

Original Article

Agreement of a National Routing Scale with Validated Measures of Change in Pediatric Rehabilitation

Syed Gufran Sadiq Zaidi¹, Ihteshamul Hayat², Saule Karibzhanova¹, Zeeshan¹¹School of Medicine, Nazarbayev University, Astana, Kazakhstan²Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan

Received: Oct 11, 2026

Accepted: Oct 20, 2026

Corresponding author's email:

sgs.zaidi@nu.edu.kz

Citation: Zaidi SGS, Hayat I, Karibzhanova S, Zeeshan Z. Agreement of a National Routing Scale with Validated Measures of Change in Pediatric Rehabilitation. *Epidemiology and Health Data Insights*. 2026;2(5):ehdi061.
<https://doi.org/10.63946/ehdi/19690>

Copyright: By the author(s).

License: A non-exclusive license by the publisher. Published by Australasia Publishing Group LLP.

Open Access: This article is an open access article distributed under the terms and conditions of the CC-BY Creative Commons Attribution license
<https://creativecommons.org/licenses/by/4.0>



ABSTRACT

Introduction: Routine indicators of rehabilitation effectiveness are rarely checked against validated measures in children. Kazakhstan mandates the Rehabilitation Routing Scale, whose properties in children are unknown. We tested whether its change agrees with change on the Gross Motor Function Measure and the Barthel Index.

Methods: We analyzed 318 consecutive admissions to a national pediatric rehabilitation center (August to September 2025) from anonymized discharge records. Change scores were available for 306 children (Routing Scale), 215 (Gross Motor Function Measure) and 150 (Barthel Index). We estimated the magnitude of change, Spearman correlations of change scores (hypothesis: at least 0.30) with bootstrap 95% confidence intervals, tertile agreement and classification agreement.

Results: Most children improved on every scale (88%, 92% and 96%); the Routing Scale never recorded a decline. Admission scores agreed moderately (correlation 0.69 with the Gross Motor Function Measure and 0.52 with the Barthel Index). Change scores did not: 0.04 (95% confidence interval -0.10 to 0.18; 208 children) and -0.05 (-0.21 to 0.12; 143 children), while change on the two validated scales correlated weakly, at 0.29 (0.08 to 0.48; 90 children). Among 86 children with all three scales, 40 improved on all three and 38 improved on both validated scales but not by a ratio of at least 1.10 on the Routing Scale. Sensitivity analyses gave similar results, except that the Barthel Index correlation changed sign between age groups.

Conclusion: The national scale tracked baseline severity, but its change was not interchangeable with change on validated measures. Effectiveness reports should use more than one instrument. Limitations include a single center, a short stay, no test-retest data and no external anchor.

Keywords: Rehabilitation; Child; Outcome Assessment, Health Care; Reproducibility of Results; Routinely Collected Health Data

Introduction

Rehabilitation needs are high, frequently unmet, and is often not as high in lower resource settings [1]. Health systems therefore need to regularly check to see if inpatient rehab serves its function using indicators they gather regularly. They may appear useful because they do not require extra data collection, but they have to function like the validated measure, they are intended to represent to be informative [2].

The Rehabilitation Routing Scale needs to be documented at the time of admission and dismissal in Kazakhstan. It was created to achieve two goals: (1) staging rehabilitation and (2) moving patients between levels of care and was benchmarked against the modified Rankin Scale in an adult intensive care unit [3,4]. There is no description of its measurement properties in children. In reality the ratio of admission to discharge score is presented as an effectiveness indicator and this becomes significant for the evaluation of services, as it is the same children for whom this change is detected as in the established instruments.

Two internationally adopted instruments are present along with it within the same records: Gross Motor Function Measure validated in children with

cerebral palsy [5] and the Barthel Index of activities of daily living developed for an adult population [6,7] and modified for children with rare disorders [8]. Children who gain on one instrument should gain on the other and agreement could be seen both at time of admission and change. The greater the agreement in change, the better; children may have different severity measurements on admission and agree anyway despite differences in the instruments' responsiveness to the treatment [9].

We therefore asked three questions in a cohort of consecutive admissions to a national tertiary pediatric rehabilitation center. First, how large is the change on each instrument over a standard admission? Second, do the instruments agree at admission, and do they agree in change? Third, do they classify the same children as improved? We framed the work as a descriptive measurement study, following the COSMIN taxonomy of measurement properties [2] and published quality criteria for measurement properties [10]. In this taxonomy, the size of change describes how much children have changed, whereas responsiveness is tested by hypotheses about how change relates to change on comparator instruments [11].

Materials and Methods

Study design and setting

This was a retrospective cohort study of all consecutive inpatient admissions to a national tertiary pediatric neurorehabilitation center between August and September 2025. The center receives children with neurodevelopmental and neurological conditions from across the country and delivers a protocol-based program of about 24 days. The study is reported according to the STROBE statement [12], with the RECORD extension for routinely collected health data [13]. The completed RECORD checklist is provided in the Supplementary Material.

Data source and preparation

Electronic discharge summaries provided for every admission were anonymized. Duplicates for the same admission were removed, and 318 unique admissions were included in the final cohort for analysis. Structured variables, including the admission and discharge scores on the three instruments and the recorded effectiveness ratio, were extracted by a rule-based parser (regular expressions applied to the text layer of each summary) that left a field blank when its pattern was not found. Extracted values were screened against plausible ranges, and about half of the summaries (50%) were checked manually against the source documents. For the Gross Motor Function Measure and the Barthel Index two independent

extractions existed; they agreed exactly for 93% (admission) and 96% (discharge) of overlapping Gross Motor Function Measure values and for 96% and 82% of Barthel Index values (disagreement 7%, 4%, 4% and 18%), and the reconciled value was used. ICD-10 codes were used to assign diagnosis group (cerebral palsy, speech and language delay, autism spectrum disorder, Down syndrome, other neurological disorders, other) in the discharge summary.

The Functional Independence Measure was excluded because admission and discharge scores in part of the cohort were recorded on different item sets (an 18-item version at admission and a 13-item motor version at discharge). Two Gross Motor Function Measure change scores of -16 and -50 points were judged implausible in magnitude for a stay of this length and were set aside in the primary analysis; they were retained in a sensitivity analysis.

Instruments and change scores

The Routing Scale was introduced in 2017 by the Russian professional rehabilitation community, building on a national pilot project, to stage patients and route them between levels of rehabilitation care [14]. In the Russian adult version it has grades 0 to 6, assigned from the most pronounced feature; grades 0 to 1 need no rehabilitation, grade 2 is routed to day-hospital rehabilitation, grade 3 to a day hospital or 24-

hour department, and grades 4 to 6 to 24-hour inpatient rehabilitation [14,15]. The Russian pediatric order does not use the scored scale [15]. In Kazakhstan, Order No. 65 of the Ministry of Health (7 April 2023, as amended) [4] defines a separate pediatric version (Appendix 6 to the Standard), which grades function and activity limitation from 0 (no impairment) to 5 (extreme impairment), with descriptions specific to five age bands from 0 to 18 years. Clause 30 of the Standard defines the effectiveness of rehabilitation in children from the initial (R1) and final (R2) ICF-based assessments, where b is the rating of body functions and d the rating of activity and participation, $R = (b + d)/2$, and the effectiveness ratio is $R1:R2$; values below 1 are unsatisfactory, 1 to 1.5 satisfactory, 1.5 to 2.0 good and above 2.0 significant. Because R is a mean of two ratings, scores are not integers; lower scores indicate better function. In this cohort admission scores ranged from 1.62 to 3.75 and discharge scores from 1.40 to 3.50. Gross Motor Function Measure and Barthel Index scores were recorded as percentage or total scores on a 0-100 range, with higher scores indicating better function. For each instrument, change was defined so that positive values mean improvement: admission minus discharge for the Routing Scale and discharge minus admission for the other two. We analyzed change in each instrument's own units, and, where scales needed to be compared, in standardized form.

Improvement was classified per instrument as a Gross Motor Function Measure gain of at least 1.5 points and a Barthel Index gain of at least 5 points. These thresholds are conventional reporting values, not validated minimal important differences. The 1.5-point Gross Motor Function Measure threshold is close to the 1.3-point difference published for the 66-item version (GMFM-66) in ambulatory children with cerebral palsy [16]; the version recorded here (GMFM-88 or GMFM-66) is not stated in the discharge summaries, and the cohort includes children without cerebral palsy, so the value may not transfer. A Barthel Index gain of at least 5 points is effectively any gain: scores changed in steps of 5 points in 98.7% of children and only one child gained 1 to 4 points. Neither threshold was derived in this cohort, which has no external anchor. For the Routing Scale we used two definitions: any reduction in score, and a recorded admission-to-discharge ratio of at least 1.10. A ratio of 1.10 corresponds to a reduction of about 9.1% in the admission score ($1 - 1/1.10$), not 10%. In this cohort 1.10 is the median of the recorded ratio (25th to 75th percentile 1.04 to 1.10). It was chosen because the effectiveness bands in Order No. 65 [4] place every child in this cohort in a single band (1.0 to 1.5; recorded ratios 1.00 to 1.40), so they cannot separate children. The threshold classifies about half of the

children as improved (51% of 311) by construction, against 85% to 96% on the validated scales. Because this caps the raw agreement that can be reached, we also report the lenient definition and the largest raw agreement attainable.

Analysis sets

Change on the Routing Scale and the Gross Motor Function Measure was available for 208 children (set 1), on the Routing Scale and the Barthel Index for 143 children (set 2), on all three for 86 children (set 3), and on the two validated scales for 90 children. Missing scores were not imputed. Availability was not random (see Results), so pairs were analyzed on their own available cases, with the denominator given for every estimate. Denominators differ across analyses for the reasons shown in Figure 1. A Routing Scale change score required both an admission and a discharge score (306 of 318 admissions). The Gross Motor Function Measure was paired in 217 children, of whom 215 remained after the two implausible falls were set aside; the admission-score correlation with the Routing Scale used the 216 children with both admission scores. The classification comparisons in Table 4 that use the recorded ratio ($n = 212$ and 144) include children with a recorded ratio but no computed Routing Scale change, whereas the correlations of change scores use the computed difference ($n = 208$ and 143).

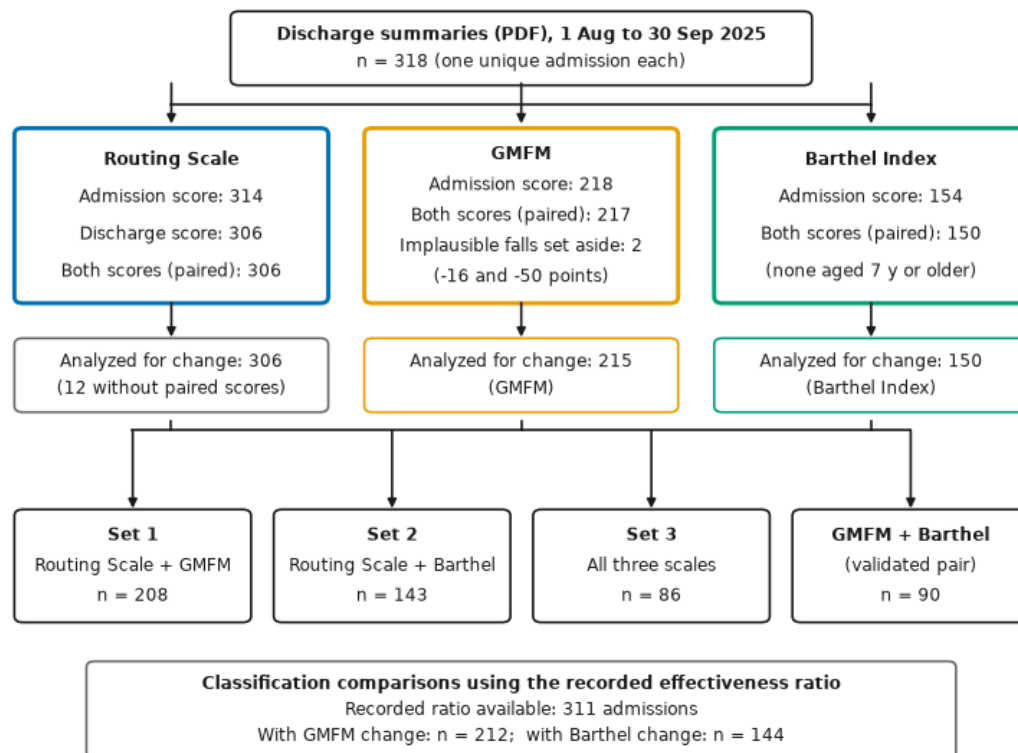


Figure 1. Participant flow. Consecutive admissions between August and September 2025 and the number of children with paired admission and discharge scores on each instrument; two Gross Motor Function Measure change scores (−16 and −50 points) were set aside as implausible. The analysis sets are complete cases for each pair of scales (no imputation); recorded-ratio comparisons in Table 4 use children with a recorded effectiveness ratio. GMFM, Gross Motor Function Measure.

Statistical analysis

Continuous variables are summarized as median and interquartile range or mean and standard deviation, as stated. The magnitude of change was described by the standardized response mean (mean change divided by the standard deviation of change) with percentile bootstrap 95% confidence intervals from 2,000 resamples [17]; the standardized response mean follows Husted and colleagues [18], and values of 0.8 or more were regarded as large [19]. Wilcoxon signed-rank test was used to test admission-to-discharge change. Following COSMIN, the standardized response mean describes how large the change was and is not by itself a test of responsiveness. Responsiveness was assessed through a hypothesis about change scores: that change on the Routing Scale would correlate with change on each validated instrument at 0.30 or more, as expected if they capture the same change, and likewise for the two validated instruments [11]. The hypothesis was specified after a descriptive review of the data and is therefore a criterion for interpretation, not a prespecified test.

Agreement was examined on three levels. The first step was to calculate Spearman correlations with

the Routing Scale score reversed (agreements changed to positive scores), and between change scores (spanning 6 pairs of scores: bootstrapped 95% confidence intervals and Benjamini-Hochberg adjusted) [20]. Partial rank correlations of change scores adjusted for both admission scores addressed regression to the mean [9]. The second attempt was to arrange changes into tertiles at 33rd and 67th percentile (ties assigned to the lower tertile) and then summaries agreement in terms of linearly weighted kappa. Third, binary improvement classifications were compared using raw agreement, Cohen's kappa, Gwet's first-order agreement coefficient and the prevalence-adjusted bias-adjusted kappa, because kappa is unstable when most children improve [21,22]. Kappa values were interpreted using conventional bands [23], and correlations were described as negligible (below 0.10), weak (0.10 to 0.39), moderate (0.40 to 0.69) or strong (0.70 or more) [24]. The same bands are used throughout the text, tables and figures.

Sensitivity analyses repeated the changing correlations in children with cerebral palsy, in other diagnoses, after excluding repeat admissions, in the 86 children with all three scales, with the two implausible

falls retained, with the recorded ratio in place of the difference, and within age groups. Age-group results were considered exploratory. The random seed was fixed at 2025. Analyses were run in Python (pandas, SciPy, statsmodels, scikit-learn, matplotlib). Additional tables and figures are in the Supplementary Material (Tables S1 to S6; Figures S1 to S3). Because the analyses were planned after a descriptive review of the data, the findings should be regarded as hypothesis-generating rather than confirmatory.

Ethics approval and consent

The study was approved by the institutional research ethics committee of the authors' university (see

Ethics) and analyzed only anonymized discharge summaries. The requirement for informed consent was waived because only anonymized, routinely collected data were analyzed.

Data and code availability

The discharge summaries contain sensitive clinical information and the data are not publicly available. De-identified data may be requested from the corresponding author, subject to approval by the research ethics committee and the data-providing center. The analysis code and supplementary materials are available from the corresponding author on request.

Results

Cohort and availability of scores

Participant flow and the number of children contributing to each analysis are shown in Figure 1. Of 318 admissions, 121 (38.1%) were children with cerebral palsy and 79 (24.8%) had speech and language delay. Median age was 5.8 years (interquartile range 3.7 to 10.0). Change on the Routing Scale was available for 306 children (96.2%), on the Gross Motor Function Measure for 215 (67.6%) and on the Barthel Index for 150 (47.2%) (Table 1). Availability differed by diagnosis and age (Supplementary Table S1; Supplementary Figure S3).

The Gross Motor Function Measure was recorded for 118 of 121 children with cerebral palsy (97.5%) but for 14 of 79 with speech and language delay (17.7%) and 10 of 36 with autism spectrum disorder (27.8%). The Barthel Index was recorded only in children younger than 7 years: it was available for none of the 127 children aged 7 years or older. Children with Barthel Index data (set 2) were younger than other children (median 3.6 versus 9.5 years). The three analysis sets therefore describe different parts of the caseload.

Table 1. Characteristics of all admissions and of children in each analysis set

Characteristic	All admissions	Routing Scale change available	Set 1: Routing Scale + GMFM	Set 2: Routing Scale + Barthel	Set 3: all three
Children	318	306	208	143	86
Age, median (IQR), years	5.8 (3.7-10.0)	5.9 (3.8-10.0)	7.0 (3.1-12.3)	3.6 (2.4-5.0)	2.6 (1.7-4.2)
Age 0 to <3 y, n (%)	56 (17.6%)	54 (17.6%)	50 (24.0%)	53 (37.1%)	49 (57.0%)
Age 3 to <7 y, n (%)	135 (42.5%)	128 (41.8%)	54 (26.0%)	90 (62.9%)	37 (43.0%)
Age 7 to <12 y, n (%)	70 (22.0%)	68 (22.2%)	51 (24.5%)	0 (0.0%)	0 (0.0%)
Age 12 y or older, n (%)	57 (17.9%)	56 (18.3%)	53 (25.5%)	0 (0.0%)	0 (0.0%)
Male, n (% of children with sex recorded)	214 (84.9%) of 252	206 (84.8%) of 243	131 (80.4%) of 163	93 (85.3%) of 109	51 (78.5%) of 65
Cerebral palsy, n (%)	121 (38.1%)	116 (37.9%)	113 (54.3%)	41 (28.7%)	41 (47.7%)
Speech/language delay, n (%)	79 (24.8%)	74 (24.2%)	14 (6.7%)	51 (35.7%)	9 (10.5%)
Autism spectrum disorder, n (%)	36 (11.3%)	36 (11.8%)	10 (4.8%)	11 (7.7%)	2 (2.3%)
Other neurological, n (%)	27 (8.5%)	26 (8.5%)	25 (12.0%)	11 (7.7%)	11 (12.8%)
Down syndrome, n (%)	18 (5.7%)	18 (5.9%)	17 (8.2%)	10 (7.0%)	10 (11.6%)
Other, n (%)	37 (11.6%)	36 (11.8%)	29 (13.9%)	19 (13.3%)	13 (15.1%)
Repeat admission, n (%)	116 (36.5%)	113 (36.9%)	89 (42.8%)	38 (26.6%)	28 (32.6%)
Admission Routing Scale score, median (IQR)	2.5 (2.0-2.8)	2.5 (2.0-2.8)	2.5 (2.0-2.9)	2.7 (2.4-2.9)	2.8 (2.5-3.0)

Abbreviations: GMFM, Gross Motor Function Measure; IQR, interquartile range. Sets are not mutually exclusive. Sex was recorded for 252 of 318 children; percentages use recorded children only.

Magnitude of change

Every instrument showed improvement over the admission (Table 2, Figure 2). The mean Routing Scale reduction was 0.18 points (standard deviation 0.13; standardized response mean 1.42, 95% confidence interval 1.29 to 1.57), with 87.9% of children improved, 12.1% unchanged and none worse. The Gross Motor Function Measure rose by 3.4 points on average (standard deviation 3.3; standardized response mean 1.04, 0.87 to 1.27); 91.6% improved, 0.5% was unchanged and 7.9% declined. The Barthel Index rose

by 9.2 points (standard deviation 5.7; standardized response mean 1.62, 1.46 to 1.85); 96.0% improved and none declined. Routing Scale changes took 36 distinct values, 75.8% were multiples of 0.05, and 12.1% were zero. At discharge, 9.2% of children scored 90 or more on the Gross Motor Function Measure and 16.0% on the Barthel Index, against 4.6% and 5.2% at admission. The magnitude of change was similar in children with cerebral palsy and in other diagnoses; whether change on the Routing Scale corresponds to change on the validated instruments is examined next.

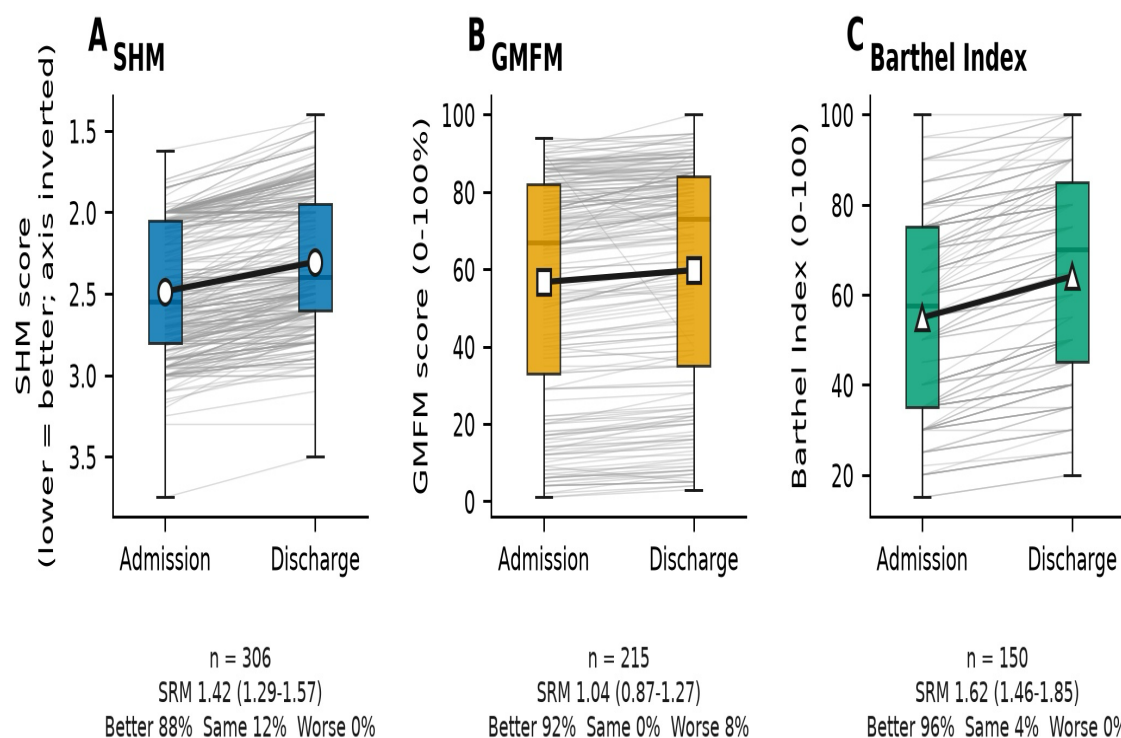


Figure 2. Admission-to-discharge change on each instrument. Thin lines are individual children; boxes show median and interquartile range; the thick line joins means. The Routing Scale axis is inverted so that upward means better in every panel. SRM, standardized response mean (95% confidence interval) from 2,000 bootstrap resamples. SHM, Rehabilitation Routing Scale; GMFM, Gross Motor Function Measure.

Table 2. Magnitude of change on the three instruments over admission

Group and instrument	n	Admission, mean	Discharge, mean	Change, mean (SD)	Standardized response mean (95% CI)	Better / same / worse, %
All: Routing Scale	306	2.49	2.31	0.18 (0.13)	1.42 (1.29 to 1.57)	87.9 / 12.1 / 0.0
All: GMFM	215	56.4	59.8	3.4 (3.3)	1.04 (0.87 to 1.27)	91.6 / 0.5 / 7.9
All: Barthel Index	150	54.9	64.1	9.2 (5.7)	1.62 (1.46 to 1.85)	96.0 / 4.0 / 0.0
Cerebral palsy: Routing Scale	116	2.49	2.31	0.18 (0.14)	1.34 (1.17 to 1.54)	84.5 / 15.5 / 0.0
Cerebral palsy: GMFM	118	52.4	55.4	3.0 (2.9)	1.03 (0.80 to 1.33)	89.0 / 0.8 / 10.2

Cerebral palsy: Barthel Index	45	45.3	52.9	7.6 (5.3)	1.43 (1.14 to 2.11)	93.3 / 6.7 / 0.0
Other diagnoses: Routing Scale	190	2.48	2.30	0.18 (0.12)	1.47 (1.29 to 1.70)	90.0 / 10.0 / 0.0
Other diagnoses: GMFM	97	61.3	65.2	3.9 (3.7)	1.07 (0.86 to 1.43)	94.8 / 0.0 / 5.2
Other diagnoses: Barthel Index	105	59.0	68.9	9.8 (5.7)	1.72 (1.52 to 2.01)	97.1 / 2.9 / 0.0

Abbreviations: CI, confidence interval; GMFM, Gross Motor Function Measure; SD, standard deviation. Change is defined so that positive values mean improvement (Routing Scale: admission minus discharge). Confidence intervals are percentile bootstrap, 2,000 resamples. $p < 0.001$ for change from admission in all rows (Wilcoxon signed-rank).

Agreement at admission and in change

At admission the three instruments agreed moderately (Table 3, Figure 3). The Spearman correlation of Routing Scale function (sign reversed) was 0.69 (95% confidence interval 0.62 to 0.76; 216

children) with the Gross Motor Function Measure and 0.52 (0.40 to 0.64; 150 children) with the Barthel Index, and the two validated scales correlated at 0.69 (0.54 to 0.80; 95 children).

Table 3. Agreement between instruments at admission and in change

Pair	Admission scores: Spearman (95% CI); n	Change scores: Spearman (95% CI); n	Adjusted p (change)	Partial Spearman of change (95% CI)	Tertile weighted kappa (95% CI)
Routing Scale vs GMFM	0.69 (0.62 to 0.76); 216	0.04 (-0.10 to 0.18); 208	0.563	0.05 (-0.09 to 0.19)	-0.06 (-0.17 to 0.04)
Routing Scale vs Barthel	0.52 (0.40 to 0.64); 150	-0.05 (-0.21 to 0.12); 143	0.563	-0.02 (-0.19 to 0.13)	-0.06 (-0.17 to 0.05)
GMFM vs Barthel	0.69 (0.54 to 0.80); 95	0.29 (0.08 to 0.48); 90	0.009	0.24 (0.03 to 0.45)	0.15 (0.01 to 0.28)

Abbreviations: CI, confidence interval. Routing Scale function (sign reversed) is used for admission scores, so positive values indicate agreement. Adjusted p is Benjamini-Hochberg across the six Spearman correlations in this analysis. Partial correlations adjust for both admission scores. Responsiveness hypothesis: correlation of change scores of at least 0.30. Tertile cut points: Routing Scale 0.10 and 0.22; Gross Motor Function Measure 2 and 4 points; Barthel Index 5 and 10 points.

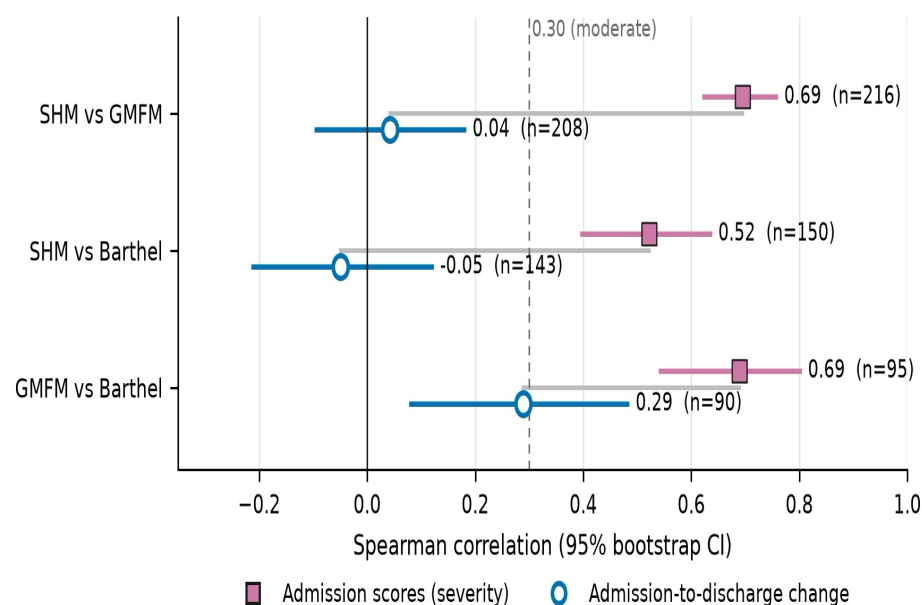


Figure 3.

Agreement between instruments at admission and in change. Spearman correlations with bootstrap 95% confidence intervals. Routing Scale function (sign reversed) is used for admission scores. The dashed line marks a correlation of 0.30, the hypothesized threshold for responsiveness. SHM, Rehabilitation Routing Scale; GMFM, Gross Motor Function Measure.

Change scores give a different story. No correlation was observed between Routing Scale change and Gross Motor Function Measure change (0.04; 95% confidence interval -0.10 to 0.18; 208 children) or Barthel Index change (-0.05; -0.21 to 0.12; 143 children), with both these scores adjusted for age and length of stay in the clinic. The upper confidence limits exclude correlations of 0.30 in both pairs. The hypothesis of a correlation of at least 0.30 was therefore not supported for the Routing Scale against either validated instrument. Change on the Gross Motor

Function Measure and the Barthel Index correlated weakly (0.29; 0.08 to 0.48; 90 children; adjusted $p = 0.009$). Adjusting for both admission scores left these estimates essentially unchanged (partial correlations 0.05, -0.02 and 0.24). Tertile agreement was no better than chance for the Routing Scale against either validated scale (weighted kappa -0.06 for both) and slight for the two validated scales (0.15, 0.01 to 0.28) (Figure 4). Of 208 children with both Routing Scale and Gross Motor Function Measure change, 28.8% were in opposite extreme tertiles.

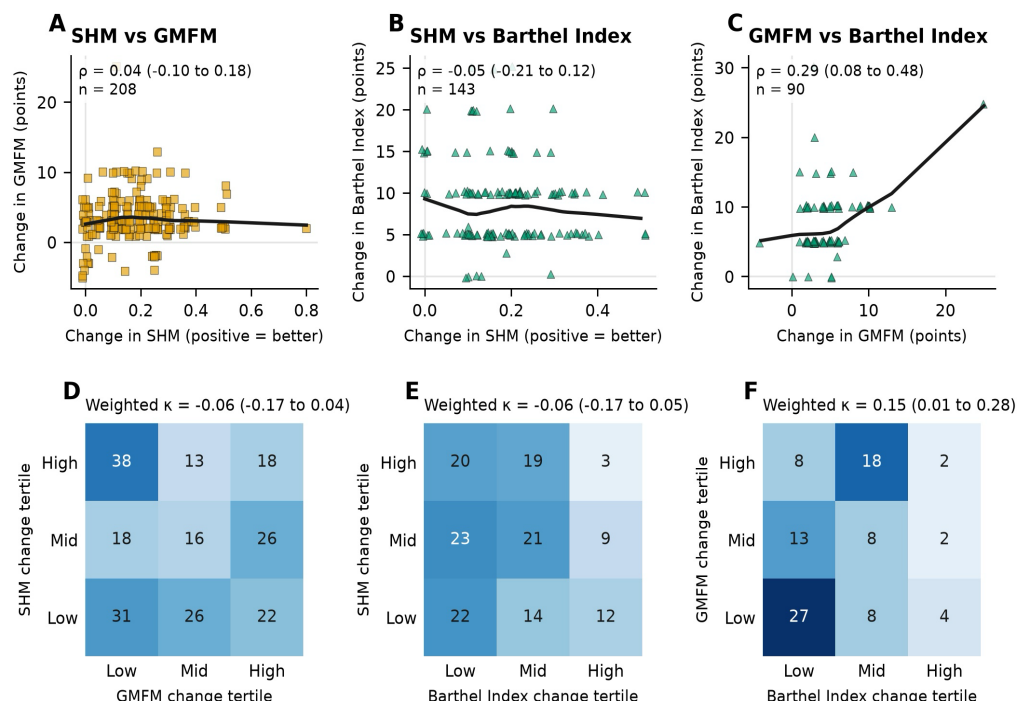


Figure 4. Agreement between change scores. (A-C) Scatter plots with small random jitter to separate tied values and a LOWESS smoother; ρ , Spearman correlation (95% confidence interval). (D-F) Cross-tabulation of change tertiles (low, mid, high) with linearly weighted kappa; ties are assigned to the lower tertile. SHM, Rehabilitation Routing Scale; GMFM, Gross Motor Function Measure.

Classification of improvement

Using any reduction in Routing Scale score as the definition, raw agreement with Gross Motor Function Measure improvement was 80.3% (kappa 0.21; Gwet coefficient 0.74) and with Barthel Index improvement 86.7% (kappa -0.06; Gwet coefficient 0.85) (Table 4). The high raw agreement reflects that 85% to 96% of children improved on every scale, which the Gwet coefficient reflects and kappa does not. With the stricter ratio of at least 1.10, agreement was 54.2% (kappa 0.04) with the Gross Motor Function Measure and 47.2% (kappa 0.00) with the Barthel Index, and the

Routing Scale classified 53.3% and 47.2% of children as improved against 84.9% and 95.8% by the validated scales. Given the proportions classified as improved, the largest raw agreement attainable with the ratio definition was 68% against the Gross Motor Function Measure and 51% against the Barthel Index (one minus the difference in the proportions improved); observed agreement was 54% and 47%. Among the 86 children with all three scales, 40 improved on all three, 38 improved on both validated scales but not by a ratio of at least 1.10 on the Routing Scale, and the remaining 8

were spread over four other patterns of two children each (Figure 5).

Table 4. Agreement of binary improvement classifications

Comparison	n	Improved, % (first / second)	Raw agreement, %	Cohen kappa	Gwet AC1	PABAK	Both / first only / second only / neither
Routing Scale ratio ≥ 1.10 vs GMFM ≥ 1.5	212	53.3 / 84.9	54.2	0.04	0.20	0.09	98 / 15 / 82 / 17
Routing Scale any gain vs GMFM ≥ 1.5	208	86.1 / 84.6	80.3	0.21	0.74	0.61	157 / 22 / 19 / 10
Routing Scale ratio ≥ 1.10 vs Barthel ≥ 5	144	47.2 / 95.8	47.2	0.00	0.11	-0.06	65 / 3 / 73 / 3
Routing Scale any gain vs Barthel ≥ 5	143	90.9 / 95.8	86.7	-0.06	0.85	0.73	124 / 6 / 13 / 0
GMFM ≥ 1.5 vs Barthel ≥ 5	90	94.4 / 94.4	91.1	0.15	0.90	0.82	81 / 4 / 4 / 1

Abbreviations: AC1, Gwet first-order agreement coefficient; PABAK, prevalence-adjusted bias-adjusted kappa. Thresholds: Gross Motor Function Measure gain ≥ 1.5 points; Barthel Index gain ≥ 5 points; Routing Scale ratio ≥ 1.10 or any reduction. Rows using the recorded ratio include children with a recorded ratio even when no discharge Routing Scale score was available, so *n* differs from the correlations of change scores (212 versus 208; 144 versus 143). The largest raw agreement attainable for the ratio rows, given the proportions improved, is 68% and 51%. Exact McNemar *p* for the ratio-based comparisons was < 0.001 .

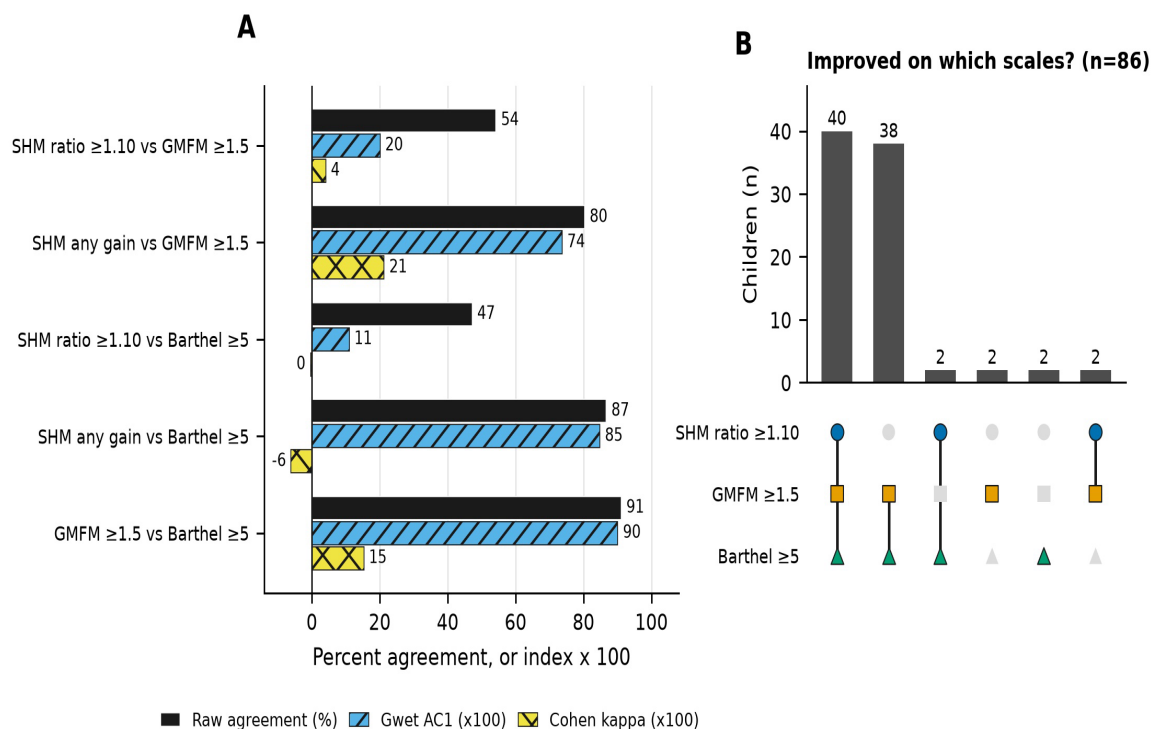


Figure 5. Agreement on who improved. (A) Raw agreement, Gwet first-order agreement coefficient (AC1) and Cohen kappa for five definitions of improvement; indices multiplied by 100. (B) Combinations of improvement among the 86 children with all three instruments (Routing Scale ratio ≥ 1.10 , Gross Motor Function Measure gain ≥ 1.5 points, Barthel Index gain ≥ 5 points); filled symbols mark instruments on which children improved; the six most common combinations are shown. SHM, Rehabilitation Routing Scale; GMFM, Gross Motor Function Measure.

Sensitivity analyses

The near-zero correlation of Routing Scale change with Gross Motor Function Measure change was similar in children with cerebral palsy (0.09; 95%

confidence interval -0.11 to 0.28; 113 children), in other diagnoses (-0.02; -0.21 to 0.17; 95 children), after excluding repeat admissions (0.01; -0.18 to 0.18; 119 children), in the 86 children with all three scales (-0.09;

-0.30 to 0.13), and when the two implausible falls were retained (0.06; -0.08 to 0.20; 210 children) (Figure 6). Using the recorded ratio instead of the difference gave -0.01 (-0.15 to 0.13). Routing Scale change against Barthel Index change varied by age in exploratory

analysis: it was positive in children aged 0 to 3 years (0.27; 0.02 to 0.51; 53 children) and negative at 3 to 7 years (-0.23; -0.42 to -0.02; 90 children). These two intervals do not account for the number of subgroup tests and should be read with caution.

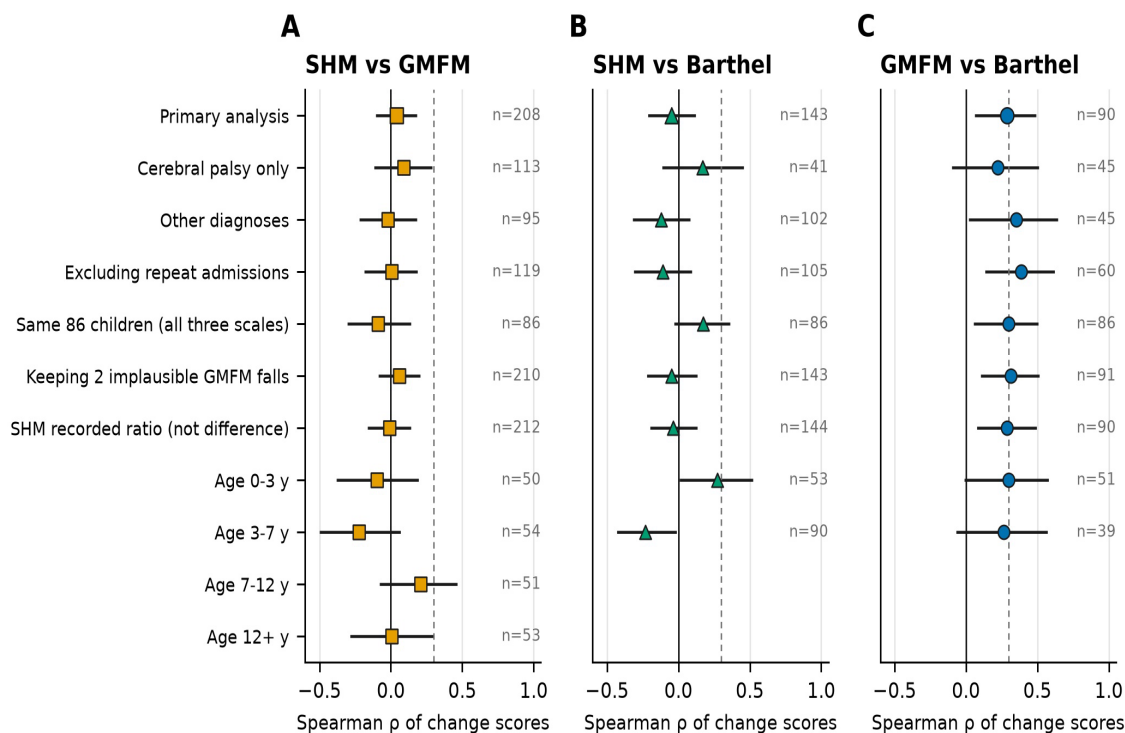


Figure 6. Sensitivity analyses. Spearman correlations of change scores with bootstrap 95% confidence intervals across subgroups and alternative analytic choices; n at right. The dashed line marks 0.30, the hypothesized threshold for responsiveness. SHM, Rehabilitation Routing Scale; GMFM, Gross Motor Function Measure.

Discussion

In 318 consecutive admissions to a national pediatric rehabilitation center, three instruments agreed about how severe children were at admission but did not agree about how much they changed. Almost every child improved on every scale, yet the children who gained most on the Routing Scale were not those who gained most on the Gross Motor Function Measure or the Barthel Index. The two validated instruments, which measure different constructs, were weakly related in change; the national scale was related to neither. The near-zero correlation with Gross Motor Function Measure change held in children with and without cerebral palsy, after adjustment for admission scores, and across the sensitivity analyses. The correlation with Barthel Index change was also near zero overall but changed sign between children aged 0 to 3 and 3 to 7 years, so this finding is less secure.

Answers to the three questions. First, all three instruments showed a large average change over a 24-day admission (standardized response means 1.42 for

the Routing Scale, 1.04 for the Gross Motor Function Measure and 1.62 for the Barthel Index), but this describes the size of change, not whether the instruments detect the same change. Second, the instruments agreed moderately at admission (0.52 to 0.69) but not in change: Routing Scale change did not reach the hypothesized correlation of 0.30 with either validated instrument (0.04 and -0.05), whereas change on the two validated instruments correlated weakly (0.29). On this test the Routing Scale was not shown to detect the same change as the validated instruments. Third, classification of improvement depended on the definition: with any reduction in score, raw agreement was 80% to 87% but kappa was low (0.21 and -0.06) because 85% to 96% of children improved; with a ratio of at least 1.10, agreement was 54% and 47% (kappa 0.04 and 0.00), and part of this shortfall reflects a threshold that classifies about half of the children as improved. Under neither definition did the Routing Scale identify the same children as the validated scales.

The result has a plain interpretation for monitoring. A change in the Routing Scale in this setting should not be read as a stand-in for functional change on validated instruments, and an effectiveness ratio built from it describes something else, or the same thing measured with more noise. Several features of the data may contribute, though this study cannot separate them. The Routing Scale moves in small steps (typically 0.05 to 0.5 points over the stay), so many children share the same change and rank agreement is hard to achieve. It is a composite of two clinician ratings and was designed for routing between levels of care, not for detecting short-term change. Thresholds for meaningful change also differ between scales; in one pediatric inpatient instrument they ranged from 6 to 15 points across its scales [25], so a single ratio cut-off such as 1.10 may not correspond to a clinically anchored threshold. It recorded no decline in any of 306 children, whereas the Gross Motor Function Measure fell in 8% of children; whether this reflects the course of the children, a recording convention or a ceiling in how the scale is used cannot be determined from discharge summaries. The 1.10 ratio cut-off also sits at the cohort median and so classifies about half of the children as improved by construction; low agreement for that definition therefore partly reflects the threshold, whereas the near-zero correlations of change scores do not depend on any threshold. Finally, 24 days is short, and ratings at discharge may reflect the stage of the program rather than function. These explanations are hypotheses, not findings.

Two methodological points deserve attention. First, agreement at admission does not necessarily mean agreement in change: admission scores were correlated at 0.52-0.69, while change scores were near zero, which can be achieved or masked by regression to the mean [9]. Second, when everyone improves slightly, the chance-corrected statistics of the level of agreement change differently. Raw agreement and Gwet's coefficient were high for the lenient definition of improvement while kappa was low, so the choice of statistic could lead to opposite conclusions; we report all of them and rely on the correlation and tertile results, which are not distorted by prevalence [21,22].

The results have implications for services that rely on the routine national indicators for reporting effectiveness. Reports should include the name of the

instrument used, definition of the improvement, and at least one validated measure, if possible. As indicated by our data, validated tools are used selectively: the Gross Motor Function Measure was requested primarily for children with cerebral palsy and motor impairments related to this condition, while the Barthel Index was only filled out for children under 7. The effectiveness analysis using any of the instruments that is available will describe different children, and the comparison between one dataset, among all the different instruments used, will be misleading.

Strengths and limitations

Strengths include consecutive admissions, a standard program and length of stay, three instruments in the same records, analyses at several levels of agreement and a full set of sensitivity analyses. The limitations are material. The study is from one center over two months and is retrospective, and scores were extracted from discharge summaries. Only 86 children had all three instruments, and subgroup estimates are imprecise. There was no test-retest or inter-rater data, so a lack of correlation may partly reflect measurement error, which cannot be separated from true lack of agreement here. There was no external anchor of clinically important change, so we could not estimate minimal important change or judge which instrument is right; thresholds for the Gross Motor Function Measure and the Barthel Index were not derived in this cohort, the Barthel Index threshold of 5 points is effectively any gain, the Routing Scale ratio threshold of 1.10 lies at the cohort median, and the version of the Gross Motor Function Measure used is not recorded in the discharge summaries. The Barthel Index was developed for adults and evidence for its use in young children is limited [8], so the correlations with it may also be affected by its own limitations. A blank score was treated as not recorded, and a small share may reflect extraction failure; the two extraction streams disagreed for up to 18% of overlapping Barthel Index discharge values before reconciliation, and residual extraction error cannot be excluded. Availability of scores was related to diagnosis and age, and we did not impute paired outcomes. The analyses were planned after a descriptive review of the data. Finally, it is not known whether assessors were blinded to admission scores.

Conclusion

In this single-center cohort the national Rehabilitation Routing Scale tracked baseline severity, but its change was not related to change on the Gross Motor Function Measure or the Barthel Index, whereas the two validated instruments were weakly related to

each other. In answer to the three questions, all three instruments showed large average change over the admission; the Routing Scale agreed with the validated instruments on severity at admission but did not reach the hypothesized correlation of 0.30 in change, so its

responsiveness against them was not shown; and it classified different children as improved under both of its definitions. Effectiveness indicators based on the national scale should not be treated as interchangeable with validated measures until their responsiveness and reliability in children are examined directly. Future

work should add test-retest and inter-rater data, an external anchor such as parent-reported or goal-based change, and a longer follow-up, and should repeat the comparison in other centers where all three instruments are recorded for the same children.

Supplementary Materials

Supplementary file available via:

https://www.journalehdi.com/supfile/793/ehdi061_Supplementary_Material_Updated.pdf

Acknowledgements

Author contributions (CRediT): **Syed Gufran Sadiq Zaidi:** conceptualization; methodology; validation; formal analysis; investigation; resources; data curation; writing – original draft; writing – review and editing; revisions; visualization; supervision. **Ihteshamul Hayat:** data curation; writing – review and editing. **Saule Karibzhanova:** validation; writing – review and editing. **Zeeshan:** writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Dataset declaration: The cohort analyzed here, 318 pediatric rehabilitation admissions, was also used in other manuscripts by the same research group (submitted or in preparation). Each manuscript addresses a different research question and reports different analyses.

Funding and acknowledgements: The research work presented in this paper was supported by Grant BR27199433 from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan. We are sincerely thankful to the National Center for Children's Rehabilitation, Kazakhstan for their cooperation in the provision of data.

Ethics: The study was approved by the Nazarbayev University Institutional Review Ethics Committee Board (IRB No:1081/21.05.2025)

Disclosure of AI Use: Generative AI is used only for language improvement purposes using the prompt "Proofread:". We confirm that the conceptualization, data curation, analyses, and manuscript writing are original.

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10267):2006-17. doi:[10.1016/S0140-6736\(20\)32340-0](https://doi.org/10.1016/S0140-6736(20)32340-0)
2. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737-45. doi:[10.1016/j.jclinepi.2010.02.006](https://doi.org/10.1016/j.jclinepi.2010.02.006)
3. Belkin AA, Rudnik EN, Belkin VA, Pinchuk EA, Lipovka NS. Optimization of step-by-step care for patients in intensive care units based on grade 6 of the Rehabilitation Routing Scale. *Phys Rehabil Med Med Rehabil*. 2021;3(1):142-8. doi:[10.36425/rehab64082](https://doi.org/10.36425/rehab64082)
4. Ministry of Health of the Republic of Kazakhstan. Order No. 65 of April 7, 2023, on approval of the standard for organizing the provision of medical rehabilitation [Internet]. Astana: Adilet Legal Information System; 2023 [cited 2026 Oct 11]. Available from: <https://adilet.zan.kz/rus/docs/V2300032263>
5. Russell DJ, Avery LM, Rosenbaum PL, Raina PS, Walter SD, Palisano RJ. Improved scaling of the Gross Motor Function Measure for children with cerebral palsy: evidence of reliability and validity. *Phys Ther*. 2000;80(9):873-85. doi:[10.1093/ptj/80.9.873](https://doi.org/10.1093/ptj/80.9.873)
6. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md State Med J*. 1965;14:61-5.

7. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. *Int Disabil Stud.* 1988;10(2):61-3. doi:[10.3109/09638288809164103](https://doi.org/10.3109/09638288809164103)
8. Lim Y, Pritchard KT, Nam S, Hong I. Psychometric properties of the Modified Barthel Index for children with rare disorders. *J Korean Soc Occup Ther.* 2020;28(4):111-20. doi:[10.14519/kjot.2020.28.4.09](https://doi.org/10.14519/kjot.2020.28.4.09)
9. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol.* 2005;34(1):215-20. doi:[10.1093/ije/dyh299](https://doi.org/10.1093/ije/dyh299)
10. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60(1):34-42. doi:[10.1016/j.jclinepi.2006.03.012](https://doi.org/10.1016/j.jclinepi.2006.03.012)
11. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1147-57. doi:[10.1007/s11136-018-1798-3](https://doi.org/10.1007/s11136-018-1798-3)
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med.* 2007;4(10):e296. doi:[10.1371/journal.pmed.0040296](https://doi.org/10.1371/journal.pmed.0040296)
13. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al.; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* 2015;12(10):e1001885. doi:[10.1371/journal.pmed.1001885](https://doi.org/10.1371/journal.pmed.1001885)
14. Ivanova GE, Melnikov EV, Belkin AA, Beliaev AT, Bodrova RA, Builova TV, et al. How to organize medical rehabilitation? *Bulletin of Rehabilitation Medicine.* 2018;17(2):2-12. (In Russ.)
15. Blinov DV, Solopova AG, Sandzhieva LN, Achkasov EE, Korabelnikov DI, Ampilogova DM, et al. Strengthening medical rehabilitation services in health system: a situation analysis. *FARMAKOEKONOMIKA Modern Pharmacoeconomics and Pharmacoepidemiology.* 2022;15(2):237-49. (In Russ.) doi:[10.17749/2070-4909/farmakoeconomika.2022.140](https://doi.org/10.17749/2070-4909/farmakoeconomika.2022.140)
16. Oeffinger D, Bagley A, Rogers S, Gorton G, Kryscio R, Abel M, et al. Outcome tools used for ambulatory children with cerebral palsy: responsiveness and minimum clinically important differences. *Dev Med Child Neurol.* 2008;50(12):918-25. doi:[10.1111/j.1469-8749.2008.03150.x](https://doi.org/10.1111/j.1469-8749.2008.03150.x)
17. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci.* 1986;1(1):54-75. doi:[10.1214/ss/1177013815](https://doi.org/10.1214/ss/1177013815)
18. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol.* 2000;53(5):459-68. doi:[10.1016/S0895-4356\(99\)00206-1](https://doi.org/10.1016/S0895-4356(99)00206-1)
19. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995;57(1):289-300. doi:[10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)
21. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* 2008;61(Pt 1):29-48. doi:[10.1348/000711006X126600](https://doi.org/10.1348/000711006X126600)
22. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol.* 1993;46(5):423-9. doi:[10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-74. doi:[10.2307/2529310](https://doi.org/10.2307/2529310)
24. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. *Anesth Analg.* 2018;126(5):1763-8. doi:[10.1213/ANE.0000000000002864](https://doi.org/10.1213/ANE.0000000000002864)

25. Iyer LV, Haley SM, Watkins MP, Dumas HM. Establishing minimal clinically important differences for scores on the Pediatric Evaluation of Disability Inventory for inpatient rehabilitation. *Phys Ther.* 2003;83(10):888-98. doi:[10.1093/ptj/83.10.888](https://doi.org/10.1093/ptj/83.10.888)