



Comparison of patient- and clinician-reported outcome measures in lower back rehabilitation: introducing a new integrated performance measure (t2D)

Andrej Zdravkovic¹ · Vincent Grote^{1,2,5} · Michael Pirchl² · Martin Stockinger² · Richard Crevenna¹ · Michael J. Fischer^{2,3,4}

Accepted: 4 June 2021 / Published online: 15 June 2021
© The Author(s) 2021

Abstract

Purpose Patient- and clinician-reported outcome measures (PROMs, CROMs) are used in rehabilitation to evaluate and track the patient's health status and recovery. However, controversy still exists regarding their relevance and validity when assessing a change in health status.

Methods We retrospectively analyzed the changes in a CROM (Fingertip-To-Floor Test – FTF) and PROMs (ODI, HAQ-DI, NPRS, EQ5D) and the associations between these outcomes in 395 patients with lower back pain (57.2 ± 11.8 years, 49.1% female). We introduced a new way to measure and classify outcome performance using a distribution-based approach (t2D). Outcome measures were assessed at baseline and after 21 days of inpatient rehabilitation.

Results Overall, the rehabilitation (Cohens $d=0.94$) resulted in a large effect size outcome. Medium effect sizes were observed for FTF ($d=0.70$) and PROMs ($d>0.50$). Best performance rating was observed for pain (NPRS). We found that 13.9% of patients exhibited a deterioration in the PROMs, but only 2.3%, in the FTF. The correlation between the PROMs and FTF were low to moderate, with the highest identified for HAQ-DI ($\rho=0.30-0.36$); no significant correlations could be shown for changes. High consistency levels were observed among the performance scores (t2D) in 68.9% of the patients.

Conclusions Different and complementary assessment modalities of PROMs and CROMs can be used as valuable tools in the clinical setting. Results from both types of measurements and individual performance assessments in patients provide a valid basis for the meaningful interpretation of the patients' health outcomes.

Trial registration. This clinical study was entered retrospectively on August 14, 2020 into the German Clinical Trials Register (DRKS, registration number: DRKS00022854).

Keywords Patient-reported outcome measures · Clinician-reported outcome measures · Performance score · Orthopedic rehabilitation · Inpatient · Lower back pain

Abbreviations

CI	Confidence interval
CROM	Clinician-reported outcome measure
d	Cohen's d (effect size)

These authors Andrej Zdravkovic and Vincent Grote have contributed equally to this work.

✉ Vincent Grote
vincent.grote@rehabilitation.lbg.ac.at

✉ Michael J. Fischer
Michael.Fischer@reha-kitz.at

¹ Department of Physical Medicine, Rehabilitation and Occupational Medicine, Medical University of Vienna, Vienna, Austria

² Ludwig Boltzmann Institute for Rehabilitation Research, Reizenpfeninggasse 1, 1140 Vienna, Austria

³ Vamed Rehabilitation Center Kitzbühel, Hornweg 32, 6370 Kitzbühel, Austria

⁴ Hannover Medical School MHH, Clinic for Rehabilitation Medicine, Hannover, Germany

⁵ Otto Loewi Research Center, Division of Physiology, Medical University of Graz, Graz, Austria

EQ5D	European Quality of Life-5 Dimensions (questionnaire)
EQ-VAS	VAS valuations of EQ-5D (self-rated health in %; 0–100)
HAQ-DI	Health Assessment Questionnaire Disability Index
MQO	Medical quality outcome
$N(n)$	Sample size
NPRS	Numerical pain rating scale
p	Significance level (risk of error)
PROM	Patient-reported outcome measure
r	Correlation coefficient (Pearson)
ρ	Spearman's rank correlation
SD	Standard deviation
SMD	Standardized mean difference
$t1$	Pre-test (baseline) score
$t2$	Post-test (end of rehabilitation) score
$t2D$	Performance score ($t2 + \Delta$)
TTO	Time trade-off (EQ5D valuation technique)
VAS	Visual analogue scale
z	z -Value, standard score (scale: 0 ± 1 ; mean \pm SD)
η_p^2	Partial eta ² (effect size)

Introduction

Low back pain (LBP) is highly prevalent in the general population [1] and has been ranked sixth globally in terms of overall disease burden [2]. Exercise therapy, which is often prescribed for LBP, has been found to positively affect pain levels, as well as physical functioning in LBP patients [3]. However, it is still unclear whether subjective, patient-reported outcome measures (PROMs) reflect the actual course of convalescence in LBP rehabilitation, if this is better achieved by applying objective, clinician-reported outcome measures (CROMs) alone, or if a combination of PROMs and CROMs should be applied. The responsiveness and validity of different PROMs have been studied in LBP patients, although usually not in relation to changes in CROMs [4–6]. Although several PROMs may exhibit certain methodological limitations [7], they represent a necessary tool that can be used to involve patients in therapy [6] and help predict the socioeconomic cost of LBP [8].

Despite some controversy regarding the optimal methods and techniques for the measurement of lumbar function, lumbar flexion tests form the cornerstone of assessments in impairment due to LBP [9]. Apart from the historical reasons for their use, measurements of spinal flexion have been shown to correlate highly with the degree of disability [10]. The Fingertip-to-Floor test (FTF) and the Schoeber test represent commonly used CROMs which have been validated for the LBP patient population [11, 12].

In this and in other patient populations, however, researchers have observed a divergence between self-reported outcome measurements and performance outcomes. They have also identified multiple factors that can influence this discrepancy in outcomes, including gender, education and mood [13–16]. This discrepancy needs to be considered in light of the different health conditions of the patients, in order to find a performance evaluation that also adequately considers those patients who already show satisfactory results (ceiling effects). These findings indicate that further research is warranted to ascertain the relative changes in PROMs and CROMs, as well as their mutual coherence. Furthermore, the simplicity of reporting outcomes in the clinical setting should not be neglected, and user-friendly options should be developed, in order to facilitate the implementation of research findings into clinical practice. To this end, in this paper, we describe a new way to evaluate outcomes and the results of an analysis of the PROMs and CROMs of LBP patients before and after a course of multidisciplinary inpatient rehabilitation at a center in Tyrol, Austria.

Methods

Study aim, design and setting

LBP patients were treated at a specialized orthopedic rehabilitation center in Austria. The most frequent International Classification of Diseases, 10th Revision (ICD-10), diagnoses noted at the study center were M51.1, M53.9, M54.9, M51.2 and M54.4 (all $n > 30$). The PROMs consisted of the following instruments: Oswestry Disability Index (ODI), Numeric Pain Rating Scale (NPRS), Health Assessment Questionnaire Disability Index (HAQ-DI) and the Five-Level EuroQol-5D (EQ5D-5L). The CROM used was the Fingertip-to-Floor test (FTF). In a retrospective cohort study, the changes and correlations between the CROM and PROMs were analyzed at the beginning ($t1$) and end ($t2$) of orthopedic rehabilitation. We empirically tested and present a new method that can be used to measure and stratify outcome performance. This method uses a distribution-based approach and is based on two measurements at the beginning and end of rehabilitation, the “performance score ($t2D$)”. The patients were fully informed about the study content and purpose and gave their written consent to participate.

Intervention

The inpatient program lasted 21 days, as defined in the service portfolio of the Austrian social security institutions [17]. The medical treatments last on average 2–3 h per day, including exercise therapy, electrotherapy, lymphatic drainage and massage as well as hydrotherapy. These treatments

amounted to at least 1800 therapy minutes during the three-week rehabilitation program. The amount of individual therapy depends on the medical history if the rehabilitation program is classified as a follow-up treatment procedure after surgery.

Ethics approval

The Ethics Committee of the Medical University of Innsbruck approved the study protocol on August 23, 2019 (Ref: EC Nr: 1158/2019) in accordance with the current version of the Declaration of Helsinki.

Clinician-reported outcome measure

The CROM used in the assessments was the FTF. While different techniques can be used to measure lumbar flexion, including the FTF and the Schoeber Test [12], the method used should be safe, user-friendly and time-efficient in clinical practice. We preferred the FTF, as it fulfills the mentioned criteria. The FTF tests combine spinal and hip flexion and correlate highly with radiographic measurements of lumbar flexion [11].

Patient-reported outcome measures

PROMs are standardized, validated questionnaires that are completed by patients in order to measure perceptions of their functional status and wellbeing [18]. Outcomes reflect the overall care for a patient's medical condition, in which professionals in multiple specialties are usually involved [19]. Professionals using PROMs as clinical tools need to be sensitive to the situation of the individual patients. PROMs can provide insights that support direct clinical decision-making and enhance experiences of care [20].

Numeric pain rating scale (NPRS)

Pain intensity and impairments in physical functions are associated in patients with chronic pain, and improvement in pain has been associated with improvements in functioning [21]. There are two aspects of pain, which can be evaluated independently. Firstly, the intensity or how strong the pain feels and, secondly, the affective dimension of pain or how unpleasant the pain feels [22]. Self-report measures provide the 'gold standard' in assessing pain, as they reflect the subjective nature of pain. The commonly used methods of rating pain include the visual analogue scale, verbal rating scales and – the method used in our study – numerical rating scales [21].

ODI 2.1a

The Oswestry Disability Index is an instrument used to quantify disability in patients with low back pain, which was originally described in 1980 [23]. The questionnaire has since been revised, with the current version being 2.1a [24]. It encompasses ten dimensions of disability involving pain, personal care, lifting, walking, sitting, standing, sleeping, sex life, social life and travelling [25]. The score is rated on a percentage scale, with 0% representing no disability and 100% representing the highest degree of disability. The German version used in the current study has been validated in a German-speaking population [26].

HAQ-DI

The Health Assessment Questionnaire was first proposed in 1980 as a comprehensive measure of a patient's health status and patient-centered care in rheumatoid arthritis [27]. The disability dimension of the original questionnaire is widely used, as it addresses common activities of daily living. These are scored on a scale of 0–3, corresponding to "do without difficulty" and "unable to do", respectively [27–29].

EQ5D-5L

The EuroQol-5D is a generic instrument used for assessing health-related quality of life. It was designed as a self-complete questionnaire. The EuroQol-5D was originally introduced with three levels of severity in 1990 by the EuroQol Group [30–32] and was subsequently expanded to include five levels to improve its responsiveness and reduce the ceiling effect [33]. The EQ5D-5L is applied to measure five dimensions of health status [34]: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. In addition, the subjective overall health status is estimated using a visual analogue scale (EQ-VAS [0–100]). The five dimensions are rated on five severity levels, with 1 corresponding to "no problem" and 5 to "unable to do / extreme problems". The value set of the EQ5D-5L (EQ5D TTO [-0.66–1.00]) for a German population has been published elsewhere [35].

Performance score ($t2D: t2 + \Delta$)

Taking objective measurements of physical mobility can lead to ceiling effects. For example, if a patient has good lumbar flexion (FTF test; spinal and hip flexion) at the beginning of rehabilitation, no strong increase (change) is expected during the course of the rehabilitation. In this case, although the patient's overall performance may be good, only a slight or no increase can be measured statistically. Thus, to assess the performance of patients using each score, a new method was introduced to account for the fact that the change in scores

depends on the patient’s initial functional status. The simple formula $t2 + (t2 - t1)$ best reflects the performance and considers the functional status at the end of rehabilitation and improvements (changes from $t1$ to $t2$; Δ). It was possible to interpret “performance scores” using a distribution-based approach, in which the $t2 + \Delta$ were transformed into standardized scores for $t1$ and $t2$ with z -transformation.

Statistical analysis

SPSS for Windows (version 27) was used for data analysis. For each outcome measure, score differences (Δ , changes) between the beginning ($t1$, pre-test score) and the end ($t2$, post-test score) of rehabilitation were calculated and tested for significant changes using t -tests. For multiple comparisons, 2×2 MANOVA for repeated measurements was used. Z -values and effect sizes for within-subjects designs were calculated (Cohen’s d and partial Eta-squared, η_p^2). Effect sizes were interpreted according to Cohen [36], while correlations between CROM and PROMs were determined using Spearman’s rank correlation coefficients (ρ), Pearson product-moment correlations (r) or linear regression models for $t1$, $t2$, changes (Δ) and performance scores ($t2 + \Delta$).

The difference between the standardized performance (z -)scores of PROMs and CROMs was calculated to show the level of consistency. Performance scores ($t2D$; $t2_z + \Delta_z$) for each outcome measure were classified as high consistent (within one/same tertile; cut-off: normalized z -difference between absolute value of $t2_z + \Delta_z$ of CROM and PROMs < 0.43), moderate consistent (if the scores ranged between one and two tertiles; z -difference within 0.43–0.97),

or as low consistent and discrepant (more than two tertiles of difference between performance scores; z -difference > 0.97). By chance, this would result in an equivalence of 33.3% in each category, if no correlation existed between the different measured performance outcomes.

Results

A total of 395 LBP patients who underwent a standardized rehabilitation program between January and December 2018 were included in this study. All LBP patients suffered from afflictions of the lower back and either had or had not experienced recent surgical treatment. The average age of the patients was 57.2 years with a standard deviation of 11.8 years. Among these patients, 49.1% were women. Post-intervention, statistically significant changes in the EQ5D, NPRS, HAQ, ODI and FTF were detected (all $p < 0.001$) with medium effect sizes (Cohen’s $d = 0.55$ – 0.70). The changes in PROMs and the FTF are shown in Table 1.

The percentage of patients who showed an improvement, deterioration, or no change of the outcome measures is shown in Table 2. The overall Medical Quality Outcome (MQO; mean of PROMs and CROM-FTF) revealed improvements in 68.9% of patients immediately at the end of the rehabilitation (cut-off: z -difference ($t2 - t1$) < -0.20 ; [16]). The status of 24.8% of the patients remained unchanged (0.00 ± 0.20), and the conditions of 6.3% worsened (> 0.20) between the beginning and the end of the inpatient rehabilitation. The overall improvement in

Table 1 Patient’s health status and changes in PROMs and the FTF

Quality-of-outcome measures		$t1$	$t2$	Δ	p	Cohen’s d_z
CROM	FTF	17.9 ± 15.1	13.2 ± 12.4	-4.71 ± 6.74	***	0.70
	FTF [z]	0.17 ± 1.08	-0.17 ± 0.88	-0.34 ± 0.48	***	0.70
PROMs	EQ5D Health (EQ-VAS)	61.1 ± 19.3	66.7 ± 22.7	5.63 ± 22.57	***	0.25
	EQ5D TTO	0.81 ± 0.17	0.87 ± 0.15	0.06 ± 0.14	***	0.39
	NPRS	4.58 ± 2.07	3.50 ± 1.98	-1.09 ± 1.88	***	0.58
	HAQ	0.27 ± 0.29	0.23 ± 0.31	-0.04 ± 0.17	***	0.25
	ODI	22.6 ± 14.2	17.7 ± 14.2	-4.84 ± 8.86	***	0.55
	Mean PROMs [z]	0.21 ± 0.97	-0.21 ± 0.99	-0.41 ± 0.63	***	0.66
Overall MQO	Medical Outcome [z]	0.23 ± 1.02	-0.23 ± 0.93	-0.45 ± 0.48	***	0.94

CROM-FTF, mean of PROMs and the overall medical outcome (MQO; mean of PROMs and CROM-FTF) are highlighted in bold

Quality-of-outcome measures were documented in the discharge report at the beginning ($t1$) and at the end ($t2$) of the 21-day inpatient rehabilitation program. The PROMs consisted of Oswestry Disability Index (ODI), Numeric Pain Rating Scale (NPRS), the Health Assessment Questionnaire Disability Index (HAQ-DI) and the Five-Level EuroQol-5D (EQ5D-5L). The CROM was the Fingertip-to-Floor test (FTF), where the optimal value in this study has been defined as 0 [cm]. Differences between those measurements (difference: $t2-t1$) and effect sizes (Cohen’s d_z) were used to evaluate recovery in rehabilitation. The level of statistical significance was reached for all outcome measures (all $p < 0.001$ ***; η_p^2 multivariate = 0.490)

n (m/f): 395 (201/194)

Table 2 Improvements of outcome measurements

SMD (<i>z</i> -differences) classified	<i>N</i> [%]	Better (+) (%)	Equal (=) (%)	Worse (–) (%)
CROM (1)	FTF ⁺	50.9	46.8	2.3
PROMs (6)	EQ5D Health (EQ-VAS)	54.2	23.3	22.5
	EQ5D TTO	52.9	33.2	13.9
	NPRS	61.8	20.8	17.5
	HAQ ⁺⁺	31.6	54.7	13.7
	ODI	57.2	28.1	14.7
	Mean PROMs	62.3	23.8	13.9
Overall MQO (PROMs and CROM-FTF)		68.9	24.8	6.3

CROM-FTF, mean of PROMs and the overall medical outcome (MQO; mean of PROMs and CROM-FTF) are highlighted in bold

Changes between admission and discharge (categorical presentation: better, equal, worse); The threshold used was an average *z*-difference (SMD) of > 0.20. + ... Results for subsample without 73 LBP Patients who had optimal FTF values (0) for *t*₁ and *t*₂: 62.4% better, 34.8% equal and 2.8% worse (*n*₁ = 322); + + ... Results for the subsample also without (not the same) 73 LBP patients who had optimal HAQ scores (0) for *t*₁ and *t*₂: 38.8% better, 44.4% equal and 16.8% worse (*n*₂ = 322)

Normalized changes between admission (*t*₁) to discharge (*t*₂) are revealed by examining the effect sizes (*z*, SMD) and the number of patients (*n* [%]), which could be improved in clinically relevant ways [16]. Based on the value distributions, the individual outcome parameters were transformed into *z*-values. By means of *z*-standardization, differently scaled quantities were summarized, and the changes were uniformly quantified. *Z*-differences from 0.00 ± 0.20 were classified as equal (no changes) [16, 36]

MQO was around 13.12 ± 14.24 percentile points or a SMD of 0.45 ± 0.48 (*r*_{*t*₁, *t*₂} = 0.89; Cohen's *d* = 0.94, 95% CI [0.83, 1.06]). The majority of the measures improved following the intervention; however, 13.7–22.5% of patients exhibited a deterioration in the individual PROMs, whereas only 2.3% experienced a deterioration in the FTF.

The correlations among the different outcome measures are shown in Table 3a–d. Although multiple measures reached the level of statistical significance (*p* < 0.001; η_p^2 _{multivariate} = 0.490), the correlation between the FTF and the PROMs was weak (*rho* = 0.16–0.36). The strongest relation to CROM was observed for HAQ-DI (*rho* = 0.30–0.36), which was confirmed by linear regression models. The relationships at the beginning and end of rehabilitation are similar, but no correlations between changes (differences from *t*₁ to *t*₂) or between FTF and PROMs could be found (all *rho* < 0.05). In contrast, the performance scores in the PROMs showed significant correlations with the CROM-FTF performance score (R^2 _{linear regression} = 0.09; *p* < 0.001).

Figure 1 shows changes in the results of a Fingertip-To-Floor (FTF) test as compared to baseline values (*t*₁) and the categorized performance score. Based on this example of an objective measurement of physical mobility, a patient's performance could still be classified as good or medium, even if only a small or no increase could be measured during rehabilitation. Classification on improvements in Fig. 2 are based on the value distributions (changes; SMD; Fig. 2a; [16]) and tertiles for FTF performance score (Fig. 2b right). A cubic relationship (*r*² = 0.31, *p* < 0.001) could be observed between the different perspectives of an outcome assessment. Based

on the method of performance evaluation (*t*_{2D}) a distinction was made between the high, moderate, or discrepant consistency of CROM-FTF and PROM outcomes (Table 4 and Fig. 3).

In most cases (68.9%), the performance scores between CROM-FTF and PROMs pointed in the same direction (Table 4). Specifically, 27.3% (18.2–34.7%) of results in PROM performance scores showed a moderate agreement in terms of performance with CROM-FTF. In 3.8% of the cases, the results between CROM vs PROMs were contrary to each other (discrepant). The strongest descriptive performance score was observed for NPRS (*t*_{2_z} + Δ_z = -0.78 ± 1.56).

The stochastic test distribution of a performance score shows that a dependence exists between the *t*_{2_z} + Δ_z with a *t*_{2_z} of *r* = 0.90 and with a *t*₁ of *r* = -0.45, respectively. This is quite similar to the dependence measured between *t*₁ or *t*₂ scores and the difference score (Δ_z , *r* = -0.70). A correlation of independent, normally distributed random variables between *t*_{2_z} + Δ_z and changes (*t*₂ - *t*₁) would result in a correlation coefficient of 0.95.

For each outcome measure, score differences (Δ_z , changes) between the beginning (*t*₁, pre-score) and the end (*t*₂, post-score) of rehabilitation were calculated. These differences were tested to detect significant interactions with moderating factors like sex, age, the BMI, or ICD diagnoses (Table 5). The test results show that only BMI (η_p^2 = 0.041, *p* < 0.001) and baseline values (η_p^2 = 0.182, *p* < 0.001) served as critical success (between) factors that contributed to significant changes in outcome measurements. Obese LBP patients with a BMI > 30 and patients with poorer FTF-initial values (*t*₁)

Table 3 **abcd** Intercorrelation of the measurements for $t1$, $t2$, differences (Δ) and $t2 + \Delta$

(a) Spearman correlations (rho) for $t1$; $n = 395$						
$t1$	FTF-CROM	EQ-VAS	EQ5D TTO	NPRS	HAQ	ODI
FTF-CROM	–					
EQ5D health (EQ-VAS)	–0.16**	–				
EQ5D TTO	–0.19**	0.39**	–			
NPRS	0.18**	–0.36**	–0.58**	–		
HAQ	0.30**	–0.37**	–0.66**	0.44**	–	
ODI	0.22**	–0.43**	–0.68**	0.58**	0.69**	–
Mean PROMs	0.27**	0.64**	0.83**	0.77**	0.77**	0.85**
Overall MQO	0.86**	0.44**	0.55**	0.52**	0.62**	0.59**
Linear regression with constant (beta coefficients) for $t1$ scores (predictor FTF $t1$; $R^2 = 0.098$; $p = 0.000$ ***; $c = 18.371$): –0.073*EQ-VAS $t1$ + –0.029*EQ5D TTO $t1$ + 0.074*NPRS $t1$ + 0.252*HAQ $t1$ + –0.036*ODI $t1$						
(b) Spearman correlations (rho) for $t2$; $n = 395$						
$t2$	FTF-CROM	EQ-VAS	EQ5D TTO	NPRS	HAQ	ODI
FTF-CROM	–					
EQ5D health (EQ-VAS)	–0.17**	–				
EQ5D TTO	–0.29**	0.45**	–			
NPRS	0.21**	–0.42**	–0.67**	–		
HAQ	0.36**	–0.41**	–0.73**	0.55**	–	
ODI	0.27**	–0.46**	–0.77**	0.64**	0.73**	–
Mean PROMs	0.32**	0.73**	0.84**	0.80**	0.80**	0.85**
Overall MQO	0.84**	0.53**	0.65**	0.57**	0.68**	0.65**
Linear regression with constant (beta coefficients) for $t2$ scores (predictor FTF $t2$; $R^2 = 0.139$; $p = 0.000$ ***; $c = 16.653$): –0.029*EQ-VAS $t2$ + –0.082*EQ5D TTO $t2$ + 0.022*NPRS $t2$ + 0.254*HAQ $t2$ + 0.037*ODI $t2$						
(c) Spearman correlations (rho) for differences ($t2 - t1$; Δ); $n = 395$						
Difference; Δ ($D = t2 - t1$)	FTF-CROM	EQ-VAS	EQ5D TTO	NPRS	HAQ	ODI
FTF-CROM	–					
EQ5D health (EQ-VAS)	0.01	–				
EQ5D TTO	0.02	0.11*	–			
NPRS	0.05	–0.12*	–0.38**	–		
HAQ	–0.00	–0.08	–0.34**	0.24**	–	
ODI	0.02	–0.17**	–0.40**	0.37**	0.34**	–
Mean PROMs	0.05	0.54**	0.67**	0.66**	0.49**	0.62**
Overall MQO	0.60**	0.42**	0.49**	0.55**	0.36**	0.49**
Linear regression with constant (beta coefficients) for difference scores (predictor FTF difference; $R^2 = 0.003$; $p = 0.964$; $c = -4.528$): –0.006*EQ-VAS D + 0.026 EQ5D TTO D + 0.023*NPRS D + 0.006*HAQ D + 0.037*ODI D						
(d) Spearman correlations (rho) for performance score ($t2 + \Delta$; $t2D$); $n = 395$						
Performance scores	FTF-CROM	EQ-VAS	EQ5D TTO	NPRS	HAQ	ODI
FTF-CROM	–					
EQ5D health (EQ-VAS)	–0.11*	–				
EQ5D TTO	–0.15**	0.22**	–			
NPRS	0.13**	–0.26**	–0.48**	–		
HAQ	0.28**	–0.26**	–0.36**	0.40**	–	
ODI	0.23**	–0.33**	–0.50**	0.53**	0.24**	–
Mean PROMs	0.23**	0.68**	0.64**	0.73**	0.63**	0.75**
Overall MQO	0.74**	0.52**	0.51**	0.56**	0.56**	0.63**
Linear regression with constant (beta coefficients) for perf. scores (predictor FTF $t2D$; $R^2 = 0.093$; $p = 0.000$ ***; $c = 7.709$): 0.001*EQ-VAS $t2D$ + –0.028*EQ5D TTO $t2D$ + –0.020*NPRS $t2D$ + 0.191*HAQ $t2D$ + 0.140*ODI $t2D$						

CROM-FTF, mean of PROMs and the overall medical outcome (MQO; mean of PROMs and CROM-FTF) are highlighted in bold

Although the level of significant correlations (*, ** or bold) was reached for multiple measures (all $p < 0.05$ *)—with the exception of the difference values—the Spearman's rho of the FTF with the PROMs was very weak to weak (rho = 0.00–0.36). A rho = 0.10 corresponds to a small effect, rho of 0.30 correspond to a middle effect, and rho values > 0.50, to a large effect size [36]

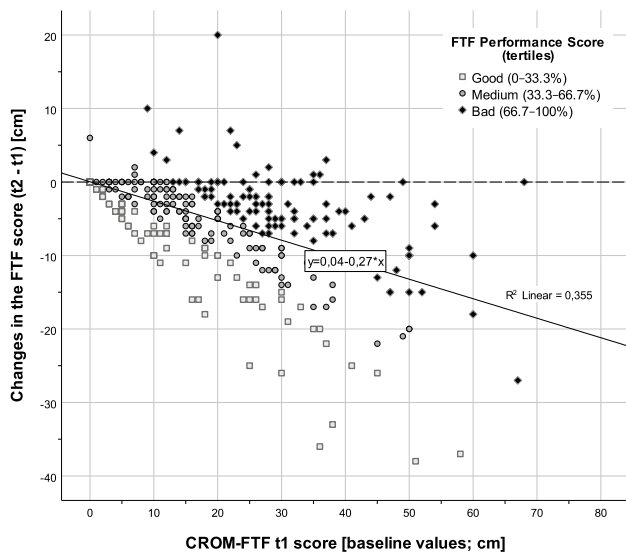


Fig. 1 Baseline values (t_1) and changes of CROM-FTF ($t_2 - t_1$). Changes of a Fingertip-To-Floor (FTF) test in relation to baseline values (t_1) and the categorized performance score ($t_2 + \Delta$; tertiles highlighted in from and color). Tertiles were chosen because the expected and observed improvements in outcome measures (see Table 2) in inpatient rehabilitation are clearly visible in around 2/3 of patients [16]. Among the LBP patients, 73 had optimal values (0) for t_1 and t_2

showed the greatest improvements in CROM-FTF. The latter finding contrasts with the CROM performance score, where most patients who already had good initial values were rated as good "performers" ($\eta_p^2 = 0.222, p < 0.001$). As shown in Table 5, the method of performance evaluation (t_2D) was much more sensitive to the individual factors as compared to the difference scores, since this method also depended on the actual functional status of the patient and rehabilitative clinical practice. Patients with poor functional status in PROMs received more individualized therapies ($\eta_p^2 = 0.043, p < 0.01$; not valid for CROM-FTF: $p > 0.05$). Younger male LBP patients could be expected to perform better in HAQ (age: $\eta_p^2 = 0.040, p < 0.001$; sex: $\eta_p^2 = 0.011, p < 0.05$), whereas female patients showed better performance in the FTF ($\eta_p^2 = 0.016, p < 0.05$). In addition to already good initial values (all $p < 0.001$; overall MQO: $\eta_p^2 = 0.290$), the BMI had the greatest influence on the performance evaluation ($\eta_p^2 = 0.027, p < 0.01$), because overweight patients still showed the worst performance in PROMs after orthopedic inpatient rehabilitation.

Discussion

In this retrospective cohort study, we analyzed the changes observed in outcome measures during LBP rehabilitation and introduced a new performance outcome measure. In terms of overall medical quality outcome, the rehabilitation

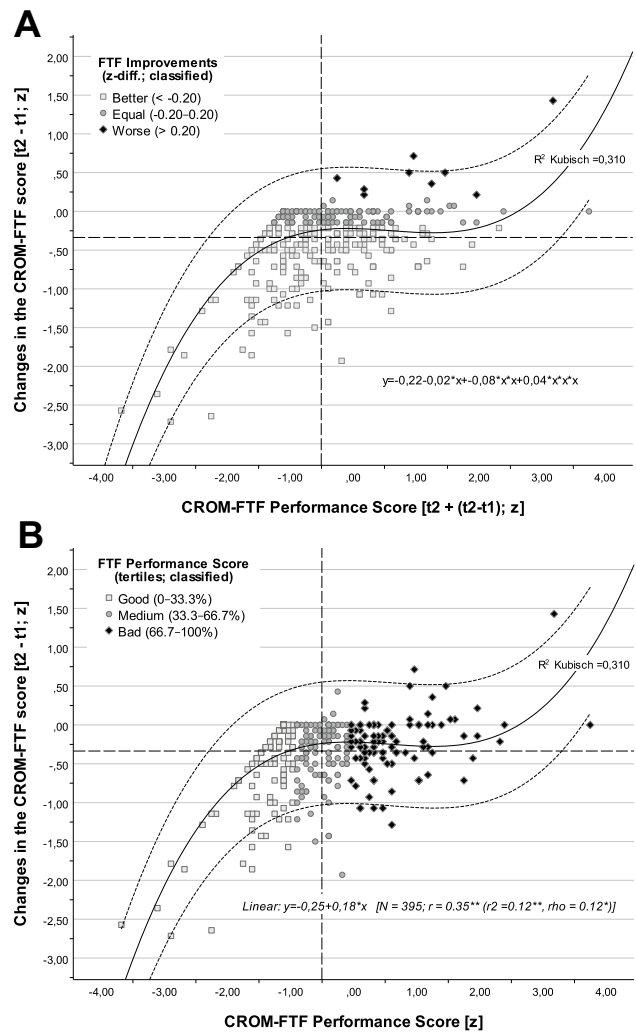


Fig. 2 ab FTF improvements vs. performance score and changes of FTF. Classification for improvements is based on the value distributions (z-differences; SMD; a top). Difference values with no significant changes normally range randomly from 0.00 ± 0.20 (1/5 SD) [36]. Tertiles for FTF performance score were chosen, because the expected and observed improvements are around 2/3 (b bottom) [16]

resulted in a large effect size (Cohens $d = 0.94$). Medium effect sizes were observed for CROM-FTF ($d = 0.70$) and PROMs ($d > 0.50$). PROMs deteriorated in 13.9% of all LBP patients, while only 2.3% showed a deterioration in CROMs (Table 2). The correlations between PROMs and CROM-FTF were low to moderate, with the highest identified for HAQ-DI ($\rho = 0.30-0.36$); regarding changes, no significant correlations could be shown (Table 3).

In previous studies, a number of factors were identified that influence the PROMs. The BMI has been reported to correlate with PROMs, including pain, in LBP patients, whereas no such correlation was found for the Timed Up-and-Go Test [37]. However, a systematic review of twin

Table 4 Consistency of performance scores between outcome measurements

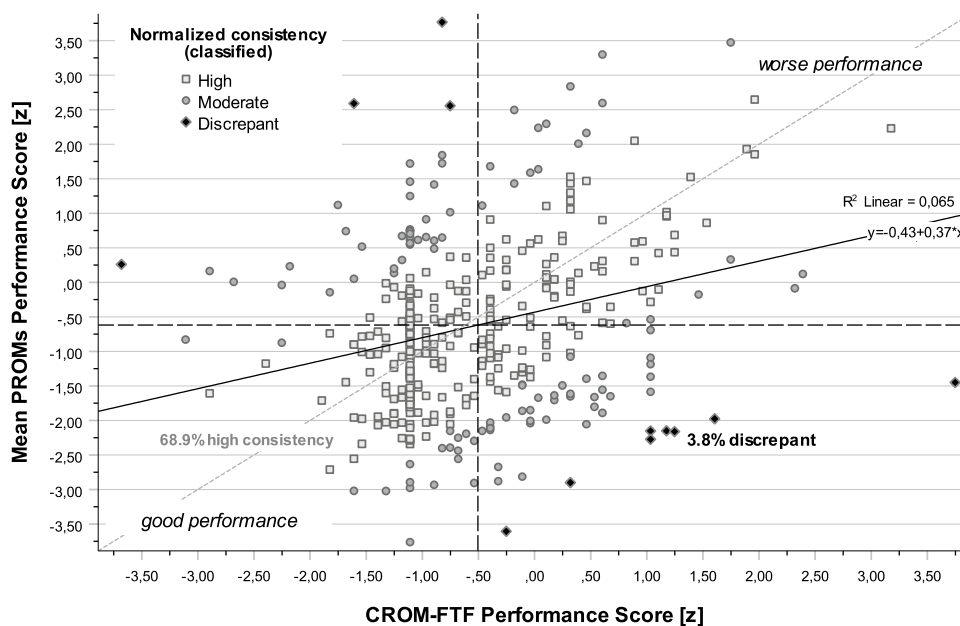
Consistency (z-differences; classified within/ between tertiles)	CROM-FTF performance score ($t2 + \Delta$) normalized with SD [z]			Performance score \pm SD [$t2z + \Delta z$, $t2D$]		Diff. $t2D$ (CROM – PROM)	
	High consistency (%)	Moderate consistency (%)	Discrepant (%)	Mean	SD	Mean	SD
CROM-FTF $t2 + \Delta$	100.0	0.0	0.0	-0.50	\pm 0.93	0.00	\pm 0.00
Changes FTF ($t2 - t1$)	61.3	36.5	2.3	NaN	\pm NaN	NaN	\pm NaN
Random variable $t2 + \Delta$ ($0 \pm \sqrt{5}$) no consistency	33.3	33.3	33.3	0.00	\pm 2.24	0.00	\pm 3.16
EQ5D health (EQ-VAS) $t2 + \Delta$	59.2	27.3	13.4	-0.40	\pm 1.93	-0.11	\pm 2.10
EQ5D TTO $t2 + \Delta$	72.9	22.3	4.8	-0.50	\pm 1.45	0.00	\pm 1.55
NPRS $t2 + \Delta$	57.5	34.7	7.8	-0.78	\pm 1.56	0.27	\pm 1.70
HAQ $t2 + \Delta$	76.7	18.2	5.1	-0.22	\pm 1.34	-0.29	\pm 1.39
ODI $t2 + \Delta$	71.4	24.8	3.8	-0.51	\pm 1.32	0.00	\pm 1.40
Mean PROMs $t2 + \Delta$	68.9	27.3	3.8	-0.62	\pm 1.34	0.11	\pm 1.42

CROM-FTF, mean of PROMs and the overall medical outcome (MQO; mean of PROMs and CROM-FTF) are highlighted in bold

Consistency. In most cases (68.9%), the results in CROM-FTF and PROMs point in the same direction (high consistency). Specifically, 27.3% of results in PROM scores showed a moderate agreement in performance with CROM-FTF (moderate consistency). In 3.8% of the cases, the results between CROM-FTF vs PROMs were contrary to each other (discrepant)

Consistency of performance ($t2 + \Delta$; Differences of $t2D/\sqrt{10}$) using normal scores in three categories: (high consistency = $t2D$ within one/same tertile. moderate consistency = between one and two tertiles. discrepant = more than two tertiles difference in normalized performance scores); n = 395

Fig. 3 Consistency within CROM vs. PROMs performance scores. In most cases (68.9%), the performance scores between CROM-FTF and PROMs point in the same direction. Specifically, 27.3% of results in mean PROM performance scores showed moderate agreement in terms of performance with CROM-FTF. In 3.8% of the cases, the results between CROM-FTF vs PROMs were contrary to each other (discrepant)



studies revealed that the association of LBP and obesity seems to be weak [38]. In LBP, females seem to report more pain, higher levels of disability and lower quality of life than males, even though the values of CROMs do not seem to differ [14]. Similar findings were reported in a recent study, suggesting that female gender, a lower education level and higher ODI scores predict worse outcomes after surgery for LBP [39]. Socioeconomic status

and depression have also been identified as possible factors that predict a poor rehabilitation outcome [40].

Concerning the correlation between PROMs and CROMs, Melzer et al. [41] reported poor to moderate associations between performance-based measures and self-reported functional status in older patients. Obvious interdependencies exist between patient-reported outcomes and performance measures; however, these are used

Table 5 Effect sizes for outcome measurements (changes, $f2D$) and moderating factors

η^2_p	Unifactorial part. Eta^2 for changes (interaction)*						Main effect			Unifactorial part. Eta^2 for performance scores**						
	Between-factors			Sex	Age	ICD	BMI	IT	Pre MQO	Time	Sex	Age	ICD	BMI	IT	Pre MQO
Outcome measures	f/m	3-stage	6-stage	f/m	3-stage	6-stage	3-stage	3-stage	3-stage	2-stage	f/m	3-stage	6-stage	3-stage	3-stage	3-stage
CROM (1)	0.000	0.001	0.010	0.000	0.041***	0.010	0.008	0.182***	0.328***	0.016*	0.009	0.008	0.033***	0.004	0.222***	
PROMs (6)	0.000	0.006	0.009	0.001	0.002	0.009	0.001	0.007	0.059***	0.001	0.017*	0.015	0.001	0.005	0.044***	
EQ5D Health (EQ-VAS)	0.003	0.008	0.001	0.009	0.007	0.010	0.007	0.010	0.133***	0.001	0.004	0.003	0.024***	0.035**	0.066***	
EQ5D TTO	0.001	0.003	0.012	0.006	0.003	0.019*	0.003	0.019*	0.251***	0.002	0.004	0.012	0.013	0.020*	0.045***	
NPRS	0.000	0.003	0.007	0.015	0.008	0.005	0.008	0.005	0.058***	0.011*	0.040***	0.008	0.018*	0.030**	0.187***	
HAQ	0.011*	0.003	0.001	0.004	0.001	0.006	0.001	0.006	0.230***	0.002	0.006	0.004	0.012	0.030**	0.140***	
ODI	0.003	0.002	0.004	0.009	0.005	0.008	0.005	0.008	0.302***	0.005	0.016*	0.009	0.021*	0.043**	0.150***	
Mean PROMs	0.001	0.000	0.007	0.017*	0.001	0.124***	0.001	0.124***	0.472***	0.001	0.017*	0.010	0.027***	0.024*	0.290***	
PROMs and CROM-FTF	0.002	0.003	0.006	0.029**	0.013	0.042***	0.013	0.042***	0.009	0.022***	0.008	0.006	0.024**	0.034**	0.003***	
Difference CROM – PROMs																

Sex (female, male); age (<= 50, 51–60, 61+); ICD...six main-diagnoses ($n > 30$; M51.1, M51.2, M53.9, M54.4, M54.9, other LBG); BMI...Body Mass Index (<= 25, 25–30, 30+); IT... Individual therapy minutes (<= 360, 361–450, 451+); pre...pre-rehabilitation value of MQO (tertile)

A part. Eta^2 (η^2_p) between 0.01 and 0.06 corresponds to a small effect, occurrences of 0.06–0.14 a middle effect and values > 0.14 a large effect; level of significance (bold): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; $N = 395$ LBP patients

*Differences (improvements; post – pre) from admission to discharge (corresponds to the interaction: time x between-factor)

**Performance Scores: $f2D = f2_z + (f2_z - f1_z)$

Performance scores were much more sensitive to the individual factors compared to the difference scores. The factors sex and age showed significant effects in performance scores. In addition to initial values (all $p < 0.001$; overall MQO: $\eta^2_p = 0.290$), the BMI had the greatest influence on the performance evaluation ($\eta^2_p = 0.027$, $p < 0.01$). The amount of individual therapy (IT) depends on the medical history, if the rehabilitation program is classified as follow-up treatment procedure after surgery (IT > 450 min)

to evaluate different outcomes and to assess the effects of complementary therapeutic modalities in orthopedic rehabilitation. This finding is consistent with that of Stratford et al. [42], who proposed that self-reported outcomes and performance measures can be used to evaluate different aspects of physical functioning. The authors concluded that self-report measures provide information about the experience associated with the execution of the task, while performance measures contain information about the ability to complete the task [42]. It is, therefore, necessary to carry out performance-based tests to fully characterize the changes in the patients' physical functions [43]. Performance-based tests like the FTF provide objective information about how the patients actually function, information that cannot be captured by PROMs alone. These CROMs of physical function allow healthcare staff to evaluate what individuals can actually do rather than what they perceive they can do; the latter is then assessed using PROMs.

Outcome measures and endpoints are often systematically associated with influences other than treatment and can be interpreted as both causes and responses, which may elicit different responses in different individuals. This leads to large variability and confounds the observed outcomes. Therefore, the level of agreement between PROMs and CROMs cannot be taken for granted. In addition to human factors and differences in baseline conditions between patients, methodological and conceptional issues such as ceiling effects, nonlinearity and reliability (cf. state vs. trait) play moderating roles in observable associations between PROMs and CROMs. Outcome measures are influenced by daily activities, underlying conditions and personal factors such as educational level, mood [13], patient hospital experience, overall satisfaction, personal expectations [44] and gender [14, 45]. Temporal aspects should also not be neglected, such as retest reliability and the characteristics of the methods and outcome variables used. It is reasonable to assume that an aggregate outcome, such as a generic measure used to quantify patient disability, may have more stable characteristics or be subject to less variation than some other disease-specific measures, such as self-reports of pain conditions. This is true for both CROMs and PROMs. In this study, the FTF ($r_{t1\text{ to }t2} = 0.90$) showed high reliability, as did the ODI ($r = 0.80$) and HAQ-DI ($r = 0.83$), whereas the EQ-VAS ($r = 0.43$) and NPRS ($r = 0.57$) showed lower stability over time. In addition to the associations between outcomes, this obviously has implications for the responsiveness of the measures. Each approach has its strengths in this regard and should be considered when designing clinical trials. Lower test–retest correlations can also be observed for "objective" physiological measures such as heart rate or diastolic blood pressure [46]. Hamilton et al. [47, 48] indicated that lower confidence in 'subjective' PROMs as

compared to 'objective' clinical measures is not justified, stating "... we would expect to see a similar direction of change in the respective scores when measuring the effect of an intervention, but to expect the same result misunderstands that PROMs capture a different aspect of outcome than a performance test does. The relationship between assessment of performance and report of performance improved as the patient's report of pain diminished, suggesting that patients' reporting of functional outcome after TKR is influenced more by their pain level than their ability to accomplish tasks."

Unlike the commonly used methods, the performance scores in the PROMs showed significant correlations with the FTF performance scores (Table 3). This method of performance evaluation ($t2D$) was much more sensitive to the individual factors as compared to the difference scores (Table 5), since the medical evaluation also depends on the patient's actual functional status and the rehabilitative clinical practice (Fig. 1). In a theoretical sample with independent, normally distributed random variables, a regression analysis of $t2_z + \Delta_z$ and $t2 - t1$ would result in a maximum correlation coefficient of 0.95. In the present sample, this coefficient could not be achieved (e.g., for mean PROMs: $\rho = 0.70$), possibly due to the interdependency of outcome measures within a subject, smaller variances and the desired effects of the intervention.

In a minority (13.9%) of patients, the PROMs worsened significantly over time. Although this finding cannot be readily explained, it may have been caused in part by ceiling effects, as these patients usually displayed relatively good outcome values at the beginning of rehabilitation (Table 5). In the future, research should be carried out to identify the cause of these differences by more thoroughly comparing the consistent and discrepant results within a patient.

In more than half of the cases, high consistency levels were observed among the newly introduced performance scores between CROM-FTF and PROMs (Table 4). Further studies would be needed to identify critical success factors and non-responders in the rehabilitation process. Simply looking at the changes in outcome measures does not seem to allow healthcare professionals to detect such factors. The calculation of the performance score presented in this paper provides a promising alternative approach, as it takes into account a classification of the patient's health status after rehabilitation ($t2$), on the one hand, and the patient's progress made during the rehabilitation process (changes; Δ), on the other.

Limitations

For ethical, practical and economic reasons, it was not possible to include a randomized control group in this study. Due to the study design, causal conclusions must be drawn with caution. The beneficial effects resulting from participation in an inpatient rehabilitation may not be sustainable once the patients return to their usual everyday lives [49]. However, even minor changes in lifestyle can lead to functional adaptations and the normalization of physiological functions, which help the patient to recover from chronic inflammatory or degenerative diseases [50].

Conclusions

Rehabilitation plays a vital role in preventing and minimizing the functional limitations associated with ageing and chronic conditions. Strong evidence supports the argument that inpatient rehabilitation is a necessary part of the treatment of inflammatory and degenerative diseases, as well as functional limitations after surgery. Despite the large international differences observed in terms of the variety of the composition of teams involved in rehabilitation and the implemented treatment measures, the observed strong effect sizes can support individual evaluation. These effect sizes allow medical professionals and researchers to compare health programs and developments in prevention, healthcare and rehabilitation. More importantly, they allow them to more effectively improve the wellbeing of patients with chronic conditions. Due to the use of standardized service portfolios, external reviews and the fact that insurers centrally control the assignment of modalities, we assume that the initial values and outcomes are representative for an orthopedic inpatient rehabilitation program in Austria.

An improvement or decline in the CROM did not serve as an indicator for what patients reported about their perceived functioning or pain. A deterioration in self-reported outcomes in 13.9% of all LBP patients during rehabilitation was observed. These changes alone do not reflect clinical evaluation practice, because they may not detect non-responders and the respective critical factors. They are usually not significantly influenced by known moderating (critical individual) factors. The new method presented in this work to assess individual “performance scores” within a patient can be used effectively to identify critical success factors and non-responders in the rehabilitation process, in a simple and user-friendly way. Further research is warranted, in order to ascertain the usefulness of this new method in other patient populations and treatments,

as well as its utility in predicting long-term success and optimizing current rehabilitative practice.

Acknowledgements The authors sincerely thank the team of the Rehabilitation Center in Kitzbühel and their colleagues and the Medical University of Vienna as well as the Ludwig Boltzmann Society for their support in the development and execution of this study.

Author contributions Each author of our work was significantly involved in the conception, design, data acquisition, data analysis and interpretation. All authors contributed to the writing of the manuscript and have released the final version for publication. All authors take responsibility for the accuracy and integrity of all aspects of research. The individual authors had the following respective main tasks: Study concept and design: MJF, MP, MS, VG. Acquisition of data: MJF, MP, MS. Analysis and interpretation of data: VG, AZ, MJF, MP, RC. Drafting of the manuscript: AZ, VG. Final revision and addenda: VG, AZ, MJF, RC.

Funding This study was supported by the Ludwig Boltzmann Society, Vienna. The researchers did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval The study was approved by the Ethics Committee of the Medical University Innsbruck (EK Nr: 1158/2019) in accordance with the current version (2013) of the Declaration of Helsinki.

Consent to participate The patients were fully informed about the extent and purpose of the study and gave their written consent.

Consent for publication All authors provided their consent to submit and publish the final version of this manuscript. The views expressed in this publication are those of the author(s) and not necessarily those of the Ludwig Boltzmann Society and VAMED AG.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Hoy, D., Brooks, P., Blyth, F., & Buchbinder, R. (2010). The epidemiology of low back pain. *Best Practice & Research Clinical Rheumatology*, 24(6), 769–781. <https://doi.org/10.1016/j.berh.2010.10.002>
- Hoy, D., March, L., Brooks, P., Blyth, F., Woolf, A., Bain, C., et al. (2014). The global burden of low back pain: Estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases*, 73(6), 968–974. <https://doi.org/10.1136/annrheumdis-2013-204428>
- Chou, R., Deyo, R., Friedly, J., Skelly, A., Hashimoto, R., Weimer, M., et al. (2017). Nonpharmacologic therapies for low back pain: A systematic review for an American College of Physicians Clinical Practice Guideline. *Annals of Internal Medicine*, 166(7), 493–505. <https://doi.org/10.7326/M16-2459>
- Ramasamy, A., Martin, M. L., Blum, S. I., Liedgens, H., Argoff, C., Freynhagen, R., et al. (2017). Assessment of patient-reported outcome instruments to assess chronic low back pain. *Pain Medicine*, 18(6), 1098–1110. <https://doi.org/10.1093/pm/pnw357>
- McGirt, M. J., Bydon, M., Archer, K. R., Devin, C. J., Chotai, S., Parker, S. L., et al. (2017). An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: predicting likely individual patient outcomes for shared decision-making. *J Neurosurg Spine*, 27(4), 357–369. <https://doi.org/10.3171/2016.11.SPINE16526>
- Stamm, T. A., Boesendorfer, A., Omara, M., Ritschl, V., Stefanac, S., & Mosor, E. (2019). Outcomes research in non-specific low back pain: Knowledge transfer in clinical practice. *Wiener Klinische Wochenschrift*, 131(21–22), 550–557. <https://doi.org/10.1007/s00508-019-1523-4>
- Chiarotto, A., Ostelo, R. W., Boers, M., & Terwee, C. B. (2018). A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. *Journal of Clinical Epidemiology*, 95, 73–93. <https://doi.org/10.1016/j.jclinepi.2017.11.005>
- Hansson-Hedblom, A., Jonsson, E., Fritzell, P., Hagg, O., & Borgstrom, F. (2019). The association between patient reported outcomes of spinal surgery and societal costs: A register based study. *Spine (Phila Pa 1976)*, 44(18), 1309–1317. <https://doi.org/10.1097/BRS.0000000000003050>
- Waddell, G. (1987). Clinical assessment of lumbar impairment. *Clin Orthop Relat Res* (221), 110–120.
- Atya, A. M. (2013). The validity of spinal mobility for prediction of functional disability in male patients with low back pain. *Journal of Advanced Research*, 4(1), 43–49. <https://doi.org/10.1016/j.jare.2012.01.002>
- Perret, C., Poiraudou, S., Fermanian, J., Colau, M. M., Benhamou, M. A., & Revel, M. (2001). Validity, reliability, and responsiveness of the fingertip-to-floor test. *Archives of Physical Medicine and Rehabilitation*, 82(11), 1566–1570. <https://doi.org/10.1053/apmr.2001.26064>
- Macrae, I. F., & Wright, V. (1969). Measurement of back movement. *Annals of the Rheumatic Diseases*, 28(6), 584–589. <https://doi.org/10.1136/ard.28.6.584>
- van Delden, A. L., Peper, C. L., Beek, P. J., & Kwakkel, G. (2013). Match and mismatch between objective and subjective improvements in upper limb function after stroke. *Disability and Rehabilitation*, 35(23), 1961–1967. <https://doi.org/10.3109/09638288.2013.768303>
- Gautschi, O. P., Corniola, M. V., Smoll, N. R., Joswig, H., Schaller, K., Hildebrandt, G., et al. (2016). Sex differences in subjective and objective measures of pain, functional impairment, and health-related quality of life in patients with lumbar degenerative disc disease. *Pain*, 157(5), 1065–1071. <https://doi.org/10.1097/j.pain.0000000000000480>
- Kennedy, D. M., Stratford, P. W., Wessel, J., Gollish, J. D., & Penney, D. (2005). Assessing stability and change of four performance measures: A longitudinal study evaluating outcome following total hip and knee arthroplasty. *BMC Musculoskeletal Disorders*, 6, 3. <https://doi.org/10.1186/1471-2474-6-3>
- Grote, V., Unger, A., Bottcher, E., Muntean, M., Puff, H., Markt, W., et al. (2020). General and disease-specific health indicator changes associated with inpatient rehabilitation. *Journal of the American Medical Directors Association*, 21(12), 2017.e2010–2017.e2027. <https://doi.org/10.1016/j.jamda.2020.05.034>
- Bachner, F., Bobek, J., Habimana, K., Ladurner, J., Lepuschutz, L., Ostermann, H., et al. (2018). Austria: Health system review. *Health Syst Transit*, 20(3), 1–254.
- Wilson, J., Arshad, F., Nnamoko, N., Whiteman, A., Ring, J., & Roy, B. (2014). Patient-reported outcome measures: An on-line system empowering patient choice. *Journal of the American Medical Informatics Association*, 21(4), 725–729. <https://doi.org/10.1136/amiainl-2012-001183>
- Porter, M. E., Larsson, S., & Lee, T. H. (2016). Standardizing patient outcomes measurement. *New England Journal of Medicine*, 374(6), 504–506. <https://doi.org/10.1056/NEJMp1511701>
- Wolpert, M. (2014). Uses and abuses of patient reported outcome measures (PROMs): Potential iatrogenic impact of PROMs implementation and how it can be mitigated. *Administration and Policy In Mental Health*, 41(2), 141–145. <https://doi.org/10.1007/s10488-013-0509-1>
- Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*, 113(1–2), 9–19. <https://doi.org/10.1016/j.pain.2004.09.012>
- Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, 17(1), 45–56. [https://doi.org/10.1016/0304-3959\(83\)90126-4](https://doi.org/10.1016/0304-3959(83)90126-4)
- Fairbank, J. C., Couper, J., Davies, J. B., & O'Brien, J. P. (1980). The Oswestry low back pain disability questionnaire. *Physiotherapy*, 66(8), 271–273.
- Fairbank, J. C. (2007). Use and abuse of Oswestry disability index. *Spine (Phila Pa 1976)*, 32(25), 2787–2789. <https://doi.org/10.1097/BRS.0b013e31815b9e9d>
- Fairbank, J. C., & Pynsent, P. B. (2000). The Oswestry disability index. *Spine (Phila Pa 1976)*, 25(22), 2940–2952. <https://doi.org/10.1097/00007632-200011150-00017> discussion 2952.
- Mannion, A. F., Junge, A., Fairbank, J. C., Dvorak, J., & Grob, D. (2006). Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reliability, and validity. *Eur Spine J*, 15(1), 55–65. <https://doi.org/10.1007/s00586-004-0815-0>
- Fries, J. F., Spitz, P., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism*, 23(2), 137–145. <https://doi.org/10.1002/art.1780230202>
- Aletaha, D., Stamm, T., & Smolen, J. (2006). Measuring disease activity for rheumatoid arthritis. *Z Rheumatol*, 65(2), 93–96, 98–102. <https://doi.org/10.1007/s00393-006-0041-8>
- Bruce, B., & Fries, J. (2004). Longitudinal comparison of the Health Assessment Questionnaire (HAQ) and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). *Arthritis and Rheumatism*, 51(5), 730–737. <https://doi.org/10.1002/art.20695>
- Rabin, R., & de Charro, F. (2001). EQ-5D: A measure of health status from the EuroQol Group. *Annals of Medicine*, 33(5), 337–343. <https://doi.org/10.3109/07853890109002087>

31. Janssen, M. F., Bonsel, G. J., & Luo, N. (2018). Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. *Pharmacoeconomics*, 36(6), 675–697. <https://doi.org/10.1007/s40273-018-0623-8>
32. Whynes, D. K., & Group, T. (2008). Correspondence between EQ-5D health state classifications and EQ VAS scores. *Health and Quality of Life Outcomes*, 6, 94. <https://doi.org/10.1186/1477-7525-6-94>
33. Janssen, M. F., Birnie, E., & Bonsel, G. J. (2008). Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. *Quality of Life Research*, 17(3), 463–473. <https://doi.org/10.1007/s11136-008-9318-5>
34. Johnsen, L. G., Hellum, C., Nygaard, O. P., Storheim, K., Brox, J. I., Rossvoll, I., et al. (2013). Comparison of the SF6D, the EQ5D, and the Oswestry disability index in patients with chronic low back pain and degenerative disc disease. *BMC Musculoskeletal Disorders*, 14(1), 148. <https://doi.org/10.1186/1471-2474-14-148>
35. Ludwig, K., Graf von der Schulenburg, J. M., & Greiner, W. (2018). German Value Set for the EQ-5D-5L. *Pharmacoeconomics*, 36(6), 663–674. <https://doi.org/10.1007/s40273-018-0615-8>
36. Cohen, J. (2016). Statistical power analysis. *Current Directions in Psychological Science*, 1(3), 98–101. <https://doi.org/10.1111/1467-8721.ep10768783>
37. Stienen, M. N., Joswig, H., Smoll, N. R., Corniola, M. V., Schaller, K., Hildebrandt, G., et al. (2016). Influence of body mass index on subjective and objective measures of pain, functional impairment, and health-related quality of life in lumbar degenerative disc disease. *World Neurosurg*, 96(570–577), e571. <https://doi.org/10.1016/j.wneu.2016.09.070>
38. Dario, A. B., Ferreira, M. L., Refshauge, K. M., Lima, T. S., Ordona, J. R., & Ferreira, P. H. (2015). The relationship between obesity, low back pain, and lumbar disc degeneration when genetics and the environment are considered: A systematic review of twin studies. *Spine J*, 15(5), 1106–1117. <https://doi.org/10.1016/j.spinee.2015.02.001>
39. Kim, G. U., Park, J., Kim, H. J., Shen, F., Cho, J., Chang, B. S., et al. (2020). Definitions of unfavorable surgical outcomes and their risk factors based on disability score after spine surgery for lumbar spinal stenosis. *BMC Musculoskeletal Disorders*, 21(1), 288. <https://doi.org/10.1186/s12891-020-03323-0>
40. Sanchez-Santos, M. T., Garriga, C., Judge, A., Batra, R. N., Price, A. J., Liddle, A. D., et al. (2018). Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery. *Science and Reports*, 8(1), 3381. <https://doi.org/10.1038/s41598-018-21714-1>
41. Melzer, I., Kurz, I., Sarid, O., & Jette, A. M. (2007). Relationship between self-reported function and disability and balance performance measures in the elderly. *Journal of Rehabilitation Research and Development*, 44(5), 685–691. <https://doi.org/10.1682/jrrd.2006.10.0133>
42. Stratford, P. W., & Kennedy, D. M. (2006). Performance measures were necessary to obtain a complete picture of osteoarthritic patients. *Journal of Clinical Epidemiology*, 59(2), 160–167. <https://doi.org/10.1016/j.jclinepi.2005.07.012>
43. Mizner, R. L., Petterson, S. C., Clements, K. E., Zeni, J. A., Jr., Irrgang, J. J., & Snyder-Mackler, L. (2011). Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: A longitudinal analysis of outcomes. *Journal of Arthroplasty*, 26(5), 728–737. <https://doi.org/10.1016/j.arth.2010.06.004>
44. Hamilton, D. F., Lane, J. V., Gaston, P., Patton, J. T., Macdonald, D., Simpson, A. H. R. W., et al. (2013). What determines patient satisfaction with surgery? A prospective cohort study of 4709 patients following total joint replacement. *British Medical Journal Open*, 3(4), e002525. <https://doi.org/10.1136/bmjopen-2012-002525>
45. Choong, A. L. C., Shadbolt, C., Dowsey, M. M., & Choong, P. F. M. Sex-based differences in the outcomes of total hip and knee arthroplasty: A narrative review. *ANZ Journal of Surgery*, n/a(n/a), doi:<https://doi.org/10.1111/ans.16299>
46. Hahn, E. A., Cella, D., Chassany, O., Fairclough, D. L., Wong, G. Y., Hays, R. D., et al. (2007). Precision of health-related quality-of-life data compared with other clinical measures. *Mayo Clinic Proceedings*, 82(10), 1244–1254. <https://doi.org/10.4065/82.10.1244>
47. Hamilton, D. F., Giesinger, J. M., & Giesinger, K. (2017). It is merely subjective opinion that patient-reported outcome measures are not objective tools. *Bone & Joint Research*, 6(12), 665–666. <https://doi.org/10.1302/2046-3758.612.BJR-2017-0347>
48. Hamilton, D. F., Gaston, P., & Simpson, A. H. (2012). Is patient reporting of physical function accurate following total knee replacement? *Journal of Bone and Joint Surgery British*, 94(11), 1506–1510. <https://doi.org/10.1302/0301-620x.94b11.30081>
49. Oliveira, C. B., Franco, M. R., Maher, C. G., Christine Lin, C. W., Morelhaio, P. K., Araujo, A. C., et al. (2016). Physical activity interventions for increasing objectively measured physical activity levels in patients with chronic musculoskeletal pain: A systematic review. *Arthritis Care Res (Hoboken)*, 68(12), 1832–1842. <https://doi.org/10.1002/acr.22919>
50. Grote, V., Levnjajic, Z., Puff, H., Ohland, T., Goswami, N., Fruhwirth, M., et al. (2019). Dynamics of Vagal Activity Due to Surgery and Subsequent Rehabilitation. [Original Research]. *Front Neurosci*, 13(1116), 1116. <https://doi.org/10.3389/fnins.2019.01116>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.