Eur Spine J (2009) 18 (Suppl 3):S386–S394
DOI 10.1007/s00586-009-1028-3

ORIGINAL ARTICLE

Ratings of global outcome at the first post-operative assessment after spinal surgery: how often do the surgeon and patient agree?

Friederike Lattig · Dieter Grob · Frank S. Kleinstueck · François Porchet · Dezsö Jeszenszky · Viktor Bartanusz · David O'Riordan · Anne F. Mannion

Received: 22 December 2008 / Revised: 21 April 2009 / Accepted: 1 May 2009 / Published online: 22 May 2009

Abstract Patient-orientated questionnaires are becoming increasingly popular in the assessment of outcome and are considered to provide a less biased assessment of the surgical result than traditional surgeon-based ratings. The present study sought to quantify the level of agreement between patients' and doctors' global outcome ratings after spine surgery. 1,113 German-speaking patients (59.0 ± 16.6 years; 643 F, 470 M) who had undergone spine surgery rated the global outcome of the operation 3 months later, using a 5-point scale: operation helped a lot, helped, helped only little, didn't help, made things worse. They also rated pain, function, quality-of-life and disability, using the Core Outcome Measures Index (COMI), and their satisfaction with treatment (5-point scale). The surgeon completed a SSE Spine Tango Follow-up form, blind to the patient's evaluation, rating the outcome with the McNab criteria as excellent, good, fair, and poor. The data were compared, in terms of (1) the correlation between surgeons' and patients' ratings and (2) the proportions of identical ratings, where the doctor's "excellent" was considered equivalent to the patient's "operation helped a lot", "good" to "operation helped", "fair" to "operation helped only little" and "poor" to "operation didn't help/made things worse". There was a significant correlation (Spearman Rho = 0.57, $p < 0.0001$) between the surgeons' and patients' ratings. Their ratings were identical in 51.2% of the cases; the surgeon gave better ratings than the patient ("overrated") in 25.6% cases and worse ratings ("underrated") in 23.2% cases. There were significant differences between the six surgeons in the degree to which their ratings matched those of the patients, with senior surgeons "overrating" significantly more often than junior surgeons ($p < 0.001$). "Overrating" was significantly more prevalent for patients with a poor self-rated outcome (measured as global outcome, COMI score, or satisfaction with treatment; each $p < 0.001$). In a multivariate model controlling for age and gender, "low satisfaction with treatment" and "being a senior surgeon" were the most significant unique predictors of surgeon "overrating" ($p < 0.0001$; adjusted $R^2 = 0.21$). Factors with no unique significant influence included comorbidity (ASA score), first time versus repeat surgery, one-level versus multilevel surgery. In conclusion, approximately half of the patient's perceptions of outcome after spine surgery were identical to those of the surgeon. Generally, where discrepancies arose, there was a tendency for the surgeon to be slightly more optimistic than the patient, and more so in relation to patients who themselves declared a poor outcome. This highlights the potential bias in outcome studies that rely solely on surgeon ratings of outcome and indicates the importance of collecting data from both the patient and the surgeon, in order to provide a balanced view of the outcome of spine surgery.

Keywords Spine surgery · Satisfaction · Global outcome · Self-assessment · Registry

Introduction

The evaluation and documentation of symptoms, functional outcomes and health after orthopaedic surgery is becoming

F. Lattig · D. Grob · F. S. Kleinstueck · F. Porchet · D. Jeszenszky · V. Bartanusz
Spine Center, Schulthess Klinik, Lengghalde 2,
8008 Zurich, Switzerland

D. O'Riordan · A. F. Mannion (✉)
Spine Center Division, Department of Research
and Development, Schulthess Klinik,
Lengghalde 2, 8008 Zurich, Switzerland
e-mail: anne.mannion@kws.ch

more and more important, not only for assessing patients' progress but also for evaluating the effectiveness and quality of the ever-increasing number of treatments available. For these purposes, standardised subjective and objective outcome measures are typically used.

For both the patient and the surgeon, the patient's overall satisfaction with the treatment and how much it helped his back problem is of the utmost importance. However, this criterion is somewhat subjective, and likely depends not only on the technical success of the operation but also on a number of factors such as the age and gender of the patient and the surgeon, their relationship to each other, their own individual self-esteem and personality, the general health status of the patient, the appropriateness of the preoperative information and the expectations it elicited, the extent of the treatment, and probably many more factors too.

Despite this, a number of studies in the area of hip, knee and shoulder arthroplasty have shown a substantial degree of agreement between patient and physician assessments of outcome [1, 12, 13, 16], although not without fail [14]. Smith et al. [16] reported a high level of agreement between patients' and surgeons' ratings of pain, function and satisfaction in a group of patients who were followed-up at least 6 months after shoulder arthroplasty. Another study compared the results from a 16-item clinical evaluation questionnaire after hip arthroplasty: for 12 items, acceptable consensus was found, although for patients with other health problems, revision surgery or mild to moderate pain there was a greater likelihood of disagreement [12]. Generally, where disagreement has been recorded in these studies, it is the case that the clinician systematically overrates outcome or underrates symptoms compared with the patient [11, 12, 16] and the discrepancy is more marked the more senior the clinician [11] and the poorer the patients' outcome [1, 6]. To the authors' knowledge, no study has investigated this issue in any depth in patients undergoing spine surgery. One study reported a good match between surgeons' and patients' ratings after surgery for lumbar disc herniation, but only limited details were given, especially in relation to factors that might influence the discrepancies found [15]. To this end, the data collected in surgical registries are particularly useful, since they are acquired from large numbers of patients, and from the perspective of both the patient and the surgeon. Using such a framework, it should hence be possible to answer the following questions:

- How well-matched are the patients' and the surgeons' ratings of global outcome after spine surgery procedures? Do they depend on the type of surgery done?
- Does the concordance in ratings depend on the age, gender and comorbidity of the patient and vary with the postoperative symptom status/patient's satisfaction (i.e. are discrepancies consistent across the range of possible outcomes, or does the gap widen with "better" or "worse" outcomes)?
- Are there relevant differences between surgeons in the agreement between patient and surgeon, indicating that the personality, attitude or seniority/status of the surgeon may be of importance?

The answers to these questions should allow us to arrive at some firm conclusions as to whether it is necessary to collect subjective quality rating data from the patient *and* the surgeon; whether the preoperative information for the surgical procedure is adequate and the expectations of the patient realistic; and whether, during the post-operative check-up, the questions being asked of the patient are pertinent enough to accurately gauge his perception of the benefits derived from the operation.

Methods

Inclusion criteria

The study was carried out within the framework of the SSE Spine Tango Spine Surgery Registry. It included the data of patients undergoing spinal surgery for a range of indications by one of the six experienced spine surgeons (4 orthopaedic and 2 neurosurgeons) in the Spine Centre of a specialised orthopaedic hospital (from Jan 2005 to Dec 2007 inclusive). Patients had to be fluent in either German or English, and be at a minimum 3 months post-op, without having had any re-intervention in those 3 months.

Surgeon forms

The SSE Spine Tango Surgery form was used to document pathology, previous treatment, patient morbidity status (assessed with the American Society of Anesthesiologists Physical Status Score (ASA Score) from 1 (no disturbance) to 5 (moribund)), surgeon credentials, surgical procedures applied, duration of operation, and the occurrence and type of both general and surgical complications before discharge.

At follow-up visits, the surgeon completed an SSE Follow-Up Form, based on his/her evaluation of the patient during a 15- to 30-min consultation. The patient was asked about current pain levels and locations, use of pain medication, activity level, duration of pain-free walking and sitting, involvement in rehabilitation, and ability to work. The wound, muscle function in the neck/back, range of movement and neurological status were also checked, and an X-ray of the operated area was assessed. The SSE Follow-Up Form included a rating of the global outcome, according to the MacNab classification [7], as excellent, good, fair, or poor.

The surgeons routinely carried out their first post-operative follow-up between 6 weeks and 3 months after surgery, depending on their own preference. All the patient questionnaires were routinely completed at 3 months (see below). Hence, where patients had had a 3-month clinical follow-up (or both 6 weeks and 3 months), the 3-month rating was used for comparison; where only a 6-week form had been completed, then this was used.

Patient-orientated questionnaires

Before and 3 months after surgery, patients were requested to complete the multidimensional Core Outcome Measures Index (COMI) questionnaire. On each occasion, the questionnaires were sent to the patients to complete at home, to ensure that the information given was free of care-provider influence. The COMI is a multidimensional index consisting of validated questions covering the domains of pain (leg/buttock and back pain intensity, each measured separately on a 0–10 graphic rating scale), function, symptom-specific well-being, general quality of life, and social and work disability. The COMI was originally developed based on the recommendations for a short series of Core Outcome questions by an Expert Group in the field of Low Back Pain Outcome measurement [2] and subsequently validated as an outcome instrument by three research groups [3, 8, 9, 18]. In addition to the COMI questions, at the 3-month follow-up there were further questions inquiring about overall satisfaction with treatment of the back problem in the hospital (5 categories from “very satisfied” to “very dissatisfied”), the global outcome of surgery (“how much did the operation help your back problem?”, with five categories from “helped a lot” to “made things worse”), and patient-rated complications (yes/no; if yes, describe).

Statistical analyses

Descriptive data are presented as means ± standard deviations (SD).

The correlation between surgeons’ and patients’ global ratings was examined using Spearman Rank correlation coefficients. The ratings were also compared using contingency analysis with Chi-squared, and for these purposes the surgeon’s “excellent” (= score of 1) was considered equivalent to the patient’s “operation helped a lot”, surgeon’s “good” (= score of 2) to patient’s “operation helped”, surgeon’s “fair” (= score of 3) to patient’s “operation helped only little” and surgeon’s “poor” (= score of 4) to patient’s “operation didn’t help/made things worse”. For further analysis, the difference between the surgeon’s and the patient’s rating (“surgeon–patient discrepancy” score) was calculated, using a convention in which the patient’s rating represented the baseline or the “true” value. This yielded a score that ranged from −3 (i.e. surgeon grossly overestimated outcome compared with patient) to +3 (surgeon grossly underestimated outcome compared with patient), with 0 representing equivalent ratings. For some analyses, these scores were condensed to −1 (surgeon-rating rating better than patient-rating), 0 (equivalent ratings) and +1 (surgeon-rating worse than patient-rating).

Spearman Rank and Kendall Rank correlation coefficients were used to determine the strength of the association between the “surgeon–patient discrepancy score” and various other ordinal variables measured at 3 months. Chi-square contingency analyses were used to examine the association between variables such as gender, or main pathology, and the “surgeon–patient discrepancy” category.

Multiple linear regression analysis was used to determine the relative importance of various factors in explaining the discrepancy between surgeons’ and patients’ ratings at 3 months. Age and gender were entered as control variables in a first step, then using a forward conditional approach for variable entry ($p < 0.05$ to enter) the following potential predictors were examined for selection in the model: satisfaction with treatment (1–5), surgeon status (junior 0, senior 1), and the multidimensional COMI index score at 3 months (0–10).

Statistical significance was accepted at the $p < 0.05$ level.

Results

In the time-period of the investigation, 3,106 patients were eligible for inclusion in the study. 2,882 (93%) of these completed a pre-operative questionnaire (administrative errors, refusal to participate, and emergencies accounted for the remainder). 3,054 patients were sent a patient-orientated questionnaire to complete after 3 months (52 were not sent a questionnaire for various reasons: died, expressed a wish not to complete any questionnaires, serious illnesses, etc.). 2,875/3,054 (94.1%) patients returned a completed questionnaire.

An SSE Follow-Up Form was completed by the surgeon at the first follow-up (6 weeks or 3 months post-op) for 1180/3106 (38%) patients. This rather low rate was a reflection of the only gradual participation in the SSE Tango follow-up system of the individual surgeons over the 3-year period: for example, two of them had reached 80–90% compliance in the second year, whilst others were slow to adopt the system even in the final year (see discussion).

For 1,113 (36%) patients, forms had been completed by both the surgeon and the patient and these comprised the study group in the present report. There were 643 (57.8%) women and 470 (42.2%) men and their mean (SD) age was 59.0 (16.6) years.

The surgeons' and patients' ratings of global outcome after surgery are shown in Table 1. There was a significant correlation between the two (Spearman Rho = 0.57, $p < 0.0001$). Their ratings were identical in 51.2% of the cases; the surgeon gave better ratings than the patient in 25.6% cases and worse ratings in 23.2% cases.

There were significant differences ($p < 0.001$) between the six surgeons in the degree to which their ratings matched those of the patients (Table 2). Compared with the rest of the group, one surgeon (surgeon 4) appeared to more frequently underestimate the patient's rating of outcome, whilst one tended to more commonly overestimate it (surgeon 3). When the surgeons were dichotomised as either "senior" (3 surgeons) or "junior" (3 surgeons), in terms of their status in the hierarchy, there was a significant difference between the groups, with senior surgeons overestimating the patient's outcome in 34.3% cases and underestimating it in 14.4%, compared with 11.3% overestimation and 37.6% underestimation for the junior surgeons ($p < 0.0001$). In each group, the proportion of identical classifications was almost identical (51.3% for the senior surgeons, 51.1% for the junior surgeons).

There was no significant association between main pathology, as indicated on the SSE Spine Tango pre-operative form, and the "surgeon–patient discrepancy" (Fig. 1), although some of the group sizes were very small.

Table 1 Surgeons' and patients' ratings of global outcome at the first follow-up after surgery, shown as absolute numbers in each category followed by the proportion of surgeon ratings in each of the patient-rating categories

		Surgeon's rating at first follow-up, up to 3 months post-surgery (McNab rating of global outcome)				
		Excellent	Good	Fair	Poor	TOTAL
PATIENT'S RATING AT 3 MONTHS POST-SURGERY (How much did the operation help your back problem?)	Helped a lot	**275** (58.6%)	175 (37.3%)	18 (3.9%)	1 (0.2%)	469 (100%)
	Helped	113 (29.4%)	**222** **(57.8%)**	46 (12.0%)	3 (0.8%)	384 (100%)
	Helped only little	26 (15.0%)	85 (49.1%)	**47** (27.2%)	15 (8.7%)	173 (100%)
	Didn't help	1 (1.4%)	23 (31.5%)	23 (31.5%)	**26** **(35.6%)**	73 (100%)
	Made things worse	1 (7.1%)	1 (7.1%)	7 (50.0%)	**5** **(35.7%)**	14 (100%)
	TOTAL					1113

Bold values in dark-shaded areas highlight cases of absolute agreement between patient and surgeon ratings. Light-shaded areas indicate where the surgeon rated the outcome more positively than the patient

Table 2 Difference between individual surgeons' for the discrepancy between their ratings and the patients' ratings of global outcome (contingency analysis, $p < 0.001$)

	Surgeon overestimated outcome compared with patient %	Surgeon underestimated outcome compared with patient %	Surgeon matched outcome reported by patient %
All surgeons	25.6	23.2	51.2
Surg 1	18.8	37.5	43.7
Surg 2	11.6	30.0	58.4
Surg 3	37.4	11.4	51.2
Surg 4	10.2	47.7	42.1
Surg 5	13.2	21.1	65.8
Surg 6	21.1	15.8	63.2

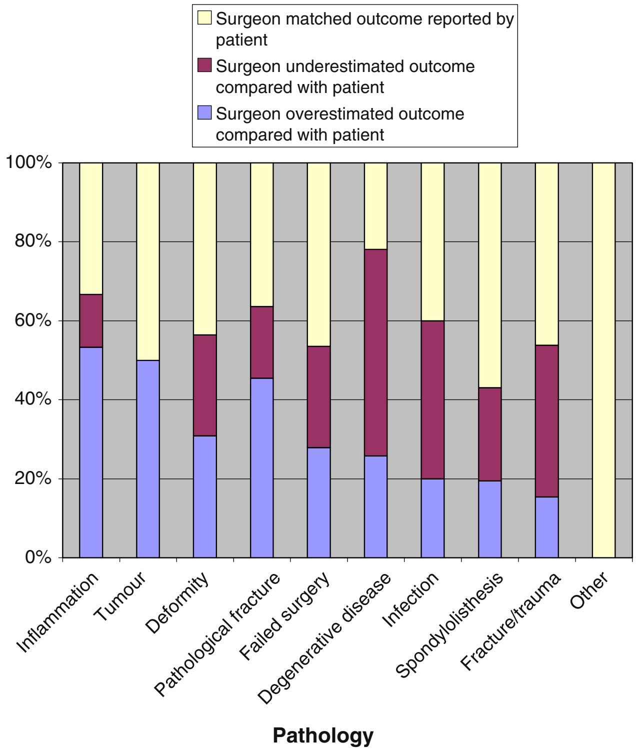

Fig. 1 Relationship between pathology and discrepancy in the surgeon–patient outcome ratings (N.B. caution by groups with small number of patients (inflammation, tumour, pathological fracture, infection, fracture/trauma and other; all <20 patients per group))

Table 3 Influence of previous surgery on the discrepancy between surgeon's ratings and the patients' ratings of global outcome (Chi-square contingency analysis, $p = 0.47$)

Previous surgery group[a]	Surgeon overestimated outcome compared with patient %	Surgeon underestimated outcome compared with patient %	Surgeon matched outcome reported by patient %
No previous surgery ($N = 641$)	25.4	22.3	52.3
Previous surgery at a different level ($N = 114$)	22.8	21.9	55.3
Previous surgery at the same level ($N = 267$)	30.3	22.1	47.6

[a] $N = 1{,}022$ since SSE surgery forms were not available for all patients

Whether the patient was undergoing spinal surgery for the first time or had previously undergone spinal surgery at either the same level or at a different level had no influence on the surgeon–patient discrepancy ($p = 0.47$) (Table 3). Similarly, whether the patient had a one-segment or multiple-segment lesion had no significant influence on the discrepancy (surgeon "overrated" in 23% cases with one-segment lesions and 28% cases with multi-segment lesions; $p = 0.14$).

Age showed a very low but, nonetheless, significant correlation with the "surgeon–patient discrepancy score" (Tau corrected for ties = −0.073, $p = 0.0003$): the greater the patient's age the more the surgeon tended to overestimate the success of their outcome, compared with their own rating.

There was a tendency for the surgeon to overestimate the global outcome (compared with the patient's rating) more often in women (27.7%) than in men (22.8%), but the difference failed to reach significance ($p = 0.13$). There was no significant association between the surgeon–patient discrepancy in rating and the patient's comorbidity (ASA) score (Rho corrected for ties = 0.048, $p = 0.12$).

The three categories describing the rating-discrepancy between the surgeon and patient (surgeon overestimates compared with patient, surgeon/patient identical ratings, surgeon underestimates compared with patient) differed significantly in their mean scores for each of the COMI domains, the COMI composite index score, the global outcome and satisfaction with treatment at 3 months post-surgery (Table 4): the patients' self-rated status was consistently worse in the group for which the outcome had been overestimated by the surgeon.

In a multivariate model to identify the most important factors explaining "a more optimistic rating by the surgeon than the patient", controlling for age and gender, the following were unique significant predictors: being a senior surgeon, lower patient satisfaction, and having a worse status as recorded by the multidimensional CORE index ($p < 0.0001$; adj $R^2 = 0.21$; Table 5).

Discussion

How well-matched are the patients' and the surgeons' ratings of global outcome after spine surgery procedures?

The present study sought to examine the comparability of surgeons' and patients' independent ratings of global outcome at the first follow-up after spine surgery. The main findings were that there was a statistically significant correlation between the ratings, with the patient and the surgeon showing exact agreement in approximately half of the cases. At first sight, this level of agreement suggests a

Table 4 Differences in satisfaction, global outcome and COMI item scores between the three "surgeon–patient score discrepancy" categories

	Surgeon overestimated outcome compared with patient	Surgeon matched outcome reported by patient	Surgeon underestimated outcome compared with patient	p value
% Satisfied/very satisfied	69.0%	91.2%	96.1%	<0.0001
% Good outcome (operation helped/helped a lot)	39.6%	87.2%	94.2%	<0.0001
Worst (of back or leg) pain intensity: 0 (none) to 10 (max)	5.2 ± 2.4	3.5 ± 2.7	3.6 ± 2.4	<0.0001
Back-related function: 1 (good) to 5 (poor)	3.3 ± 1.0	2.6 ± 1.2	2.8 ± 1.2	<0.0001
Symptom-specific well-being: 1 (good) to 5 (poor)	3.9 ± 1.2	2.7 ± 1.5	2.8 ± 1.3	<0.0001
General QoL: 1 (good) to 5 (poor)	3.2 ± 0.9	2.5 ± 1.0	2.6 ± 0.9	<0.0001
Social disability: 1 (none) to 5 (max)	3.4 ± 1.6	2.6 ± 1.6	2.9 ± 1.6	<0.0001
Work disability: 1 (none) to 5 (max)	2.9 ± 1.8	2.2 ± 1.6	2.4 ± 1.7	<0.0001
COMI composite score: 0 (best status) to 10 (worst status)	5.8 ± 2.4	3.8 ± 2.7	4.1 ± 2.4	<0.0001

Table 5 Results of the final step of the forward multiple regression analysis to identify factors accounting for the discrepancy in patient–surgeon outcome ratings (where higher scores in the dependent variable indicate that the surgeon over-rates compared with the patient)

Independent variables	Unstandardised regression coefficients		Standardised coefficients	t	Sig (p value)	95% CI for B		Correlations			% Explained variance in surgeon–patient discrepancy in outcome rating Adj R^2
	B	Std. error	Beta			CI low	CI high	Zero-order	Partial	Part	
Constant	2.179	0.106		20.522	0.000	1.971	2.387				21.0
Age (years)	0.001	0.002	0.010	0.353	0.724	−0.002	0.003	0.072	0.011	0.009	
Gender (F = 0, M = 1)	−0.021	0.046	−0.012	−0.458	0.647	−0.112	0.070	−0.048	−0.014	−0.012	
Patients' satisfaction rating (1 = best, 5 = worst)	0.223	0.026	0.253	8.548	0.000	0.172	0.274	0.337	0.250	0.228	
Surgeon status (junior = 0, senior = 1)	0.522	0.048	0.298	10.892	0.000	0.428	0.616	0.319	0.312	0.291	
COMI score at 3 months (0 = best, 10 = worst)	0.044	0.009	0.139	4.690	0.000	0.025	0.062	0.234	0.140	0.125	

Significant predictors, when all variables were considered together in the model, included the patient's rating of satisfaction (the worse the satisfaction, the greater the surgeon "over-rated"), surgeon status (senior surgeons "over-rated" more than the juniors), COMI score at follow-up (the worse the multidimensional outcome status, the greater the surgeon "over-rated")

notable difference of opinion or misunderstanding between patient and surgeon, especially when compared with the reported discrepancies in outcome ratings between patient and physician of only 13% after hip arthroplasty [6] or 31% after shoulder arthroplasty [16]. One explanation that immediately comes to mind relates to the fact that arthroplasty surgery has a long history of successful development, and is well standardised, worldwide. In contrast, in the present study, patients were undergoing all types of spinal surgery, ranging from relatively straightforward interventions, such as simple decompression, all the way up to complicated multiple revisions. In the latter cases, the quality of the result is sometimes difficult to ascertain and, in contrast to the patient, the surgeon may be satisfied with less improvement in view of his/her experience and knowledge of the evidence for the (ever reducing) success of repeated surgeries. However, in the present study the agreement between the surgeons' and the patients' ratings was no better for patient groups that were undergoing first-time surgery than for those that had undergone multiple surgeries or for patients with different pathologies. Another possible explanation for the discrepancy between studies concerns the rating-system itself and the manner in which the data are analysed. It can be expected that the more response-category options there are, the lower the perfect concordance in ratings between different individuals. The SSE Spine Tango Follow-up form offered the surgeon four outcome options using the MacNab criteria [7], with three positive (excellent, good, fair) and one negative (poor) response categories; the patients could choose between five options on a 5-point scale, with three positive (operation helped a lot, helped, only helped a little) and two negative outcomes (didn't help, made things worse; with the latter two merged for comparison with the "poor" category of the surgeons). Smith et al. [16] used a modified Neer rating system with just three global outcome categories (excellent, satisfactory, and unsatisfactory) and found exact agreement between the patient and surgeon in 69% cases; when the outcomes were dichotomised (excellent and satisfactory together versus unsatisfactory) then perfect agreement was found in 87% cases. Brokelman et al. [1] reported no significant difference between the group mean satisfaction ratings of the patient and the surgeon after total hip arthroplasty, with a correlation coefficient of 0.7 between them. However, correlation analyses are used to determine consistency in the relative ranking of individuals in a group, not agreement in their absolute values per se; when agreement was analysed on an individual basis, some large discrepancies for certain individual surgeon–patient pairs were revealed [1]. If the data from the present study are examined using different statistics and categorisation methods, then our data are not so out of keeping with the rest of the literature. First, the correlation coefficient (relative ranking) between patient and surgeon ratings of 0.6 compares reasonably favourably with the 0.7 of Brokelman et al. [1]. Further, although in the present study the absolute agreement was only 52%, the proportion of patient–surgeon ratings that differed by a maximum of one category was 92%, and when the data were dichotomised (into "positive" and "negative" outcomes, as described above) then agreement was 93%.

A further difference between the present study and others [12, 16], that would explain our more pronounced surgeon–patient discrepancies, is the fact that the patients and surgeons conducted their appraisals completely independently of each other. This contrasts with three previous studies, in which the patient completed the questionnaire in the waiting room immediately prior to the consultation [6, 12, 16]. In some of these studies, the doctor then discussed an identical questionnaire with the patient [16] and/or completed it in the patient's presence [12, 16]; in others, even if the patient's evaluation was not available for the physician to review and the physician completed his/her questionnaire *after* the patient's visit, he/she was still aware that the patient had just evaluated the result of the operation [6]. These factors, and the acute awareness that the "accuracy" of the ratings was being subject to investigation, may—whether consciously or otherwise—have biased the outcome ratings given by the surgeons, leading them to proffer less "enthusiastic" ratings than normal. Even if no such bias occurred, the mere fact that the patient and doctor expressly conferred during completion of the doctor's questionnaire, and the proximity in time of completion of the two questionnaires, likely increased the comparability of the responses given. In the present study, when completing their forms the doctors were not even aware that their ratings were going to be analysed and compared with those of the patient. Hence, they represented the closest possible portrayal of each surgeon's typical everyday practice. Not using an identical scale for the doctors and patients may have influenced the degree of agreement between them; however, obtaining an appropriate patient-rated outcome question per se was considered more important than having identical surgeon/patient options, and asking patients to rate the outcome using the MacNab criteria would have been less relevant than enquiring about the impact of the operation on their own specific spinal problem.

In summary, and given the truly blinded nature of the data collection in the present study, the observed degree of agreement between the doctors' and patients' rating must be considered reasonably good and certainly comparable with the figures found in the literature available to date for other orthopaedic interventions.

Does the concordance in ratings depend on the age, gender and comorbidity of the patient, or vary with postoperative symptom status/patient satisfaction?

The influence of the patient's age on the discrepancy between their and the surgeon's outcome ratings has not been well-studied, and still remains unclear. McGee et al. [12] found that, in relation to hip arthroplasty, agreement was less good in older patients, for whom the outcome tended to be overestimated by the surgeon. Another study, however, has shown that surgeons feel more warmth and enthusiasm for older patients [5], and this greater empathy might be expected to lead to better agreement, not to an underestimation of an individual's health problems. Our own data showed a weak but significant relationship between the patient's age and the surgeon–patient outcome discrepancy, in the same direction as reported by McGee et al. [12] (i.e. physician overrates outcome in older patients). Nonetheless, the relationship failed to acquire significance in the final multivariate model. It might be expected that, in contemplating the evaluation form, the older patient would fail to focus on the operated area only, and on the specific results of the intervention, and would instead allow the effect of other health problems (that typically accumulate with increasing age) to influence their rating of the overall result. Interestingly, however, in the present study comorbidity (ASA score) did not influence the concordance in ratings. Supporting previous findings [12], the gender of the patient was also of no consequence in explaining the patient–surgeon discrepancies in outcome ratings.

A most interesting finding in the present study was the observation that the patients' self-rated post-operative status was consistently worse in the group for which the outcome had been overestimated by the surgeon. This was true regardless of whether the status was expressed as global outcome, the COMI multidimensional outcome score, or satisfaction with treatment for the back problem. This confirms the findings of previous studies in patients after hip arthroplasty, in which the discrepancy between patient and physician assessments (physician better rating than patient) was even greater when the patient was not completely satisfied with the result [1, 6]. These authors considered various potential explanations for this phenomenon, most notably concerning the notion of the patient and surgeon having different expectations regarding what could be achieved by the surgery, and different opinions as to what constitutes success. Due to his/her clinical experience in judging the expected result, based on hundreds of preceding examples, the physician may take into consideration the number of previous surgeries, the preoperative level of function, the quality of the supporting soft tissues, and so on, and may be able to differentiate the origin of any further/existing pain—sources of information to which the patient is not necessarily privy. The worse the starting point, the more satisfied the physician may be with any reasonable improvement; the patient, in contrast, naturally hopes to be pain-free and fully functioning again, regardless. This highlights the importance of clearly discussing the expected result with the patient before the surgery, since unmet expectations are a significant cause of dissatisfaction with outcome [10]. Other factors of possible importance in explaining the larger discrepancy in patients with a poorer outcome include the inability of the patient to clearly verbalise their dissatisfaction or the surgeon being oblivious to the sometimes subtle cues revealing discontent.

The clinical consequence of overrating the outcome in poor-outcome patients is that the patient may be prematurely discharged from care, with his/her continuing treatment needs being unmet. This may then lead to the development of chronic "failed-back" problems and long-term incapacity. Hence, even if the surgeon is correct in his/her assertions about the outcome, and the patient's perception arises solely due to unrealistically optimistic expectations of surgery or "over-anxiety"/catastrophising about the current situation, this must still be acknowledged and dealt with; simple dismissal or denial of the situation can only be expected to result in feelings of abandonment, "not being understood", and further dissatisfaction.

Are there differences between surgeons in the patient–surgeon agreement, indicating that surgeon-specific factors may be of importance?

It can never be certain to what extent different individuals agree on the meanings associated with the specific adjectives used in rating scales [17]: does "excellent" or "fair" convey the same meaning to different people? The issue is even more complicated, when the matter in question is effectively "the success (or otherwise) of my own work", as in the case of surgeons rating the success of their surgery. Conceivably, in the present study, these issues may have accounted for some of the variability *between* surgeons in the surgeon–patient agreement. The personality, attitude, or seniority status of the surgeon may all have an important influence on this discrepancy. A Pubmed search of the literature reveals a paucity of information on this topic, with no studies examining how critically surgeons assess their own work or how the surgeon's personality might influence his/her own quality ratings. In the present study, significant differences between the six surgeons were observed in terms of the surgeon–patient agreement in outcome ratings. The interpretation of these results is not easy. The long duration over which the data were collected suggests that temporary conditions, such as the current health status or social and financial contentedness of the doctor, should not have influenced the results in any major way. Using a convention in which the patient's rating was considered the "true" value, outcome was more commonly "overestimated" by senior surgeons than by juniors. These findings concur with those of a previous investigation in which the underestimation of patients' pain ratings by physicians in an emergency unit setting was greater in more experienced physicians [11]. It was suggested that, when rating pain, patients are likely to have in mind the worst pain they have felt, and physicians the worst pain they have seen [11]. This difference in references points was considered likely to induce a "miscalibration", since people make judgements depending on their own reference points. The analogous interpretation in the present study would be that senior surgeons, who have likely seen more "extremely bad results" (or generally more extreme pain and disability) in their time, use this as the anchor for the "extreme end" of the scale, such that in comparison any given outcome is rated "less negatively" than it might otherwise be. Marquie et al. [11] also considered that the greater training and years in the job might allow senior physicians to better identify which cues are relevant and to neglect those they have learnt to be unreliable. This is a positive way of looking at the phenomenon; an alternative interpretation may be that the compassion for and empathy with the patient, shown by a healthy, pain-free, active, and enthusiastic young colleague, is simply higher.

Since many of the aforementioned variables expected to influence the surgeon–patient outcome discrepancy were not relevant, we should consider whether factors peculiar to the individual surgeon may play a role. In the theory of psychological types expounded by Jung [4], every individual senses and reacts differently on the basis of his/her own idiosyncrasy. The personality types described by Jung [4] may have a bearing on the interpretation of subjective rating scales, especially scales like the McNab that are associated with no clearly measurable objective criteria. The same personality typology surely also exists for the patient. This may be an area worthy of further investigation in the field of outcomes research.

Limitations of the study

Certain limitations of the present study are worthy of mention. The patient questionnaires always enjoyed a high completion rate upon first initiation of the registry in the hospital; however, at the start of the project, only few surgeons were regularly completing the surgical follow-up questionnaires. The routine adoption of any new documentation system in the clinical environment is notoriously difficult. Nonetheless, without the matched data from the surgeons, this resulted in fewer data sets being available for the current investigation.

The patients systematically completed their first follow-up 3 months after surgery, whilst the surgeons' first follow-up typically varied from between 6 weeks and 3 months post-operation (and occasionally included both). When no follow-up form had been completed by the surgeon at 3 months, the 6-week-form was taken instead (where available) and considered to represent outcome at the "early follow-up". This was done to increase the pool of surgeon-rated outcome data available. It is, however, possible that there was a true change in outcome over this short period that would consequently influence the comparability of the surgeon's and patient's ratings.

The first follow-up after surgery (3 months) may be considered too early for some patients, after some interventions, to give an accurate appraisal of their outcome. Whilst the doctor may well know what is to be expected at that stage of follow-up, for the patient it may still constitute a stage of recovery with uncertain outcome. Nonetheless, our preliminary data on approximately 200 patients show that similar trends to those reported here are also found for the 12-month post-operative follow-up: although the degree of "over-estimation" by the surgeon is slightly lower, the influence of "poor outcome" and "surgeon seniority" is still evident. Our future studies will concentrate on the agreement between ratings after longer follow-up periods in larger groups of patients.

Conclusions

The lack of agreement between the patient and the surgeon in their ratings of outcome measured at the first post-operative assessment suggests that both ratings should be taken into consideration when judging the overall success of the procedure. The inter-individual differences between surgeons in their surgeon–patient agreement may be the result of the seniority/personality type of the surgeon and the use of "open/subjective" rating scales that allow too much leeway for individual interpretation. Further investigations should be carried out to see whether the influence of these factors can be minimised or eliminated in future global outcome rating scales. The patient and surgeon may have a different view of what constitutes an excellent, good, fair or poor result as far as a given patient is concerned; as such, the adequacy of the preoperative informed consent procedures should be further investigated and optimised to provide the patient with more realistic expectations of the outcome of surgery.

Acknowledgments We are grateful to Gordana Balaban and Julian Amacker for their excellent work collecting the questionnaire data and managing our quality management system.

Conflict of interest statement None of the authors has any potential conflict of interest.

References

1. Brokelman RB, van Loon CJ, Rijnberg WJ (2003) Patient versus surgeon satisfaction after total hip arthroplasty. J Bone Joint Surg Br 85:495–498
2. Deyo RA, Battie M, Beurskens AJHM, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korff M, Waddell G (1998) Outcome measures for low back pain research. A proposal for standardized use. Spine 23:2003–2013. doi:10.1097/00007632-199809150-00018
3. Ferrer M, Pellise F, Escudero O, Alvarez L, Pont A, Alonso J, Deyo R (2006) Validation of a minimum outcome core set in the evaluation of patients with back pain. Spine 31:1372–1379. doi:10.1097/01.brs.0000218477.53318.bc discussion 1380
4. Jung DG, Baynes HG (1921) Psychological types, or, the psychology of individuation. Kegan Paul Trench Trubner, London
5. Levinson W, Frankel RM, Roter D, Drum M (2006) How much do surgeons like their patients? Patient Educ Couns 61:429–434. doi:10.1016/j.pec.2005.05.009
6. Lieberman JR, Dorey F, Shekelle P, Schumacher L, Thomas BJ, Kilgus DJ, Finerman GA (1996) Differences between patients' and physicians' evaluations of outcome after total hip arthroplasty. J Bone Joint Surg Am 78:835–838
7. Macnab I (1973) Chapter 14. Pain and disability in degnerative disc disease. Clin Neurosurg 20:193–196
8. Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Dvorak J, Jacobshagen N, Semmer NK, Boos N (2007) Predictors of multidimensional outcome after spinal surgery. Eur Spine J 16:777–786. doi:10.1007/s00586-006-0255-0
9. Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Semmer NK, Jacobshagen N, Dvorak J, Boos N (2005) Outcome assessment in low back pain: how low can you go? Eur Spine J 14:1014–1026. doi:10.1007/s00586-005-0911-9
10. Mannion AF, Junge A, Elfering A, Dvorak J, Porchet F, Grob D (2009) Great expectations: really the novel predictor of outcome after spinal surgery? Spine (in press)
11. Marquie L, Raufaste E, Lauque D, Marine C, Ecoiffier M, Sorum P (2003) Pain rating by patients and physicians: evidence of systematic pain miscalibration. Pain 102:289–296. doi:10.1016/S0304-3959(02)00402-5
12. McGee MA, Howie DW, Ryan P, Moss JR, Holubowycz OT (2002) Comparison of patient and doctor responses to a total hip arthroplasty clinical evaluation questionnaire. J Bone Joint Surg Am 84-A:1745–1752
13. McGrory BJ, Morrey BF, Rand JA, Ilstrup DM (1996) Correlation of patient questionnaire responses and physician history in grading clinical outcome following hip and knee arthroplasty. A prospective study of 201 joint arthroplasties. J Arthroplasty 11:47–57. doi:10.1016/S0883-5403(96)80160-4
14. Ragab AA (2003) Validity of self-assessment outcome questionnaires: patient–physician discrepancy in outcome interpretation. Biomed Sci Instrum 39:579–584
15. Ronnberg K, Lind B, Zoega B, Halldin K, Gellerstedt M, Brisby H (2007) Patients' satisfaction with provided care/information and expectations on clinical outcome after lumbar disc herniation surgery. Spine 32:256–261. doi:10.1097/01.brs.0000251876.98496.52
16. Smith AM, Barnes SA, Sperling JW, Farrell CM, Cummings JD, Cofield RH (2006) Patient and physician-assessed shoulder function after arthroplasty. J Bone Joint Surg Am 88:508–513. doi:10.2106/JBJS.E.00132
17. Streiner DL, Norman GR (1995) Health Measurement Scales: a practical guide to their development and use. Oxford University Press Inc., Oxford
18. White P, Lewith G, Prescott P (2004) The core outcomes for neck pain: validation of a new outcome measure. Spine 29:1923–1930. doi:10.1097/01.brs.0000137066.50291.da