doi: 10.1111/obr.12710

Obesity Surgery/Outcome Assessment

Recommendations on the most suitable quality-of-life measurement instruments for bariatric and body contouring surgery: a systematic review

C. E. E. de Vries[1], M. C. Kalff[1], C. A. C. Prinsen[2], K. D. Coulman[3], C. den Haan[4], R. Welbourn[5], J. M. Blazeby[3,6], J. M. Morton[7] and B. A. van Wagensveld[1]

[1]Department of Surgery, OLVG West, Amsterdam, The Netherlands, [2]Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands, [3]Centre for Surgical Research, School of Social and Community Medicine, University of Bristol, Bristol, UK, [4]Medical Library, OLVG, Amsterdam, The Netherlands, [5]Department of Bariatric and Upper GI Surgery, Musgrove Park Hospital, Taunton and Somerset NHS Foundation Trust, Taunton, UK, [6]Division of Surgery, Head and Neck, University Hospitals Bristol National Health Service Foundation Trust, Bristol, UK, and [7]Section of Bariatric and Minimally Invasive Surgery, Stanford University School of Medicine, Stanford University, Stanford, CA, USA

Received 7 March 2018; accepted 19 April 2018

Address for correspondence: CEE de Vries, MD, Obesity Center AmsterdamlOLVG West, Jan Tooropstraat 164, 1061 AE Amsterdam, The Netherlands.
E-mail: c.e.e.devries@olvg.nl; devries.cee@gmail.com

Summary

Objective: The objective of this study is to systematically assess the quality of existing patient-reported outcome measures developed and/or validated for Quality of Life measurement in bariatric surgery (BS) and body contouring surgery (BCS).

Methods: We conducted a systematic literature search in PubMed, EMBASE, PsycINFO, CINAHL, Cochrane Database Systematic Reviews and CENTRAL identifying studies on measurement properties of BS and BCS Quality of Life instruments. For all eligible studies, we evaluated the methodological quality of the studies by using the COnsensus-based Standards for the selection of health Measurement INstruments checklist and the quality of the measurement instruments by applying quality criteria. Four degrees of recommendation were assigned to validated instruments (A–D).

Results: Out of 4,354 articles, a total of 26 articles describing 24 instruments were included. No instrument met all requirements (category A). Seven instruments have the potential to be recommended depending on further validation studies (category B). Of these seven, the BODY-Q has the strongest evidence for content validity in BS and BCS. Two instruments had poor quality in at least one required quality criterion (category C). Fifteen instruments were minimally validated (category D).

Conclusion: The BODY-Q, developed for BS and BCS, possessed the strongest evidence for quality of measurement properties and has the potential to be recommended in future clinical trials.

Keywords: Bariatric surgery, body contouring surgery, PRO measurement, quality of Life.

Abbreviations: BCS, body contouring surgery; BS, bariatric surgery; BQL-Index, Bariatric Quality of Life Index; COS, Core Outcome Set; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; GIQLI, Gastrointestinal Quality of Life Index; HRQoL, Health Related Quality of Life; IWQoL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PRO, Patient-reported outcome; PROMs, patient-reported outcome measures; QoL, Quality of Life; RMT, Rasch Measurement Theory.

Introduction

The prevalence of severe and complex obesity is worsening in most parts of the world (1). One effective treatment is bariatric surgery (BS), and this is a rapidly growing area (2). Evaluating surgery including assessment of patients' views is important (3). Quality of Life (QoL) is a key outcome that should be reported in all clinical effectiveness trials in BS (4,5). Indeed, a recently developed Core Outcome Set (COS) for BS identified QOL as one of the nine items included in the final COS and therefore assessed in clinical trials in this area (4). The World Health Organization defined QoL as an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. QoL data offer a reliable assessment of the patients' perspectives of BS outcomes and can be useful in decision-making.

Bariatric surgery generates significant improvements in some aspects of QoL, particularly in physical domains (6). The substantial weight loss following BS, however, leaves many patients with other issues, such as excess skin, which can have a negative impact on other domains of QoL (e.g. body image, physical and sexual functioning). Subsequent body contouring surgery (BCS) has considerable potential to restore QoL (7–9). Therefore, the growth of BS has been paralleled with increasing numbers of post-bariatric body contouring procedures (10). Patient-reported outcome (PRO) measurement of QoL is highly desirable in post-bariatric BCS but is not yet routinely captured in (post-)bariatric outcome measurement. Given the growing field of BCS, the need to understand QoL and cosmetic issues related to surgical outcomes over the entire weight loss journey is important and requires information about the most appropriate PRO measurement instrument for this purpose.

A limitation in the pursuit of such evidence is the lack of standardization of the measurement of QoL and instruments available to assess it. The wide variety of generic and disease-specific questionnaires has made meta-analyses and interpretation across studies difficult (11–14). This makes it difficult to decide what treatments are best as perceived by the patient and, hence, hampers evidence-based clinical decision making. To address the lack of standardized PRO measurement instruments (PROMs) in both bariatric and post-bariatric BCS, recommendations about the most appropriate measurement instrument(s) of QoL should be made based on quality standards and criteria. In this systematic review, we aim to provide validation evidence about the most appropriate PROMs of QoL in bariatric and post-bariatric BCS. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology was used as guidance to assess and provide recommendations on the most appropriate measurement instrument(s) (15). This review is a continuation of a project to develop a COS for bariatric and metabolic surgery clinical trials (4).

Material and methods

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement was used as a guidance for reporting this systematic review (16). This review has been registered in the International Prospective Register of Systematic Reviews: CRD42017059783.

Eligibility criteria

Studies were selected if published as full-text papers and if their purpose was the development ('development paper') and/or evaluation ('validation paper') of the measurement properties of instruments that measure QoL in BS and/or BCS patients. Studies that reported indirect evidence, such as clinical trials measuring QoL, were not considered eligible. Eligible instruments included all PROMs that are specifically designed and/or validated to measure QoL in BS and/or BCS patients. Articles were excluded if (i) a different construct than QoL was measured, (ii) PROMs were developed for children or adolescents (age < 18 years) that underwent BS, and (iii) PROMs were solely developed to measure the QoL of patients with obesity not related to BS or BCS.

Literature search

On 25 January 2017, we conducted a systematic literature search in PubMed, EMBASE, Ebsco/PsycINFO, Ebsco/CINAHL, Cochrane Database Systematic Reviews and CENTRAL identifying studies on measurement properties of PROMs of QoL or Health-Related Quality of Life (HRQoL) in BS and BCS patients. We interpreted QoL and HRQoL interchangeably. We involved a clinical librarian to optimize the search strategy. The main search strategy consisted of four blocks of search terms: (i) target population (BS and BCS patients); (ii) construct of interest (QoL); (iii) type of instrument; and (iv) measurement properties. The search filter Patient Reported Outcomes Measures (PROMs), developed by Oxford University and available through the COSMIN website (www.cosmin.nl) and a highly sensitive search filter for finding studies on properties of measurement instruments (17), were used for these purposes. We did not limit our search to year of publication, study design or language. The entire search can be found in Data S1. The search results were handled within Reference Manager. The reference lists of included studies were hand searched for additional articles. We searched the Patient-Reported Outcome and Quality of Life Instruments Database database (http://www.proqolid.org) for additional information.

Study selection and data extraction

Two reviewers (C. V. and M. K.) independently screened titles and abstracts and, at a second stage, assessed the full-text articles retrieved by the literature search to identify studies evaluating measurement properties. Conflicts were resolved by consensus of the two reviewers and, if necessary, a third reviewer (C. P.). For all eligible studies, the same two independent reviewers (C. V. and M. K.) extracted data from the selected studies. Data extracted included the characteristics of included studies and instruments, and results on measurement properties. Evidence tables were used to summarize data.

Evaluation of the methodological quality of studies on measurement properties

The COSMIN study developed a consensus-based checklist to evaluate the methodological quality of studies on measurement properties (15). The COSMIN checklist describes standards for design requirements and preferred statistical methods. We evaluated the methodological quality of studies on their measurement properties using the COSMIN checklist. The COSMIN taxonomy was used to select which measurement properties of an instrument were evaluated (18). According to the COSMIN taxonomy, three domains can be distinguished: reliability, validity and responsiveness (18). The measurement properties internal consistency, reliability and measurement error fall within the domain validity, the measurement properties content validity, construct validity and criterion validity fall within the domain validity and the measurement property responsiveness falls within the domain responsiveness. We assessed internal consistency, reliability, measurement error, content validity (including face validity), structural validity, hypotheses testing (i.e. for convergent and divergent validity) and cross-cultural validity (these three are aspects of construct validity), criterion validity and responsiveness. As, in general, there is no gold standard for PROMs, 'criterion validity' was not considered. Each measurement property can be rated as excellent, good, fair or poor (19). The 'lowest score counts' principle was used, which means that the overall rating for the study was determined by the lowest rated measurement property.

The same two reviewers (C. V. and M. K.) independently evaluated the methodological quality of included studies. Discrepancies were discussed with a third reviewer (C. P.) to reach consensus. For the measurement property 'hypothesis testing', we evaluated convergent/divergent validity and discriminative validity. Convergent/divergent validity is the correlation between a comparator instrument that measures a similar or different construct. Discriminative validity is the ability of a measurement instrument to make a distinction between different subgroups. Data on the interpretability and generalizability were collected when data were available. Interpretability and generalizability do not refer to the quality of an instrument and are therefore not measurement properties. These characteristics are included in the COSMIN checklist, as they provide important information on the suitability of a measurement instrument. Interpretability describes whether it is clear what the scores or change scores of the instrument of interest mean. Interpretability includes an assessment of the distribution of scores, floor and ceiling effects (the percentage of patients with the lowest or highest scores; e.g. a high percentage of patients with the highest scores [ceiling effect] limits the ability to measure changes) and minimal clinically important difference (the smallest difference in construct [QoL scores] between patients that is considered clinically important). Generalizability describes whether the patient population in which the measurement instrument was evaluated was adequately described to generalize the results. We evaluated different language versions of QoL measurement instruments as distinct instruments. We assumed that different language versions possess different measurement properties.

Quality of the measurement properties

We independently evaluated the quality of the measurement properties of the included measurement instruments by applying the Terwee criteria for good measurement properties (20) on which international consensus was obtained (21) (Table 1) (22). The quality of each measurement property was rated as positive (+), negative (−) or indeterminate (?).

Quality of the instruments: best evidence synthesis

Taking both the quality of the studies and the quality of the measurement instruments into account, the overall evidence on a measurement property includes the number and the methodological quality of the included studies and the consistency of their results. The overall rating of the quality of a measurement property was based on a levels of evidence approach (23). The results of studies of poor methodological quality were not included in the best evidence synthesis. The criteria of best evidence synthesis are shown in Table 2.

Recommendations for the selection of the most suitable Quality of Life measurement instruments

Recommendations on the most suitable QoL instruments were based on the methodological quality of included studies and on the adequacy of the instrument. As previously described by the Harmonizing Outcome Measures for Eczema initiative, the three criteria of the Outcome Measures in Rheumatology filter had to be met by a measurement instrument to be recommended for use. Outcome measures

Table 1 Quality criteria for measurement properties adapted from Terwee *et al.* (20) and PROMIS Methodology (22)

Property	Rating	Adequacy criteria
Reliability		
Internal consistency	+	Cronbach's alpha(s) ≥ 0.70
(CTT methods applied)	?	Cronbach's alpha not determined
	–	Cronbach's alpha(s) < 0.70
Internal consistency	+	Person separation index ≥ 0.70
(IRT methods applied)	?	Person separation index not determined
	–	Person separation index < 0.70
Measurement error	+	MIC > SDC OR MIC outside the LoA
	?	MIC not defined
	–	MIC ≤ SDC OR MIC equals or inside LoA
Reliability	+	ICC/weighted Kappa ≥ 0.70, OR Pearson's *r* ≥ 0.80
	?	Neither ICC/weighted Kappa nor Pearson's *r* determined
	–	ICC/weighted Kappa < 0.70 OR Pearson's *r* < 0.80
Validity		
Content validity	+	All items are considered to be relevant for the construct to be measured, for the target population and for the purpose of the measurement AND the questionnaire is considered to be comprehensive
	?	Not enough information available
	–	Not all items are considered to be relevant for the construct to be measured, for the target population and for the purpose of the measurement OR the questionnaire is considered not to be comprehensive
Construct validity		
Structural validity	+	Factors should explain at least 50% of the variance
(CTT methods applied)	?	Explained variance not mentioned
	–	Factors explain <50% of the variance
Structural validity	+	Residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37, item scalability > 0.30, IRT model fit: G2 > 0.01, no DIF for important subject characteristics (such as age, gender and education): McFadden's R2 < 0.02, OR no non-uniform DIF
(IRT methods applied)	?	Important statistics not reported
	–	Residual correlations among the items after controlling for the dominant factor ≥ 0.20 OR Q3's ≥ 0.37, item scalability ≤ 0.30, IRT model fit: G2 ≤ 0.01, important DIF for important subject characteristics (such as age, gender and education): McFadden's R2 ≥ 0.02, OR non-uniform DIF
Hypothesis testing	+	Correlations with instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses AND correlation with related constructs is higher than with unrelated constructs
(convergent/divergent validity)	?	Solely correlations determined with unrelated constructs
	–	Correlations with instruments measuring the same construct <0.50 OR <75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Hypothesis testing	+	Differences in scores on the measurement instrument for all evaluated patient subgroups are statistically significant OR ≥75% of results in accordance with hypotheses
(discriminative validity)	?	Some differences statistically significant, others not
	–	Differences in scores on the measurement instrument for all evaluated patient subgroups are not statistically significant OR <75% of results in accordance with hypotheses
Cross-cultural validity	+	No differences in factor structure OR no important DIF between language versions
	?	Multiple group factor analysis not applied AND DIF not assessed
	–	Differences in factor structure OR important DIF between language versions
Responsiveness		
Responsiveness	+	Correlation with changes on instruments measuring the same construct ≥0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70 AND correlations with changes in related constructs are higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	–	Correlations with changes on instruments measuring the same construct <0.50 OR <75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlations with changes in related constructs are lower than with unrelated constructs

AUC, area under the curve; CTT, classical test theory; DIF, differential item functioning; ICC, intraclass correlation coefficient; LoA, limits of agreement; MIC, minimal important change; SDC, smallest detectable change; +, positive rating; ?, indeterminate rating; –, negative rating.

should adequately meet the criteria of (i) truth (i.e. face, content, construct and criterion validity; measure what it is intended to measure), (ii) discrimination (i.e. reliability and sensitivity to change; discriminate between situations of interest) and (iii) feasibility (i.e. be applied and interpreted easily) in order to be meaningful and relevant (24). Four degrees of recommendation were assigned to validated instruments included in this review (A–D), adopted from the Harmonizing Outcome Measures for Eczema initiative and applied in previous reviews of this initiative (25):

Table 2 Levels of evidence for the overall adequacy of a measurement property adapted from Schellingerhout *et al.* (23)

Level	Rating	Criteria
Strong	+++, ? (strong) or −−−	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++, ? (moderate) or −−	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+, ? (limited) or −	One study of fair methodological quality
Conflicting	+/−	Conflicting findings
Unknown	?	Only studies of poor methodological quality

+, positive rating; ?, indeterminate rating; −, negative rating.

A. QoL measurement instrument meets all requirements and is recommended for use.
B. QoL measurement instrument meets two or more quality criteria, but performance in all other required quality criteria is unclear, so that the PRO measurement instrument has the potential to be recommended in the future depending on the results of further validation studies.
C. QoL measurement instrument has low quality in at least one required quality criterion (≥1 rating of 'minus') and therefore is not recommended for usage.
D. QoL measurement instrument has (almost) not been validated. Its performance in all or most relevant quality criteria is unclear, so it is not recommended to be used until further validation studies clarify its quality.

Results

Out of the 4,354 articles, 26 articles described development and validation of a measurement instrument and were considered eligible for assessment of the methodological quality (Data S2). These described 24 measurement instruments in 26 studies. Of the measurement instruments included, 21 were developed for BS patients, two for both BS and BCS patients and one for BCS patients. The characteristics of the different instruments are displayed in Table 3. Important characteristics of the included development and validation studies are shown in Table 4. Information on the content (i.e. on domain level) of the different instruments is shown in Table 5. Social functioning was captured in all measurement instruments except the EQ-5D-5L, physical functioning was included in all instruments except the bariatric and obesity-specific survey and sexual functioning was included in all instruments except the Bariatric Quality of Life Index (BQL-Index), the Quality of Life for Obesity Surgery, the EQ-5D-5L and the Gastrointestinal Quality of Life Index (GIQLI). Data S3 shows the detailed results of interpretability of the QoL measurement instruments. Values for the minimal clinically important difference were only available for the Laval Questionnaire and the EQ-5D-5L. The BQL-Index, the BODY-Q, the Spanish Obesity-related Problems scale (OP-scale), the Greek Moorehead-Ardelt Quality of Life Questionnaire II (M-A QoLQII) and the Portugese M-A QoLQII showed no evidence for floor or ceiling effects. The EQ-5D-5L demonstrated ceiling effects and the Danish BODY-Q demonstrated ceiling effects in the experience scales.

Quality assessment and results of the studies

The methodological quality of the included studies is presented in Table 6, and the quality assessment of the measurement properties is presented in Table 7. Data S4 describes detailed results of the different measurement properties of every single instrument and study.

Quality of the instruments: best evidence synthesis

The results of the best evidence synthesis and recommendations of QoL instruments (Table 8) are described in the succeeding texts according to the category of recommendation (A–D) (Table 9):

Category A instruments

No instrument met all required quality criteria to be recommended for the measurement of QoL in BS and/or BCS.

Category B instruments

Seven measurement instruments have the potential to be recommended for BS and/or BCS in the future depending on further validation studies.

Body-Shape-Related Quality of Life (body contouring surgery). The measurement properties of the Body-Shape-Related Quality of Life were evaluated in two studies. Strong evidence was found for good content validity (27), moderate evidence for good internal consistency, limited evidence for good reliability and moderate evidence for indeterminate structural validity and hypotheses testing (26). The evidence for responsiveness remained unclear due to poor methodological quality of the study (26). For interpretability, the mean absolute change in score from baseline was 21.9 (SD 16.9) (26).

Table 3 Characteristics of the different instruments

QoL instrument		Target population	Number of items	Number of subscales	Number/types of response categories	Scoring algorithm	Feasibility (completion time)	Recall period in the items	Administration costs	Available translations
BCS	Body-QoL (26,27)	BCS patients	20	4	Ordinal score (1–5)	Sum score (range 20 [worst]–100 [best])	4.3 ± 2.3 min	2–4 weeks	Free for clinicians, researchers, academic and non-profit organizations	Spanish, English
BCS and BS	BODY-Q (28–30)	Weight loss and BCS patients	138	18	4 point Likert scale	Sum Calculated Rasch score of sum score (range 0 [worst]–100 [best]) per subscale		1 week	Free for non-profit purposes. For profit users licencing fees	English, Danish, Dutch, Swedish, Polish, Finnish, French, Norwegian, Italian, German
BS	M-A QoLQ (BAROS) (31)	Bariatric surgery patients	5	5	Ordinal score (divided in 0.25/0.5)	Sum score (−3 [worst]–3 [best])	<1 min	ND	Lifetime Licence is $800.00 per hospital or single surgical office	Many available translations
	BOSS (32)	Bariatric surgery patients	42	6	5 point Likert scale	Standardized scores (0% [worst]–100% [best])		>2 weeks	ND	English
	The Laval Questionnaire (33)	Patients with morbid obesity	44	6	7 point Likert scale	Domain scores (1 [worst]–7 [best])		2 weeks	Available on request	French
	BQL-index (34,35)	Bariatric surgery patients	19	5	Ordinal score (1–5)/dichotomous (yes-or-no)	Sum score (range 0 [worst]–78 [best])		48 h	Available without licence fees	German, English, Spanish, Italian
	GIQLI (36)	Bariatric surgery patients	36	5	4 point Likert scale	Sum score (range 0 [worst]–144 [best])		ND	Free for not funded academic users	English, Spanish, Chinese, Dutch, French, Italian, Kazakh, Russian, Swedish
	IWQOL-Lite (37)	Persons being treated for obesity	31	5	3 point Likert scale	Sum score and one for each domain (range 0 [worst]–100 [best])		ND	Licencing fee varies according to use	76 translations
	M-A QoLQII (38)	Bariatric surgery patients	6	6	10 point Likert scale	Sum score (−50 [worst]–50 [best])		ND	Lifetime Licence is $1,000.00 per hospital or single surgical office	English, Czech, German, Italian, Spanish, Portuguese, Greek (many available translations)
	PBOT (39)	Massive weight loss (BCS) patients	27	ND	4–5 point Likert scale	Sum score (range 20 [worst]–132 [best])	10–15 min	ND	ND	English
	OP-scale (40)	Patients with obesity	8	1	4 point Likert scale	Sum score (range 0 [best]–100 [worst])		ND	ND	Swedish, Norwegian, Spanish, Korean, Finnish

(*Continues*)

Table 3 (Continued)

QoL instrument	Target population	Number of items	Number of subscales	Number/types of response categories	Scoring algorithm	Feasibility (completion time)	Recall period in the items	Administration costs	Available translations
QOLOS (41)	Bariatric surgery patients	36 (section 1), 20 (section 2)	7 (section 1), 4 (section 2)	5 point Likert scale	Average score per section (range 1 [worst]–5 [best])		ND	ND	German, English
EQ-5D-5L (42)	Bariatric surgery patients	5	5	Five response levels: no problems (Level 1); slight; moderate; severe; and extreme problems (Level 5)	A single index 'utility' score (−0.281 [worst] to 1 [best])	A few minutes	ND	Licencing fees are dependent upon the type of study/trial/ project, funding source, sample size and number of requested languages.	176 translations

BAROS, Bariatric Analysis and Reporting Outcome System; BCS, body contouring surgery; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; BS, bariatric surgery; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery.

BODY-Q (bariatric surgery and body contouring surgery). Two studies evaluated the measurement properties of the BODY-Q. There was strong evidence for good content validity (28), moderate evidence for good internal consistency and structural validity, limited evidence for good reliability and moderate evidence for indeterminate hypotheses testing (29). The evidence for responsiveness remains unknown because of poor methodological quality of the study (29). Interpretability yielded a mean absolute change ranging from 0.2 (SD 18.7) to 12.4 (SD 19.6) for appearance scales, from 0.5 (SD 17.0) to 9.9 (SD 19.4) for HRQoL scales and from −2.0 (SD 18.8) to −6.0 (SD 16.3) for experience scales (29). Participants improved on BODY-Q scales following BS measuring appearance (of abdomen, back, body, buttocks, hips/outer thighs, inner thigh), body image and physical function and social function (moderate to large effect sizes [0.60 to 2.29] and standardized response means [0.47 to 1.35]) (30).

Bariatric and obesity-specific survey (bariatric surgery). The measurement properties of the bariatric and obesity-specific survey were evaluated in one study. Moderate evidence was found for good internal consistency and reliability (32). There was moderate evidence for indeterminate content validity and structural validity (32). The evidence of hypotheses testing remains unknown because of poor methodological quality of the study (32).

Quality of Life for Obesity Surgery (bariatric surgery). The measurement properties of the Quality of Life for Obesity Surgery were evaluated in one study. There was strong evidence for good content validity, moderate evidence for good internal consistency and structural validity and limited evidence for indeterminate hypotheses testing (41).

Norwegian Obesity-related Problems scale (bariatric surgery). The measurement properties of the Norwegian OP-scale were evaluated in one study. Moderate evidence was found for good internal consistency and structural validity, and limited evidence was found for good hypotheses testing and responsiveness (45). The evidence for cross-cultural validity remains unknown because of poor methodological quality of the study (45). Interpretability yielded a mean absolute change of 42.29 (45).

Spanish Obesity-related Problems scale (bariatric surgery). The measurement properties of the Spanish OP-scale were evaluated in one study. Moderate evidence

Table 4 Important characteristics of the included development and validation studies

		Study characteristics					Study population			
QoL instrument		Number of studies	Year of publication	Geographic location(s)	Language(s)	Setting(s)	Number of participants per study	Age range (years)	Proportion of women (%)	BMI range (kg m^{-2})
BCS	Body-QoL (26,27)	*2*	2014; 2016	Chile	Spanish, English	Community, tertiary care	1,200	15–75 (20–70)	71.6	16.9–41.8
BCS and BS	BODY-Q (28–30)	3	2014; 2016; 2017	USA, UK, Canada	English	Secondary, tertiary and private care	705	18–75	88.1–94	17.8–75.8
	Danish BODY-Q (43,44)	2	2017	Denmark	Danish	Secondary and tertiary care	495	17–75	67–84	20–68
BS	M-A QoLQ (BAROS) (31)	1	1998	USA	English	ND	ND	ND	ND	
	BOSS (32)	1	2014	UK	English	ND	236	18–65	63.5–77.1	ND (mean 29.6–48.4)
	The Laval Questionnaire (33)	2	2011	Canada	French	Tertiary care	112	ND (mean 45.0 [10.2])	73–79	ND (mean 52.6–54.4)
	BQL-index (34,35)	2	2005; 2009	Germany	German, English	Secondary care	133 (446)	ND (mean 38.8 [11.0])	81	ND (mean 47.2 [7.6])
	GIQLI (36)	1	2005	Spain	Spanish, English	ND	190	18–50	ND	ND (mean 24.3–43.7)
	IWQOL-Lite (37)	1 (6)*	2010	–	–	–	1,635	ND	ND	ND (mean 47.5 [8.4])
	M-A QoLQII (38)	1	2003	USA	English	Secondary care	110	19–65	82	32–92
	PBOT (39)	1	2014	UK	English	ND	30	24–68	60	17.75–49.7
	OP-scale (40)	1	2003	Sweden	Swedish, English	Primary care	6,683	37–57	68	ND (mean 41.2–42.7)
	QOLOS (41)	1	2017	Germany	German	Secondary care	220 + 219	ND (mean 40.51 [11.24], mean 43.71 [10.82])	72.3–77.6	ND (mean 35.28–48.12)
	EQ-5D-5L (42)	1	2017	UK	English	Secondary and tertiary care	189	23 ± 70	75	35–52
	Norwegian OP-scale (45)	1	2015	Norway	Norwegian	Secondary care	181	ND (mean 43.1 [12.5])	68	ND (mean 45 [6.9])
	Spanish IWQoL-lite (46)	1	2011	Spain	Spanish	Secondary care	109	ND (mean 42.50 [8.21])	84	ND (mean 47.2 [5.94])
	Spanish OP-scale (47)	1	2009	Spain	Spanish	Secondary care	123	ND (mean 42.93 [10.7])	85.37	ND (mean 49.27 [7.57])
	Greek M-A QoLQII (48)	1	2012	Greece	Greek	Tertiary care	175	ND (mean 38.8 [11])	65.1	ND (mean 38.8 [11])

(*Continues*)

Table 4 (Continued)

QoL instrument	Study characteristics					Study population				
	Number of studies	Year of publication	Geographic location(s)	Language(s)	Setting(s)	Number of participants per study	Age range (years)	Proportion of women (%)	BMI range (kg m^{-2})
Korean M-A QoLQII (49)	1	2014	Korea	Korean	Tertiary care	53	ND (mean 37.8 [12.2])	84.9	ND (mean 30.1 [5.9])
Portuguese M-A QoLQII (50)	1	2014	Portugal	Portuguese	Secondary care	150	23–74	86	22–55.2
Czech, German, Italian and Spanish M-A QoLQII (51)	1	2009	Czech Republic, Germany, Italy and Spain	Czech, German, Italian and Spanish	Secondary and tertiary care	893	ND	78.6	17–75

*The IWQOL-Lite was evaluated on measurement properties in one systematic review describing six studies.
BAROS, Bariatric Analysis and Reporting Outcome System; BCS, body contouring surgery; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; BS, bariatric surgery; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery.

was found for good internal consistency and structural validity, and limited evidence was found for good hypotheses testing (47). The evidence for cross-cultural validity remains unknown because of poor methodological quality of the study (47).

Danish BODY-Q (bariatric surgery and body contouring surgery). Two studies evaluated the measurement properties of the Danish BODY-Q. There was strong evidence for good content validity, moderate evidence for good cross-cultural validity and limited evidence for good internal consistency and structural validity (43,44).

Category C instruments

Two instruments (BS) had poor quality in at least one required quality criterion and were not recommended to be used in BS (category C). The BQL-Index (34,35) and the Laval Questionnaire (33) had low quality in one measurement property. Structural validity was poor for the BQL-Index (34), and conflicting evidence was found for responsiveness of the Laval Questionnaire (33).

Category D instruments

Fifteen instruments (BS) were minimally validated and are not recommended for use until adequate validation studies clarify their quality. These included Moorehead-Ardelt Quality of Life Questionnaire (M-A QoLQ)/Bariatric Analysis and Reporting Outcome System (31), GIQLI (36), Impact of Weight Quality of Life-Lite (IWQoL-Lite) (37), M-A QoLQII (38), Post Bariatric Outcome Tool (39), OP-scale (40), EQ-5D-5L (42), Greek, Korean, Portuguese, Czech, German, Italian and Spanish M-A QoLQII (48–51) and Spanish IWQoL-lite (46). The evidence of the majority of the measurement properties of these instruments could not be interpreted due to poor methodological quality of the study. The OP-scale showed moderate evidence for good internal consistency and structural validity (40), and the Spanish IWQoL-lite showed moderate evidence for good internal consistency and limited evidence for good structural validity (46). Limited evidence was found for good reliability of the Greek and Portuguese M-A QoLQII (48,50). Conflicting evidence was found for hypotheses testing of the EQ-5D-5L (42).

Discussion

This systematic review evaluated and compared the measurement properties of 21 QoL instruments designed for use in BS, one QoL instrument designed for use in BCS and two QoL instruments designed for use in both BS and BCS. None of these instruments complied with the filter requirements of truth, discrimination and feasibility, which

Table 5 Comparison of content of the different QoL instruments on content domain level

Domain	BCS	BS and BCS	BS									
	Body-QoL† (26,27)	BODY-Q* (28–30)	M-A QoLQ (BAROS)* (31)	BOSS (32)	Laval questionnaire (33)	BQL Index* (34,35)	GIQLI (36)	IWQOL-Lite (37)	M-A QoLQII* (38)	PBOT (39)	OP-scale† (40)	QOLOS (41)	EQ-5D-5L (42)
Experience health care		X								ND			
Appearance	X	X		X									
(HR)QoL		X	X			X			X				
Physical	X	(x)	(x)		X	(x)	X	X	(x)			X	X
Psychological		(x)				(x)					x		
Sexual	X	(x)	(x)	X	X			X	(x)		(&)		
Social	X	(x)	(x)	X	X	(x)	X	X	(x)		x	X	
Symptoms		(x)		X									
Body image	x (&)	(x)										X	
Self esteem	x			(x)				X	(x)				
Work			(x)	X				X	(x)				
Eating				X					(x)			X	
Incapacity				X									
Personal hygiene					X								
Emotional					X		X						
Problems and symptoms related to obesity surgery						(x)							
Comorbidity (obesity)						(x)							
Digestive symptoms						X		X					
Medical treatment								X					
Family												X	
Positive activities												X	
Partnership												X	
Excess skin												X	
Eating adjustment												X	
Dumping												X	
Satisfaction with surgery												X	
Self-care													X
Usual activities													X
Pain													X
Anxiety													X

*The scales or items of measurement instruments that overlap the content domains of other measurement instruments are listed in brackets on the right side of the column. The BODY-Q captures three domains via 18 independently functioning scales. The M-A QoLQ (BAROS) and the M-A QoLQII capture only one domain via five and six items.

†A domain that is described in the article as one domain but consists of two domains in the table are listed as lower-case letters and an ampersand in between.

BAROS, Bariatric Analysis and Reporting Outcome System; BCS, body contouring surgery; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; BS, bariatric surgery; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery.

Table 6 Methodological quality of the included development and validation studies according to the COSMIN checklist

	QoL instrument	IRT	Box A Internal consistency	Box B Reliability	Box C Measurement error	Box D Content validity	Box E Structural validity	Box F Hypotheses† testing	Box G Cross-cultural validity	Box I Responsiveness	Box J Interpretability*
BCS	Body-QoL (26,27)		Good	Fair		Excellent	Good	Poor		Poor	Yes
BCS and BS	BODY-Q (28–30)	Good	Good	Fair		Excellent	Good	Poor		Poor	Yes
	Danish BODY-Q (43,44)	Excellent	Fair			Excellent	Fair		Good		
BS	M-A QoLQ (BAROS) (31)					Poor					
	BOSS (32)		Good	Good		Good	Good	Poor		Poor	
	The Laval Questionnaire (33)		Poor	Good		Excellent	Poor	Poor		Fair	Yes
	BQL-index (34,35)		Good	Poor		Fair	Fair	Poor		Poor	
	GIQLI (36)							Poor			
	IWQOL-Lite (37)		Poor			Poor		Poor			
	M-A QoLQII (38)		Poor	Poor				Poor			
	PBOT (39)		Poor	Poor				Poor			
	OP-scale (40)		Good				Good	Poor		Poor	Yes
	QOLOS (41)		Good			Excellent	Good	Poor			
	EQ-5D-5L (42)							Fair		Poor	Yes
	Norwegian OP-scale (45)		Good				Good	Fair	Poor	Fair	Yes
	Spanish IWQoL-lite (46)		Good				Good	Poor			
	Spanish OP-scale (47)		Good				Good	Fair	Poor		
	Greek M-A QoLQII (48)		Poor	Fair				Poor	Poor		
	Korean M-A QoLQII (49)		Poor					Poor	Poor		
	Portuguese M-A QoLQII (50)		Poor	Good				Poor	Poor		Yes
	Czech, German, Italian and Spanish M-A QoLQII (51)		Poor					Poor	Poor	Poor	

*Quality rating for interpretability not according to COSMIN/Terwee *et al.* as no formal quality criteria for interpretability exist.
†Subtypes of hypothesis testing include convergent validity, divergent validity, discriminant validity and discriminative validity.
BAROS, Bariatric Analysis and Reporting Outcome System; BCS, body contouring surgery; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; BS, bariatric surgery; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery.

Table 7 Quality assessment of measurement properties according to predefined criteria proposed by Terwee *et al.* (20)

		Box A Internal consistency	Box B Reliability	Box C Measurement error	Box D Content validity	Box E Structural validity	Box F Hypotheses testing	Box G Cross-cultural validity	Box I Responsiveness	Box J Interpretability*
	QoL instrument									
BCS	Body-QoL (26,27)	+	+		+	?	?		?	+
BCS and BS	BODY-Q (28–30)	+	+		+	+	?		?	+
	Danish BODY-Q (43,44)	+			+	+		+		
BS	M-A QoLQ (BAROS) (31)				?					
	BOSS (32)	+	+		?	?	?			
	The Laval Questionnaire (33)	+	+		+	+	?		±	+
	BQL-index (34,35)	+	?		+	-	?		?	
	GIQLI (36)						±			
	IWQOL-Lite (37)	+			?		?			
	M-A QoLQII (38)	+	+				?			
	PBOT (39)	+	?				+			
	OP-scale (40)	+				+	+		?	+
	QOLOS (41)	+			+	+	?			
	EQ-5D-5L (42)						±		?	+
	Norwegian OP-scale (45)	+				+	+	?	+	
	Spanish IWQoL-lite (46)	+				+	?			
	Spanish OP-scale (47)	+				+	+	?		
	Greek M-A QoLQII (48)	+	+				+	?		
	Korean M-A QoLQII (49)	+					+	?		
	Portuguese M-A QoLQII (50)	+	+				+	?		+
	Czech, German, Italian and Spanish M-A QoLQII (51)	+					+	?	?	

*Quality rating for interpretability not according to COSMIN/Terwee *et al.* as no formal quality criteria for interpretability exist.

BAROS, Bariatric Analysis and Reporting Outcome System; BCS, body contouring surgery; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; BS, bariatric surgery; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery; +, positive rating; ?, indeterminate rating; −, negative rating.

Table 8 Best evidence synthesis and recommendations

QoL instrument		Box A Internal consistency	Box B Reliability	Box C Measurement error	Box D Content validity	Box E Structural validity	Box F Hypotheses testing	Box G Cross-cultural validity	Box I Responsiveness	Recommendation
BCS	Body-QoL (26,27)	++	+		+++	? (mod)	? (mod)		?	B
BCS and BS	BODY-Q (28–30)	++	+		+++	++	? (mod)		?	B
BS	Danish BODY-Q (43,44)	+			+++	+		++		B
BS	M-A QoLQ (BAROS) (31)				?					D
	BOSS (32)	++	++		? (mod)	? (mod)	?			B
	The Laval Questionnaire (33)	?	++		+++	?	? (lim)		±	C
	BQL-index (34,35)	++	?		? (lim)	-	?		?	C
	GIQLI (36)						?			D
	IWQOL-Lite (37)	?			?		?			D
	M-A QoLQII (38)	?	?				?			D
	PBOT (39)	?	?				?			D
	OP-scale (40)	++				++	?		?	D
	QOLOS (41)	++				++	?			B
	EQ-5D-5L (42)				+++		±		?	D
	Norwegian OP-scale (45)	++				++	+	?	+	B
	Spanish IWQoL-lite (46)	++				+	?			D
	Spanish OP-scale (47)	++				++	+	?		B
	Greek M-A QoLQII (48)	?	+				?	?		D
	Korean M-A QoLQII (49)	?					?	?		D
	Portuguese M-A QoLQII (50)	?	++				?	?		D
	Czech, German, Italian and Spanish M-A QoLQII (51)	?						?	?	D

BAROS, Bariatric Analysis and Reporting Outcome System; BCS, body contouring surgery; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; BS, bariatric surgery; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery; +++, ++, +: positive rating indicating adequate measurement property; ?, ? limited, ? moderate: indeterminate rating indicating indeterminate measurement property; –: negative rating indicating inadequate measurement property; ±: conflicting findings; n/i: not interpretable.

Table 9 Category of recommendations

Category of recommendation	instruments
A	–
B	Body-QoL, BODY-Q, BOSS, QOLOS, Danish BODY-Q, Norwegian and Spanish OP-scale
C	BQL-Index, Laval questionnaire
D	M-A QoLQ/BAROS, GIQLI, IWQoL-Lite, M-A QoLQII, PBOT, OP-scale, EQ-5D-5L, Greek, Korean, Portuguese, Czech, German, Italian and Spanish M-A QoLQII

BAROS, Bariatric Analysis and Reporting Outcome System; Body-QoL, Body-Shape-Related Quality of Life; BOSS, bariatric and obesity-specific survey; BQL Index, Bariatric Quality of Life Index; GIQLI, Gastrointestinal Quality of Life Index; IWQOL-Lite, Impact of Weight Quality of Life-Lite; M-A QoLQ, Moorehead-Ardelt Quality of Life Questionnaire; M-A QoLQII, Moorehead-Ardelt Quality of Life Questionnaire II; OP-scale, Obesity-related Problems scale; PBOT, Post Bariatric Outcome Tool; QOLOS, Quality of Life for Obesity Surgery.

indicates the need for further validation studies. Hence, no QoL instrument can currently be highly recommended. All identified instruments have gaps in their validation, and none of the instruments provided evidence of the quality of measurement error or estimated a minimally important difference. Remarkably, criterion validity was still described as a measurement property in some validation studies, even though no true gold standard is available in QoL measures. Further work is needed to validate the available tools for BS and BCS, and this should be a priority.

The most frequently used instruments in BS, such as the M-A QoLQ (Bariatric Analysis and Reporting Outcome System), the GIQLI, M-A QoLQII, EQ-5D and the IWQoL-Lite, lack adequate and methodologically good validation data (12,52). Their methodological quality was mostly *poor* across all measurement properties. The GIQLI lacks domains relevant to BS and BCS in particular. Adequate validation studies are needed to clarify the quality of these instruments before they can be recommended for use in clinical trials and prospective studies. Of the QoL instruments placed in category B, this review suggests that the BODY-Q has the most potential to be recommended in the future depending on the results of further validation studies. The BODY-Q intended for use in both BS and BCS was supported by positive evidence of internal consistency, reliability and adequate structural validity. Compared with the other instruments, the BODY-Q is unique in the application of a modern psychometric approach, Rasch Measurement Theory (RMT) analysis, in the development of the measurement instruments. RMT analysis provides more sophisticated information than the traditional approach, the classical test theory, and offers a major contribution to the concept of reliability. In the classical test theory approach, scale (IWQOL-Lite) or item (M-A QoLQ) scores can be added to create a total score for a measurement instrument. There is no evidence, however, that the summed total scores allow for meaningful interpretation of scores (53,54). This approach limits its information to identify whether or not treatment effects are influenced by some scales or items and not others (53,54). Most importantly, the BODY-Q showed excellent content validity, as the development of the items of the BODY-Q was based on a literature review, patient interviews, cognitive patient interviews and input from experts. The BODY-Q is useful for the target population (BS and BCS), particularly with the growing field of BCS in mind. A disadvantage of the BODY-Q is that not many translations are already available and the long completion time of 138 items. Nevertheless, the RMT approach provides the opportunity for Computer Adaptive Testing, which can reduce the length of the measurement instrument. Further validation studies of the BODY-Q should focus on measurement error, construct validity and interpretability.

This is the first systematic review to make recommendations based on quality standards and criteria; previous systematic reviews reported only on measurement properties (11–14). The COSMIN checklist, the levels of evidence and the four degrees of recommendations were used as quality standards and criteria. We applied two sensitive and validated search filters and used predefined eligibility criteria to identify all PROMs in BS and BCS. In the present study, we were as inclusive as possible to give a comprehensive overview of all PROMs in BS and BCS (e.g. we included the IWQoL-Lite in our review, even though the measurement properties were only described in a systematic review). On the other hand, we were strict in excluding PROMs validated solely in patients with obesity – not BS or BCS – such as the Obesity and Weight-Loss Quality of Life and the Weight-Related Symptoms measures. We gathered information on interpretability and generalizability. At least two reviewers performed all steps in our systematic review with frequent discussions to resolve conflicts.

We have used the COSMIN checklist, published in 2010, to evaluate the methodological quality of the studies. Some of the PROMs may be of higher quality than indicated by the COSMIN checklist simply because the studies were performed longer ago and measurement properties were not reported. Almost all instruments developed after publication of the COSMIN checklist performed best, which could be due to more strict standards on how to perform studies on measurement properties. Only the Post Bariatric Outcome

Tool (developed in 2014) is minimally validated; however, they described in their discussion some measurement properties that still have to be evaluated in further validation studies. Furthermore, the most important components for QoL in BS and BCS have yet to be established internationally and therefore could not be integrated into our recommendations. Additionally, in the COSMIN checklist, the benchmark for missing data is weighted in almost all measurement properties. Studies that did not clearly report how missing data were handled were rated as 'fair'. We felt this would outweigh the overall quality of the studies, and, although we strongly recommend developers to report on missing data, we therefore decided to deviate from this guidance. We rated the overall quality of the study as *good*, if only the item 'description of how missing items were handled' was rated as inadequate.

In this systematic review, we identified validation evidence for PROMs used in BS and BCS and evaluated the methodological quality of all outcome measure validity studies using the COSMIN checklist. The final step of how QoL should be measured is the development of an evidence-based consensus over the preferred PROMs for measuring QoL in BS and BCS. We propose an international and multi-professional consensus meeting to define generic recommendations on the selection of PROMs for QoL in BS and BCS. We aim to achieve global consensus over the key components of QoL and the preferred PROMs to capture this information. The consensus meeting will produce guidelines for researchers undertaking BS and BCS research and for clinicians considering incorporation of QoL assessment in the evaluation of treatments for severe and complex obesity. A secondary goal is to stimulate further discussion and research ideas in the interpretation of QoL assessment in BS and BCS.

Improving the consistency of reporting QoL will reduce heterogeneity between trials and outcome reporting bias, which will improve the quality of data to undertake meta-analyses and inform clinical decision-making. Additionally, it may not be feasible for individual clinical trials to perform all steps in the selection of PROMs. This underlines the importance of worldwide consensus on the selection of PROMs for QoL in bariatric and metabolic surgery.

Conflict of interest statement

No conflict of interest was declared.

Acknowledgements

This work was undertaken with the support of the MRC ConDuCT-II Hub (Collaboration and innovation for Difficult and Complex randomised controlled Trials In Invasive procedures – MR/K025643/1) for Trials Methodology Research. J. M. B. is an NIHR Senior Investigator.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article. https://doi.org/10.1111/obr.12710

Data S1 Search strategy 'Recommendations on the most suitable quality-of-life measurement instruments for bariatric and body contouring surgery: a systematic review.'

Data S2 *Figure 1*: Flow diagram of identification and selection process (according to the PRISMA statement)
Data S3. Interpretability of the included measurement instruments
Data S4. Rating of measurement properties of quality of life outcome instruments for bariatric and body contouring surgery, and assessment of the methodological quality of the included studies, pertaining to "Recommendations of quality-of-life measurement instruments for bariatric and body contouring surgery: a systematic review."

References

1. Ng M, Fleming T, Robinson M *et al*. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2014; **384**(9945): 766–781.
2. Maggard MA, Shugarman LR, Suttorp M *et al*. Meta-analysis: surgical treatment of obesity. *Ann Intern Med* 2005; **142**(7): 547–559.
3. Coulman KD, Howes N, Hopkins J *et al*. A comparison of health professionals' and patients' views of the importance of outcomes of bariatric surgery. *Obes Surg* 2016; **26**(11): 2738–2746.
4. Coulman KD, Hopkins J, Brookes ST *et al*. A Core Outcome Set for the benefits and adverse events of bariatric and metabolic surgery: the BARIACT project. *PLoS Med* 2016; **13**(11): e1002187.
5. Brethauer SA, Kim J, el Chaar M *et al*. Standardized outcomes reporting in metabolic and bariatric surgery. *Surg Obes Relat Dis* 2015; **11**(3): 489–506.
6. Song P, Patel NB, Gunther S *et al*. Body image & quality of life: changes with gastric bypass and body contouring. *Ann Plast Surg* 2016; **76** Suppl 3: S216–S221.
7. van der Beek ES, Geenen R, de Heer FA, van der Molen AB, van RB. Quality of life long-term after body contouring surgery following bariatric surgery: sustained improvement after 7 years. *Plast Reconstr Surg* 2012; **130**(5): 1133–1139.
8. Song AY, Rubin JP, Thomas V, Dudas JR, Marra KG, Fernstrom MH. Body image and quality of life in post massive weight loss body contouring patients. *Obesity (Silver Spring)* 2006; **14**(9): 1626–1636.
9. Modarressi A, Balague N, Huber O, Chilcott M, Pittet-Cuenod B. Plastic surgery after gastric bypass improves long-term quality of life. *Obes Surg* 2013; **23**(1): 24–30.
10. Kitzinger HB, Abayev S, Pittermann A *et al*. The prevalence of body contouring surgery after gastric bypass surgery. *Obes Surg* 2012; **22**(1): 8–12.
11. Raaijmakers LCH, Pouwels S, Thomassen SEM, Nienhuijs SW. Quality of life and bariatric surgery: a systematic review of short- and long-term results and comparison with community norms. *Eur J Clin Nutr* 2016.

12. Coulman KD, Abdelrahman T, Owen-Smith A, Andrews RC, Welbourn R, Blazeby JM. Patient-reported outcomes in bariatric surgery: a systematic review of standards of reporting. *Obes Rev* 2013; **14**(9): 707–720.
13. Hachem A, Brennan L. Quality of life outcomes of bariatric surgery: a systematic review. *Obes Surg* 2016; **26**(2): 395–409.
14. Jumbe S, Bartlett C, Jumbe SL, Meyrick J. The effectiveness of bariatric surgery on long term psychosocial quality of life – a systematic review. *Obes Res Clin Pract* 2016; **10**(3): 225–242.
15. Mokkink LB, Terwee CB, Patrick DL *et al*. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010; **19**(4): 539–549.
16. Moher D, Shamseer L, Clarke M *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; **4**(1).
17. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009; **18**(8): 1115–1123.
18. Mokkink LB, Terwee CB, Patrick DL *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; **63**(7): 737–745.
19. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012; **21**(4): 651–657.
20. Terwee CB, Bot SD, de Boer MR *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; **60**(1): 34–42.
21. Prinsen CA, Vohra S, Rose MR *et al*. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials* 2016; **17**(1): 449.
22. Methodology P. Patient Reported Outcomes Measurement Information System (PROMIS) Standards 2015.. Available at: http://wwwnihpromisorg/science/methodology (last accessed 14/07/2015). 2015.
23. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. *Qual Life Res* 2012; **21**(4): 659–670.
24. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. *J Rheumatol* 1998; **25**(2): 198–199.
25. Schmitt J, Langan S, Deckert S *et al*. Assessment of clinical signs of atopic dermatitis: a systematic review and recommendation. *J Allergy Clin Immunol* 2013; **132**(6): 1337–1347.
26. Danilla S, Cuevas P, Aedo S *et al*. Introducing the Body-QoL(R): a new patient-reported outcome instrument for measuring body satisfaction-related quality of life in aesthetic and post-bariatric body contouring patients. *Aesthet Plast Surg* 2016; **40**(1): 19–29.
27. Danilla S, Dominguez C, Cuevas P *et al*. The Body-QoL((R)): patient reported outcomes in body contouring surgery patients [corrected]. *Aesthet Plast Surg* 2014; **38**(3): 575–583.
28. Klassen AF, Cano SJ, Scott A, Tsangaris E, Pusic AL. Assessing outcomes in body contouring. *Clin Plast Surg* 2014; **41**(4): 645–654.
29. Klassen AF, Cano SJ, Alderman A *et al*. The BODY-Q: a patient-reported outcome instrument for weight loss and body contouring treatments. *Plastic and reconstructive surgery Global open* 2016; **4**(4): e679.
30. Klassen AF, Cano SJ, Kaur M, Breitkopf T, Pusic AL. Further psychometric validation of the BODY-Q: ability to detect change following bariatric surgery weight gain and loss. *Health Qual Life Outcomes* 2017; **15**(1).
31. Oria HE, Moorehead MK. Bariatric analysis and reporting outcome system (BAROS). *Obes Surg* 1998; **8**(5): 487–499.
32. Tayyem RM, Atkinson JM, Martin CR. Development and validation of a new bariatric-specific health-related quality of life instrument "bariatric and obesity-specific survey (BOSS)". *J Postgrad Med* 2014; **60**(4): 357–361.
33. Therrien F, Marceau P, Turgeon N, Biron S, Richard D, Lacasse Y. The laval questionnaire: a new instrument to measure quality of life in morbid obesity. *Health Qual Life Outcomes* 2011; **9**: 66.
34. Weiner S, Sauerland S, Weiner R, Cyzewski M, Brandt J, Neugebauer E. Validation of the adapted Bariatric Quality of Life Index (BQL) in a prospective study in 446 bariatric patients as one-factor model. *Obes Facts* 2009; **2**(Suppl 1): 63–66.
35. Weiner S, Sauerland S, Fein M, Blanco R, Pomhoff I, Weiner RA. The bariatric quality of life index: a measure of well-being in obesity surgery patients. *Obes Surg* 2005; **15**(4): 538–545.
36. Poves PI, Macias GJ, Cabrera FM, Situ L, Ballesta LC. Quality of life in morbid obesity. *Rev Esp Enferm Dig* 2005; **97**(3): 187–195.
37. Forhan M, Vrkljan B, MacDermid J. A systematic review of the quality of psychometric evidence supporting the use of an obesity-specific quality of life measure for use with persons who have class III obesity: diagnostic in obesity and complications. *Obes Rev* 2010; **11**(3): 222–228.
38. Moorehead MK, Ardelt-Gattinger E, Lechner H, Oria HE. The validation of the Moorehead-Ardelt quality of life questionnaire II. *Obes Surg* 2003; **13**(5): 684–692.
39. Al-Hadithy N, Welbourn R, Aditya H, Stewart K, Soldin M. A preliminary report on the development of a validated tool for measuring psychosocial outcomes for massive weight loss patients. *Journal of plastic, reconstructive & aesthetic surgery : JPRAS* 2014; **67**(11): 1523–1531.
40. Karlsson J, Taft C, Sjostrom L, Torgerson JS, Sullivan M. Psychosocial functioning in the obese before and after weight reduction: construct validity and responsiveness of the Obesity-related Problems scale. *Int J Obes Relat Metab Disord* 2003; **27**(5): 617–630.
41. Muller A, Crosby RD, Selle J *et al*. Development and Evaluation of the Quality of Life for Obesity Surgery (QOLOS) Questionnaire. *Obes Surg* 2017.
42. Fermont JM, Blazeby JM, Rogers CA, Wordsworth S. The EQ-5D-5L is a valid approach to measure health related quality of life in patients undergoing bariatric surgery. *PLoS One* 2017; **12**(12).
43. Poulsen L, Klassen A, Rose M *et al*. Psychometric validation of the BODY-Q in Danish patients undergoing weight loss and body contouring surgery. *Plastic and reconstructive surgery Global open* 2017; **5**(10): e1529.
44. Poulsen L, Rose M, Klassen A, Roessler KK, Sorensen JA. Danish translation and linguistic validation of the BODY-Q: a description of the process. *Eur J Plast Surg* 2017; **40**(1): 29–38.
45. Aasprang A, Andersen JR, Vage V, Kolotkin RL, Natvig GK. Psychosocial functioning before and after surgical treatment for morbid obesity: reliability and validation of the Norwegian version of obesity-related problem scale. *PeerJ* 2015; **3**: e1275.
46. Andres A, Saldana C, Mesa J, Lecube A. Psychometric evaluation of the IWQOL-Lite (Spanish version) when applied to a sample of obese patients awaiting bariatric surgery. *Obes Surg* 2012; **22**(5): 802–809.

47. Bilbao A, Mar J, Mar B, Arrospide A, Martinez de AG, Quintana JM. Validation of the Spanish translation of the questionnaire for the obesity-related problems scale. *Obes Surg* 2009; **19**(10): 1393–1400.
48. Charalampakis V, Daskalakis M, Bertsias G, Papadakis JA, Melissas J. Validation of the Greek translation of the obesity-specific Moorehead-Ardelt quality-of-life questionnaire II. *Obes Surg* 2012; **22**(5): 690–696.
49. Lee YJ, Song HJ, Heo Y *et al*. Validation of the Korean version Moorehead-Ardelt quality of life questionnaire II. *Ann Surg Treat Res* 2014; **87**(5): 265–272.
50. Maciel J, Infante P, Ribeiro S *et al*. Translation, adaptation and validation of a Portuguese version of the Moorehead-Ardelt Quality of Life Questionnaire II. *Obes Surg* 2014; **24**(11): 1940–1946.
51. Sauerland S, Weiner S, Hausler E *et al*. Validity of the Czech, German, Italian, and Spanish version of the Moorehead-Ardelt II questionnaire in patients with morbid obesity. *Obes Facts* 2009; **2**(Suppl 1): 57–62.
52. Tayyem R, Ali A, Atkinson J, Martin CR. Analysis of health-related quality-of-life instruments measuring the impact of bariatric surgery: systematic review of the instruments used and their content validity. *The patient* 2011; **4**(2): 73–87.
53. Cano SJ, Hobart JC. The problem with health measurement. *Patient Prefer Adherence* 2011; **5**: 279–290.
54. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009; **13**(12):iii, ix-x,): 1–177.