



A composite measure for patient-reported outcomes in orthopedic care: design principles and validity checks

Lukas Schöner¹ · David Kuklinski² · Alexander Geissler² · Reinhard Busse¹ · Christoph Pross¹

Accepted: 8 March 2023
© The Author(s) 2023

Abstract

Background The complex, multidimensional nature of healthcare quality makes provider and treatment decisions based on quality difficult. Patient-reported outcome (PRO) measures can enhance patient centricity and involvement. The proliferation of PRO measures, however, requires a simplification to improve comprehensibility. Composite measures can simplify complex data without sacrificing the underlying information.

Objective and methods We propose a five-step development approach to combine different PRO into one composite measure (PRO-CM): (i) theoretical framework and metric selection, (ii) initial data analysis, (iii) rescaling, (iv) weighting and aggregation, and (v) sensitivity and uncertainty analysis. We evaluate different rescaling, weighting, and aggregation methods by utilizing data of 3145 hip and 2605 knee replacement patients, to identify the most advantageous development approach for a PRO-CM that reflects quality variations from a patient perspective.

Results The comparison of different methods within steps (iii) and (iv) reveals the following methods as most advantageous: (iii) rescaling via z-score standardization and (iv) applying differential weights and additive aggregation. The resulting PRO-CM is most sensitive to variations in physical health. Changing weighting schemes impacts the PRO-CM most directly, while it proves more robust towards different rescaling and aggregation approaches.

Conclusion Combining multiple PRO provides a holistic picture of patients' health improvement. The PRO-CM can enhance patient understanding and simplify reporting and monitoring of PRO. However, the development methodology of a PRO-CM needs to be justified and transparent to ensure that it is comprehensible and replicable. This is essential to address the well-known problems associated with composites, such as misinterpretation and lack of trust.

Keywords Patient reported outcome measures · Composite measure · Outcome quality measurement · Value-based health care · Composite measure development validation

Introduction

The complex, multidimensional nature of healthcare quality makes quality measurement and transparency as well as provider and treatment decisions difficult for patients [1–5]. Patient participation in healthcare decision making presupposes that patients can understand quality information, which requires suitable quality measurement and reporting

instruments [3, 6–9]. Patient-reported outcome measures (PROMs) are promising instruments that, in contrast to clinical indicators, measure patients' own assessment of their current health status and enhance patient engagement [1, 4, 10–13]. PROMs are used to determine patient-reported outcomes (PRO), which are results of longitudinal comparison of individual PROM-scores, i.e., the change in individual PROM-scores attributable to a particular treatment. Despite their potential, the growing number of PROM makes it difficult to easily and comprehensively evaluate outcome quality [2, 12, 14–16]. Composite measures (CMs) can simplify complex, multidimensional data without sacrificing the underlying power of information [17–19].

A CM is a combination of two or more individual measures into one index, which captures multidimensional aspects that cannot be reflected by solely either of

✉ Lukas Schöner
Lukas.schoener@tu-berlin.de

¹ Department of Health Care Management, Technical University Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

² Department of Health Care Management, University of St. Gallen, St. Gallen, Switzerland

the individual measures [18]. In healthcare, CM provides a holistic picture of healthcare quality and can enhance ease of interpretation and comparability [20–22]. Next to benchmarking hospital or countries' health system performance, CM can facilitate monitoring recovery paths and outcome quality as well as enhancing public accountability and quality transparency [23–26]. CM also plays an important role for the emerging value-based healthcare (VBHC) movement and allow researchers to better evaluate the results of clinical studies with several different PRO by a single outcome measure [27, 28]. Due to their advantages, healthcare CMs have already been widely applied in many different areas with different purposes [21, 22, 29–34]. However, there are also important downsides and challenges with CMs, which are controversially discussed in the literature [6, 17, 25]. Poorly constructed or opaque CMs can be particularly alarming as they have the potential to mask poor quality or deceive those who use them to make important policy and treatment decisions.

It is thus essential that the development methodology is clear and transparent to ensure that the CM is comprehensible and replicable. The chosen methodology is well justified and plausible and represents the relevant quality dimensions without losing or disguising important information [6, 35–37]. The development of CM, however, is often controversial, neither is there a gold-standard approach. Some guidelines for CM development are provided, e.g., by the OECD [35, 37] or, in a healthcare context, by Shwartz et al. [19]. However, so far CMs are mostly used to aggregate clinical outcomes. Furthermore, there is still a lack of studies that put these guidelines into practice.

In the present study, we combine the different considerations of OECD and Shwartz et al. to develop a patient-reported outcome CM (PRO-CM) applicable in routine orthopedic care and clinical studies. We propose a five-step development approach and highlight the need of transparency and justification of decisions in each step. We evaluate advantages and disadvantages of different rescaling, weighting, and aggregation methods, by utilizing PRO-data of primary hip and knee arthroplasty (PHA and PKA) patients. Due to the increasing case volume of hip and knee arthroplasty worldwide [38, 39] and since PROMs are already widely used in this field [40], the orthopedic setting provides a good example for illustrating development and benefits of a PRO-CM. Finally, we identify the most advantageous development approach for a multidimensional orthopedic PRO-CM that is transparent and replicable, combines all relevant sub-dimensions of PHA and PKA, and captures the relative differences and quality variations among these sub-dimensions. It is more sensitive to variations in the sub-dimensions that are

most relevant for patients and partly compensates poorer outcomes in one dimension.

Methods

Data

We use data from the PROMoting Quality study [41], which provides PRO-data of 3,145 PHA- and 2,605 PKA patients of nine participating German hospitals between 2019 and 2021. Participants were adults undergoing an elective and primary hip or knee arthroplasty with pre-specified surgery codes (including total and partial arthroplasties) between 2019 and 2020. Exclusion criteria were emergency and life-threatening cases, ASA classification 4–6, and patients without direct or indirect access to an e-mail account or without a relative supporting the survey PROM response. The randomized-controlled trial was registered at the German Clinical Trials Register under trial number DRKS00019916 and examined the benefit of PROM-based patient follow-up based on the ICHOM standard set for Hip and Knee Osteoarthritis with minor modifications [11, 42]: EQ-5D-5L captures Health-related Quality of Life (HRQoL) [43], Hip or Knee Osteoarthritis Outcome Score Physical Function Shortform (HOOS-PS or KOOS-PS) joint-associated problems and functionality [44], analogue pain scales assess pain in hip (left and right), knee (left and right), and lower back [42]. PROMIS Depression Shortform (PROMIS-D-SF) and Fatigue Shortform (PROMIS-F-SF) are included to capture mental health [45]. For a detailed description of the PROM, see Appendix I.

Stepwise method for developing a composite measure

The study was preceded by a literature review on CM in general and in the healthcare context. The development approaches presented here are mainly based on current standards as provided by the OECD [35] and Shwartz et al. [19]. While we consider the OECD guidelines as a general toolkit for relevant technical and methodological issues (e.g., rescaling- and weighting- and aggregation-methods), the framework of Shwartz et al. provides relevant considerations in a healthcare context for creating hospital-level composites aggregating clinical outcomes. For the PRO-CM, we merge these considerations, adjust them to fit a patient-level orthopedic purpose and propose five PRO-CM development steps: (i) theoretical framework and metric selection, (ii) initial data analysis, (iii) rescaling, (iv) weighting and aggregation, and (v) sensitivity

and uncertainty analysis [18, 19, 35]. Assessing risks and benefits of the different options we consider in step (iii) and (iv), we select a priori the most advantageous option with respect to the data structure and theoretical framework (i.e., “Model 1”) and compare the results to the other options (Model 2–5).

Theoretical framework and metric selection

The theoretical framework lays the foundation for a CM. It defines the quality construct (i.e., the phenomenon to be measured) and identifies its sub-dimensions [18, 35, 37]. Relevant quality indicators are identified so as to conform to the quality construct [46]. We select validated and well-established generic and disease-specific PROMs that align to the sub-dimensions of the quality construct.

Initial data analysis

We examine the PRO individually to analyze the underlying data structure (e.g., outliers and scale), which guides subsequent rescaling and weighting decisions. We plot descriptive statistics and compute Spearman’s rank correlations to check for collinearity [19, 24, 35]. Following similar studies [24, 37, 46], we consider indicators correlated higher than $r=0.7$ to be merged into one variable to avoid redundancy or preponderance of one particular dimension [18, 36].

Rescaling

When indicators have different units of scale, rescaling on a common scale is required to allow comparison and aggregation. Different methods may produce different CM [19, 35] and it is not clear which method is favorable. Following Schwartz et al., we compare the two most widely used approaches for healthcare CM, i.e., z-score standardization and min–max normalization [19]. A priori we use z-score standardization (Model 1), as it preserves the relative differences, and extreme values and outliers don’t distort the mean but are recognized as exceptional performance. The z-score standardization transforms all individual measures on a dimensionless scale with mean = 0 and standard deviation (SD) = 1. Z-scores express how many SD an individual’s outcome is above or below the average of the population and is calculated as:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x is the observed PRO of an individual, μ is the PRO-mean, and σ is the SD. See Appendix II for an exemplary rescaling calculation.

Weighting and aggregation

Weights determine the contribution of each PRO to the CM [19, 35]. We consider three different weighting options: Equal weighting (EW), differential weighting (DW), and factor analysis (FA). Literature suggests that, without strong justification to use DW (e.g., not all sub-dimensions have the same importance in the quality construct), EW should be applied [19, 47]. EW assigns the same weight to all PRO, yielding a CM to which all PROs contribute equally. However, since orthopedic care primarily addresses joint functionality and HRQoL [48], we select a priori DW for Model 1, where physical dimensions and HRQoL receive higher weighting than mental dimensions. Ideally, DW perfectly reflects patient preferences which could be determined in a patient survey [19]. Since this exceeds the scope of this study, we approximate importance by each PROM-score’s improvement: The more a PROM-score has improved 12 months post-surgery, the higher its importance. The corresponding weights are determined by measuring the improvement of each sub-dimension in standard deviation units and calculating its proportion of the total sum of all improvements. Appendix III entails more detailed considerations of different weighting methods.

Aggregation combines the weighted individual PRO into the final PRO-CM. We consider a compensatory and a non-compensatory aggregation method. A priori we use additive aggregation (Model 1), a compensatory method where worse outcomes can be counterbalanced by better outcomes. Since both surgery and recovery process differ between PKA and PHA, two treatment-specific composites are generated. They are computed as:

$$CM_i = \sum_{j=1}^n w_j I_j \quad (2)$$

where CM_i is the CM for treatment i , w_j is the weight of the j th rescaled PRO I_j .

Sensitivity and uncertainty analysis

In the sensitivity analysis, we calculate Pearson’s correlations between the resulting CM and the individual PRO to determine the PRO-CM’s sensitivity to quality variations among the sub-dimensions, i.e., the responsiveness of the PRO-CM to changes in its sub-components. In the uncertainty analysis, we compare the results of models 1–5 to examine the impact of decisions in the chosen development approach and to analyze the associated uncertainties. For this, we convert the results of each model, in each of which we alter one decision, into patient rankings to illustrate the impact of altering a decision in the development process on the final result of a patient. The patient with the highest CM value gets assigned rank 1,

Table 1 Development approaches

CM	Rescaling	Weighting	Aggregation
Model 1	z-Score	DW	Additive
Model 2	Min–max	DW	Additive
Model 3	z-Score	EW	Additive
Model 4	z-Score	FA	Additive
Model 5	Min–max	DW	Geometric

CM Composite measure development approach, *z-Score* z-Score standardization, *min–max* min–max normalization, *DW* differential weighting, *EW* equal weighting, *FA* factor analysis

the second highest rank 2, and so on. Patient rankings of our selected approach (Model 1) are compared to four alternative models (see Table 1). The greater the scatter between two compared models, i.e., the more the rankings of patients change depending on the model, the greater the impact of the corresponding changed development method. Models 2–5 are constructed as follows:

Model 2 Rescaling PRO with min–max normalization method. Min–max normalization transforms the data's original range to a common range from 0 to 1. It is calculated as:

$$m = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3)$$

where x is a PRO of an individual, $\min(x)$ is the minimum PRO, and $\max(x)$ is the maximum PRO. Min–max normalization is more sensitive to outliers and can distort relative differences and mean values. However, due to a clearly defined boundary range, it has an intuitive appeal and strong interpretative power [19]. Also, when PROs are within a small interval, the range can be expanded to increase the effect on the CM [35].

Model 3 Applying EW where all PROs contribute to the CM with the same importance. It is considered as the easiest strategy to implement, and it is not subject to any special interests and easily replicable by others [36, 49].

Model 4 Using FA to derive weights statistically. The weight of each PRO is relative to the amount of variance in common with other PRO. An approach which is resistant to potentially intentional manipulation and often applied when a great amount of indicators exist [50–52].

Model 5 Using geometric aggregation, a non-compensatory multiplicative approach that prevents poor outcomes from being compensated by good outcomes. It is computed as:

$$CM_i = \prod_{j=1}^n I_j^{w_j} \quad (4)$$

where CM_i is the CM for treatment i , w_j is the weight of the j th rescaled PRO I_j [35, 49, 53].

Results

Theoretical framework and metric selection

The PRO-CM is specific to PHA and PKA. It aims to reflect a multi-faceted picture of post-arthroplasty improvement in health as reported by patients, hence, does not entail clinical outcomes. Improvement in health (i.e., the PRO) is defined as PROM-score difference between hospital admission (HA) and the 12-month follow-up (12FU). To capture all patient-relevant aspects of post-arthroplasty improvement, we outline three main sub-dimensions of the PRO-CM. Those are general HRQoL (EQ-5D-5L) [43, 54], physical health (HOOS-PS, KOOS-PS, pain scales) [42, 44, 54], and mental health (PROMIS-D-SF, PROMIS-F-SF), as practical experience of healthcare experts and literature suggests that, although arthroplasty primarily addresses physical health, also mental health has a significant influence on patient recovery and is not sufficiently covered by EQ-5D-5L [41, 45, 48, 55, 56]. See Table 1 in Appendix I (Electronic Supplementary Material) for the PRO-CM dimensions and its sub-components.

Initial data analysis

Table 2 shows summary statistics for hip and knee PROM-scores at HA and 12FU. EQ-5D-5L has mean of 0.62 (0.60) for PKA (PHA) patients at HA and 0.84 (0.87) at 12FU, with higher scores indicating better HRQoL. Scores range between -0.661 and 1, which covers the possible total range of EQ-5D-5L. All remaining PROM-scores have opposite directionality, with higher scores indicating worse outcomes. KOOS-PS (HOOS-PS) is at 43.3 (47.6) at HA and 26.0 (14.8) at 12FU with values between 0 and 100.

While Pain-OJ shows relatively high improvement for PKA (PHA) from 6.8 (6.5) at HA to 1.9 (1.1) at 12FU with a possible range from 0 to 10, Pain-Other is at a comparatively low level at HA and barely shows change during the recovery. Since neither PKA nor PHA appears to influence Pain-Other, this score is excluded. For mental health, PKA (PHA) patients have a mean level of depression of 49.4 (49.8) at HA and 47.7 (47.3) at 12FU with scores between 41 and 79.4, and a mean level of fatigue of 48.4 (49.4) at HA and 45.9 (45.2) at 12FU with values between 33.7 and 75.8.

Computing the PRO shows that physical health dimensions improved the most during recovery. On average, Pain-OJ was reduced by 1.55 (1.61) SD for PKA (PHA) patients, followed by an improvement in KOOS-PS (HOOS-PS) of 1.08 (1.45) SD. HRQoL improved by

Table 2 Summary statistics of selected metrics for the PRO-CM

	PKA patients				PHA patients			
	HA		12FU		HA		12FU	
	Mean (SD)	Min (max)	Mean (SD)	Min (max)	Mean (SD)	Min (max)	Mean (SD)	Min (max)
Number of cases	2605				3145			
Female, n (%)	1396 (53.6)				1781 (56.6)			
Age, mean (SD)	66.1 (9.1)				66.0 (10.4)			
PROM	HA		12FU		HA		12FU	
	Mean (SD)	Min (max)	Mean (SD)	Min (max)	Mean (SD)	Min (max)	Mean (SD)	Min (max)
HRQoL								
EQ-5D-5L	0.623 (0.253)	− 0.661 (1)	0.842 (0.201)	− 0.576 (1)	0.600 (0.257)	− 0.485 (1)	0.876 (0.173)	− 0.661 (1)
Physical health								
KOOS-PS	43.2 (12.9)	0 (100)	26.0 (14.2)	0 (91.8)	–	–	–	–
HOOS-PS	–	–	–	–	47.6 (16.3)	0 (100)	14.8 (14.6)	0 (100)
Pain-OJ	6.8 (2.0)	0 (10)	1.9 (2.0)	0 (10)	6.5 (2.1)	0 (10)	1.1 (1.7)	0 (10)
Pain-other	1.6 (1.4)	0 (9)	1.5 (1.4)	0 (8.5)	1.8 (1.5)	0 (9)	1.3 (1.4)	0 (10)
Mental health								
PROMIS- depression	49.4 (8.2)	41.0 (79.4)	47.7 (8.2)	41.0 (79.4)	49.8 (8.3)	41.0 (73.3)	47.3 (7.8)	41.0 (79.4)
PROMIS-fatigue	48.4 (9.8)	33.7 (75.8)	45.9 (9.4)	33.7 (75.8)	49.4 (9.9)	33.7 (75.8)	45.2 (9.1)	33.7 (75.8)

PKA Primary knee arthroplasty, PHA primary hip arthroplasty, SD standard deviation, HA hospital admission, 12FU 12-month follow-up, *min* minimum value, *max* maximum value, PROM patient-reported outcome measure, HRQoL Health-related quality of life, Pain-OJ Pain in operated joint, Pain-Other Pain in non-operated joint

0.86 (1.04) SD for PKA (PHA) patients. Less variation is seen in mental health, with an average improvement in fatigue symptoms of 0.25 (0.43) SD and an improvement in depression symptoms of 0.21 (0.30) SD for PKA (PHA) patients [See Table 1 in Appendix III (Electronic Supplementary Material)]. Compared to PKA, PHA patients improve more during recovery in either dimension as they report worse PROM-scores at HA and better PROM-scores at 12FU. This is most evident in physical health, but also visible in HRQoL and mental health. Outliers exist for all PROM, with most extreme values of KOOS-PS (HOOS-PS). We found correlations between PROM albeit weak ones. EQ-5D-5L, which comprises mental health and pain

sub-dimensions, is only weakly correlated ($r \leq 0.5$) with mental health and pain. Since none of the correlations is > 0.7 , each PROM has sufficient independent explanatory power to the purposes of this study.

Rescaling

As a third step, we rescale via z-score standardization and compare it to min–max normalization (see Table 3). After z-score standardization, each PRO has mean = 0 and SD = 1. For equal directionality and an intuitive interpretation, each PRO is rescaled so that a higher value indicates more improvement. Values above 0 indicate more improvement

Table 3 Rescaling of patient-reported outcomes (PRO)

	PKA				PHA			
	z-score standardization		Min–max normalization		z-score standardization		Min–max normalization	
	Mean (SD)	Min (max)	Mean (SD)	Min (max)	Mean (SD)	Min (max)	Mean (SD)	Min (max)
Δ EQ5D	0 (1)	− 4.9 (3.7)	0.6 (0.1)	0 (1)	0 (1)	− 5.5 (4.1)	0.6 (0.1)	0 (1)
Δ KOOS	0 (1)	− 4.8 (5.6)	0.5 (0.1)	0 (1)	–	–	–	–
Δ HOOS	–	–	–	–	0 (1)	− 4.0 (3.6)	0.5 (0.1)	0 (1)
Δ Pain	0 (1)	− 4.6 (1.9)	0.7 (0.2)	0 (1)	0 (1)	− 4.4 (1.8)	0.7 (0.2)	0 (1)
Δ Dep	0 (1)	− 5.0 (3.5)	0.6 (0.1)	0 (1)	0 (1)	− 5.0 (3.4)	0.6 (0.1)	0 (1)
Δ Fat	0 (1)	− 4.0 (4.2)	0.5 (0.1)	0 (1)	0 (1)	− 4.8 (4.0)	0.6 (0.1)	0 (1)

PKA Primary knee arthroplasty, PHA primary hip arthroplasty, Δ patient-reported outcomes (i.e., PROM-Score Changes between Hospital Admission and 12-month Follow-up), SD standard deviation, *min* minimum value, *max* maximum value, Pain Pain in operated joint, Dep PROMIS depression, Fat PROMIS fatigue

than average in units of SD and vice versa. Upper and lower bounds can take (theoretically) infinite values, with values beyond ± 3 usually considered to be outliers.

Min–max normalization transforms all PRO onto the same scale from 0 to 1 (Model 2). Since especially negative outliers are present, most normalized PROs have mean values greater than 0.5, indicating how min–max normalization is affected by outliers. Caution must be exceeded in interpretation as the worst PRO defines the lower boundary and a normalized value of 0 can indicate PRO-deterioration.

Weighting and aggregation

The initial data analysis shows physical health dimensions to improve the most, followed by HRQoL and mental health dimensions. Consequently, for Model 1, estimated weights are 0.3 for each physical health sub-dimension, 0.2 for HRQoL, and 0.1 for each mental health sub-dimension [for a more detailed description, see Table 1 in Appendix III (Electronic Supplementary Material)]. This is in line with our assumption that physical health should be assigned more importance than mental health. Contrarily, EW assigns the same weight to each PRO, i.e., 0.2 (Model 3), while FA (Model 4) derives the weights statistically and assigns more weight to mental health. Figure 1 shows the boxplots of the five resulting PRO-CM models after aggregation of the weighted indicators.

The PRO-CM in Model 1 has a mean of 0 and SD of 0.73 for both PHA and PKA patients. Like the z-scores, it can take theoretically infinite values. Patients take values between ± 2 while PHA patients show more negative outliers with less than -3. Model 2 yields a CM with mean of 0.57 (0.60), SD of 0.09 (0.1), and a range from 0.25 to 0.89 (0.1

to 0.95) for PKA (PHA) patients. Model 3 shows a similar mean and SD as in Model 1, however, slightly contracts the range for PKA patients while expanding the range for PHA patients. Model 4 in general yields a higher SD and larger range and more extreme outliers for PKA and PHA patients with both having a mean of 0. Lastly, Model 5 has mean of 0.56 (0.59) and SD 0.09 (0.1) for PKA (PHA) patients with minimum values of 0, where at least one PRO was equal to 0.

Sensitivity and uncertainty analysis

The sensitivity analysis shows that, although in Model 1 the weights for pain-OJ and KOOS-PS (HOOS-PS) are equal, there are minimal differences in the sensitivity of the PRO-CM to variation in these PRO. Correlations (see Table 4) show the highest sensitivity in PKA (PHA) to changes in physical functionality measured by KOOS-PS (HOOS-PS) with $r=0.81$ ($r=0.82$). Thus, a change in KOOS-PS (HOOS-PS) contributes most to a change in the PRO-CM compared to other PRO. In contrast, Pain-OJ is weakly correlated with PRO-CM and has a similar level of correlation as HRQoL assessed by EQ-5D-5L. The least sensitivity is shown to change in both mental health dimensions with correlations around $r=0.5$.

This is similar in Model 2. However, the min–max normalization leads to pain-OJ becoming the largest contributor for changes in the PRO-CM, whereas it becomes somewhat less sensitive to KOOS-PS (but remains stable for HOOS-PS). Yet, this CM remains most sensitive to changes in physical health dimensions, followed by changes in HRQoL and finally in mental health dimensions. The correlations are more balanced in Model 3, with slightly higher sensitivity

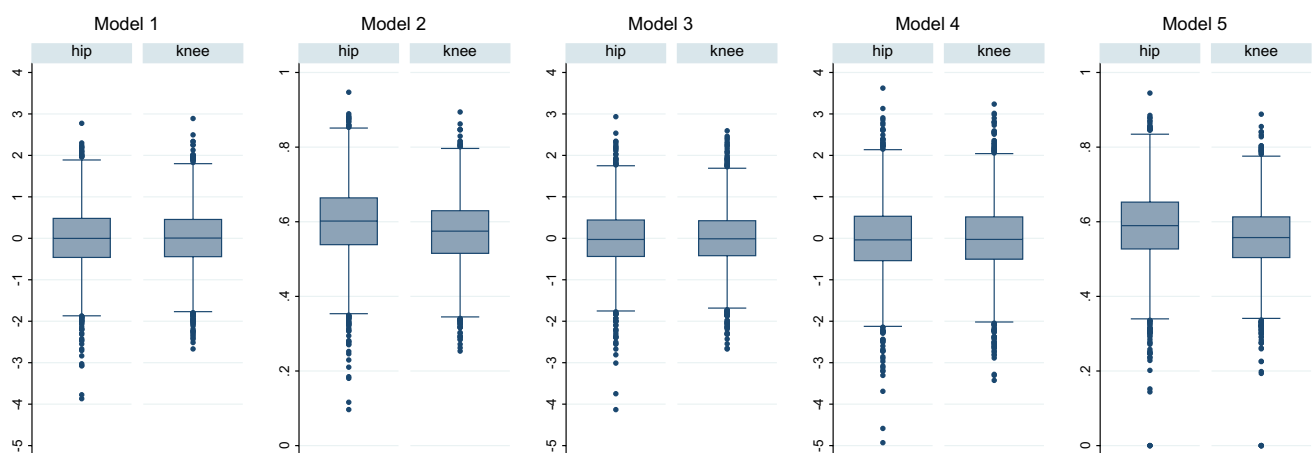


Fig. 1 Distribution of the PRO-CM Models. Figure 1 shows the PRO-CM (Model 1) and the four alternative development models: Model 1: z-score standardization, differential weights, additive aggregation; Model 2: min–max normalization, differential weights, additive aggregation; Model 3: z-score standardization, equal weights, additive aggregation; Model 4: z-score standardization, factor analysis,

additive aggregation; Model 5: min–max normalization, differential weights, geometric aggregation. The left box of each panel shows results for PHA, the right side for PKA patients. Z-Score standardized CM have a scale from -5 to 4 while minmax normalized CM have a scale from 0 to 1

Table 4 Sensitivity of the PRO-CM and alternatives to patient-reported outcomes (PRO)

	PKA (PHA)					
	Δ EQ-5D-5L	Δ KOOS-PS	Δ HOOS-PS	Δ Pain-OJ	Δ PROMIS-D	Δ PROMIS-F
Model 1	0.75 (0.77)	0.81	0.82	0.74 (0.77)	0.54 (0.50)	0.55 (0.54)
Model 2	0.73 (0.73)	0.76	0.81	0.81 (0.82)	0.52 (0.47)	0.53 (0.51)
Model 3	0.76 (0.77)	0.74	0.74	0.62 (0.65)	0.69 (0.67)	0.70 (0.69)
Model 4	0.78 (0.81)	0.76	0.77	0.56 (0.62)	0.70 (0.63)	0.71 (0.68)
Model 5	0.73 (0.73)	0.79	0.84	0.73 (0.77)	0.54 (0.49)	0.57 (0.53)

PKA primary knee arthroplasty, PHA primary hip arthroplasty, Δ patient-reported outcomes, i.e., PROM-Score changes between hospital admission and 12-month follow-up

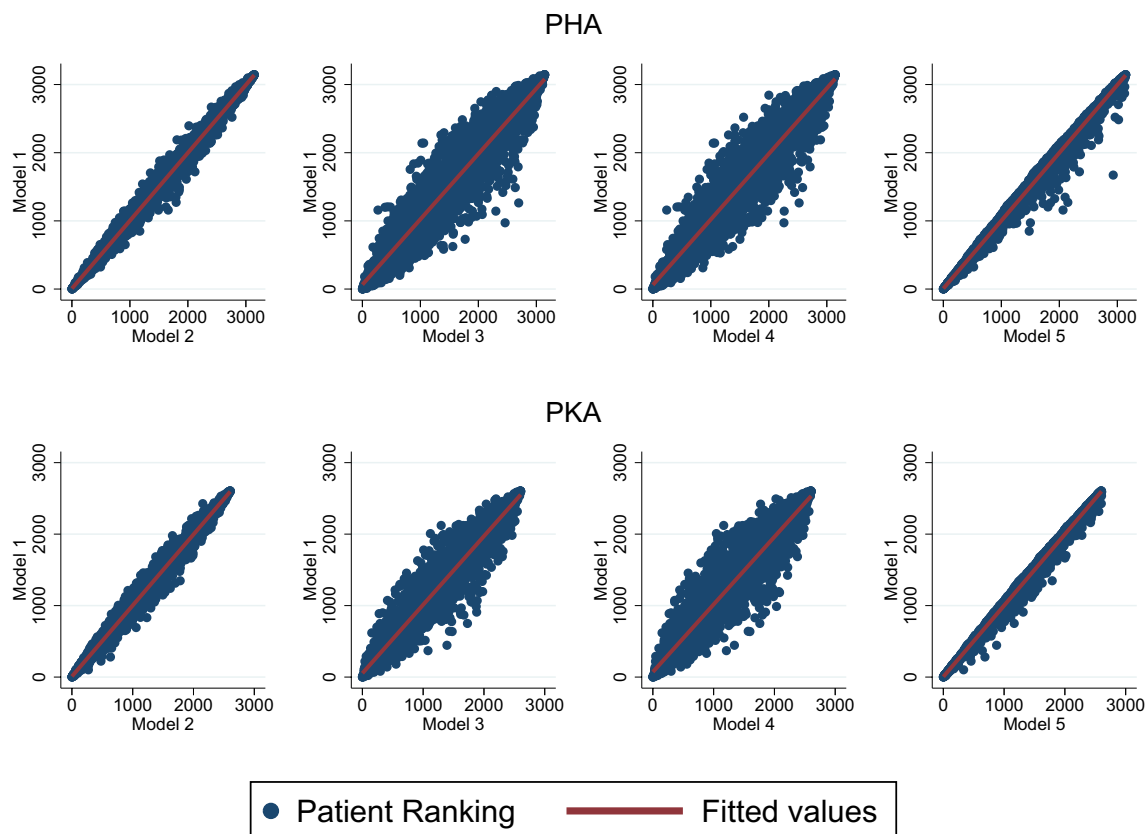


Fig. 2 Relation between Model 1 and alternative Models. Figure 2 shows the relation between the PRO-CM (Model 1) and the four alternative development models: Model 1: z-score standardization, differential weights, additive aggregation; Model 2: min-max normalization, differential weights, additive aggregation; Model 3: z-score standardization, equal weights, additive aggregation; Model 4: z-score standardization, factor analysis, additive aggregation; Model 5: min-max normalization, differential weights, geomet-

ric aggregation. The results of the respective development models are converted into patient rankings to show the impact of altering a decision in the development process can have on the final result of a patient in comparison to other patients. The closer the scattering of patient rankings is to the fitted line, the more strongly the two models compared are correlated and accordingly smaller is the impact of their different calculation method on the final result

to changes in KOOS-PS (HOOS-PS) and HRQoL than to changes in pain-OJ and mental health dimensions. Model 4 results in a CM that is most sensitive to changes in HRQoL and KOOS-PS (HOOS-PS). Mental health dimensions gain importance, while pain-OJ has the weakest correlation.

Lastly, Model 5 shows very similar results to the additive approach in Model 1.

Results of the uncertainty analysis are illustrated in Fig. 2, which shows the relation between the PRO-CM in Model 1 and the four alternative models. The y-axis represents Model 1 patient rankings and the x-axis patient rankings of

the respective alternative approach. Correlations between Model 1 and the alternative approaches are generally high, with values between $r=0.95$ and $r=0.99$. In particular, there are only minor changes in patient ranking between z-score standardization and min–max normalization ($r=0.99$), when the same weighting scheme is applied (Model 1 vs. Model 2). Altering the rescaling method does not lead to any significant distortions in our case. Also altering between additive and geometric aggregation has no significant effect on the resulting PRO-CM (Model 1 vs. Model 5). The biggest discrepancies arise when applying different weighting schemes, i.e., EW (Model 1 vs. Model 3; $r=0.96$) and FA (Model 1 vs. Model 4; $r=0.95$). Hence, while aggregation and rescaling approaches play a negligible role for the PRO-CM, it is most sensitive to the weighting methods.

Discussion

In this study, we have proposed a development approach of a patient-centered PRO-CM for PKA and PHA patients and compared it to four alternative models. The PRO-CM is robust towards different aggregation and rescaling methods, while applying different weighting schemes can have a greater impact on the final result. We consider the approach with z-scores, DW, and additive aggregation as most advantageous with respect to the data properties and the theoretical framework (Model 1). Z-scores do not distort the mean by preserving the relative differences and extreme values are acknowledged as exceptional performance, while min–max normalization (Model 2) is heavily affected by outliers [35]. DW assigns more importance to physical health dimensions that play an important role in PKA and PHA recovery [48]. EW (Model 3) should be applied when there is no strong justification to apply DW, while FA (Model 4) is rather suitable when a great number of different indicators are combined to one score [50, 52]. Additive aggregation allows, to some extent, poor outcomes to be compensated by good outcomes. In some cases, depressive symptoms were already at a low level and thus an improvement of 0 took place. With non-compensatory aggregation (Model 5), this would lead to a final CM value of 0 despite a very large improvement in physical dimensions.

As shown in the sensitivity analysis, the PRO-CM is capable of measuring relevant quality variations among sub-dimensions. The information from the individual PRO is still contained, but for outcome comparisons, only one metric must be considered instead of many different metrics. The PRO-CM can therefore empower patients, as it simplifies the monitoring of their recovery and enables them to make meaningful provider and treatment choices through enhanced comprehensibility [21, 23]. Physicians can track their patients' recovery and quickly respond to

health deteriorations with treatment adjustments [25, 26]. It is also eligible for public reporting, since assessing and ranking provider performance is facilitated [2, 3]. Reducing the outcome-side of any cost–benefit consideration to one-multidimensional metric also might aid health policy decisions, whether to calculate and present the cost-effectiveness of new forms of treatments, or to determine patient-value in the emerging VBHC considerations [27, 28, 58].

As with any CM, there are some specific and some more general limitations [6]. First, since z-scores have no clear boundaries, interpretation of z-score-based CM is difficult and not intuitive. Interpretability and comprehensibility can be enhanced by transforming the PRO-CM, e.g., to a scale from 0 to 100 (T-score transformation). Other possible approaches, such as ranking or 5-star classification, have been excluded in advance, as these methods entail a loss of information [19, 35]. However, intuitive visualization formats are highly relevant for the presentation of health data, such as the PRO-CM, and need to be discussed in a separate study [57]. Next, ideally DW perfectly reflects the preferences of patients [19]. Approximating preferences from PRO is a strong assumption and is certainly not the same for all patients. However, without knowing the true preferences, it is difficult to evaluate otherwise. Further, in this study, a complete dataset without missing values from a clinical study was used. However, in most datasets, missing values are present for which appropriate imputation methods must be applied to avoid selection bias [35]. Lastly, we illustrated the benefits of a PRO-CM with available data from the PROMoting Quality study. For broad application and realizing full potential, cross-clinic PRO-data must be available nationwide. This underlines the urgency of advancing broader PRO-measurement and usage along the patient pathway, which, at least in Germany, is still in its infancy [5]. As is, the PRO-CM developed here will primarily be applied in the evaluation of clinical trials.

Generally, opaque construction methods or individual components of poor quality can cause misinterpretation and, hence, mislead patients or trigger overly simplistic treatment, management, or policy decisions [19, 24]. When the construction methodology and its robustness are not transparently displayed, CM can easily and intentionally be skewed [6]. They can be misused for individual goals and purposes if intentionally formed for specific desired policies. It can lead to disguising very poor performance in one dimension by better performance in another and, hence, complicates the task of making targeted interventions to improve individual dimensions [6, 17]. Since a specific weighting of the underlying indicators is applied, conflicts might appear with different preferences of patients and admitting physicians [3, 59]. Although the threats and problems are widely known, CMs are often presented without going into more

detail about the development process [6]. In this study, we addressed these problems and enable replicability by justifying each step in the development.

Conclusion

We provide a transparent, stepwise development approach for a multidimensional PRO-CM that can effectively capture quality variations in orthopedic surgery. Combining multiple PRO provides a simplified but holistic picture of patients' health status while single PRO only provides information about a specific dimension. By reducing information overload, using a PRO-CM can enhance the benefits of quality transparency. However, to avoid misleading of policy, treatment, or provider decisions, the development methodology of a PRO-CM, as presented here, needs to be justified and transparent to ensure that the composite is comprehensible and replicable. Only in this way can the known problems of CM be counteracted and their full potential unfolded, which should serve one thing above all else, the promotion of quality in healthcare.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11136-023-03395-0>.

Acknowledgements The authors express their gratitude toward the following people for their contribution to this work: Viktoria Steinbeck, Laura Oschmann, Benedikt Langenberger, and Julia Silzle at the Technical University of Berlin for their valuable feedback and support, which has helped to improve the quality of the manuscript. In addition, the authors would like to thank all the partners in the PROMoting Quality research consortium, which contributed to the success of the PROMoting Quality study and thus also to this paper. Finally, the authors would also like to express great gratitude to all patients who participated in the study, without whom neither PROMoting Quality nor this article could have been completed.

Author contribution LS contributed to concept and design and drafting the manuscript. LS, DK, and CP contributed to data analysis and interpretation. RB, AG, DK, and CP contributed to critical revision of the manuscript. CP contributed to supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. PROMoting Quality is funded by the Federal Joint Committee (G-BA) Innovations fund. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declarations

Conflict of interest All authors report receiving support from the Federal Joint Committee (G-BA) Innovations fund during the conduct of the study. CP reported receiving grants from the German Research Foundation (Deutsche Forschungsgemeinschaft) and personal fees from Stryker Corporation Stryker GmbH outside the submitted work. RB reported receiving consulting fees from Dresden hospitals and Paracelsus hospitals outside the submitted work. Further RB reports receiving honoraria from Lilly, Abbvie, and Barmer sickness fund out-

side the submitted work and reports being part of the German Government Commission on Hospital Reform.

Ethical approval PROMoting Quality was conducted in accordance with the Declaration of Helsinki.

Informed consent Informed consent was obtained from all individual participants included in the study. The authors also affirm that human research participants provided informed consent for publication of analysis results of the collected data.

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

1. Gutacker, N., Siciliani, L., Moscelli, G., & Gravelle, H. (2016). Choice of hospital: Which type of quality matters? *Journal of Health Economics*, 2016(50), 230–246.
2. Pross, C., Averdunk, L.-H., Stjepanovic, J., Busse, R., & Geissler, A. (2017). Health care public reporting utilization: User clusters, web trails, and usage barriers on Germany's public reporting portal Weisse-Liste.de. *BMC Medical Informatics and Decision Making*, 17, 48. <https://doi.org/10.1186/s12911-017-0440-6>
3. Pross, C., Geissler, A., & Busse, R. (2017). Measuring, reporting, and rewarding quality of care in 5 nations: 5 Policy levers to enhance hospital quality accountability. *Milbank Quarterly*, 95, 136–183. <https://doi.org/10.1111/1468-0009.12248>
4. Pross, C., Schöner, L., Geissler, A., & Busse, R. (2021). Qualitätstransparenz im Gesundheitswesen: Eine gesundheitsökonomische Modellbetrachtung. *Gesundheitsökonomie & Qualitätsmanagement*. <https://doi.org/10.1055/a-1543-4831>
5. Ernst, S.-C.K., Steinbeck, V., Busse, R., & Pross, C. (2022). Toward system-wide implementation of patient-reported outcome measures: A framework for countries, states, and regions. *Value in Health*, 20, 22. <https://doi.org/10.1016/j.jval.2022.04.1724>
6. Barclay, M., Dixon-Woods, M., & Lyrtatzopoulos, G. (2019). The problem with composite indicators. *BMJ Quality and Safety*, 28, 338–344. <https://doi.org/10.1136/bmjqs-2018-007798>
7. Vahdat, S., Hamzehgardeshi, L., Hessam, S., & Hamzehgardeshi, Z. (2014). Patient involvement in health care decision making: a review. *Iran Red Crescent Medical Journal*, 16, e12454. <https://doi.org/10.5812/ircmj.12454>
8. Hofstede, S. N., Ceyisakar, I. E., Lingsma, H. F., Kringos, D. S., & Marang-van de Mheen, P. J. (2019). Ranking hospitals: do we gain reliability by using composite rather than individual indicators? *BMJ Quality & Safety*, 28, 94–102. <https://doi.org/10.1136/bmjqs-2017-007669>
9. Eapen, Z. J., Fonarow, G. C., Dai, D., O'Brien, S. M., Schwamm, L. H., Cannon, C. P., et al. (2011). Comparison of composite measure methodologies for rewarding quality of care: An analysis from the American Heart Association's get with the guidelines

- program. *Circulation. Cardiovascular Quality and Outcomes*, 4, 610–618. <https://doi.org/10.1161/CIRCOUTCOMES.111.961391>
10. Kelley, T. A. (2015). International consortium for health outcomes measurement (ICHOM). *Trials*, 16, 1. <https://doi.org/10.1186/1745-6215-16-S3-O4>
 11. ICHOM. Standard Sets. 19.03.2021. Retrieved October 13, 2021, from <https://www.ichom.org/standard-sets/>.
 12. Kuklinski D, Vogel J, Geissler A. The impact of quality on hospital choice Which information affects patients' behavior for colorectal resection or knee replacement? Health Care Management Science. 2021. doi:<https://doi.org/10.1007/s10729-020-09540-2>.
 13. Steinbeck V, Ernst S-C, Pross C. Patient-Reported Outcome Measures (PROMs): ein internationaler Vergleich: Bertelsmann Stiftung; 2021.
 14. Emmert, M., Kast, K., & Sander, U. (2019). Characteristics and decision making of hospital report card consumers: Lessons from an onsite-based cross-sectional study. *Health Policy*, 123, 1061–1067. <https://doi.org/10.1016/j.healthpol.2019.07.013>
 15. Hibbard, J. H. (2017). Patient activation and the use of information to support informed health decisions. *Patient Education and Counseling*, 100, 5–7. <https://doi.org/10.1016/j.pec.2016.07.006>
 16. Hibbard, J. H., Greene, J., & Daniel, D. (2010). What is quality anyway? Performance reports that clearly communicate to consumers the meaning of quality of care. *Medical Care Research and Review*, 67, 275–293. <https://doi.org/10.1177/1077558709356300>
 17. Friebe, R., & Steventon, A. (2019). Composite measures of healthcare quality: Sensible in theory, problematic in practice. *BMJ Quality and Safety*, 28, 85–88. <https://doi.org/10.1136/bmjqs-2018-008280>
 18. National Quality Forum. Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety: Composite measures: a consensus report 2009.
 19. Shwartz, M., Restuccia, J. D., & Rosen, A. K. (2015). Composite Measures of Health Care Provider Performance: A Description of Approaches. *Milbank Quarterly*, 93, 788–825. <https://doi.org/10.1111/1468-0009.12165>
 20. Jensen, M. P., Turner, J. A., & Romano, J. M. (1992). Chronic pain coping measures: individual vs. composite scores. *The Journal of Pain*, 51, 273–280. [https://doi.org/10.1016/0304-3959\(92\)90210-3](https://doi.org/10.1016/0304-3959(92)90210-3)
 21. Agniel, D., Haviland, A., Shekelle, P., Scherling, A., & Damberg, C. L. (2020). Distinguishing high-performing health systems using a composite of publicly reported measures of ambulatory care. *Annals of Internal Medicine*, 173, 791–798. <https://doi.org/10.7326/M20-0718>
 22. Tyner, C. E., Boulton, A. J., Sherer, M., Kisala, P. A., Glutting, J. J., & Tulskey, D. S. (2020). Development of composite scores for the TBI-QOL. *Archives of Physical Medicine and Rehabilitation*, 101, 43–53. <https://doi.org/10.1016/j.apmr.2018.05.036>
 23. Institute of Medicine. (2006). *Performance measurement: Accelerating improvement*. National Academies Press.
 24. Saisana, M. (2002). State-of-the-art report on current methodologies and practices for composite indicator development. Retrieved from <http://bookshop.europa.eu/en/state-of-the-art-report-on-current-methodologies-and-practices-for-composite-indicator-development-pbEUNA20408/>.
 25. McKenna, S. P., & Heaney, A. (2020). Composite outcome measurement in clinical research: The triumph of illusion over reality? *Journal of Medical Economics*, 23, 1196–1204. <https://doi.org/10.1080/13696998.2020.1797755>
 26. Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *JAMA*, 289, 2554–2559. <https://doi.org/10.1001/jama.289.19.2554>
 27. Walraven, J., Jacobs, M. S., & Uyl-de Groot, C. A. (2021). Leveraging the similarities between cost-effectiveness analysis and value-based healthcare. *Value Health*, 24, 1038–1044. <https://doi.org/10.1016/j.jval.2021.01.010>
 28. Porter, M. E. (2010). What is value in health care? *New England Journal of Medicine*, 363, 2477–2481. <https://doi.org/10.1056/NEJMp1011024>
 29. Tiftikçioğlu, B. İ. (2018). Multiple sclerosis functional composite (MSFC): Scoring instructions. *Noro Psikiyatr Ars.*, 55, S46–S48. <https://doi.org/10.29399/npa.23330>
 30. Fischer, J. S., Rudick, R. A., & Cutter, G. R. (1999). Reingold SC (1999) The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment, National MS Society Clinical Outcomes Assessment Task Force. *Multiple Sclerosis Journal*, 5, 244–250. <https://doi.org/10.1177/135245859900500409>
 31. Rajaram, R., Barnard, C., & Bilimoria, K. Y. (2015). Concerns about using the patient safety indicator-90 composite in pay-for-performance programs. *JAMA*, 313, 897–898. <https://doi.org/10.1001/jama.2015.52>
 32. DHHS. PSI90_Factsheet_FAQ_v1. Retrieved June 22, 2021, from https://www.qualityindicators.ahrq.gov/News/PSI90_Factsheet_FAQ_v1.pdf.
 33. Campione, J. R., Smith, S. A., & Mardon, R. E. (2017). Hospital-level factors related to 30-day readmission rates. *American Journal of Medical Quality*, 32, 48–57. <https://doi.org/10.1177/1062860615612158>
 34. Schmitt, J., & Wozel, G. (2005). The psoriasis area and severity index is the adequate criterion to define severity in chronic plaque-type psoriasis. *Dermatology*, 210, 194–199. <https://doi.org/10.1159/000083509>
 35. OECD. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD.
 36. Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). Tools for Composite Indicators Building: Ispra.
 37. Wilhelm, D., Lohmann, J., de Allegri, M., Chinkhumba, J., Muula, A. S., & Brenner, S. (2019). Quality of maternal obstetric and neonatal care in low-income countries: Development of a composite index. *BMC Medical Research Methodology*, 19, 154. <https://doi.org/10.1186/s12874-019-0790-0>
 38. Price, A. J., Alvand, A., Troelsen, A., Katz, J. N., Hooper, G., Gray, A., et al. (2018). Knee replacement. *Lancet*, 392, 1672–1682. [https://doi.org/10.1016/S0140-6736\(18\)32344-4](https://doi.org/10.1016/S0140-6736(18)32344-4)
 39. Ferguson, R. J., Palmer, A. J. R., Taylor, A., Porter, M. L., Malchau, H., & Glyn-Jones, S. (2018). Hip replacement. *Lancet*, 392, 1662–1671. [https://doi.org/10.1016/S0140-6736\(18\)31777-X](https://doi.org/10.1016/S0140-6736(18)31777-X)
 40. Harris, K., Dawson, J., Gibbons, E., Lim, C. R., Beard, D. J., Fitzpatrick, R., & Price, A. J. (2016). Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. *Patient Related Outcome Meas.*, 7, 101–108. <https://doi.org/10.2147/PROM.S97774>
 41. Kuklinski, D., Oschmann, L., Pross, C., Busse, R., & Geissler, A. (2020). The use of digitally collected patient-reported outcome measures for newly operated patients with total knee and hip replacements to improve post-treatment recovery: Study protocol for a randomized controlled trial. *Trials*, 21, 322. <https://doi.org/10.1186/s13063-020-04252-y>
 42. ICHOM. (2017). hip & knee osteoarthritis Data Collection reference guide.
 43. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, 20, 1727–1736. <https://doi.org/10.1007/s11136-011-9903-x>
 44. Roos, E. M., & Lohmander, L. S. (2003). The Knee injury and Osteoarthritis Outcome Score (KOOS): From joint injury to

- osteoarthritis. *Health and Quality of Life Outcomes*, 1, 64. <https://doi.org/10.1186/1477-7525-1-64>
45. PROMIS: (2013). Patient-reported outcomes measurement information system: Home page. Retrieved October 13, 2021, from <https://commonfund.nih.gov/promis/index>.
 46. Obrien, S. M., Shahian, D. M., DeLong, E. R., Normand, S.-L.T., Edwards, F. H., Ferraris, V. A., et al. (2007). Quality measurement in adult cardiac surgery: part 2—Statistical considerations in composite measure scoring and provider rating. *The Annals of Thoracic Surgery*, 83, S13–S26. <https://doi.org/10.1016/j.athoracsur.2007.01.055>
 47. Babbie, E. R. (2021). *The practice of social research*. Cengage.
 48. Ray, G. S., Ekelund, P., Nemes, S., Rolfson, O., & Mohaddes, M. (2020). Changes in health-related quality of life are associated with patient satisfaction following total hip replacement: An analysis of 69,083 patients in the Swedish Hip Arthroplasty Register. *Acta Orthopaedica*, 91, 48–52. <https://doi.org/10.1080/17453674.2019.1685284>
 49. Gan, X., Fernandez, I. C., Guo, J., Wilson, M., Zhao, Y., Zhou, B., & Wu, J. (2017). When to use what: Methods for weighting and aggregating sustainability indicators. *Ecological Indicators*, 81, 491–502. <https://doi.org/10.1016/j.ecolind.2017.05.068>
 50. MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99. <https://doi.org/10.1037/1082-989X.4.1.84>
 51. Rosato, R., Testa, S., Bertolotto, A., Confalonieri, P., Patti, F., Lugaresi, A., et al. (2016). Development of a short version of MSQOL-54 using factor analysis and item response theory. *PLoS ONE*, 11, e0153466. <https://doi.org/10.1371/journal.pone.0153466>
 52. Tucker LR MRC. Exploratory factor analysis; 1997.
 53. Talukder, B., Hipel, K., & vanLoon, G. (2017). Developing composite indicators for agricultural sustainability assessment: Effect of normalization and aggregation techniques. *Resources*, 6, 66. <https://doi.org/10.3390/resources6040066>
 54. Rolfson, O., Wissig, S., van Maasackers, L., Stowell, C., Ackerman, I., Ayers, D., et al. (2016). Defining an international standard set of outcome measures for patients with hip or knee osteoarthritis: Consensus of the International Consortium for Health Outcomes Measurement Hip and Knee Osteoarthritis Working Group. *Arthritis Care Res (Hoboken)*, 68, 1631–1639. <https://doi.org/10.1002/acr.22868>
 55. Singh, J. A., & Lewallen, D. G. (2014). Depression in primary TKA and higher medical comorbidities in revision TKA are associated with suboptimal subjective improvement in knee function. *BMC Musculoskeletal Disorders*, 15, 127. <https://doi.org/10.1186/1471-2474-15-127>
 56. National Institutes of Health. (2008). PROMIS domain framework/definitions. Retrieved June 16, 2021, from <https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis>.
 57. Albers, E. A. C., Fraterman, I., Walraven, I., Wilthagen, E., Schagen, S. B., van der Ploeg, I. M., et al. (2022). Visualization formats of patient-reported outcome measures in clinical practice: A systematic review about preferences and interpretation accuracy. *J Patient Rep Outcomes*, 6, 18. <https://doi.org/10.1186/s41687-022-00424-3>
 58. Tsevat, J., & Moriates, C. (2018). Value-based health care meets cost-effectiveness analysis. *Annals of Internal Medicine*, 169, 329–332. <https://doi.org/10.7326/M18-0342>
 59. Dixon, A., Robertson, R., Appleby, J., Burge, P., & Devlin, N. J. (2010). Patient choice: how patients choose and how providers respond.

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