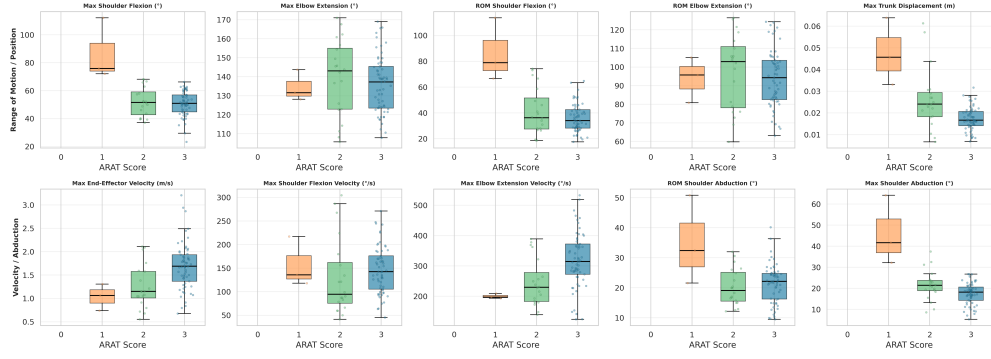
Fig. A3: Distribution of kinematic metrics across clinical ARAT scores (0–3) for pick-and-place tasks per analysis group. Individual trials shown as jittered points. Outliers removed ($IQR \times 3.0$).



(a) Hand behind head



(b) Hand on top of head



(c) Hand to mouth

Fig. A4: Distribution of kinematic metrics across clinical ARAT scores (0–3) for gross movement tasks per analysis group. Individual trials shown as jittered points. Outliers removed ($IQR \times 3.0$).

Table A2: Mean absolute deviation (MAD, in percentage points on the shared Q75-anchored 0–100 % scale) between change vectors of the clinical ARAT score and the kinematic compound scores across the sub-ceiling window of each patient, for all six Analysis Groups. On this scale, 100 % corresponds to the per-metric 75th percentile of Score 3 cohort sessions and values are not clipped at the upper bound; the Compensation domain is inverted during normalisation (Q25 used as the 100 % anchor instead of Q75) so that higher values consistently indicate less compensatory movement. Smaller MAD indicates closer agreement between the kinematic compound and the clinical trajectory; larger MAD indicates divergence in either magnitude or direction. S_{Overall} = overall compound score; S_{ROM} , S_{Vel} , S_{Comp} = domain compound scores. Reference threshold: 15 % = minimal clinically important difference (MCID) [4], expressed in pp on the same Q75-anchored scale. Sessions n is the number of sub-ceiling sessions per task group.

Task	Sessions	n	S_{Overall} (% pts)	S_{ROM} (% pts)	S_{Vel} (% pts)	S_{Comp} (% pts)
<i>P10 — Ischaemic stroke (ICD I63.1), right side (affected)</i>						
Grasp	1-4	4	5.6	5.4	28.4	6.8
Grip	1-4	4	18.2	10.2	27.0	26.6
Pinch	1-4	4	8.5	6.2	12.5	11.1
Hand Behind Head	1-4	4	17.4	44.0	24.3	38.6
Hand on Top of Head	1-3	3	18.1	24.5	20.6	9.1
Hand to Mouth	1-3	3	7.1	8.1	10.9	13.3
<i>P20 — Stroke (right hemisphere), left side (affected)</i>						
Grasp	1-3	3	5.9	3.8	16.2	31.3
Grip	1-3	3	6.3	3.9	2.1	14.4
Pinch	1-3	3	12.1	13.4	4.0	21.8
Hand Behind Head	1-3	3	13.6	12.3	39.2	1.4
Hand on Top of Head	1-3	3	8.0	17.7	17.9	12.8
Hand to Mouth	1-3	3	7.2	22.5	10.6	15.3

Characteristic	Value
Age (years), mean $\pm$ SD	71.2 $\pm$ 9.9
Age range (years)	55–90
Sex: male	15 (75%)
Height (cm), mean $\pm$ SD	174.1 $\pm$ 6.5
Weight (kg), mean $\pm$ SD	74.9 $\pm$ 12.5
Diagnosis: stroke	10 (50%)
Diagnosis: Parkinson’s disease	6 (30%)
Diagnosis: other	4 (20%)

Table A3: Participant characteristics ($N=20$).

References

- [1] GBD 2021 Nervous System Disorders Collaborators: Global, regional, and national burden of disorders affecting the nervous system, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet. Neurology* **23**(4), 344–381 (2024) [https://doi.org/10.1016/S1474-4422\(24\)00038-3](https://doi.org/10.1016/S1474-4422(24)00038-3)
- [2] Feigin, V.L., Brainin, M., Norrving, B., Martins, S.O., Pandian, J., Lindsay, P., F Grupper, M., Rautalin, I.: World Stroke Organization: Global Stroke Fact Sheet 2025. *International Journal of Stroke* **20**(2), 132–144 (2025) <https://doi.org/10.1177/17474930241308142> . Accessed 2026-03-18
- [3] Nakayama, H., Jørgensen, H.S., Raaschou, H.O., Olsen, T.S.: Recovery of upper extremity function in stroke patients: The Copenhagen stroke study. *Archives of Physical Medicine and Rehabilitation* **75**(4) (1994) [https://doi.org/10.1016/0003-9993\(94\)90161-9](https://doi.org/10.1016/0003-9993(94)90161-9)
- [4] Kwakkel, G., Van Wegen, E., Burridge, J., Winstein, C., Dokkum, L., Alt Murphy, M., Levin, M., Krakauer, J.: Standardized measurement of quality of upper limb movement after stroke: Consensus-based core recommendations from the Second Stroke Recovery and Rehabilitation Roundtable. *International Journal of Stroke* **14**(8), 783–791 (2019) <https://doi.org/10.1177/1747493019873519> . Accessed 2026-01-05
- [5] Kwakkel, G., Lannin, N.A., Borschmann, K., English, C., Ali, M., Churilov, L., Saposnik, G., Winstein, C., Wegen, E.E., Wolf, S.L., Krakauer, J.W., Bernhardt, J.: Standardized measurement of sensorimotor recovery in stroke trials: Consensus-based core recommendations from the Stroke Recovery and Rehabilitation Roundtable. *International Journal of Stroke* **12**(5), 451–461 (2017) <https://doi.org/10.1177/1747493017711813> . Accessed 2026-01-05

- 708 [6] Lee, J.H.: Action Research Arm Test (ARAT). In: Occupational Therapy
709 Assessments for Older Adults. Occupational Therapy Assessments for Older
710 Adults: 100 Instruments for Measuring Occupational Performance, pp. 213–
711 214. Taylor and Francis, ??? (2024). <https://doi.org/10.4324/9781003525288-87>
712 . <https://www.scopus.com/pages/publications/85199034662> Accessed 2026-02-16
- 713 [7] Demers, M., Levin, M.F.: Do Activity Level Outcome Measures Commonly
714 Used in Neurological Practice Assess Upper-Limb Movement Quality? Neurore-
715 habilitation and Neural Repair **31**(7), 623–637 (2017) [https://doi.org/10.1177/](https://doi.org/10.1177/1545968317714576)
716 [1545968317714576](https://doi.org/10.1177/1545968317714576)
- 717 [8] Pike, S., Lannin, N.A., Wales, K., Cusick, A.: A systematic review
718 of the psychometric properties of the Action Research Arm Test in
719 neurorehabilitation. Australian Occupational Therapy Journal **65**(5),
720 449–471 (2018) <https://doi.org/10.1111/1440-1630.12527> . _eprint:
721 <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1440-1630.12527>. Accessed
722 2026-03-09
- 723 [9] Nordin, A., Alt Murphy, M., Danielsson, A.: Intra-rater and inter-rater reliability
724 at the item level of the Action Research Arm Test for patients with stroke. Jour-
725 nal of Rehabilitation Medicine **46**(8), 738–745 (2014) [https://doi.org/10.2340/](https://doi.org/10.2340/16501977-1831)
726 [16501977-1831](https://doi.org/10.2340/16501977-1831)
- 727 [10] Prange-Lasonder, G.B., Murphy, M.A., Lamers, I., Hughes, A.M., Buurke, J.H.,
728 Feys, P., Keller, T., Klamroth-Marganska, V., Tarkka, I.M., Timmermans, A.,
729 Burridge, J.H.: European evidence-based recommendations for clinical assess-
730 ment of upper limb in neurorehabilitation (CAULIN): data synthesis from
731 systematic reviews, clinical practice guidelines and expert consensus. Journal
732 of NeuroEngineering and Rehabilitation **18**(1) (2021) <https://doi.org/10.1186/>

- [11] Yozbatiran, N., Der-Yeghiaian, L., Cramer, S.C.: A Standardized Approach to Performing the Action Research Arm Test. *Neurorehabilitation and Neural Repair* **22**(1), 78–90 (2008) <https://doi.org/10.1177/1545968307305353> . Accessed 2026-06-02
- [12] Burridge, J., Alt Murphy, M., Buurke, J., Feys, P., Keller, T., Klamroth-Marganska, V., Lamers, I., McNicholas, L., Prange, G., Tarkka, I., Timmermans, A., Hughes, A.-M.: A Systematic Review of International Clinical Guidelines for Rehabilitation of People With Neurological Conditions: What Recommendations Are Made for Upper Limb Assessment? *Frontiers in Neurology* **10**, 567 (2019) <https://doi.org/10.3389/fneur.2019.00567> . Accessed 2026-06-02
- [13] Hervé-Colas, J., Newton, S.P., Engelter, S.T., Hayward, K.S., Held, J.P.O., Interling, N., Kwakkel, G., Pohl, J., Reisman, D.S., Schwarz, A., Sunnerhagen, K.S., Veerbeek, J.M., Wiesner, K., Zandvliet, S.B., Alt Murphy, M.: Standardized International Manual of the Fugl-Meyer Assessment of Motor Function After Stroke. *Neurorehabilitation and Neural Repair* **40**(4), 306–319 (2026) <https://doi.org/10.1177/15459683251412300>
- [14] Wilson, N., Howel, D., Bosomworth, H., Shaw, L., Rodgers, H.: Analysing the Action Research Arm Test (ARAT): a cautionary tale from the RATULS trial. *International Journal of Rehabilitation Research. Internationale Zeitschrift Fur Rehabilitationsforschung. Revue Internationale De Recherches De Readaptation* **44**(2), 166–169 (2021) <https://doi.org/10.1097/MRR.0000000000000466> . Accessed 2026-02-16
- [15] Alt Murphy, M., Murphy, S., Persson, H.C., Bergström, U.-B., Sunnerhagen, K.S.: Kinematic Analysis Using 3D Motion Capture of Drinking Task in People With

- 758 and Without Upper-extremity Impairments. Journal of Visualized Experiments :
759 JoVE (133), 57228 (2018) <https://doi.org/10.3791/57228> . Accessed 2026-01-05
- 760 [16] Murphy, M.A., Willén, C., Sunnerhagen, K.S.: Responsiveness of upper extremity
761 kinematic measures and clinical improvement during the first three months after
762 stroke. Neurorehabilitation and Neural Repair **27**(9), 844–853 (2013) <https://doi.org/10.1177/1545968313491008>
763
- 764 [17] Cotton, R.J.: Differentiable Biomechanics Unlocks Opportunities for Mark-
765 erless Motion Capture. In: 2025 International Conference On Rehabilitation
766 Robotics (ICORR), pp. 44–51 (2025). <https://doi.org/10.1109/ICORR66766.2025.11063174> . ISSN: 1945-7901. <https://ieeexplore.ieee.org/document/11063174>
767
768 Accessed 2026-07-23
- 769 [18] Pagnon, D., Domalain, M., Reveret, L.: Pose2Sim: An End-to-End Workflow
770 for 3D Markerless Sports Kinematics—Part 2: Accuracy. Sensors **22**(7) (2022)
771 <https://doi.org/10.3390/s22072712>
- 772 [19] D’Souza, S., Siebert, T., Fohanno, V.: A comparison of lower body gait kinematics
773 and kinetics between Theia3D markerless and marker-based models in healthy
774 subjects and clinical patients. Scientific Reports **14**(1), 29154 (2024) <https://doi.org/10.1038/s41598-024-80499-8> . Accessed 2026-07-13
775
- 776 [20] Matthis, J.: FreeMoCap - Free Motion Capture for Everyone (2026). <https://freemocap.org> Accessed 2026-05-29
777
- 778 [21] Uhlich, S.D., Falisse, A., Kidziński, L., Muccini, J., Ko, M., Chaudhari, A.S.,
779 Hicks, J.L., Delp, S.L.: OpenCap: Human movement dynamics from smartphone
780 videos. PLOS Computational Biology **19**(10), 1011462 (2023) <https://doi.org/10.1371/journal.pcbi.1011462> . Accessed 2026-07-13
781

- 782 [22] Cuevas-Velasquez, H., Yiannakidis, A., Shin, S., Becherini, G., Höschle, M.,
783 Tesch, J., Obersat, T., Alexiadis, T., Halilaj, E., Black, M.J.: MAMMA: Marker-
784 less & Automatic Multi-Person Motion Action Capture. arXiv. arXiv:2506.13040
785 [cs.CV] (2026). <https://doi.org/10.48550/arXiv.2506.13040> . [http://arxiv.org/](http://arxiv.org/abs/2506.13040)
786 [abs/2506.13040](http://arxiv.org/abs/2506.13040) Accessed 2026-07-13
- 787 [23] Unger, T., Moslehian, A.S., Peiffer, J.D., Ullrich, J., Gassert, R., Lambercy,
788 O., Cotton, R.J., Easthope, C.A.: Differentiable Biomechanics for Markerless
789 Motion Capture in Upper Limb Stroke Rehabilitation: A Comparison With Opti-
790 cal Motion Capture. IEEE Transactions on Medical Robotics and Bionics, 1–1
791 (2025) <https://doi.org/10.1109/TMRB.2025.3605962> . Accessed 2025-12-17
- 792 [24] Unger, T., Yao, X., Schmitt, A., Cebulla, L., Easthope Awa, C.: Towards routine
793 biomechanical data collection in stroke rehabilitation: a usability comparison of
794 IMU and markerless motion capture systems for functional upper-limb assess-
795 ments. Journal of NeuroEngineering and Rehabilitation (2026) [https://doi.org/](https://doi.org/10.1186/s12984-026-02075-7)
796 [10.1186/s12984-026-02075-7](https://doi.org/10.1186/s12984-026-02075-7) . Accessed 2026-07-24
- 797 [25] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression,
798 3rd edn. Wiley series in probability and statistics. Wiley, Hoboken, New Jersey
799 (2013)
- 800 [26] Schwarz, A., Kanzler, C.M., Lambercy, O., Luft, A.R., Veerbeek, J.M.: Sys-
801 tematic Review on Kinematic Assessments of Upper Limb Movements After
802 Stroke. Stroke **50**(3), 718–727 (2019) [https://doi.org/10.1161/STROKEAHA.](https://doi.org/10.1161/STROKEAHA.118.023531)
803 [118.023531](https://doi.org/10.1161/STROKEAHA.118.023531)
- 804 [27] Frykberg, G.E., Grip, H., Murphy, M.A.: How many trials are needed in kine-
805 matic analysis of reach-to-grasp?—A study of the drinking task in persons with
806 stroke and non-disabled controls. Journal of NeuroEngineering and Rehabilitation

- 807 **18**(1), 101 (2021) <https://doi.org/10.1186/s12984-021-00895-3>
- 808 [28] Vet, H.C.W., Terwee, C.B., Mokkink, L.B., Knol, D.L.: Measurement in
809 Medicine: A Practical Guide. Practical Guides to Biostatistics and Epidemiol-
810 ogy. Cambridge University Press, Cambridge (2011). [https://doi.org/10.1017/](https://doi.org/10.1017/CBO9780511996214)
811 [CBO9780511996214](https://doi.org/10.1017/CBO9780511996214) . [https://www.cambridge.org/core/books/measurement-in-](https://www.cambridge.org/core/books/measurement-in-medicine/8BD913A1DA0ECCBA951AC4C1F719BCC5)
812 [medicine/8BD913A1DA0ECCBA951AC4C1F719BCC5](https://www.cambridge.org/core/books/measurement-in-medicine/8BD913A1DA0ECCBA951AC4C1F719BCC5) Accessed 2026-06-03
- 813 [29] Unger, T., Kühnis, B., Sauerzopf, L., Spiess, M.R.R., Spindler, A., Luft, A.R.R.,
814 Easthope Awai, C., Schönhammer, J.G.G., Gavagnin, E.: Using deep learning to
815 detect upper limb compensation in individuals post-stroke using consumer-grade
816 webcams—A feasibility study. *Frontiers in Medicine* **12** (2025) [https://doi.org/](https://doi.org/10.3389/fmed.2025.1645369)
817 [10.3389/fmed.2025.1645369](https://doi.org/10.3389/fmed.2025.1645369) . Accessed 2026-07-13
- 818 [30] Sauerzopf, L., Panduro, C.G.C., Luft, A.R., Kühnis, B., Gavagnin, E., Unger,
819 T., Awai, C.E., Schönhammer, J.G., Degenfellner, J., Spiess, M.R.: Evaluating
820 inter- and intra-rater reliability in assessing upper limb compensatory move-
821 ments post-stroke: creating a ground truth through video analysis? *Journal of*
822 *NeuroEngineering and Rehabilitation* **21**(1), 217 (2024) [https://doi.org/10.1186/](https://doi.org/10.1186/s12984-024-01506-7)
823 [s12984-024-01506-7](https://doi.org/10.1186/s12984-024-01506-7) . Accessed 2025-01-07
- 824 [31] Pohl, J., Mayrhuber, L., Lamercy, O., Awai, C.E.: From Metrics to Meaning in
825 Neurological Rehabilitation: Clinicians’ Perspectives on Digital Metrics of Upper
826 Limb Functioning—A Focus Group Study. *JMIR Rehabilitation and Assistive*
827 *Technologies* **13**(1), 87339 (2026) <https://doi.org/10.2196/87339> . Accessed 2026-
828 07-14
- 829 [32] Adans-Dester, C., Hankov, N., O’Brien, A., Vergara-Diaz, G., Black-Schaffer, R.,
830 Zafonte, R., Dy, J., Lee, S.I., Bonato, P.: Enabling precision rehabilitation inter-
831 ventions using wearable sensors and machine learning to track motor recovery. *npj*

- 832 Digital Medicine **3**(1), 121 (2020) <https://doi.org/10.1038/s41746-020-00328-w> .
833 Accessed 2026-01-05
- 834 [33] Pohl, J., Verheyden, G., Held, J.P.O., Luft, A.R., Easthope Awai, C., Veerbeek,
835 J.M.: Construct validity and responsiveness of clinical upper limb measures and
836 sensor-based arm use within the first year after stroke: a longitudinal cohort
837 study. *Journal of NeuroEngineering and Rehabilitation* **22**(1), 14 (2025) <https://doi.org/10.1186/s12984-024-01512-9> . Accessed 2026-06-02
838
- 839 [34] Schwartz, M.H., Rozumalski, A.: The Gait Deviation Index: a new comprehensive
840 index of gait pathology. *Gait & Posture* **28**(3), 351–357 (2008) <https://doi.org/10.1016/j.gaitpost.2008.05.001>
841
- 842 [35] Zaidi, K.F., Harris-Love, M.: Upper extremity kinematics: development of a quan-
843 titative measure of impairment severity and dissimilarity after stroke. *PeerJ* **11**,
844 16374 (2023) <https://doi.org/10.7717/peerj.16374> . Accessed 2026-06-02
- 845 [36] Yang, R., Kennedy, A., Cotton, R.J.: BiomechGPT: Towards a Biomechanically
846 cally Fluent Multimodal Foundation Model for Clinically Relevant Motion Tasks.
847 arXiv. arXiv:2505.18465 [cs] (2025). <https://doi.org/10.48550/arXiv.2505.18465>
848 . <http://arxiv.org/abs/2505.18465> Accessed 2025-06-04
- 849 [37] Maggioni, V., Azevedo-Coste, C., Durand, S., Bailly, F.: Optimisation and Com-
850 parison of Markerless and Marker-Based Motion Capture Methods for Hand
851 and Finger Movement Analysis. *Sensors (Basel, Switzerland)* **25**(4), 1079 (2025)
852 <https://doi.org/10.3390/s25041079> . Accessed 2026-01-05
- 853 [38] Firouzabadi, P., Murray, W., Sobinov, A.R., Peiffer, J.D., Shah, K., Miller, L.E.,
854 Cotton, R.J.: Biomechanical Arm and Hand Tracking with Multiview Markerless
855 Motion Capture. In: 2024 10th IEEE RAS/EMBS International Conference for

- Biomedical Robotics and Biomechatronics (BioRob), pp. 1641–1648. IEEE, Heidelberg, Germany (2024). <https://doi.org/10.1109/BioRob60516.2024.10719940> .
<https://ieeexplore.ieee.org/document/10719940/> Accessed 2024-11-07
- [39] Donahue, S., Peiffer, J.D., Richardson, R.T., Zhong, Y., Tan, S.Q.Y., Marteau, B.L., Russo, S.A., Wang, M.D., Cotton, R.J., Chafetz, R.: Monocular Markerless Motion Capture Enables Quantitative Assessment of Upper Extremity Reachable Workspace. *Sensors* **26**(11), 3421 (2026) <https://doi.org/10.3390/s26113421> . Accessed 2026-06-02
- [40] Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Shin, S., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., Kitani, K.: SAM 3D Body: Robust Full-Body Human Mesh Recovery. *arXiv*. arXiv:2602.15989 [cs.CV] (2026). <https://doi.org/10.48550/arXiv.2602.15989> . <http://arxiv.org/abs/2602.15989> Accessed 2026-07-13
- [41] Cotton, R.J., Firouzabadi, P., Murray, W.: Monocular Biomechanical Tracking of Fingers with Inverse Kinematics to Foundation Models. *arXiv* (2026). <https://doi.org/10.48550/arXiv.2605.09258> . <http://arxiv.org/abs/2605.09258> Accessed 2026-06-02