Validation of the SF-36 as a measure of postoperative recovery after colorectal surgery

Ioana Antonescu · Francesco Carli · Nancy E. Mayo · Liane S. Feldman

Received: 14 January 2014 / Accepted: 19 April 2014 / Published online: 31 May 2014

Abstract

Introduction Surgery is evolving, and new techniques are introduced to improve "recovery." Postoperative recovery is complex, and evaluating the effectiveness of surgical innovations requires assessment of patient-reported outcomes. The Short-Form-36 (SF-36), a generic health-related quality of life questionnaire, is the most commonly used instrument in this context. The objective of this study was to contribute evidence for the validity of the SF-36 as a metric of postoperative recovery.

Methods Data from 128 patients undergoing planned colorectal surgery at one university hospital between 2005 and 2010 were analyzed. In the absence of a gold standard, the responsiveness and construct validity (known groups and convergent) of the SF-36 were evaluated. Standardized response means were computed for the former and non-parametric tests were used to assess the statistical significance of the changes observed. Multiple linear regression was used to determine whether the SF-36 discriminates between patients with versus without complications and between laparoscopic and open surgery (known groups); correlations between the SF-36 and the 6-min walk test, a measure of functional walking capacity (convergent) was investigated with Spearman's rank correlation.

Results The SF-36 was sensitive to clinically important changes. Scores on six of eight domains and the physical component summary score deteriorated postoperatively (SRM 0.86 for the PCS, $p < 0.01$) and improved to baseline thereafter. Patients with complications had significantly lower scores on five SF-36 domains (with differences from −9 (−18, −1), $p = 0.04$ to −18 (−32, −2), $p = 0.03$), and scores on all subscales were lower than those in a healthy population ($p < 0.01$ to $p = 0.04$). The SF-36 did not differentiate between laparoscopic and open surgery. Physical functioning scores correlated with 6MWT distance at 1 and 2 months (Spearman's $r = 0.31$ and 0.36, $p < 0.01$).

Conclusions The SF-36 is responsive to expected physiological changes in the postoperative period, demonstrates construct validity, and thus constitutes a valid measure of postoperative recovery after planned colorectal surgery. The SF-36 did not, however, discriminate between recovery after laparoscopic and open surgery.

Keywords SF-36 · Responsiveness · Validity evidence · Postoperative recovery · Colorectal surgery

Surgery is evolving. The number of procedures is increasing, while hospital stay and the number and severity of complications are decreasing [1]. Patients' expectations have shifted from merely surviving the operation with manageable complications to recovering their quality of

Presented as a poster at SAGES April 2014, Salt Lake City (Utah).

I. Antonescu · L. S. Feldman (✉)
Division of General Surgery, Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, 1650 Cedar Ave, L9.300, Montreal,
QC H3G 1A4, Canada
e-mail: liane.feldman@mcgill.ca

I. Antonescu
e-mail: ioana.antonescu@mail.mcgill.ca

F. Carli
Department of Anesthesia, McGill University Health Centre, Montreal, QC, Canada

N. E. Mayo
Division of Clinical Epidemiology, McGill University Health Centre, Montreal, QC, Canada

life (QoL) and returning to their baseline level of functioning [2, 3]. Much of this evolution is the result of surgical innovation, including the widespread adoption of laparoscopic surgery [4, 5], the emergence of robotic surgery [6], and the introduction of surgical checklists and care pathways [7, 8]. Nearly all new techniques and processes of care are advocated on the basis that they "improve recovery." Yet postoperative recovery is a poorly defined and even less well-measured construct [9]. A rapid but transient deterioration in physical capacity is expected immediately after surgery, followed by a more gradual return towards and occasionally beyond baseline [10]. Though this anticipated trajectory is defined, there is no single instrument that has been validated as the gold standard measure of recovery after abdominal surgery [11]. This may reflect the fact that postoperative recovery is in fact a complex and multi-dimensional construct that requires assessment of several interrelated and increasingly complex dimensions [12, 13].

Health-related quality of life (HRQL) instruments are frequently used in research in an effort to capture this complexity and to operationalize the construct of recovery. The Short-Form-36 (SF-36) is one of the most common tools used to measure postoperative recovery [14]. The SF-36 is a generic HRQL questionnaire designed in the early 1990s by the RAND Corporation, and abundant evidence exists to support its validity in a variety of medical contexts [15, 16]. There is a willingness to extrapolate the available validity evidence from the medical to the surgical context. For example, guidelines recommend the use of the SF-36 to measure QoL after laparoscopic surgery [17]. Yet this practice is questionable, and the choice of which instrument to use in a trial of comparative effectiveness as well as the interpretation of the results obtained should be based on the measure's psychometric properties specifically in the context of interest [18–22]. It should not be assumed that because an instrument is valid in one context (e.g., asthma) it would have similar properties in another context (e.g., recovery after surgery).

The objective of this study was therefore to contribute evidence for the longitudinal (sensitivity to change) and construct (cross-sectional convergent and known groups) validity of the SF-36 as an indicator of postoperative recovery in patients undergoing planned colorectal surgery.

Materials and methods

Participants and setting

Data collected prospectively within the frame of two separate studies approved by the Institutional Research Ethics Board (ethics approval codes REC#02-053 and GEN06-023) and previously reported were used [23–25]. The study sample thus consisted of adult patients scheduled to undergo colorectal surgery at one university-affiliated teaching hospital in 2005–2006 or 2009–2010. Exclusion criteria within the frame of these studies were: the presence of a psychiatric condition significantly limiting the patients' ability to understand and complete the SF-36, baseline mobility restricted by a pre-existing condition, metastatic cancer, contraindications for neuraxial anesthesia, and chronic opioid use. Eligible patients were approached by a research assistant at the time of their visit to the preoperative clinic, at which point written informed consent was obtained. Participants were evaluated 1 week preoperatively and at 1 then 2 months postoperatively. At each of these times, they completed the SF-36 and their walking capacity was assessed with the 6-min walk test (6MWT). Baseline demographic characteristics were recorded, and data were also collected on intra- and post-operative parameters, including the occurrence and severity of complications. For the purpose of this validation study, complete case analysis was performed, resulting in the analysis of a subgroup of patients from the combined original datasets.

Measures

The SF-36 (http://www.rand.org/health/surveys_tools/mos/mos_core_36item_survey.html) is a generic self-reported HRQL questionnaire that defines and evaluates three principal health attributes namely functional status, well-being, and general health perceptions and overall QoL [15, 16]. The SF-36 is an instrument used to measure patient-reported outcomes, and will interchangeably be called a questionnaire, an instrument, a measure, or a profile throughout the text. It was developed by the RAND Corporation within the context of the Medical Outcomes Study [26]. This was a 2-year prospective observational study investigating determinants of health outcomes in patients with chronic disease and/or depression and the impact of the health care system on these outcomes. The SF-36 consists of 35 individual questions (items) divided into eight subscales that represent eight domains of health. An additional independent health transition item is present but not included in the scoring algorithm. Patients answer each question using an ordinal scale (0–3 or 0–6, depending on the question). These numerical answers are then recoded according to a pre-specified algorithm to yield scores ranging from 0 (worst health state) to 100 (best health state). The scores on items pertaining to the same dimension are then aggregated to generate a score for each of the eight domains of health (physical functioning, role physical, pain, social functioning, role emotional, vitality, mental health, and general health perception). These subscale-specific scores are then further combined to produce a physical and a mental component summary score (PCS and MCS) [15].

The 6MWT is a test of performance that evaluates patients' fitness to sustain an intermediate level of physical activity for a given period of time. Patients are invited to walk along a hospital hallway for 6 min at a pace that should tire them by the end of the 6 min, and the distance covered (in meters) is recorded [27]. This level of fitness is reflective of patients' ability to perform more strenuous activities of daily living [28]. The 6MWT has been validated as a measure of postoperative recovery after colorectal surgery [25].

Validity evidence

In the context of measurement, validity (or construct validity) is the extent to which a given instrument actually measures what it is intended to measure (the relevant construct). Validity is not absolute, but rather depends on the intended use of the instrument as well as on the target population. In assessing outcomes of an intervention for example, longitudinal validity, or a tool's sensitivity to expected clinically important changes over time, is also of critical importance. A valid measure of postoperative recovery should reflect the anticipated trajectory of initial deterioration followed by improvement that occurs after an operation. Construct validity, defined above, can be divided into cross-sectional convergent and known-groups validity. The former is the degree of correlation between scores on the instrument and another measure of the same construct. In this case, determining whether scores on the physical functioning domain of the SF-36 are correlated with the distance covered in 6 min would be appropriate. Cross-sectional convergent validity is particularly relevant in the absence of a gold standard metric of postoperative recovery against which to compare the SF-36. Finally, establishing known-groups validity involves determining whether the instrument behaves in a predictable way, allowing differentiation between groups that are expected to be different on substantive grounds. This includes patients with versus without complications, patients versus healthy individuals, and patients undergoing open versus laparoscopic surgery [29].

Our a priori hypotheses for evaluating longitudinal and construct validity were as follows: (1) Longitudinal validity: Scores on selected domains of health will decline significantly from baseline to 3–5 weeks, and improve near baseline at 8–9 weeks. (2) Construct—cross-sectional convergent: At each assessment time, scores on the physical functioning domain of the SF-36 will correlate with scores on the 6MWT. (3) Construct—known groups: Scores on selected domains of health will be lower at 3–5 weeks in patients with complications compared to patients without complications; scores on selected domains of health will be lower at 3–5 weeks in patients compared to a healthy population (Canadian norms [30]); and scores on selected domains of health will be lower at 3–5 weeks in patients undergoing open compared to those having laparoscopic surgery.

Statistical analyses

Standardized response means (SRM), defined as the change in scores divided by the standard deviation of this change, were calculated to determine the evolution of scores on subscales of the SF-36 over time (longitudinal validity). Values between 0.5 and 0.8 are considered moderate, and the sign of the SRM reflects the direction of change. The Wilcoxon signed-rank test was used for significance testing. The magnitude of the change was also considered in relation to the context-specific minimal clinically important difference (MCID) for each of the eight domains of health, which represents the smallest change in an outcome measure that would influence patient management [31, 32]. In previous work, we estimated the MCID for domains of health of the SF-36 to range between 8 (6–9) and 15 (12–18) and between 15 (12–19) and 32 (28–36) points (on a scale of 0–100), depending on the patient's baseline level of functioning [33]. Spearman's rank correlation was used to test the cross-sectional convergent validity hypothesis. Known-groups validity was investigated by determining the effect of complications on domain-specific scores, adjusting for age, gender, American Society of Anesthesiologists grade (ASA), and laparoscopic approach. The one sample test for the median was used to compare patients' scores to Canadian norms. Fisher's exact test was used to compare the proportion of patients having returned to baseline among individuals with versus without complications.

Allowing a 5 % probability of committing a type I error ($\alpha = 0.05$), 126 patients would have been required to detect a MCID of 15 points with 80 % power.

Statistical significance was defined a priori as $p < 0.05$. Data analysis was conducted using the statistical program STATA (Version 11.2, StataCorp, College Station, TX, USA). Results are presented as mean (95 % confidence interval), median [25th; 75th percentile], and *n* (%) where appropriate.

Results

A total of 128 (of 194 available) patients with data at all three time points were included in the analysis. After generating missing PCS and MCS scores using accepted algorithms, 66 patients (34 % of the original sample) were

Table 1 Patient and operative characteristics of the study sample

	$n = 128$
Age (years)	63 (52; 73)
Male/female	76 (59 %)/52 (41 %)
BMI (kg/m^2)	27 (20, 37)
ASA	
I	22 (17 %)
II	86 (67 %)
III	20 (16 %)
Surgery	
Segmental colectomy	53 (41 %)
Sigmoid resection	13 (10 %)
Low anterior resection	41 (32 %)
Abdomino-perineal resection	8 (6 %)
Proctectomy	11 (9 %)
Small bowel resection	2 (2 %)
Stoma creation (yes/no)	28 (22 %)/100 (78 %)
Laparoscopic approach (yes/no)	69 (54 %)/59 (46 %)
Complication grade [34]	
None	77 (60 %)
I	14 (11 %)
II	14 (11 %)
IIIa	5 (4 %)
IIIb	2 (2 %)
IV	2 (2 %)
Missing	14 (11 %)
Length of stay (days)	4 (3, 6)

Results are presented as *n* (%), median (25th; 75th percentile), mean (95 % CI)

excluded from the complete case analysis. The demographic and operative characteristics were similar between the included and excluded patients. Baseline characteristics of the study sample are presented in Table 1. Patients were mostly men, with a median age of 63 years old [52; 73], mildly overweight with a mean BMI of 27 kg/m^2 [20, 37], and in relatively good health (with 84 % having ASA I or II). They had predominantly undergone segmental colectomies (41 %) or low anterior resections (32 %). Less than 25 % of patients had received a stoma, and a laparoscopic approach was used in 54 % of all operations. Follow-up clinic visits occurred between 3 and 5 weeks and between 8 and 9 weeks, primarily as a consequence of scheduling conflicts.

Longitudinal validity

Compared with baseline, scores on six of the eight domains of health (physical functioning, role physical, pain, social functioning, role emotional, and vitality) and the PCS had decreased significantly by the first postoperative appointment ($p < 0.01$). Decreases in scores on these domains ranged from −7 (−11, −3) to −42 (−52, −32). The same six domains of health had subsequently improved significantly between the first and second postoperative visits ($p < 0.01$), with changes ranging between +6 (2, 9) and +32 (24, 40). Moreover, for the physical functioning, role physical, pain, and social functioning domains, which represent biophysical constructs, changes in scores were equal to or greater than the corresponding MCID, both during the deterioration and recovery phases. SRMs for deterioration were between 0.26 and 0.86 (small to large); those for recovery were between 0.32 and 0.79 (small to moderate).

The change from baseline to 1 month was minimal and not significant for the mental health ($p = 0.99$) and general health perception ($p = 0.13$) subscales and for the MCS ($p = 0.53$).

At 2 months, patients had mostly recovered to baseline, but continued to report some limitations on the role physical subscale, with scores 10 points (0.34, 19) below baseline.

Median scores illustrating this trajectory and changes over time are presented in Table 2.

Defining “return to baseline” as a score within 10 % of the baseline score, the percentage of patients that were below baseline, at baseline, or above baseline is presented in Table 3 for each subscale of the SF-36 at 1 and 2 months. The analysis was repeated using Canadian norms rather than baseline with no substantive changes in the results obtained.

Construct validity: cross-sectional convergent

Significant positive correlations were found between physical functioning scores and 6MWT distance at 1 and 2 months (Spearman’s $r = 0.31$ and 0.36, respectively, $p < 0.01$). This correlation only approached significance 1 week pre-operatively (Spearman’s $r = 0.16$, $p = 0.07$). A positive correlation was also identified between the PCS and the 6MWT distance at 1 month (Spearman’s $r = 0.22$, $p = 0.02$). Small non-significant correlations were observed between the seven other subscales and 6MWT distance.

Within each domain of health, higher baseline scores were found to correlate with higher subsequent scores (Spearman’s r between 0.20 and 0.66, $p < 0.01$). This was also true for both the physical and mental PCS and MCS (Spearman’s r between 0.29 and 0.55, $p < 0.01$).

Construct validity: known groups

Twenty-eight patients experienced Clavien-Dindo I or II [34] complications in the postoperative period, with an

Table 2 Perioperative changes in scores on measures used to evaluate recovery after colorectal surgery

	Preoperatively	3–5 weeks postoperatively	8–9 weeks postoperatively
SF-36[a]			
Physical function	90 (70; 95)	75 (50; 85)	100 (0; 100)
Δ vs. baseline[b]		−15 (−20, −10)*	−4 (−9, 1)
Δ vs. 3–5 weeks[b]			10 (6, 14)*
Role physical	100 (0; 100)	0 (0; 25)	75 (0; 100)
Δ vs. baseline		−42 (−52, −32)*	−10 (−19, −0.34)[c]
Δ vs. 3–5 weeks			32 (24, 40)*
Bodily pain	78 (61; 100)	55 (41; 80)	84 (55; 100)
Δ vs. baseline		−17 (−22, −11)*	3 (−3, 9)
Δ vs. 3–5 weeks			18 (13, 22)*
Social function	75 (63; 100)	63 (38; 88)	88 (63; 100)
Δ vs. baseline		−14 (−20, −7)*	2 (−4, 8)
Δ vs. 3–5 weeks			15 (10, 20)*
Role emotional	100 (33; 100)	67 (0; 100)	100 (0; 100)
Δ vs. baseline		−14 (−23, −4)*	3 (−6, 13)
Δ vs. 3–5 weeks			16 (7, 25)*
Vitality	60 (40; 75)	50 (35; 65)	55 (40; 75)
Δ vs. baseline		−7 (−11, −3)*	−1 (−6, 4)
Δ vs. 3–5 weeks			6 (2, 9)*
Mental health	76 (60; 88)	76 (60; 92)	80 (64; 92)
Δ vs. baseline		−0.08 (−3, 3)	5 (1, 8)
Δ vs. 3–5 weeks			4 (1, 6)
General health perception	70 (55; 82)	72 (52; 85)	72 (55; 87)
Δ vs. baseline		2 (−1, 5)	4 (−0.14, 8)
Δ vs. 3–5 weeks			2 (−0.20, 5)
PCS	48 (42; 54)	41 (31, 45)	48 (40; 54)
Δ vs. baseline		−9 (−11, −7)*	−2 (−4, 0.33)
Δ vs. 3–5 weeks			7 (5, 8)*
MCS	49 (34; 57)	47 (35; 56)	50 (36; 57)
Δ vs. baseline		−0.21 (−2, 2)	3 (1, 5)
Δ vs. 3–5 weeks			3 (1, 5)
6MWT[a]			
	511 (485, 536)	461 (433, 490)	486 (457, 515)
Δ vs. baseline		−30 (−40, −19)*	−27 (−43, −11)[c]
Δ vs. 3–5 weeks			25 (11, 40)*

Data expressed as median (25th; 75th percentile) for the SF-36 and as mean (95 % CI) for the 6MWT

[a] SF-36 subscales scored on a scale from 0 to 100, with higher scores indicating better HRQL; 6MWT distance recorded in meters

[b] Mean difference (95 % CI)

[c] Score (0–100) or distance walked (meters) still below baseline at 8–9 weeks

* Significant difference ($p < 0.01$)

additional nine experiencing grade III and higher complications.

Baseline scores on seven of the SF-36 domains did not differ between patients with versus without complications (p values from 0.08 to 0.90). Baseline PCS and MCS were also similar between the two groups ($p = 0.14$ for both PCS and MCS). The only difference was found in baseline general health perception, with lower scores in patients who subsequently developed a complication (−10 (−18, −2), $p = 0.01$).

At 1 month and after adjusting for age, gender, ASA class, and laparoscopic approach, scores on all eight subscales were lower in patients with complications by 7–18 points. Although this difference was not statistically significant for three of the eight domains, the lower bound of the 95 % confidence interval was nevertheless highly suggestive of a possible clinically relevant negative effect of complications (Table 4). In addition, a greater proportion of patients without complications had recovered to baseline at 2 months when compared to patients with complications. This was statistically significant for three of the four biophysical domains (role physical 47 vs. 26 %, $p = 0.03$; pain 74 vs. 45 %, $p < 0.01$; social functioning 81 vs. 45 %, $p < 0.01$; physical functioning 62 vs. 45 %, $p = 0.07$) as well as for the PCS (73 vs. 50 %, $p = 0.02$).

Table 3 Percentage of patients who had returned to baseline at each assessment time

	At 3–5 weeks (%)	At 8–9 weeks (%)	p[c]
SF-36			
Physical function			
Below (81)[a]	66	47	
At baseline ±10 %	25	34	0.041
Above (99)[b]	8	19	
Role physical			
Below (90)	85	61	
At baseline ±10 %	15	39	0.358
Above (110)	–	–	
Bodily pain			
Below (70)	64	37	
At baseline ±10 %	12	9	<0.001
Above (85)	24	54	
Social function			
Below (68)	54	32	
At baseline ±10 %	11	15	0.002
Above (85)	35	54	
Role emotional			
Below (90)	50	34	
At baseline ±10 %	50	66	0.433
Above (110)	–	–	
Vitality			
Below (54)	58	40	
At baseline ±10 %	19	24	0.019
Above (66)	23	37	
Mental health			
Below (68)	39	29	
At baseline ±10 %	17	23	0.269
Above (84)	44	48	
GHP			
Below (63)	32	33	
At baseline ±10 %	23	24	0.554
Above (77)	39	39	
PCS			
Below (45)	74	37	
At baseline ±10 %	22	46	0.007
Above (55)	4	22	
MCS			
Below (45)	45	31	
At baseline ±10 %	23	30	0.508
Above (55)	32	39	
6MWT			
Below (460)	68	43	
At baseline ±10 %	27	38	<0.001
Above (562)	6	19	

[a] Score in parentheses is baseline score −10 %

[b] Score in parentheses is baseline score +10 %

[c] Where the p value corresponds to the comparison between the proportion of patients above versus below baseline at 3–5 weeks and the proportion of patients above versus below baseline at 8–9 weeks

Table 4 Scores on the SF-36 domains at 3–5 weeks postoperatively in patients without complications, and differences in scores in patients with complications

	Score in patients without complications (95 % CI)[a]	Differences in scores in patients with complications (95 % CI)[b]	p
SF-36			
Physical function	69 (68, 70)	−10 (−22, 2)	0.092
Role physical	27 (26, 29)	−18 (−33, −2)	0.027
Bodily pain	61 (60, 62)	−7 (−18, 4)	0.195
Social function	65 (65, 66)	−14 (−26, −1)	0.034
Role emotional	58 (57, 60)	−14 (−34, 6)	0.173
Vitality	55 (54, 56)	−14 (−23, −5)	0.003
Mental health	76 (75, 77)	−9 (−18, −1)	0.035
GHP	72 (71, 73)	−9 (−17, −1)	0.027
PCS	40 (39, 41)	−8 (−14, −2)	0.014
MCS	48 (41, 55)	−7 (−12, −2)	0.007

[a] Adjusted for age, gender, ASA, laparoscopic approach

[b] Baseline scores on each of the SF-36 domains did not differ between patients with and without complications (p values from 0.08 to 0.90), except for "general health perception": −10 (−18, −2) $p = 0.01$

After adjusting for complications, age, and gender, no significant differences were identified between patients having had laparoscopic versus open surgery, with scores in the laparoscopic group ranging from 11 points higher (−4, 26) on the role physical domain to 7 points lower (−15, 1) on the general health perception domain. Similarly, except for a marginal benefit on the role physical domain (in bold in Tables 5, 6), no significant differences were identified between a laparoscopic and an open approach among patients without complications, or among patients with no or with Clavien I or II complications. This lack of a significant difference between laparoscopic and open surgery was demonstrated at both follow-up times (Tables 5, 6).

One month after surgery, scores on all subscales were significantly lower in patients when compared to corresponding Canadian norms. The differences between patients and healthy individuals were moderate to large on six of the eight domains and on the PCS and MCS. This is shown in Table 7.

Discussion

The SF-36 is widely used in the clinical setting and in research studies to operationalize the construct of postop-

Table 5 Difference in scores on the SF-36 domains at 3–5 weeks in patients undergoing laparoscopic versus open surgery

	Difference in scores in patients undergoing lap versus open surgery (95 % CI)[a]	Difference in scores in patients without complications undergoing lap versus open surgery (95 % CI)[a]	Difference in scores in patients without or with Clavien I or II complications undergoing lap versus open surgery (95 % CI)[a]
SF-36			
Physical function	−4 (−16, 7)	−4 (−17, 8)	−4 (−15, 8)
Role physical	11 (−4, 26)	**22** (**2**, **43**)	13 (−2, 29)
Bodily pain	−1 (−11, 10)	−1 (−12, 10)	0 (−10, 10)
Social function	−5 (−17, 7)	−7 (−22, 7)	−5 (−18, 8)
Role emotional	8 (−12, 27)	17 (−6, 39)	13 (−7, 32)
Vitality	4 (−5, 13)	4 (−7, 15)	4 (−6, 14)
Mental health	−1 (−9, 8)	2 (−8, 12)	0 (−9, 9)
GHP	−7 (−15, 1)	−6 (−15, 3)	−8 (−16, 0)
PCS	−1 (12, 9)	0 (−14, 15)	1 (−11, 14)
MCS	−9 (−23, 5)	−1 (−22, 19)	−1 (−18, 16)

[a] Adjusted for age, gender, and ASA

Table 6 Difference in scores on the SF-36 domains at 8–9 weeks in patients undergoing laparoscopic versus open surgery

	Difference in scores in patients undergoing lap versus open surgery (95 % CI)[a]	Difference in scores in patients without complications undergoing lap versus open surgery (95 % CI)[a]	Difference in scores in patients without or with Clavien I or II complications undergoing lap versus open surgery (95 % CI)[a]
SF-36			
Physical function	−4 (−13, 6)	−5 (−15, 6)	−3 (−14, 7)
Role physical	**19** (**3**, **35**)	16 (−3, 36)	**17** (**0**, **35**)
Bodily pain	1 (−8, 11)	−1 (−11, 10)	2 (−8, 12)
Social function	−1 (−11, 9)	−1 (−13, 10)	0 (−11, 10)
Role emotional	12 (−5, 29)	14 (−5, 37)	11 (−7, 29)
Vitality	−1 (−10, 7)	−1 (−11, 10)	−2 (−11, 8)
Mental health	−5 (−12, 1)	−1 (−9, 6)	−5 (−12, 2)
GHP	−6 (−13, 1)	−7 (−15, 1)	−7 (−14, 0)
PCS	−3 (−16, 10)	0 (−18, 18)	3 (−11, 18)
MCS	−1 (−13, 11)	7 (−11, 25)	4 (−11, 19)

[a] Adjusted for age, gender, and ASA

erative recovery [14, 35]. Despite extensive evidence of its validity in multiple settings, including orthopedic and spine surgery [36, 37], no studies have specifically investigated its performance in the context of recovery after digestive surgery. This study contributes evidence for the longitudinal and construct (known groups and cross-sectional convergent) validity of the SF-36 as it was applied to a cohort recovering from planned colorectal surgery. The SF-36 was responsive to clinically meaningful changes and discriminated between patients and healthy individuals as well as between patients with versus without complications. Scores on the physical functioning domain correlated with the 6MWT, a measure of submaximal exercise capacity. However, it did not differentiate recovery after laparoscopic and open surgery. These findings support the use of the SF-36 as a metric of postoperative recovery, but also underscore the limitations inherent to using generic measures of HRQL in this context.

The SF-36 was responsive to physiological postoperative changes, with scores on all subscales but mental health and general health perception being significantly lower than baseline at 1 month and having returned to baseline at 2 months. At 2 months, scores on all domains were back to baseline except for role physical, which remained significantly below baseline. These findings are consistent with the substantive difference between biophysical and emotional parameters. Indeed, patients' state of mind after surgery may reflect the relief associated with "being

Table 7 Differences in scores on the SF-36 domains in patients 3–5 weeks postoperatively compared to Canadian norms

	Difference in scores compared to Canadian norms (95 % CI) [30]	|SRM|	p
SF-36			
Physical function	−17 (−22, −12)	0.60	<0.01
Role physical	−58 (−65, −51)	1.52	<0.01
Bodily pain	−16 (−21, −12)	0.63	<0.01
Social function	−26 (−32, −20)	0.84	<0.01
Role emotional	−34 (−42, −25)	0.71	<0.01
Vitality	−17 (−22, −13)	0.74	<0.01
Mental health	−6 (−10, −3)	0.31	0.025
GHP	−6 (−9, −2)	0.28	0.042
PCS	−9 (−11, −8)	1.01	<0.01
MCS	−8 (−10, −6)	0.65	<0.01

cured" or simply coming through the surgery itself, which dissipates over time as they return to their baseline functional level. Emotional domains are consequently not expected to follow the same trajectory of deterioration and improvement as physical function parameters. Previous work reports the MCID for four biophysical domains of the SF-36 [33]. It is interesting to note that the magnitude of the changes observed in the current study was equal to or exceeded these subscale-specific MCIDs, though this may in part be related to the partial overlap between the datasets used in these two studies.

As the SF-36 is a multidimensional generic HRQL questionnaire, it should not be expected that all domains would meet construct validity criteria, as the SF-36 was not designed to specifically target postoperative recovery. Physical functioning at 1 and 2 months and 1-month PCS correlated with distance walked in 6 min. Though the correlations were not strong, they support the substantively relevant hypothesis that subscales that most closely reflect physical performance would correlate with a more objective measure of the same construct. Furthermore, the SF-36 discriminated between patients and healthy individuals on all domains, as well as between patients with versus without complications on five of eight domains. Interestingly, the ability to capture the emotional burden and the impact on general health perception associated with experiencing any degree of complication supports the validity of the SF-36 as a HRQL measure. However even after appropriate adjustments, the SF-36 did not identify differences between the subgroups of patients undergoing laparoscopic versus open surgery, which is not unexpected given that this questionnaire was neither designed to measure postoperative recovery nor to capture differences between laparoscopic and open approaches. This suggests that such a generic instrument may not be optimal to detect these. This finding is unlikely to be the result of selection bias, as the cohort of patients included in this study was representative, regarding their demographic and operative characteristics, of a typical colorectal surgery population.

A large number of HRQL instruments are currently being used to evaluate patient-centered outcomes during the postoperative recovery period. Strengths of the SF-36 include its generic nature, allowing recovery to be assessed and comparisons to be made between interventions and settings, as well as its ability to capture multiple dimensions of patient outcomes by targeting eight relevant HRQL domains [13, 17]. The SF-36 is also simple to complete, either independently or as administered by an interviewer, in person or by telephone [17]. It was developed based on rigorous measurement methodology, and has since then been shown to be reliable, valid, and sensitive to change in many contexts. It has consequently been broadly translated and adapted to many cultural frameworks. Guidelines from the European Association for Endoscopic Surgery recommend the SF-36 as appropriate in a variety of contexts [17]. Yet these guidelines also highlight the fact that the validity evidence on which some recommendations are based is extrapolated from robust studies that were nonetheless not conducted in surgical populations. Even after adjusting for potential confounders that may overwhelm more subtle differences, the SF-36 fails to detect differences in instances where one might expect them, such as between laparoscopic and open colorectal resections [35, 38–41]. Although the apparent absence of a difference in a trial may have several explanations, including true equivalence, lack of power, and use of an inappropriate instrument [42, 43], generating evidence to rule out that the latter two factors are a necessary step toward the useful interpretation of research results. Thus, if researchers use the SF-36 in a surgical population and power comparative studies based on the context-specific MCIDs provided, it is presumed that true differences would be identified. This being said, if the research aims to compare recovery after laparoscopic and open colorectal resection, the SF-36 may not be sufficiently sensitive to discriminate between the two in this context.

Confirmation of the validity of the SF-36 as a measure of recovery after colorectal surgery supports its widespread use in practice, both at ours and at other institutions. Nevertheless, its inability to discriminate outcomes in patients undergoing laparoscopic versus open surgery is concerning and will steer us away from this instrument when comparing these two approaches. An awareness of this limitation is important when planning and interpreting trials, though a gold standard metric of recovery to replace the SF-36 in this context does not yet exist. Work is therefore required towards the development of such a measure, in addition to larger prospective validation studies for the SF-36, as detailed below.

Strengths and limitations

Strengths of this study include its power to detect clinically meaningful differences. Moreover, a relatively homogenous and representative patient population was used, and validity was assessed using several approaches.

Nevertheless, the principal limitation of this study is the absence of a gold standard measure of postoperative recovery. Criterion validity could consequently not be established, and surrogate validity standards had to be used. This limitation may be partially addressed, however, by the fact that sensitivity to change is perhaps the most important aspect of validity for a measure of outcomes after an intervention [29]. Further studies will be required in other surgical populations to assess the generalizability of these results.

Another limitation is the use of data collected prospectively within the frame of studies other than the current one. The inclusion and exclusion criteria as well as the timeframe for the follow-up visits were selected specifically for the purpose of these other previously published studies. Importantly, though we do not suspect that the overall validity evidence supporting the SF-36 as a measure of postoperative recovery would be affected, data at 2 weeks for example may have revealed a difference between laparoscopic and open surgery. Studies [44, 45] have shown that the benefits of laparoscopic surgery, namely decreased pain and length of stay and a faster return to work, are most pronounced in the immediate postoperative period. These differences tend to disappear over time, as patients return to their baseline functional status and activities. Thus, in designing a prospective study to assess the validity and discriminative properties of a given recovery metric, we would deliberately include a follow-up visit within 2 weeks of surgery in addition to considering slightly different inclusion and exclusion criteria. We would also schedule a follow-up visit further downstream (6 months to a year) to determine the evolution of patients' function and QoL and whether they have returned to Canadian norms or not.

These limitations underscore the need for a large prospective study specifically designed to assess the validity of the SF-36 as a measure of recovery after abdominal surgery. Such a study would be adequately powered to allow subgroup analyses (by severity of complications, for example) as well as analyses at multiple time points and comparisons between several commonly used patient-reported outcome metrics.

Conclusion

Postoperative recovery is a complex construct for which a gold standard measure is not yet available. The SF-36, a generic HRQL questionnaire, is the most widely used metric in this context. The present study provides evidence of the validity of this instrument to quantify recovery after colorectal surgery in general. It also emphasizes the importance of being aware of the psychometric properties of each instrument in the specific context in which it is used. Only when a valid measure is used in an adequately powered study can the results be interpreted as truly in favor or against the presence of a true difference in effectiveness.

Acknowledgments Dr. IA is supported by a scholarship from the Quebec Research Fund for Health Sciences (FRSQ), and the McGill Surgeon Scientist Program. The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation is supported by an unrestricted educational grant from Covidien Canada. Dr. LSF is a paid consultant for Covidien and is the recipient of an investigator-initiated research grant from Ethicon and an unrestricted educational grant from ConMed.

Disclosures Dr. IA is supported by a scholarship from the Quebec Research Fund for Health Sciences (FRSQ) and the McGill Surgeon Scientist Program, but has no conflicts of interest or financial ties to disclose. Dr. FC and Dr. NEM have no conflicts of interests or financial ties to disclose. Dr. LSF is a paid consultant for Covidien and is the recipient of an investigator-initiated research grant from Ethicon and an unrestricted educational grant from ConMed. The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation is supported by an unrestricted educational grant from Covidien Canada.

References

1. CIHI. Trends in acute inpatient hospitalizations and day surgery visits in Canada, 1995–1996 to 2005–2006. Updated January

2007; September 2012. http://www.cihi.ca/CIHI-ext-portal/pdf/internet/BUL_10JAN07_FIG1_EN

2. Cheema FN, Abraham NS, Berger DH, Albo D, Taffet GE, Naik AD (2011) Novel approaches to perioperative assessment and intervention may improve long-term outcomes after colorectal cancer resection in older adults. Ann Surg 253(5):867–874
3. Colavita PD, Tsirline VB, Belyansky I, Walters AL, Lincourt AE, Sing RF et al (2012) Prospective, long-term comparison of quality of life in laparoscopic versus open ventral hernia repair. Ann Surg 256(5):714–722 Discussion 22–3
4. Soper NJ, Brunt LM, Kerbl K (1994) Laparoscopic general surgery. N Engl J Med 330(6):409–419
5. Tsui C, Klein R, Garabrant M (2013) Minimally invasive surgery: national trends in adoption and future directions for hospital strategy. Surg Endosc 27(7):2253–2257
6. Halabi WJ, Kang CY, Jafari MD, Nguyen VQ, Carmichael JC, Mills S et al (2013) Robotic-assisted colorectal surgery in the United States: a nationwide analysis of trends and outcomes. World J Surg 37(12):2782–2790
7. Fierens J, Wolthuis AM, Penninckx F, D'Hoore A (2012) Enhanced recovery after surgery (ERAS) protocol: prospective study of outcome in colorectal surgery. Acta Chir Belg 112(5):355–358
8. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AH, Dellinger EP et al (2009) A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 360(5):491–499
9. Lee L, Tran T, Mayo NE, Carli F, Feldman LS (2014) What does it really mean to "recover from an operation"? Surgery 155(2):211–216
10. Carli F, Zavorsky GS (2005) Optimizing functional exercise capacity in the elderly surgical population. Curr Opin Clin Nutr Metab Care 8(1):23–32
11. Kluivers KB, Riphagen I, Vierhout ME, Brolmann HA, de Vet HC (2008) Systematic review on recovery specific quality-of-life instruments. Surgery 143(2):206–215
12. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL (2005) Conceptual model of health-related quality of life. J Nurs Sch 37(4):336–342
13. Wilson IB, Cleary PD (1995) Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. J Am Med Assoc 273(1):59–65
14. Velanovich V (2001) The quality of quality of life studies in general surgical journals. J Am Coll Surg 193(3):288–296
15. Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T et al (1992) Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 305(6846):160–164
16. Ware JE Jr, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30(6):473–483
17. Korolija D, Sauerland S, Wood-Dauphinee S, Abbou CC, Eypasch E, Caballero MG et al (2004) Evaluation of quality of life after laparoscopic surgery: evidence-based guidelines of the European Association for Endoscopic Surgery. Surg Endosc 18(6):879–897
18. Avery KN, Gujral S, Blazeby JM (2008) Patient-reported outcomes to evaluate surgery. Expert Rev Pharmacoecon Outcomes Res 8(1):43–50
19. Beaton DE, Boers M, Wells GA (2002) Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol 14(2):109–114
20. Crosby RD, Kolotkin RL, Williams GR (2003) Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 56(5):395–407
21. Urbach DR (2005) Measuring quality of life after surgery. Surg Innov 12(2):161–165
22. Velanovich V (2007) Using quality-of-life measurements in clinical practice. Surgery 141(2):127–133
23. Wongyingsinn M, Baldini G, Charlebois P, Liberman S, Stein B, Carli F (2011) Intravenous lidocaine versus thoracic epidural analgesia: a randomized controlled trial in patients undergoing laparoscopic colorectal surgery using an enhanced recovery program. Reg Anesth Pain Med 36(3):241–248
24. Wongyingsinn M, Baldini G, Stein B, Charlebois P, Liberman S, Carli F (2012) Spinal analgesia for laparoscopic colonic resection using an enhanced recovery after surgery programme: better analgesia, but no benefits on postoperative recovery: a randomized controlled trial. Br J Anaesth 108(5):850–856
25. Moriello C, Mayo NE, Feldman L, Carli F (2008) Validating the 6-min walk test as a measure of recovery after elective colon resection surgery. Arch Phys Med Rehabil 89(6):1083–1089
26. Tarlov AR, Ware JE Jr, Greenfield S, Nelson EC, Perrin E, Zubkoff M (1989) The Medical Outcomes Study. An application of methods for monitoring the results of medical care. J Am Med Assoc 262(7):925–930
27. Laboratories ATSCoPSfCPF (2002) ATS statement: guidelines for the 6-min walk test. Am J Respir Crit Care Med 166(1):111–117
28. Feldman LS, Kaneva P, Demyttenaere S, Carli F, Fried GM, Mayo NE (2009) Validation of a physical activity questionnaire (CHAMPS) as an indicator of postoperative recovery after laparoscopic cholecystectomy. Surgery 146(1):31–39
29. White EAB, Saracci R (2008) Principles of exposure measurement in epidemiology. Oxford University Press, London
30. Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C et al (2008) Canadian normative data for the SF-36 health survey. Canadian Multicentre Osteoporosis Study Research Group. Can Med Assoc J 163(3):265–271
31. Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 7(5):541–546
32. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 10(4):407–415
33. Antonescu I, Scott S, Tran TT, Mayo NE, Feldman LS (2014) Measuring postoperative recovery: what are clinically meaningful differences? Surgery. doi:10.1016/j.surg.2014.03.005
34. Dindo D, Demartines N, Clavien PA (2004) Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Ann Surg 240(2):205–213
35. Dowson H, Cowie A, Ballard K, Gage H, Rockall T (2008) Systematic review of quality of life following laparoscopic and open colorectal surgery. Colorectal Dis. doi:10.1111/j.1463-1318.2008.01603.x
36. Busija L, Osborne RH, Nilsdotter A, Buchbinder R, Roos EM (2008) Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. Health Qual Life Outcomes 6:55
37. Grevitt M, Khazim R, Webb J, Mulholland R, Shepperd J (1997) The short form-36 health survey questionnaire in spine surgery. J Bone Joint Surg Br 79(1):48–52
38. Dunker MS, Bemelman WA, Slors JF, van Duijvendijk P, Gouma DJ (2001) Functional outcome, quality of life, body image, and cosmesis in patients after laparoscopic-assisted and conventional restorative proctocolectomy: a comparative study. Dis Colon Rectum 44(12):1800–1807
39. Maartense S, Dunker MS, Slors JF, Cuesta MA, Gouma DJ, van Deventer SJ et al (2004) Hand-assisted laparoscopic versus open

restorative proctocolectomy with ileal pouch anal anastomosis: a randomized trial. Ann Surg 240(6):984–991 Discussion 91–2

40. Maartense S, Dunker MS, Slors JF, Cuesta MA, Pierik EG, Gouma DJ et al (2006) Laparoscopic-assisted versus open ileocolic resection for Crohn's disease: a randomized trial. Ann Surg 243(2):143–149 Discussion 50–3
41. Andersen MH, Mathisen L, Veenstra M, Oyen O, Edwin B, Digernes R et al (2007) Quality of life after randomization to laparoscopic versus open living donor nephrectomy: long-term follow-up. Transplantation 84(1):64–69
42. Urbach DR (2002) Laparoscopic-assisted surgery for colon cancer. J Am Med Assoc 287(15):1938 Author reply 9
43. Urbach DR, Harnish JL, McIlroy JH, Streiner DL (2006) A measure of quality of life after abdominal surgery. Qual Life Res 15(6):1053–1061
44. Gervaz P, Mugnier-Konrad B, Morel P, Huber O, Inan I (2011) Laparoscopic versus open sigmoid resection for diverticulitis: long-term results of a prospective, randomized trial. Surg Endosc 25(10):3373–3378
45. Schwenk W, Haase O, Neudecker J, Muller JM (2005) Short-term benefits for laparoscopic colorectal resection. Cochrane Database Syst Rev. 3:CD003145