


Agreement and reliability of repeated bedside respiratory muscle strength measurements in acute and subacute stroke

Agnieszka Lewko¹  | Marta Sidaway^{2,3} | Stefan Tino Kulnik¹  |
Maciej Krawczyk^{2,3} 

¹Faculty of Health, Social Care and Education, Kingston University and St. George's University of London, London, UK

²The Institute of Psychiatry and Neurology, Warszawa, Poland

³Faculty of Rehabilitation, Józef Piłsudski University of Physical Education, Warszawa, Poland

Correspondence

Stefan Tino Kulnik, Centre for Health and Social Care Research, Faculty of Health, Social Care and Education, Kingston University and St. George's, University of London, Cranmer Terrace, London SW17 0RE, UK.
Email: s.t.kulnik@sgul.kingston.ac.uk

Funding information

Santander Mobility Grant; Private Physiotherapy Educational Foundation

Abstract

Background and Purpose: Many stroke trials include maximal inspiratory pressure (MIP), maximal expiratory pressure (MEP), and sniff nasal inspiratory pressure (SNIP) outcome measurements. However, data on agreement and reliability of repeated MIP, MEP, and SNIP measurements in acute and subacute stroke patients are scarce.

Methods: This study employed a test–retest design. Eighteen patients (seven female) with mean (SD) age 59 (14.5) years were recruited from neurological wards. Median (range) time since first stroke was 50.5 (21–128) days. MIP, MEP, and SNIP were measured repeatedly in three testing sessions (S1–3) conducted within 24 h and following international standards. Intra-rater agreement between testing sessions was analyzed using the Bland–Altman method. Test–retest reliability was analyzed using intra-class correlation coefficient (ICC). Association between individual measurement variability, time poststroke, and level of stroke impairment was analyzed using Spearman's rho.

Results: Mean difference and 95% limits of agreement for MIP were -0.40 (-23.02 , 22.22) cmH_2O between S1 and S2, and 2.14 (-12.79 , 16.99) cmH_2O between S2 and S3; for MEP, -4.56 (-29.01 , 19.90) cmH_2O between S1 and S2, and 0.29 (-24.28 , 24.87) cmH_2O between S2 and S3; and for SNIP, -10.56 (-38.48 , 17.37) cmH_2O between S1 and S2, and -6.06 (-27.32 , 15.20) cmH_2O between S2 and S3. ICCs for MIP, MEP, and SNIP were ≥ 0.9 throughout. There were no strong correlations between individual measurement variability and time poststroke or level of stroke impairment.

Discussion: MIP, MEP, and SNIP in acute and subacute stroke patients show good test–retest reliability for group averages; however, absolute agreement can vary considerably for some individuals.

KEYWORDS

agreement, maximal mouth pressure, reliability, respiratory muscle strength, smallest detectable change, standard error of measurement

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. Physiotherapy Research International published by John Wiley & Sons Ltd.

1 | INTRODUCTION

It is well-established that stroke can lead to significant and lasting respiratory muscle weakness (Harraf et al., 2008; Luvizutto et al., 2017; Pollock, Rafferty, Moxham, & Kalra, 2013; Ward et al., 2010). This may in turn impact on stroke survivors' respiratory function and cough effectiveness (Ward et al., 2010, 2017), cardio-respiratory endurance (Lista Paz et al., 2016), and performance of daily activities (Kulnik, 2015a; Xiao, Luo, Wang, & Luo, 2012). In recent years, investigators have increasingly targeted stroke-related respiratory muscle impairment, by trialing various respiratory training methods based on differing underlying rationales and with various hypothesized patient benefits and outcomes (Gomes-Neto et al., 2016; Martin-Valero, De La Casa Almeida, Casuso-Holgado, & Heredia-Madrado, 2015; Menezes, Nascimento, Avelino, Alvarenga, & Teixeira-Salmela, 2016, 2018; Xiao et al., 2012). Patient assessments in these stroke trials often include mouth pressure measurements, that is, maximal inspiratory pressure (MIP), maximal expiratory pressure (MEP), and sniff nasal inspiratory pressure (SNIP). These measurements are taken noninvasively and can be performed conveniently at the patient bedside in any clinical or community setting (Laveneziana et al., 2019). The validity of MIP, MEP, and SNIP measurements as indicators of respiratory muscle function (or impairment) has been established in detailed physiological studies of acute stroke patients and matched healthy control participants (Harraf et al., 2008; Ward et al., 2010) and in observational studies of chronic stroke survivors with differing levels of physical ability (e.g., Pinheiro et al., 2014). Longitudinal MIP and MEP measurements are often required to adjust the intensity of respiratory muscle training to individuals' baseline levels and any incremental improvements during the training period (McConnell, 2013). In stroke trials, MIP, MEP, and SNIP data are often collected longitudinally as indicators of change in respiratory muscle strength over time, or even as the primary outcome (e.g., Parreiras de Menezes et al., 2019).

Longitudinal measurements require an understanding of agreement between repeated tests, so that differences can be compared against expected measurement variability in the absence of change. This allows a judgement on whether differences in repeated measurements are likely to represent true change, or whether differences may lie within the range of expected measurement variability (Hernaes, 2015; de Vet, Terwee, Knol, & Bouter, 2006; Domholdt, 2005; Taylor, 1997). Surprisingly, while many stroke investigators have included repeated measurements of MIP, MEP, and SNIP in their studies, published data on agreement of these measurements in stroke patients are scarce.

Consistency in these repeated measurements principally relies on the test subject acquiring the correct technique and exerting a consistent maximal volitional effort (Laveneziana et al., 2019; American Thoracic Society/European Respiratory Society, 2002). Stroke survivors can often exhibit fluctuations in alertness, cognition, fatigue, and co-ordination, but also pronounced learning effects over short time periods, particularly during the acute and subacute phases poststroke. This may disproportionately affect agreement of repeated measurements. It is, therefore, important to examine agreement in this patient group, so that longitudinal data can be interpreted in the

context of expected variability and minimal detectable difference. In addition, reliability is important when groups of individuals are being assessed. Reliability refers to the ability of a measurement to detect "real" variability between subjects, and it has been recommended that studies examining measurement properties should report both agreement and reliability (Hernaes, 2015).

The primary aim of this study was to examine agreement and reliability of repeated MIP, MEP, and SNIP measurements in acute and subacute stroke patients. The secondary aim was to examine whether measurement variability might be related to stroke impairment, by evaluating associations between agreement and stroke-related neurological and functional impairment.

2 | METHODS

2.1 | Study design

A test-retest design was used to evaluate agreement and reliability of three respiratory muscle measures (MIP, MEP, and SNIP). The study was approved by the bio-ethical committee of the Institute of Psychiatry and Neurology, Warsaw, Poland (reference number 29/2015).

2.2 | Subjects

Patients up to 5 months after first-ever stroke were recruited from neurological wards at the study site. Data were collected from September 2015 to June 2018. Inclusion criteria were cognitive ability to follow the test instructions and informed consent. Exclusion criteria were National Institutes of Health Stroke Scale (NIHSS) score <5, blood pressure >180/100 mmHg, recent acute cardiac episode, and co-existing chronic conditions affecting lung function (such as respiratory disease, neurological disorders, or chest wall deformities). All participants provided written informed consent.

2.3 | Materials

The Micro Respiratory Pressure Meter (RPM; Micro Medical) was used to measure maximal mouth pressures. The device was calibrated by the manufacturer. Spirometry was measured using the MicroLoop spirometer (Micro Medical). The spirometer was calibrated using a 3 L syringe as per manufacturer recommendation.

2.4 | Procedure

All patients attended three assessment sessions (S1, S2, and S3) within 24 h with at least 2 h between sessions. Spirometry was assessed at the beginning of S1. Assessments were performed before, or at least 1 h after meals. In each session, patients performed MIP, MEP, and SNIP maneuvers according to American Thoracic Society/

European Respiratory Society recommendations (2002). A minimum of 10 efforts were made until reaching the maximum value of three maneuvers that varied by less than 20%. The highest value was used for analysis. Patients were given at least 30 s rest between efforts. Individual feedback and strong verbal encouragement were given throughout. Each patient was assessed by the same trained investigator (A.L. or M.S.) in all three sessions. Figure 1 presents the assessment sequence.

2.5 | Maximum inspiratory/expiratory pressure

MIP was assessed from residual volume and MEP from total lung capacity using the inspiratory and expiratory pressure valve assembly, respectively. Bacterial filters were applied to prevent cross contamination between users. For optimal lip seal, a sterilized rubber-flanged mouthpiece was used. Patients were seated upright, wore nose clips, and were encouraged to make maximal inspiratory or expiratory efforts sustained for 1 s.

2.6 | Sniff nasal inspiratory pressure

SNIP tests were performed from functional residual capacity (FRC) with patients in the sitting position. Manufacturer's nasal probes in three sizes (small, medium, or large) were used together with the nasal probe adapter. The size of the probe was chosen to best fit individuals' left nostril. Patients were asked to perform a strong, sharp sniff with encouragement from the investigator.

2.7 | Lung function

Baseline spirometry was measured according to American Thoracic Society/European Respiratory Society standards (Miller et al., 2005), and repeated measurements were taken until 10% variation between three efforts was achieved. Best results for forced expiratory volume in one second (FEV_1) and forced vital capacity were recorded as actual and percent predicted values.

2.8 | Baseline characteristics and stroke-related impairment

The following participant characteristics were recorded from hospital medical records: age, sex, height, body mass index, time since stroke, stroke etiology, and lesion site. Stroke-related impairment was measured using the NIHSS (Goldstein, Bertels, & Davis, 1989), Modified Rankin Scale (Banks & Marotta, 2007), Barthel Index (BI; Mahoney & Barthel, 1965), Rivermead Motor Assessment Scale (MAS) consisting of three subscales Gross Function, Arm (A), and Leg and Trunk (Carr, Shepherd, Nordholm, & Lynne, 1985), and the Trunk Impairment Scale (TIS; Verheyden et al., 2004).

2.9 | Sample size calculation

A sample size calculation was based on data from repeated MIP and MEP measurements in subacute stroke patients (Kulnik, 2015b). Assuming an intra-class correlation coefficient (ICC) of 0.90, the sample size required to estimate this parameter with the precision of a 95% confidence interval (CI) of ± 0.1 was 12 subjects (Bonett, 2002). Accounting for a conservative estimate of 33% attrition, the target sample size for this study was 18.

2.10 | Data analysis

Participant characteristics were summarized using descriptive statistics. To examine agreement, the differences in measurements between subsequent testing sessions (S1–S2, S2–S3) were calculated. Differences were summarized using descriptive statistics and examined for normal distribution using histograms, Kolmogorov–Smirnov and Shapiro–Wilk tests. The method by Bland and Altman (1999) was used to describe mean differences and 95% limits of agreement between subsequent testing sessions, including visualization in Bland–Altman plots. Upper and lower 95% limits of agreement were calculated by mean difference $\pm 1.96 \times$ standard deviation (SD) of the differences (Bland & Altman, 1999). Systematic measurement bias was assessed through linear regression of the difference between two measurements on the mean of two measurements.

Reliability was evaluated using an ICC with 95% CI for two consecutive measurements. The type of ICC was selected according to the relevant assumptions (Hernaes, 2015), that is, using a random effects model, comparing absolute agreement, and comparing individual measurements. The standard error of measurement (SEM) was calculated as SD of mean differences/ $\sqrt{2}$ (de Vet et al., 2006, p. 1037). The smallest detectable change (also termed minimal detectable change/difference) was calculated as $1.96 \times \sqrt{2} \times$ SEM (de Vet et al., 2006, p. 1038).

To examine associations between variability in agreement of repeated MIP, MEP, and SNIP and stroke-related neurological and functional impairment, the range of measurements across all three testing sessions was calculated for all measures, indicating the magnitude of variability in measurements for each individual subject. Associations between the magnitude of variability and measures of stroke impairment (time since stroke, NIHSS, BI, Rivermead MAS, and TIS) were assessed using nonparametric correlation (Spearman's rho). All analyses were conducted in SPSS 24 (IBM, 2018).

3 | RESULTS

3.1 | Participants

Summary for participant characteristics is presented in Table 1. Data for MIP, MEP, and SNIP are presented in Table 2. All participants were naive to the investigation.

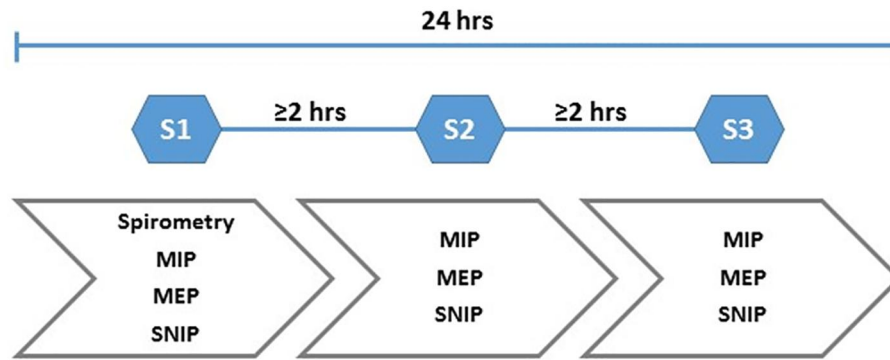


FIGURE 1 Assessment session schedule (S1, S2, S3—Session 1, 2, 3; MIP, maximal inspiratory pressure; MEP, maximal expiratory pressure; SNIP, sniff nasal inspiratory pressure)

3.2 | Agreement of repeated measurements

Mean (SD) differences and 95% limits of agreement are presented in Table 3. The distribution was sufficiently normal. Bland–Altman plots for MIP, MEP, and SNIP are given in Figures 2–4, respectively. Assessment of systematic bias showed a statistically significant result only for agreement in MIP between S1 and S2 (coefficient 0.192, $p = 0.023$), but not for the remaining comparisons.

3.3 | Reliability of repeated measurements

ICCs, SEM, and smallest detectable change for repeated MIP, MEP, and SNIP are presented in Table 4.

3.4 | Associations between variability in agreement and stroke-related impairment

Correlation coefficients ranged from -0.487 to 0.439 , indicating weak-to-moderate correlations. Correlations were nonsignificant, except for the correlation between variability in SNIP and the Arm component sub-score of the Rivermead MAS (Spearman's rho -0.487 , $p = 0.047$). The correlation matrix and statistical significance levels are presented in Table 5.

4 | DISCUSSION

This study has provided data to estimate test–retest agreement and reliability of MIP, MEP, and SNIP during the acute and subacute phases of stroke. Agreement for MIP and MEP showed small mean differences, with 95% limits of agreement spanning 30–50 cmH₂O. Agreement for SNIP showed larger mean differences and 95% limits of agreement spanning up to 55 cmH₂O. Reliability for MIP, MEP, and SNIP was very high, with ICCs ≥ 0.90 and narrow 95% CIs throughout. Agreement and reliability improved for all three measures from S1–S2 to S2–S3, indicating improved consistency in test

performance with increasing number of repetitions. This could be an indication of learning effect due to test familiarization. Individual variability in measurements was not explained by time since stroke or measures of stroke impairment (NIHSS, BI, Rivermead MAS, and TIS). Only the Arm subscale of the Rivermead MAS showed a moderate statistically significant correlation with measurement variability, and this correlation would lose statistical significance if adjusted for multiple testing.

Test–retest reliability of MIP, MEP, and SNIP has been investigated in healthy subjects (Dimitriadis, Kapreli, Konstantinidou, Oldham, & Strimpakos, 2011; Maillard, Burdet, van Melle, & Fitting, 1998) and patients with respiratory conditions (Larson et al., 1993; Nikoletou et al., 2014). ICCs for MEP and MIP in healthy subjects have been reported as 0.88 and 0.90, respectively (Dimitriadis et al., 2011). In patients with COPD, test–retest reliability for all three measures showed good reliability, with ICCs of 0.89 for MIP, 0.96 for MEP, and 0.94 for SNIP (Larson et al., 1993; Nikoletou et al., 2014). While these ICCs are overall comparable to our results, SNIP in patients with COPD showed somewhat better reliability than MIP, which is contrary to our findings in stroke patients.

Published data on agreement and reliability of MIP, MEP, and SNIP in stroke patients are scarce, but there have been studies in other neurological patient groups. Repeatability of SNIP measurements was investigated in patients with various neuromuscular nonstroke conditions and lung diseases (Lofaso et al., 2006). There was a significant difference between the first and second session, but only a marginal improvement after 20 maneuvers. Bland–Altman analysis demonstrated a mean between-session difference of 3.5 ± 7.7 cmH₂O. Although our study also demonstrated improvement after the first session, indicating similar learning effect, the mean difference for SNIP was larger, which may be due to differences in testing protocols.

Intra-rater reliability of MIP and MEP between four testing sessions has been investigated in multiple sclerosis (Smeltzer & Lavietes, 1999). The results indicated that two practice sessions were needed for each measure. Test–retest reliability was also tested for patients with Huntington's (Reyes, Cruickshank, Ziman, & Nosaka, 2014) and Parkinson's disease (Reyes, Castillo, Castillo, &

TABLE 1 Participant characteristics

Participant characteristics		Sample (n = 18)
Age (years)		59 (14.5)
Sex	Female	7 (38.9%)
	Male	11 (61.1%)
Height (cm)		175.9 (10.0)
Body mass index		27.4 (5.1)
Time since stroke (days)		50.5 (21–128)
Stroke etiology	Ischemic	12 (66.7%)
	Hemorrhagic	6 (33.3%)
Stroke site	Cortical	10 (55.6%)
	Subcortical	5 (27.8%)
	Brainstem	3 (16.7%)
Stroke side	Left	9 (50.5%)
	Right	8 (44.4%)
	Bilateral	1 (5.6%)
NIHSS (0–42) ^a		7.4 (2.1)
Modified Rankin Scale (0–5) ^a	3–Moderate disability	8 (44.4%)
	4–Moderately severe disability	8 (44.4%)
	5–Severe disability	2 (11.1%)
Barthel Index (0–20) ^b		9.2 (5.1)
Rivermead MAS–Gross Function (0–13) ^b		6.4 (3.3)
Rivermead MAS–Leg and Trunk (0–10) ^b		4.8 (2.1)
Rivermead MAS–Arm (0–15) ^b		2.4 (3.1)
Rivermead MAS–Combined score (0–38) ^b		13.5 (6.8)
Trunk Impairment Scale (0–23) ^b		13.9 (4.2)
Respiratory function	FEV1 (L)	2.1 (0.8)
	FEV1 % predicted	88.2 (12.2)
	FVC (L)	3.5 (1.0)
	FVC % predicted	87.6 (14.6)
	FEV1/FVC ratio	0.82 (0.07)
	MIP predicted (cmH ₂ O)	105.8 (42.5)
	MEP predicted (cmH ₂ O)	99.8 (27.9)

Note: Values are mean (SD), median (range), or frequencies (percentages).

Abbreviations: FEV1, forced expiratory volume in one second; FVC, forced vital capacity; MAS, Motor Assessment Scale; MEP, maximal expiratory mouth pressure; MIP, maximal inspiratory mouth pressure; NIHSS, National Institutes of Health Stroke Scale.

^aHigher scores indicate greater stroke-related impairment or disability.

^bHigher scores indicate better function or performance.

Cornejo, 2018). Although reliability for both MIP and MEP between three sessions was good for patients with Huntington's disease (ICCs of 0.94 and 0.92, respectively), it showed greater variability than a healthy control group (Reyes et al., 2014). The results for patients with Parkinson's disease also showed acceptable reliability (ICCs of 0.95 and 0.92, respectively), but there were significant differences

between the first and second session (Reyes et al., 2018). Additionally, these studies indicate learning effects in the performance of these tests, which are similar to our results and in line with studies in healthy volunteers and in respiratory patients; however, detailed comparisons between studies are difficult due to differing testing protocols and data analysis methods.

TABLE 2 Repeated MIP, MEP, and SNIP measurements at three time points (S1, S2, and S3)

Participant number	MIP (cmH ₂ O)			MEP (cmH ₂ O)			SNIP (cmH ₂ O)		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
1	58	51	^a	80	81	^a	49	52	^a
2	72	70	71	84	95	99	81	86	105
3	32	42	38	50	52	50	27	23	28
4	31	41	32	36	71	40	12	24	19
5	75	67	85	145	138	138	53	56	70
6	^b	^b	^b	80	115	112	75	96	93
7	33	^b	^b	101	96	84	43	55	73
8	60	60	57	115	113	112	39	42	47
9	131	117	121	100	110	111	74	73	96
10	108	108	105	117	124	124	119	142	142
11	^b	^b	^b	56	46	50	47	53	42
12	74	72	67	135	135	135	82	82	70
13	63	61	57	61	61	92	25	45	64
14	138	119	121	161	160	152	148	132	146
15	82	94	93	121	121	122	88	124	121
16	93	93	77	111	115	132	73	75	84
17	24	53	49	131	125	118	57	93	103
18	148	147	141	159	167	168	123	152	153
Sample mean (SD)	76.4 (38.9)	79.7 (31.8)	79.6 (33.1)	102.4 (37.2)	106.9 (34.5)	108.2 (35.9)	67.5 (36.9)	78.1 (39.7)	85.7 (40.1)

Note: S1, S2, and S3 denote testing sessions 1, 2, and 3.

Abbreviations: MEP, maximal expiratory pressure; MIP, maximal inspiratory pressure; SNIP, sniff nasal inspiratory pressure.

^aParticipant declined assessment session S3.

^bParticipant unable to perform the measurement.

Repeated outcome measurement can affect the internal validity of a longitudinal study, when an observed change reflects familiarization with the testing procedure, rather than “true” change in the underlying body function (Domholdt, 2005). This concern has been previously raised with respect to studies of respiratory muscle training interventions (Polkey & Moxham, 2004). A strategy to control for test familiarization is the inclusion of a control group, which completes all assessment procedures in the same way as the intervention group (Domholdt, 2005). Alternatively, in a controlled study design, it is possible to conduct a postintervention test only. The latter removes the aspect of repeated testing and test familiarization but also eliminates the possibility of comparing outcome parameters between study groups at baseline. Some study designs include an initial test familiarization period to achieve stable repeat test performances before the intervention is introduced. This strategy is suitable for assessments, which rely on participants acquiring the correct technique and exerting a maximal volitional effort such as maximal strength tests (Phillips, Batterham, Valenzuela, & Burkett, 2004; Phillips, Benton, Wagner, & Riley, 2006). It is perhaps more straightforward to achieve test familiarization in populations, which are stable within their condition and able to attend

familiarization sessions over several days or weeks. Our study demonstrates that it is possible to incorporate repeated assessments of MIP, MEP, and SNIP within 24 h at the beginning of a study in acute and subacute stroke. Even if in this short time period, stable test performance cannot be achieved, this at least gives an indication of individual subjects' test–retest variability to inform the interpretation of observed change in measurements, or to be incorporated in a statistical analysis model. Since our data at group level describe relatively large values for the smallest detectable change in this patient population, ranging from 15 to 28 cmH₂O (Table 4), methods which take into account smallest detectable change at individual patient level could potentially add value to the analysis of respiratory muscle training trials in stroke.

Lastly, putting the magnitude of an observed change into context, the minimal (clinically) important difference (MID/MCID) may be described. It is noteworthy that, to our knowledge, MID has not been reported for SNIP or MEP, and we are unable to interpret the clinical relevance of differences in repeated tests for these measures. There is literature describing reference values for MIP, MEP, and SNIP in the general population (Laveneziana et al., 2019), and an individual's measurements may be compared with the range

TABLE 3 Mean (SD) differences and 95% limits of agreement for repeated MIP, MEP, and SNIP measurements

Measure		Mean difference	SD of differences	95% confidence interval of mean difference	Upper limit of agreement ^a	95% confidence interval of upper limit of agreement	Lower limit of agreement ^a	95% confidence interval of lower limit of agreement
MIP (cmH ₂ O)	S1-S2 (n = 15)	-0.40	11.54	-6.79 to 5.99	22.22	11.29 to 33.15	-23.02	-33.95 to -12.09
	S2-S3 (n = 14)	2.14	7.57	-2.23 to 6.52	16.99	9.53 to 24.45	-12.70	-20.16 to -5.24
MEP (cmH ₂ O)	S1-S2 (n = 18)	-4.56	12.48	-10.76 to 1.65	19.90	9.29 to 30.51	-29.01	-39.62 to -18.40
	S2-S3 (n = 17)	0.29	12.54	-6.15 to 6.74	24.87	13.85 to 35.89	-24.28	-35.30 to -13.26
SNIP (cmH ₂ O)	S1-S2 (n = 18)	-10.56	14.25	-17.64 to -3.47	17.37	5.25 to 29.49	-38.48	-50.60 to -26.36
	S2-S3 (n = 17)	-6.06	10.85	-11.64 to -0.48	15.20	5.67 to 24.73	-27.32	-36.85 to -17.79

Note: S1, S2, and S3 denote testing sessions 1, 2, and 3.

Abbreviation: MEP, maximal expiratory pressure; MIP, maximal inspiratory pressure; SNIP, sniff nasal inspiratory pressure.

^aUpper and lower limits of agreement = mean difference \pm 1.96 \times SD of differences (Bland & Altman, 1999).

TABLE 4 Intraclass correlation coefficients, standard error of measurement, and smallest detectable change for repeated MIP, MEP, and SNIP measurements

Measure		ICC ^a	95% Confidence interval of ICC	SEM ^b	Smallest detectable change ^c
MIP	S1, S2 (n = 15)	0.950	0.857 to 0.983	8.16	22.62
	S2, S3 (n = 14)	0.973	0.920 to 0.991	5.35	14.84
MEP	S1, S2 (n = 18)	0.935	0.835 to 0.975	8.82	24.46
	S2, S3 (n = 17)	0.941	0.844 to 0.978	8.87	24.58
SNIP	S1, S2 (n = 18)	0.896	0.611 to 0.966	10.08	27.93
	S2, S3 (n = 17)	0.954	0.854 to 0.984	7.67	21.27

Note: S1, S2, and S3 denote testing sessions 1, 2, and 3.

Abbreviations: ICC, intra-class correlation coefficient; MEP, maximal expiratory pressure; MIP, maximal inspiratory pressure; SEM, standard error of measurement; SNIP, sniff nasal inspiratory pressure.

^aRandom effects model for comparison of absolute agreement and individual measurements.

^bStandard deviation of mean differences/ $\sqrt{2}$ (de Vet et al., 2006, p. 1037).

^c1.96 \times $\sqrt{2}$ \times SEM (de Vet et al., 2006, p. 1038).

TABLE 5 Correlation matrix for associations between variability in agreement (range of three repeated measurements) and stroke-related impairment. Shown are Spearman's rho and statistical significance level

Measure	NIHSS	Barthel Index	Rivermead MAS—Gross Function	Rivermead MAS—Leg and Trunk	Rivermead MAS—Arm	Rivermead MAS—combined score	Trunk Impairment Scale	Time since stroke
MIP (n = 14)	0.325 (p = 0.257)	0.377 (p = 0.184)	0.107 (p = 0.716)	0.030 (p = 0.920)	-0.449 (p = 0.107)	-0.130 (p = 0.659)	-0.051 (p = 0.861)	0.257 (p = 0.375)
MEP (n = 17)	0.203 (p = 0.435)	-0.018 (p = 0.944)	0.178 (p = 0.494)	0.019 (p = 0.942)	-0.178 (p = 0.495)	-0.011 (p = 0.966)	-0.061 (p = 0.815)	0.096 (p = 0.713)
SNIP (n = 17)	0.084 (p = 0.748)	0.439 (p = 0.078)	-0.207 (p = 0.424)	-0.209 (p = 0.421)	-0.487 (p = 0.047)	-0.331 (p = 0.195)	-0.261 (p = 0.311)	-0.130 (p = 0.618)

Abbreviations: MAS, Motor Assessment Scale; MEP, maximal expiratory pressure; MIP, maximal inspiratory pressure; NIHSS, National Institutes of Health Stroke Scale; SNIP, sniff nasal inspiratory pressure.

of norm values for a person of the same sex and age; although to our knowledge there has not been any research to validate this approach in the stroke population. A small number of studies have described MID for MIP in people with COPD, indicating clinically relevant

improvements of 13 cmH₂O (Gosselink et al., 2011) and a distribution-based MID of approximately 17 cmH₂O (Iwakura et al., 2020). In our study, the smallest detectable change for MIP from S1 to S2 was greater than these MID values, while for MIP from S2 to S3, it was

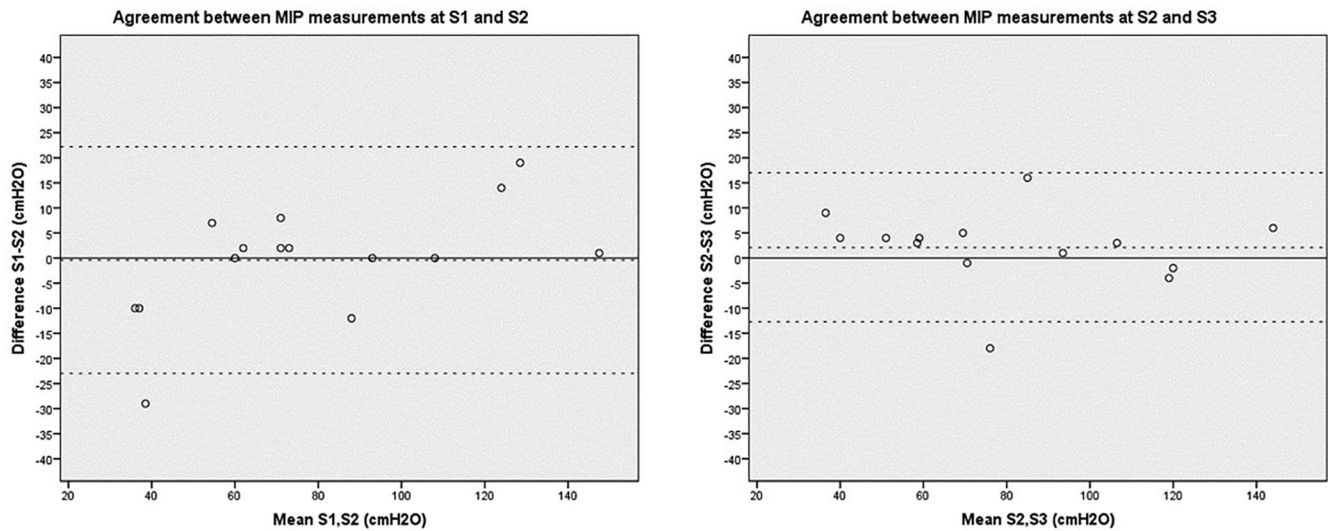


FIGURE 2 Agreement of repeated maximal inspiratory mouth pressure (MIP) measurements between first (S1) and second (S2) testing session, and between second (S2) and third (S3) testing session. Solid lines indicate the lines of equality (no difference between measurements). Three dashed lines indicate the mean difference between measurements (bias) and the upper and lower 95% limits of agreement ($\text{bias} \pm 1.96 \times \text{SD}$)

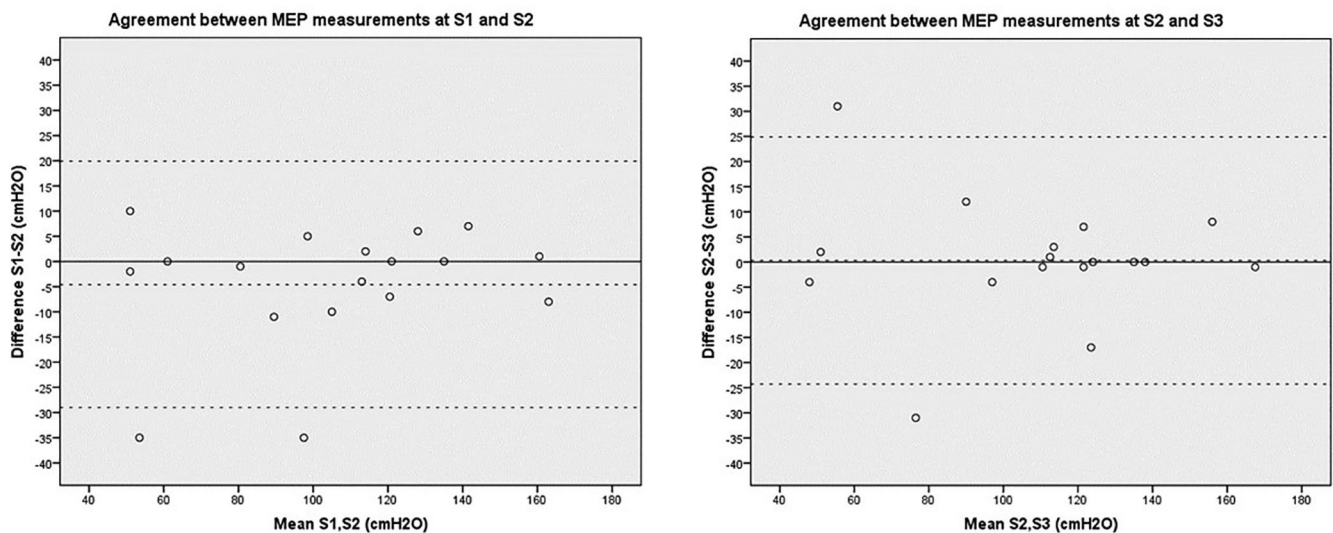


FIGURE 3 Agreement of repeated maximal expiratory mouth pressure (MEP) measurements between first (S1) and second (S2) testing session, and between second (S2) and third (S3) testing session. Solid lines indicate the lines of equality (no difference between measurements). Three dashed lines indicate the mean difference between measurements (bias) and the upper and lower 95% limits of agreement ($\text{bias} \pm 1.96 \times \text{SD}$)

similar in magnitude (Table 4). However, as there are no studies to indicate MID/MCID for MIP, MEP, or SNIP in stroke populations, these reference values should be interpreted with caution.

4.1 | Strengths and limitations of the study

A study strength was the experience of the investigators in clinical measurement, giving confidence that patients were instructed, and tests performed according to the relevant standards. Stroke may

affect individuals' cognitive function or coordination, which can lead to difficulties in performing specific respiratory maneuvers. In our study, we could see that three patients struggled with performing the MIP maneuver but were able to adequately perform MEP and SNIP. This underscores the importance of the investigator providing adequate guidance and support to patients, so that their performance of MIP, MEP, and SNIP maneuvers meets requirements, and recognizing when a patient's difficulties invalidate these measurements (Lavezziana et al., 2019). Another strength was that repeated measurements were conducted within 24 h, so that measurement

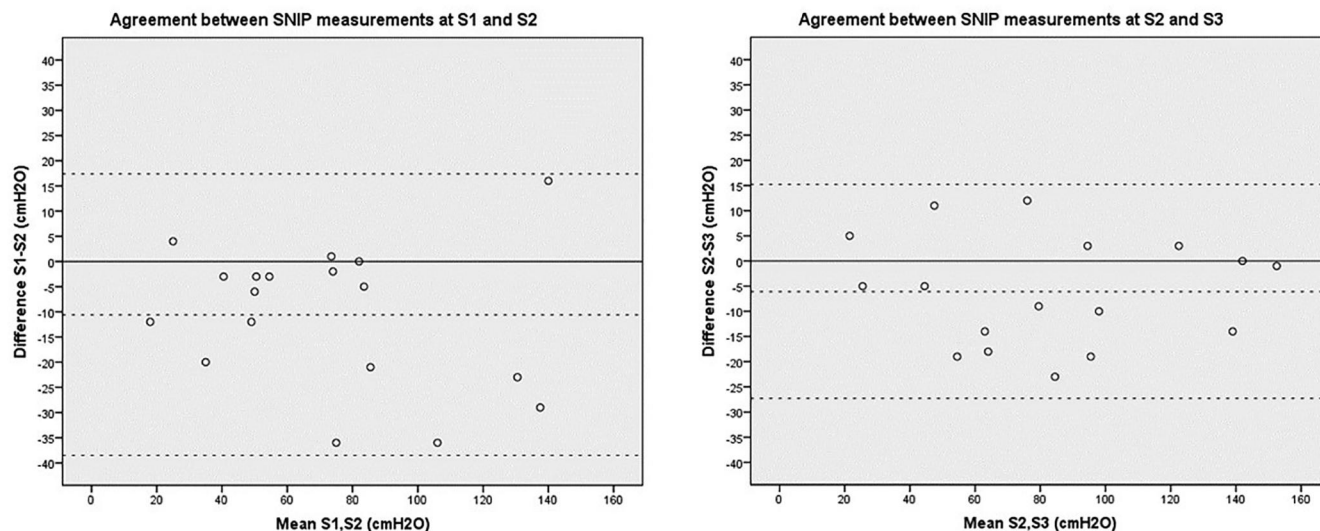


FIGURE 4 Agreement of repeated sniff nasal inspiratory pressure (SNIP) measurements between first (S1) and second (S2) testing session, and between second (S2) and third (S3) testing session. Solid lines indicate the lines of equality (no difference between measurements). Three dashed lines indicate the mean difference between measurements (bias) and the upper and lower 95% limits of agreement ($\text{bias} \pm 1.96 \times \text{SD}$)

variability is likely to represent factors other than recovery from stroke, for example, diurnal fluctuations in fatigue, attention/concentration, and motivation. Data analysis was comprehensive, using both reliability (ICC) and Bland–Altman analyses. The latter identify individual variability and - importantly - allow an appreciation of measurement variability in the unit of measurement.

The study was limited in that data represent intrarater but not inter-rater agreement and reliability. Investigators employing several assessors for repeated measurements of any one participant are advised to establish study-specific reference values for inter-rater agreement and reliability. It is acknowledged that spirometry, which was conducted at the beginning of testing session S1 but not in sessions S2 and S3, introduced added demand on participants' respiratory muscles in S1. While our data do not indicate that this may have led to systematically lower readings of MIP, MEP, and SNIP at S1 (e.g., due to spirometry causing respiratory muscle fatigue), it is acknowledged that a protocol with three identical testing sessions S1–S3 would be preferable. Lastly, the study was powered to achieve a certain level of precision of the ICC. CIs for bias and limits of agreement in Bland–Altman analyses were relatively wide, spanning between 9 and 24 cmH₂O. It may be warranted to replicate this study with a sample size powered to achieve greater precision in limits of agreement and methods for estimating required sample sizes for this have been described (Lu et al., 2016).

5 | CONCLUSION

In conclusion, this study has demonstrated that repeated measurements of MIP, MEP, and SNIP in the acute and subacute phases of stroke show consistent group averages and variances, and good test–retest reliability. However, in some participants, there were

considerable discrepancies in absolute agreement of up to 23, 29, and 38 cmH₂O for MIP, MEP, and SNIP, respectively. The magnitude of these discrepancies was not correlated with time since stroke or with measures of stroke-related impairment, leading us to conclude that test–retest variability should be considered across stroke populations, and individually assessed to inform interpretation of change over time.

6 | IMPLICATIONS FOR PHYSIOTHERAPY PRACTICE

It is feasible to conduct repeated measurements of MIP, MEP, and SNIP during the acute and subacute phases of stroke. It may be recommended that investigators incorporate repeated baseline measurements of MIP, MEP, and SNIP into study designs, to establish study-specific intra- and inter-rater reliability, and to identify individuals with high variability of repeated measurements. This will facilitate the interpretation of (change in) longitudinal measurements in the context of expected variability and minimal detectable difference.

ACKNOWLEDGMENTS

We thank Robert Grant for his assistance with the sample size estimate.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICAL APPROVAL

The study was approved by the bio-ethical committee of the Institute of Psychiatry and Neurology, Warsaw, Poland (reference number 29/2015).

ORCID

Agnieszka Lewko  <https://orcid.org/0000-0001-5688-0762>

Stefan Tino Kulnik  <https://orcid.org/0000-0001-5419-6713>

Maciej Krawczyk  <https://orcid.org/0000-0003-2230-098X>

REFERENCES

- American Thoracic Society/European Respiratory Society. (2002). ATS/ERS statement on respiratory muscle testing. *American Journal of Respiratory and Critical Care Medicine*, 166(4), 518–624. <https://doi.org/10.1164/rccm.166.4.518>
- Banks, J. L., & Marotta, C. A. (2007). Outcomes validity and reliability of the modified Rankin scale: Implications for stroke clinical trials: A literature review and synthesis. *Stroke*, 38(3), 1091–1096. <https://doi.org/10.1161/01.STR.0000258355.23810.c6>
- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2), 135–160. <https://doi.org/10.1177/096228029900800204>
- Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, 21(9), 1331–1335. <https://doi.org/10.1002/sim.1108>
- Carr, J. H., Shepherd, R. B., Nordholm, L., & Lynne, D. (1985). Investigation of a new motor assessment scale for stroke patients. *Physical Therapy*, 65(2), 175–180. <https://doi.org/10.1093/ptj/65.2.175>
- de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, 59(10), 1033–1039. <https://doi.org/10.1016/j.jclinepi.2005.10.015>
- Dimitriadis, Z., Kapreli, E., Konstantinidou, I., Oldham, J., & Strimpakos, N. (2011). Test/retest reliability of maximum mouth pressure measurements with the MicroRPM in healthy volunteers. *Respiratory Care*, 56(6), 776–782. <https://doi.org/10.4187/respcare.00783>
- Domholdt, E. (2005). *Rehabilitation research: Principles and applications*. St. Louis, MO: Elsevier Saunders.
- Goldstein, L. B., Bertels, C., & Davis, J. N. (1989). Interrater reliability of the NIH stroke scale. *Archives of Neurology*, 46(6), 660–662. <https://doi.org/10.1001/archneur.1989.00520420080026>
- Gomes-Neto, M., Saquetto, M. B., Silva, C. M., Carvalho, V. O., Ribeiro, N., & Conceicao, C. S. (2016). Effects of respiratory muscle training on respiratory function, respiratory muscle strength, and exercise tolerance in patients poststroke: A systematic review with meta-analysis. *Archives of Physical Medicine and Rehabilitation*, 97(11), 1994–2001. <https://doi.org/10.1016/j.apmr.2016.04.018>
- Gosselink, R., De Vos, J., van den Heuvel, S. P., Segers, J., Decramer, M., & Kwakkel, G. (2011). Impact of inspiratory muscle training in patients with COPD: What is the evidence? *European Respiratory Journal*, 37(2), 416–425. <https://doi.org/10.1183/09031936.00031810>
- Harraf, F., Ward, K., Man, W., Rafferty, G., Mills, K., Polkey, M., & Kalra, L. (2008). Transcranial magnetic stimulation study of expiratory muscle weakness in acute ischemic stroke. *Neurology*, 71(24), 2000–2007. <https://doi.org/10.1212/01.wnl.0000336927.30977.56>
- Hernaes, R. (2015). Reliability and agreement studies: A guide for clinical investigators. *Gut*, 64(7), 1018–1027. <https://doi.org/10.1136/gutjnl-2014-308619>
- Iwakura, M., Okura, K., Kobuta, M., Sugawara, K., Kawagoshi, A., Takahashi, H., & Shioya, T. (2020). Estimation of minimal clinically important difference for quadriceps and inspiratory muscle strength in older outpatients with chronic obstructive pulmonary disease: A prospective cohort study. *Physical Therapy Research Article ID E10049* <https://doi.org/10.1298/ptr.E10049>
- Kulnik, S. T. (2015a). Should we train respiratory muscles after stroke? *Neurology*, 85(7), 560–561. <https://doi.org/10.1212/WNL.0000000000001846>
- Kulnik, S. T. (2015b). *Pilot studies to develop and evaluate a muscle strengthening programme to reduce the risk of aspiration and improve outcome in stroke patients*. London, UK: King's College London. Retrieved from [https://kclpure.kcl.ac.uk/portal/en/theses/pilot-studies-to-develop-and-evaluate-a-muscle-strengthening-programme-to-reduce-the-risk-of-aspiration-and-improve-outcome-in-stroke-patients\(3e917f5c-153f-4fcd-94ae-709331d8ba58\).html](https://kclpure.kcl.ac.uk/portal/en/theses/pilot-studies-to-develop-and-evaluate-a-muscle-strengthening-programme-to-reduce-the-risk-of-aspiration-and-improve-outcome-in-stroke-patients(3e917f5c-153f-4fcd-94ae-709331d8ba58).html)
- Larson, J. L., Covey, M. K., Vitalo, C. A., Alex, C. G., Patel, M., & Kim, M. J. (1993). Maximal inspiratory pressure. learning effect and test-retest reliability in patients with chronic obstructive pulmonary disease. *Chest*, 104(2), 448–453. <https://doi.org/10.1378/chest.115.6.1546>
- Laveneziana, P., Albuquerque, A., Aliverti, A., Babb, T., Barreiro, E., Dres, M., & Verges, S. (2019). ERS statement on respiratory muscle testing at rest and during exercise. *European Respiratory Journal*, 53(6), 1801214. <https://doi.org/10.1183/13993003.01214-2018>
- Lista Paz, A., Gonzalez Doniz, L., Ortigueira Garcia, S., Saleta Canosa, J. L., & Moreno Couto, C. (2016). Respiratory muscle strength in chronic stroke survivors and its relation with the 6-minute walk test. *Archives of Physical Medicine and Rehabilitation*, 97(2), 266–272. <https://doi.org/10.1016/j.apmr.2015.10.089>
- Lofaso, F., Nicot, F., Lejaille, M., Falaize, L., Louis, A., Clement, A., & Faroux, B. (2006). Sniff nasal inspiratory pressure: What is the optimal number of sniffs? *European Respiratory Journal*, 27(5), 980–982. <https://doi.org/10.1183/09031936.06.00121305>
- Lu, M., Zhong, W., Liu, Y., Miao, H., Li, Y., & Ji, M. (2016). Sample size for assessing agreement between two methods of measurement by Bland-Altman method. *The International Journal of Biostatistics*, 12(2), 20150039. <https://doi.org/10.1515/ijb-2015-0039>
- Luvizutto, G. J., Dos Santos, M. R. L., Sartor, L. C. A., da Silva Rodrigues, J. C., da Costa, R. D. M., Braga, G. P., & Bazan, R. (2017). Evaluation of respiratory muscle strength in the acute phase of stroke: The role of aging and anthropometric variables. *Journal of Stroke and Cerebrovascular Disease*, 26(10), 2300–2305. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.05.014>
- Mahoney, F. I., & Barthel, D. W. (1965). Functional evaluation: The Barthel Index: A simple index of independence useful in scoring improvement in the rehabilitation of the chronically ill. *Maryland State Medical Journal*, 14, 61–65.
- Maillard, J. O., Burdet, L., van Melle, G., & Fitting, J. W. (1998). Reproducibility of twitch mouth pressure, sniff nasal inspiratory pressure, and maximal inspiratory pressure. *European Respiratory Journal*, 11(4), 901–905. <https://doi.org/10.1183/09031936.98.11040901>
- Martin-Valero, R., De La Casa Almeida, M., Casuso-Holgado, M. J., & Heredia-Madrado, A. (2015). Systematic review of inspiratory muscle training after cerebrovascular accident. *Respiratory Care*, 60(11), 1652–1659. <https://doi.org/10.4187/respcare.03981>
- McConnell, A. (2013). *Respiratory muscle training: Theory and practice*. Edinburgh, UK: Elsevier/Churchill Livingstone.
- Menezes, K. K., Nascimento, L. R., Ada, L., Polese, J. C., Avelino, P. R., & Teixeira-Salmela, L. F. (2016). Respiratory muscle training increases respiratory muscle strength and reduces respiratory complications after stroke: A systematic review. *Journal of Physiotherapy*, 62(3), 138–144. <https://doi.org/10.1016/j.jphys.2016.05.014>
- Menezes, K. K., Nascimento, L. R., Avelino, P. R., Alvarenga, M. T. M., & Teixeira-Salmela, L. F. (2018). Efficacy of interventions to improve respiratory function after stroke. *Respiratory Care*, 63(7), 920–933. <https://doi.org/10.4187/respcare.06000>
- Miller, M. R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., ... ATS/ERS Task Force (2005). Standardisation of spirometry. *European Respiratory Journal*, 26(2), 319–338. <https://doi.org/10.1183/09031936.05.00034805>
- Nikoloutou, D., Rafferty, G., Man, W. D., Mustafa, N., Donaldson, N., Grant, R. L., & Moxham, J. (2014). Sniff nasal inspiratory pressure in patients with moderate-to-severe chronic obstructive pulmonary

- disease: Learning effect and short-term between-session repeatability. *Respiration*, 88(5), 365–370. <https://doi.org/10.1159/000365998>
- Parreiras de Menezes, K. K., Nascimento, L. R., Ada, L., Avelino, P. R., Polese, J. C., Mota Alvarenga, M. T., & Teixeira-Salmela, L. F. (2019). High-intensity respiratory muscle training improves strength and dyspnea poststroke: A double-blind randomized trial. *Archives of Physical Medicine and Rehabilitation*, 100(2), 205–212. <https://doi.org/10.1016/j.apmr.2018.09.115>
- Phillips, W. T., Batterham, A. M., Valenzuela, J. E., & Burkett, L. N. (2004). Reliability of maximal strength testing in older adults. *Archives of Physical Medicine and Rehabilitation*, 85(2), 329–334. <https://doi.org/10.1016/j.apmr.2003.05.010>
- Phillips, W. T., Benton, M. J., Wagner, C. L., & Riley, C. (2006). The effect of single set resistance training on strength and functional fitness in pulmonary rehabilitation patients. *Journal of Cardiopulmonary Rehabilitation*, 26(5), 330–337. <https://doi.org/10.1097/00008483-200609000-00011>
- Pinheiro, M. B., Polese, J. C., Faria, C. D., Machado, G. C., Parreira, V. F., Britto, R. R., & Teixeira-Salmela, L. F. (2014). Inspiratory muscular weakness is most evident in chronic stroke survivors with lower walking speeds. *European Journal of Physical and Rehabilitation Medicine*, 50(3), 301–307 PMID: 24525623.
- Polkey, M. I., & Moxham, J. (2004). Improvement in volitional tests of muscle function alone may not be adequate evidence that inspiratory muscle training is effective. *European Respiratory Journal*, 23(1), 5–6. <https://doi.org/10.1183/09031936.03.00107403>
- Pollock, R. D., Rafferty, G. F., Moxham, J., & Kalra, L. (2013). Respiratory muscle strength and training in stroke and neurology: A systematic review. *International Journal of Stroke*, 8(2), 124–130. <https://doi.org/10.1111/j.1747-4949.2012.00811.x>
- Reyes, A., Cruickshank, T., Ziman, M., & Nosaka, K. (2014). Pulmonary function in patients with Huntington's disease. *BMC Pulmonary Medicine*, 14, 89. <https://doi.org/10.1186/1471-2466-14-89>
- Reyes, A., Castillo, A., Castillo, J., & Cornejo, I. (2018). The effects of respiratory muscle training on peak cough flow in patients with Parkinson's disease: A randomized controlled study. *Clinical Rehabilitation*, 32(10), 1317–1327. <https://doi.org/10.1177/0269215518774832>
- Smeltzer, S. C., & Lavietes, M. H. (1999). Reliability of maximal respiratory pressures in multiple sclerosis. *Chest*, 115(6), 1546–1552. <https://doi.org/10.1378/chest.115.6.1546>
- Taylor, J. R. (1997). *An introduction to error analysis : The study of uncertainties in physical measurements*. Sausalito, CA: University Science.
- Verheyden, G., Nieuwboer, A., Mertin, J., Preger, R., Kiekens, C., & De Weerd, W. (2004). The trunk impairment scale: A new tool to measure motor impairment of the trunk after stroke. *Clinical Rehabilitation*, 18(3), 326–334. <https://doi.org/10.1191/0269215504cr733oa>
- Ward, K., Seymour, J., Steier, J., Jolley, C. J., Polkey, M. I., Kalra, L., & Moxham, J. (2010). Acute ischaemic hemispheric stroke is associated with impairment of reflex in addition to voluntary cough. *European Respiratory Journal*, 36(6), 1383–1390. <https://doi.org/10.1183/09031936.00010510>
- Ward, K., Rao, P., Reilly, C. C., Rafferty, G. F., Polkey, M. I., Kalra, L., & Moxham, J. (2017). Poor cough flow in acute stroke patients is associated with reduced functional residual capacity and low cough inspired volume. *BMJ Open Respiratory Research*, 4(1), e000230. <https://doi.org/10.1136/bmjresp-2017-000230>
- Xiao, Y., Luo, M., Wang, J., & Luo, H. (2012). Inspiratory muscle training for the recovery of function after stroke. *Cochrane Database of Systematic Reviews*, 5, CD009360. <https://doi.org/10.1002/14651858.CD009360.pub2>

How to cite this article: Lewko A, Sidaway M, Kulnik ST, Krawczyk M. Agreement and reliability of repeated bedside respiratory muscle strength measurements in acute and subacute stroke. *Physiother Res Int*. 2021;26:e1892. <https://doi.org/10.1002/pri.1892>