Eur Spine J (2010) 19:297–306
DOI 10.1007/s00586-009-1194-3

ORIGINAL ARTICLE

A comparison of outcomes of cervical disc arthroplasty and fusion in everyday clinical practice: surgical and methodological aspects

Dieter Grob · Francois Porchet · Frank S. Kleinstück · Friederike Lattig · Dezsoe Jeszenszky · Andrea Luca · Urs Mutter · Anne F. Mannion

Received: 13 July 2009 / Revised: 27 August 2009 / Accepted: 12 October 2009 / Published online: 31 October 2009

Abstract Randomised controlled trials (RCTs) of cervical disc arthroplasty vs fusion generally show slightly more favourable results for arthroplasty. However, RCTs in surgery often have limited external validity, since they involve a select group of patients who fit very rigid admission criteria and who are prepared to subject themselves to randomisation. The aim of this study was to examine whether the findings of RCTs are verified by observational data recorded in our Spine Center in association with the Spine Society of Europe Spine Tango surgical registry. Patients undergoing fusion/stabilisation or disc arthroplasty for degenerative cervical spinal disease were selected for inclusion. They completed a questionnaire pre-operatively and at 12 and 24 months follow-up (FU). The questionnaire comprised the multidimensional Core Outcome Measures Index (COMI; 0–10 scale) and, at FU, questions on global outcome and satisfaction with treatment (5-point scales, dichotomised to "good" and "poor"), re-operation and patient-rated complications. The surgeon completed a Spine Tango Surgery form. The outcome data from 266 (208 fusion, 58 arthroplasty) out of 284 eligible patients who had reached 12 months FU, and 169 (139 fusion, 30 arthroplasty) out of 178 who had reached 24 months FU, were included. Patients with cervical disc arthroplasty were younger [46 (SD 8) years vs 56 (SD 11) years for fusion; $P < 0.05$], had less comorbidity ($P < 0.05$), more often had only mono-segmental pathology (69% arthroplasty, 47% fusion) and only one type of degenerative pathology (69% arthroplasty, 46% fusion). Surgical complication rates were similar in each group (arthroplasty, 1.5%; fusion, 2.6%). The reduction in the COMI score was significantly greater in the arthroplasty group (at 12 months, 4.8 (SD 3.0) vs 3.7 (SD 2.9) points for fusion, and at 24 months 5.1 (SD 2.8) vs 3.8 (SD 2.9) points; each $P < 0.05$). In the arthroplasty group, a "good" global outcome was recorded in 90% patients (at 12 months) and 93% (at 24 months); in the fusion group the figures were 80 and 82%, respectively (group differences at each timepoint, $P > 0.09$). Satisfaction with treatment was similar in both groups (89–93%), at each timepoint. In multiple regression analysis, treatment group was of borderline significance as a unique predictor of the change in COMI at FU ($P = 0.059$ at 12 months, $P = 0.055$ at 24 months) in a model in which known confounders (age, comorbidity, number of affected levels) were controlled for. Being in the arthroplasty group was associated with an approximately 1-point greater reduction in the COMI score at FU. The results of this observational study appear to support those of the RCTs and suggest that, in patients with degenerative pathology of the cervical spine, disc arthroplasty is associated with a slightly better outcome than fusion. However, given the small size of the difference, its clinical relevance is questionable, especially in view of the a priori more favourable outcome expected in the arthroplasty group due to the more rigorous selection of patients.

Keywords Cervical spine disc arthroplasty · Fusion/stabilisation · Patient-rated outcomes · Observational study

D. Grob · F. Porchet · F. S. Kleinstück · F. Lattig · D. Jeszenszky · A. Luca · U. Mutter · A. F. Mannion (✉)
Spine Center, Schulthess Klinik,
Lengghalde 2, 8008 Zurich, Switzerland
e-mail: anne.mannion@kws.ch; anne@annefmannion.com

Introduction

Motion-preserving techniques for the treatment of painful degenerative conditions of the spine were originally

developed to overcome the (potential) negative aspects of fusion such as pseudarthrosis, accelerated adjacent segment degeneration and morbidity associated with the bone graft harvest [14]. In view of the evidence suggesting an increased range of motion and disc pressure [16, 20] in the segment adjacent to the fused one, the replacement of the degenerate disc with a disc prosthesis, to avoid immobilisation of the spinal segment, seemed to be a logical concept. In the *lumbar* spine, disc replacement has been in existence for more than 25 years [48], but the implant never achieved the popularity that was initially anticipated. There are various possible reasons for this, one of which concerns the techniques and risks of the anterior approach to the lumbar spine, an unfamiliar approach for many spine surgeons. This would also partially explain the success of disc replacement in the cervical spine [48], where the approach is accompanied by fewer anatomical obstacles and presents a familiar anatomy to all spine surgeons.

Disc replacement in the cervical spine is currently performed on a regular basis in clinical practice, despite the fact that high-level randomised controlled trials (RCTs) examining the efficacy and safety of the new procedure have only recently been published and have a maximum follow-up (FU) of just 2 years. In fact, just four separate research groups have carried out sizeable RCTs of cervical disc arthroplasty vs fusion, although their results have been presented in multiple publications addressing different subtopics within the same trials or by individual groups of authors participating in the multicentre studies [3, 25, 29, 37, 38, 44–46]. Collectively, these and one other small RCT [42] have shown statistically superior or comparable results for arthroplasty in relation to patient-rated outcomes such as the Neck Disability Index (NDI), neck and arm pain and quality of life (SF36). The maintenance of segmental mobility with disc arthroplasty, expected to contribute to a reduced development of adjacent level change, has also been documented in three trials [29, 38, 44]. Although these studies possess the acknowledged scientific rigour of RCTs, they are not without their limitations. Firstly, in trials of this type, it is not usually considered practical to blind the patients or surgeons to the type of surgery performed. This opens up an obvious potential for bias. Any evaluation of an innovation may include both bias and the true efficacy of the new therapy; randomised controlled trials that do not use a double-blind design have a significantly higher likelihood of showing a gain for the innovation than do double-blind trials [12]. A further problem, expounded by one of the cervical disc prosthesis trial groups [3, 25], was that many patients that were originally randomly assigned to a given treatment (37 arthroplasty and 80 fusion) later withdrew their participation in the study before undergoing treatment. One of the main reasons for this, especially in the fusion group, was dissatisfaction with the group to which they had been randomised. This series of post-randomisation drop-outs contributes to a disparity between the groups, and represents another potential source of bias.

RCTs in the field of surgery are renowned for their tendency to have limited external validity, often involving only a select group of patients who fit very rigid (and sometimes relatively "atypical") eligibility criteria and who are prepared to subject themselves to randomisation [1, 2, 26, 27]. In addition, in such trials, "expectations bias" is sometimes suspected to contribute to the more favourable patient-orientated results in the novel treatment group. This is illustrated by the finding that patient satisfaction sometimes shows significant differences in favour of the new treatment that are not always reflected in the prospectively measured outcome variables such as pain and disability [9]. And, finally, single trials are usually underpowered to address adequately the absolute and relative risks of adverse events, especially uncommon ones [40]. These factors, together with the complexity of the design of randomised trials within the clinical setting [27], indicate that the results of RCTs need further confirmation by carefully conducted observational studies in daily clinical practice.

The aim of this study was to examine whether the findings of the RCTs conducted to date to compare the clinical outcome after cervical spine arthroplasty and fusion/stabilisation are verified by observational data recorded in our Spine Center's spine surgical registry.

Methods

Patients

Inclusion criteria

The study was carried out within the framework of the Spine Society of Europe Spine Tango (Spine Surgery Registry) data acquisition system. It included the data of all patients undergoing surgery by one of six experienced spine surgeons (four orthopaedic and two neurosurgeons) in the Spine Center of our specialised orthopaedic hospital (between February 2004 and April 2009). Patients had to be fluent in either German or English, and satisfy the surgical inclusion criteria. The latter were based on the data documented on the registry's "SSE Surgery Form" as follows: surgery at the mid-lower region of the cervical spine, degenerative disease as the main pathology, maximum three motion segments affected. The fusion/rigid stabilisation group included all patients who had undergone anterior interbody fusion between adjacent vertebrae using an anterior approach and/or interbody stabilisation with

cage (anterior approach); the disc arthroplasty group comprised those receiving only motion-preserving stabilisation (any make of device).

Exclusion criteria

Patients who had undergone both fusion/stabilisation and disc arthroplasty (at different levels) were excluded from the analysis.

Patient-orientated questionnaires

Before and 12 and 24 months after surgery, patients were requested to complete the multidimensional Core Outcome Measures Index (COMI) questionnaire [32]. 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 (neck and arm 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 spine outcome measurement [15] and subsequently validated as an outcome instrument by three research groups [17, 31, 32, 49]. In addition to the COMI questions, at the 12-month and 24-month FUs, there were further questions with 5-point Likert scales inquiring about satisfaction ("over the course of treatment for your neck problem how satisfied were you with the medical care in our hospital?"; response categories from "very satisfied" to "very dissatisfied") and the global outcome of surgery ("overall, how much did the operation help your neck problem?"; response categories from "helped a lot" to "made things worse"). The questionnaire also contained questions as to whether the patient had been re-operated on since the index operation and on the occurrence and nature of any complications that were experienced as a result of the surgery (e.g. problems with wound healing, paralysis, sensory disturbances, etc.) (see [23] for further details).

Surgical documentation forms

SSE Spine Tango Surgery forms were used to document information regarding the medical history [main pathology, with further indication of the specific type of pathology(ies)], number of affected levels, previous surgery, operation duration (ten categories, from <1 h to >10 h), blood loss (five categories: none, <500, 500–1,000, 1,000–2,000, >2,000 ml), comorbidity [assessed with the American Society of Anesthesiologists Physical Status Score (ASA Score), from 1 (no disturbance) to 5 (moribund)], surgical details, surgical complications and general complications.

Statistical analyses

Power calculations (MedCalc Statistical Software, Mariakerke, Belgium) revealed that, with a minimum of 31 patients in each group, the probability was 80% (with 42 patients, 90%) and that the study would detect a treatment difference at a two-sided 5.0% significance level, if the true difference between the groups was at least two points for the primary outcome measure, the reduction in the Core Outcome Measures Index (COMI) score (where two to three points are considered as the minimal clinically important difference for the COMI) [32, 34]. This was based on the assumption that the standard deviation of the response variable, i.e. the reduction in COMI from pre-operative to 12 months post-operative, was 2.8 points.

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

The significance of the difference between the fusion and disc arthroplasty groups for continuous, normally distributed data was analysed using unpaired Student's *t*-tests or repeated measures analysis of variance (for pre/post measures). Chi-square contingency analyses were used to analyse the association between surgical group and categorical variables. The global outcome was dichotomised into "good" (=operation helped or helped a lot) and "poor" (=operation only helped a little, did not help, made things worse) for the purposes of some of the subsequent analyses [33]. Multivariable linear regression analysis (with simultaneous entry of relevant variables) was used to predict the change in COMI score at FU. Age, comorbidity, number of levels affected and the baseline COMI score were entered as control variables (since they were identified as potential confounders) and treatment group (fusion 0 vs arthroplasty 1) as the independent variable of interest.

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

Results

Final study groups

In the years 2005[1] to 2008, the overall compliance rate for all surgeons' completion of SSE Surgical Forms in our Spine Center was 85%.

[1] Only one surgeon began with the registry in 2004; the remainder began participating in 2005, hence the compliance numbers for the whole Spine Center are only given for 2005–2008.

Three hundred and forty-two patients in the database satisfied the surgical admission criteria: 269 fusion/stabilisation and 73 disc arthroplasty. For the patients in the fusion/stabilisation group, 91% received autologous bone only, 1% allogenic bone, 1% both autogenic and allogenic, 1% both autogenic and bone substitute and 6% other/no fusion material. 231/269 (86%) received some sort of anterior stabilisation: 21% an interbody cage [mostly either a Harms titanium cage (dePuy) or PEEK (Medtronic) cage], 66% plates, and 13% both a cage and plates. In the disc arthroplasty group, 70% prostheses used were Prestige II (Medronic Sofamor Danek), 22% were Discover (dePuy), 5% Bryan Cervical Disc (Medtronic Sofamor Danek) and 3% Prodisc-C (Synthes-Spine). The baseline data for each treatment group are shown in Table 1. Gender distribution did not differ significantly between the groups, but the disc arthroplasty group was significantly younger than the fusion group by approximately 10 years; they also more frequently had mono-segmental pathology (69% cases) than did the fusion patients (47% cases) and only one type of degenerative pathology (69% in the arthroplasty group, compared with 46% in the fusion group) (Table 1). Comorbidity was significantly less in the arthroplasty group than in the fusion group (Table 1). The baseline status, as assessed with the multidimensional COMI, was not significantly different between the groups.

Table 1 Baseline characteristics of the two treatment groups

Variable	Fusion ($N = 269$)	Disc arthroplasty ($N = 73$)	P value
Age, mean (SD) years	56.1 (10.8)	45.8 (7.9)	**0.0001**
Gender (% M)	50.6	46.6	0.55
No. of affected segments			
With 1 segment (%)	46.5	68.5	**0.0008**
With 2–3 segments (%)	53.5	32.5	
Number of degenerative pathologies specified			
With 1 pathology (%)	45.7	68.5	**0.0006**
With 2 pathologies (%)	33.1	26.0	
With >2 pathologies (%)	21.2	5.5	
Previous surgery same level			
Yes (%)	7.4	4.1	0.31
Morbidity status			
ASA 1 (%)	29.2	68.1	**0.001**
ASA 2 (%)	58.7	30.5	
ASA 3 (%)	12.1	1.4	
COMI score, mean (SD)	6.9 (2.1)	7.2 (2.0)	0.37

P values marked bold are significant, $P < 0.05$

Table 2 Group differences in surgical details

Variable	Fusion (%)	Disc arthroplasty (%)	P value
Operation duration (%)			
<1 h	0.8	5.6	**0.0001**
1–2 h	38.3	67.6	
2–3 h	39.8	25.4	
3–4 h	14.0	1.4	
>4 h	7.1	0.0	
Blood loss (%)			
None	20.1	31.0	0.16
<500 ml	77.6	69.0	
500–1,000 ml	1.9	0.0	
1,000–2,000 ml	0.4	0.0	
General complications (intra/perioperative)	2.2[a]	0.0	0.35
Surgical complications (intra/perioperative)	2.2[b]	1.4[c]	0.99

For complications, the Fisher exact test was used to determine P values; P values in bold are significant, $P < 0.05$

[a] One cardiovascular, one pulmonary, one cardiovascular and pulmonary, one kidney, two other

[b] Two nerve root damage, one bleeding outside spinal canal, three other

[c] One nerve root damage

Operation details

Operation duration was significantly shorter in the arthroplasty group, with 73% operations lasting less than 2 h compared with only 39% lasting less than 2 h in the fusion group (Table 2). The perioperative complication rates were not significantly different between the groups (general complications: arthroplasty, 0.0%; fusion, 2.2%; surgical complications: arthroplasty, 1.4%; fusion, 2.2%).

Patient-rated outcomes

At the time of the current evaluation, FU data were available from 266 (208 fusion, 58 arthroplasty) of the 284 patients who had reached 12 months (94% FU rate) and from 169 (139 fusion, 30 arthroplasty) of the 178 patients who had reached 24 months post-operative (95% FU rate).

Figure 1 shows the change in the individual COMI domain scores and the COMI composite scores, at baseline and at 12 months FU, for each treatment group. A significant improvement ($P < 0.05$) in scores was recorded for both groups, from baseline to 12 months post-operative. There was no significant difference between the groups in the extent of the reduction in neck pain, arm pain, or "worst" pain (arm or neck) ($P > 0.05$); however, the

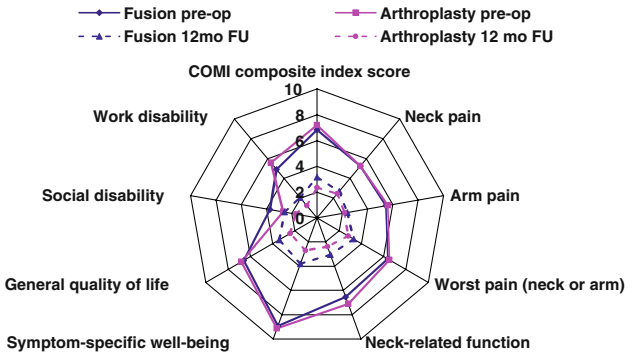

Fig. 1 Scores for each of the COMI domains, and the COMI composite score, before and 12 months after surgery in the arthroplasty and fusion groups. *0* denotes best score, *10* denotes worst score. The score reductions in all domains, except the pain domains, were significantly greater in the arthroplasty than the fusion group ($P < 0.05$). See text for further details

improvement in scores for function, symptom-specific well-being, general quality of life, and work and social disability was significantly greater in the arthroplasty group ($P < 0.05$).

At 12 months post-surgery, the reduction from baseline values in the COMI composite score was 1.1 points greater in the arthroplasty group than in the fusion group (reduction in COMI scores: arthroplasty, 4.8 (SD 3.0) points versus fusion, 3.7 (SD 2.9) points) (Table 3). The difference was statistically significant ($P = 0.007$) but less than the defined minimal clinically important difference of two points (see statistics section). A "good" global outcome was recorded in 90% patients in the arthroplasty group and 80% in the fusion group, with the difference just failing to reach significance ($P = 0.09$). Satisfaction with treatment was similar in both groups (arthroplasty 90%; fusion 89%, respectively; n.s.).

The results at 24 months FU generally mirrored those after 12 months (Table 3), with a statistically significant 1.3-point difference between the groups, favouring arthroplasty, in the reduction in COMI composite score from baseline values. There was again a tendency for better global outcome ratings in the disc arthroplasty group (93% good outcome) compared with the fusion group (82% good outcome) ($P = 0.17$); satisfaction was similarly high in both groups (90–93% satisfied; no significant group difference).

Additional analyses were carried out to examine whether outcome was influenced by various characteristics that differed significantly between the two groups at baseline, and that might act as confounders and need to be controlled for in multivariable analyses. For this, the data at 12-months FU were used, since more cases were available and the results did not change substantially between 12 and 24 months.

Table 3 Group differences with respect to outcome at 12 and 24 months post-surgery

Variable	Fusion (%)	Disc arthroplasty (%)	*P* value
12-Month global outcome (%)			
Good	80.3	89.7	*0.09*
Poor	19.7	10.3	
12-Month satisfaction (%)			
Good	89.4	89.7	0.96
Poor	10.6	10.3	
Change in COMI score 0–12 months (mean (SD))	3.7 (2.9)	4.8 (3.0)	**0.007**
24-Month global outcome (%)			
Good	82.0	93.3	0.17[a]
Poor	18.0	6.7	
24-Month satisfaction (%)			
Good	89.9	93.3	0.74[a]
Poor	10.1	6.7	
Change in COMI score 0–24 months, mean (SD)	3.8 (2.9)	5.1 (2.8)	**0.03**

Results at 12 months are from 208 fusion patients, 58 disc arthroplasty patients; at 24 months, 139 fusion patients, 30 disc arthroplasty patients; *P* values in bold are significant $P < 0.05$, in italics are borderline significant, $P < 0.10$

[a] Fisher's exact test used to determine *P* values

Gender, number of degenerative pathologies, and previous surgery at the same level had no significant influence on outcome (Table 4; similar results when analysed on a treatment group basis). An increasing number of affected segments ($P = 0.014$) and comorbidity status ($P = 0.042$) were each associated with worse outcome, and there was a tendency for age to play a role (better outcome in younger patients, $P = 0.09$; Table 4), especially in the arthroplasty group (separate group details not shown). Since the latter three baseline variables also showed significant differences between the treatment groups (Table 1), they were controlled for in the multivariable analyses when examining the unique influence of treatment group on outcome.

The multivariable model explained 23% variance in the change in COMI score from pre-operative to 12 months FU, with treatment group representing a unique predictor with borderline significance ($P = 0.059$); other important predictors were the baseline COMI score ($P = 0.0001$), comorbidity score ($P = 0.042$) and number of affected segments ($P = 0.075$) (Table 5). The regression coefficient for the variable "treatment group" (0.801) indicated that, compared with the fusion group, the arthroplasty group showed an approximately 0.8-point greater reduction in COMI score after 12 months, when the other potential confounders were controlled for.

Table 4 Association between potential confounders and outcome for patients from both groups together: reduction in COMI scores (pre-operative to 12 months post-operative) in different categories of the suspected confounders

Variable	Reduction in COMI score, pre-surgery to 12 months post-surgery[a]	*P* value
Age (years)		
<50	4.3 ± 3.0	*0.09*
>50	3.7 ± 2.9	
Gender		
Male	3.7 ± 2.9	0.21
Female	4.2 ± 3.0	
No. of affected segments		
One segment	4.4 ± 3.1	**0.014**
Two to three segments	3.5 ± 2.8	
Number of degenerative pathologies specified		
One	4.1 ± 3.1	0.72
Two	3.9 ± 2.9	
Three or more	3.7 ± 2.7	
Previous surgery same level (%)		
No	4.0 ± 2.9	0.71
Yes	3.7 ± 3.2	
Comorbidity, ASA score (%)		
I (no disturbance)	4.5 ± 2.7	**0.042**
II (mild/moderate)	3.7 ± 3.1	
III (severe)	3.0 ± 2.8	

P values in bold are significant $P < 0.05$, in italics are borderline significant, $P < 0.10$

[a] The greater the reduction in COMI score, the better the outcome

Similar findings were observed using the reduction in COMI score at 24 months as the dependent (outcome) variable, with the unique contribution of the variable "treatment group" again being of borderline statistical significance in the model ($P = 0.055$; adjusted R^2 for the whole model, 22.3%). The regression coefficient for "treatment group" was similar to that for the 12-month data—an approximately 1.2-point greater reduction in COMI score after 24 months in the arthroplasty than the fusion group, after controlling for confounders (detailed data not shown).

Patient-rated complications and re-operation rates

At 12-months FU, 26.1% patients in the fusion group and 19.0% in the arthroplasty group ($P = 0.27$) reported in their questionnaire that complications had arisen as a consequence of the index operation (most commonly sensory disturbances, general neurological, continued/new pain, problems with wound healing); at 24 months, the figures were 23 and 7% ($P = 0.045$), respectively.

At 12-months FU, the proportions of patients reporting re-operation at the same or at a different segment of the spine were, respectively, 2.4 and 2.4% for the fusion group and 1.7 and 1.7% for the arthroplasty group; at 24 months FU, the figures were, respectively, 3.6 and 5.1% for the fusion group and 0 and 3.3% for the arthroplasty group.

Discussion

Randomised controlled trials are considered to represent the pinnacle in the hierarchy of evidence, when evaluating the efficacy of therapeutic interventions [43]. However, for many medical questions of interest, a large amount of evidence is often accumulated through non-randomised studies, and these can be used to help in the interpretation of the randomised results [28]. There is substantial debate in the literature as to whether non-randomised studies deliver comparable results to those of randomised controlled trials on the same topic [28]; earlier reviews suggested that non-randomised studies may spuriously overestimate treatment benefits [12, 35], whilst more recent investigations maintain that for selected topics they generally deliver comparable results to RCTs [6, 13], especially for prospective studies [28]. When assessing trials in orthopaedic surgery—an area that does not lend itself readily to the trialling of its treatment methods—this issue is of considerable relevance.

The overall results of this observational study appear to support those of the RCTs on the same theme [25, 29, 37, 38, 45, 46] and suggest that, for degenerative pathology of the cervical spine, disc arthroplasty is associated with at least equivalent, and sometimes slightly better, patient-rated outcomes than fusion, up to 2 years after surgery. As far as superiority is concerned, interpretation of the results—both ours and those of the previous RCTs—demands that attention is paid to the clinical relevance, and not just statistical significance, of the group differences observed. The minimum clinically important difference for the primary outcome used in most of the RCTs, the NDI (0–100 scale), is 15–19 points [5, 11], yet in most of the trials the group difference for the improvement in NDI score after 2 years was just a fraction of this, between two and seven points [25, 45, 46]. Similarly, differences between the groups in the improvement of arm and neck pain scores of just 6–15 points (on the 0–100-point scale) were recorded after 2 years [45, 46], yet the minimum clinically important difference for such scales is reported to be 13–20 points [11, 24, 39]. In our own study, the adjusted group difference in the reduction in COMI score after either 12 months or 24 months was approximately one point (on the 0–10 scale), again failing to reach the minimal clinically important difference of approximately two

Table 5 Results of the multiple regression analysis to examine the influence of treatment group on the COMI score change 12 months after surgery

Independent variables	Unstandardised regression coefficients	95% CI for B		Standardised coefficients	Sig (*P* value)	% Explained variance in COMI change score pre-operative to 12 months
	B	CI low	CI high	Beta		Adj R^2
(Constant)	−0.147				0.921	23.6
Baseline COMI score	0.618	0.470	0.766	0.449	**0.0001**	
Age	0.025	−0.009	0.061	0.089	0.164	
Comorbidity (ASA score)	−0.608	−1.179	−0.008	−0.127	**0.042**	
No. affected segments	−0.581	−1.225	0.055	−0.099	*0.075*	
Treatment group (1 fusion, 2 arthroplasty)	0.801	0.041	1.694	0.113	*0.059*	

P values in bold are significant unique predictors in the multivariate model, $P < 0.05$; in italics are borderline significant, $P < 0.10$

points [32, 34]. Hence, the most appropriate conclusion from all these studies is that the two procedures are comparable in terms of their mid-term patient-rated outcomes. This must also be viewed in the face of the a priori more favourable outcome expected in the typical arthroplasty patient due to the more rigorous selection criteria for this treatment (younger patients, with less comorbidity, less extensive degenerative changes, fewer segments affected, etc.).

The differing results for the constructs "global treatment outcome/effectiveness" (which was borderline statistically significant) and "satisfaction with care" (not significant) emphasise the importance of differentiating between these two closely related, but distinct entities. Satisfaction tends to focus more on the provision of care or treatment delivery—which is strongly influenced by factors such as the patient-provider relationship and may include an expression of appreciation for the surgeon "having done his best"—than on the *effect of treatment*, which instead focuses on therapeutic improvement (symptom or functional), in terms of how much the surgery helped the back problem [21, 33]. Since, in the present study, the surgeons and the infrastructure were identical for both groups, similar results for "satisfaction" were perhaps expected. These subtle differences should be borne in mind when interpreting the outcome results of different studies, especially when satisfaction is used to indicate "effectiveness" in the assessment of new innovations in unblinded trials [9]. Our measure of global treatment outcome is likely comparable to the "overall success" ratings used in the previous RCTs, in which similar group differences (approximately 10% more arthroplasty patients with a "good/successful" outcome [25, 37]) were reported: in the present study, 90% arthroplasty vs 80% fusion patients had a good outcome at 12 months, and 93 vs 82%, respectively, at 24 months.

Segmental fusion as the treatment for painful degenerative cervical spine disorders has been the "gold standard" for years. Consistently, good results have been reported with this approach, even in the long-term [10], and, compared with the general surgical outcomes of degenerative conditions of the *lumbar* spine, it is associated with low complication rates [3, 7, 10, 45]. Hence, the threshold for advancement with the introduction of a new implant was relatively high, and improvement on the already excellent clinical results a challenge. As always in these situations, it is crucial to discuss how the parameters of "outcome" should be defined. As opposed to focusing on segmental motion, the degree of fusion, and the size of osteophytes—each of which is subject to considerable discussion regarding its optimal method of measurement—in the present study we concentrated exclusively on the subjective outcome ratings of the patients. This was done because the ultimate goal of our surgical practice should be good patient-rated outcomes as opposed to just technical success. Whether technical improvements brought about by motion preservation using the disc prosthesis will translate into better patient-outcomes in the long-term and whether these will be tempered by any (as yet poorly investigated) factors such as the build up of wear debris, fatigue failure, the influence on facet joint biomechanics, heterotopic ossification, the need and options for revision, the influence of subsequent osteoporotic changes, etc. will require considerably longer FU investigations. Currently, the incidence of these potential late complications is not known [14]; there are only limited in vitro data on wear properties of the cervical disc prosthesis [4] and the longest clinical FU for a sizeable group of patients is just 4–6 years, with the data currently published in abstract form only [22]. The influence of the specific design of prosthesis should also be further investigated; in the present study, all types of prostheses were examined as one group and compared with

all techniques of fusion—a possible limitation from the scientific methodology point of view, but a pragmatic approach that ideally reflected the everyday practice of a large Spine Center.

Our data on patient self-assessed re-operation rates and complications after surgery—a relatively new concept in the field of outcomes research, and one that reveals some startling results [23]—suggested slightly more favourable results in the arthroplasty group, especially by 2 years FU; however, the group sizes were too small for meaningful statistical analyses of the data. Again, observational studies, using the data collected in large-scale registries such as the European Spine Tango, are well disposed to address some of these issues and to complement the findings of the RCTs. As already mentioned, single trials are usually underpowered to adequately address the risks of adverse events, especially uncommon ones [40].

RCTs are considered to be the most scientifically rigorous way of examining the efficacy of a new implant or technique. Using such a study design, it is assumed that confounders and bias are excluded in an appropriate fashion and that the results best approach "the truth". This might be true for the testing of new drugs (where the design of RCTs originated), but it is questionable in the surgical field and perhaps explains why some randomised prospective trials in spine surgery have delivered such contradictory results. For example, the introduction of intradiscal electrothermal therapy (IDET) treatment was launched with a RCT [41] that indicated the positive effect of the procedure. However, a few years later, another RCT [18, 19] compared the new procedure with placebo treatment and found absolutely no clinically relevant benefit of IDET compared with the control. This situation might indicate that RCTs are not able to circumvent all the potential co-factors that might influence a clinical result, or that sample sizes are not always adequate to achieve sufficient statistical power, or that, in the interpretation of the results, clinically relevant differences are not distinguished from statistically significant (but irrelevant) differences. Interestingly, a recent meta-analysis suggested that, even within randomised trials, the ones with greater methodological rigour showed no benefit while the ones with potential flaws had spuriously overestimated the benefit of treatment [47].

Randomised controlled trials often involve only a limited selection of the typical patient population suffering from the condition, and hence have limited external validity, i.e. the results cannot always be assumed to apply to the "patient at large" [8, 30]. Further, if surgeons are required to implement the "old" technique and are not allowed to use the "new" procedure, solely for the purposes of a randomised trial, then factors concerned with experience and expectations may influence the overall outcomes. On the patients' part, the same mechanism of expectation (having the "bad luck" to be treated with the "old" technique) might influence his or her rating of the global outcome and satisfaction with treatment.

In summary, although the study design used in the present investigation cannot boast the "scientific rigour" of a randomised controlled trial and does not represent the highest in the hierarchy of evidence, it included every single eligible patient being operated in our Spine Center, reflecting the everyday clinical reality much more accurately than an "artificial" set-up within a RCT. In observational studies, it is never possible to identify and account for all potential confounders; however, those that were identified were dealt with accordingly in the multivariable statistical analysis [36]. Matched-pair analyses of the same data (an alternative, though not necessarily superior method [36]) revealed similar findings to those presented here (data not shown), giving credence to the statistical methods used to control for the effect of confounders. The similarity of the results obtained to those of the published RCTs on cervical disc prostheses suggests that RCTs may not be the only method to evaluate new technologies in spine surgery. We suggest that it is inappropriate to (indiscriminately) discredit observational studies as a relevant source of evidence in spine surgery.

Acknowledgements We are grateful to Gordana Balaban, Julian Amacker and David O'Riordan for their excellent work collecting the questionnaire data and managing our quality management system. The study was supported by the Schulthess Klinik Research Fund.

References

1. Ahn H, Bhandari M, Schemitsch EH (2009) An evidence-based approach to the adoption of new technology. J Bone Joint Surg Am 91(Suppl 3):95–98
2. Ahn H, Court-Brown CM, McQueen MM, Schemitsch EH (2009) The use of hospital registries in orthopaedic surgery. J Bone Joint Surg Am 91(Suppl 3):68–72
3. Anderson PA, Sasso RC, Riew KD (2008) Comparison of adverse events between the Bryan artificial cervical disc and anterior cervical arthrodesis. Spine 33:1305–1312
4. Anderson PA, Sasso RC, Rouleau JP, Carlson CS, Goffin J (2004) The Bryan Cervical Disc: wear properties and early clinical results. Spine J 4:303S–309S
5. Anderson PA, Subach BR, Riew KD (2009) Predictors of outcome after anterior cervical discectomy and fusion: a multivariate analysis. Spine 34:161–166
6. Benson K, Hartz AJ (2000) A comparison of observational studies and randomized, controlled trials. N Engl J Med 342:1878–1886
7. Bhadra AK, Raman AS, Casey AT, Crawford RJ (2009) Single-level cervical radiculopathy: clinical outcome and cost-effectiveness of four techniques of anterior cervical discectomy and fusion and disc arthroplasty. Eur Spine J 18:232–237
8. Black N (1996) Why we need observational studies to evaluate the effectiveness of health care. BMJ 312:1215–1218

9. Blumenthal S, McAfee PC, Guyer RD, Hochschuler SH, Geisler FH, Holt RT, Garcia R Jr, Regan JJ, Ohnmeiss DD (2005) A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes. Spine 30:1565–1575 discussion E1387–E1591
10. Bohlman HH, Emery SE, Goodfellow DB, Jones PK (1993) Robinson anterior cervical discectomy and arthrodesis for cervical radiculopathy. Long-term follow-up of one hundred and twenty-two patients. J Bone Joint Surg Am 75:1298–1307
11. Cleland JA, Childs JD, Whitman JM (2008) Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. Arch Phys Med Rehabil 89:69–74
12. Colditz GA, Miller JN, Mosteller F (1989) How study design affects outcomes in comparisons of therapy. I: medical. Stat Med 8:441–454
13. Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342:1887–1892
14. Denaro V, Papalia R, Denaro L, Di Martino A, Maffulli N (2009) Cervical spinal disc replacement. J Bone Joint Surg Br 91:713–719
15. 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
16. Eck JC, Humphreys SC, Lim TH, Jeong ST, Kim JG, Hodges SD, An HS (2002) Biomechanical study on the effect of cervical spine fusion on adjacent-level intradiscal pressure and segmental motion. Spine 27:2431–2434
17. 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 discussion 1380
18. Freeman BJ (2006) IDET: a critical appraisal of the evidence. Eur Spine J 15(Suppl 3):S448–S457
19. Freeman BJ, Fraser RD, Cain CM, Hall DJ, Chapple DC (2005) A randomized, double-blind, controlled trial: intradiscal electrothermal therapy versus placebo for the treatment of chronic discogenic low back pain. Spine 30:2369–2377 discussion 2378
20. Fuller DA, Kirkpatrick JS, Emery SE, Wilber RG, Davy DT (1998) A kinematic study of the cervical spine before and after segmental arthrodesis. Spine 23:1649–1656
21. George SZ, Hirsch AT (2005) Distinguishing patient satisfaction with treatment delivery from treatment effect: a preliminary investigation of patient satisfaction with symptoms after physical therapy treatment of low back pain. Arch Phys Med Rehabil 86:1338–1344
22. Goffin J, van Loon J, Van Calenbergh F (2006) Cervical arthroplasty with the Bryan disc: 4-and 6-year results. 21st Annual meeting of the North American Spine Society, Seattle
23. Grob D, Mannion AF (2009) The patient's perspective on complications after spine surgery. Eur Spine J 18(Suppl 3):380–385
24. Hagg O, Fritzell P, Nordwall A, Group SLSS (2003) The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 12:12–20
25. Heller JG, Sasso RC, Papadopoulos SM, Anderson PA, Fessler RG, Hacker RJ, Coric D, Cauthen JC, Riew DK (2009) Comparison of BRYAN cervical disc arthroplasty with anterior cervical decompression and fusion: clinical and radiographic results of a randomized, controlled, clinical trial. Spine 34:101–107
26. Hoppe DJ, Schemitsch EH, Morshed S, Tornetta P III, Bhandari M (2009) Hierarchy of evidence: where observational studies fit in and why we need them. J Bone Joint Surg Am 91(Suppl 3):2–9
27. Horwitz RI (1987) Complexity and contradiction in clinical trial research. Am J Med 82:498–510
28. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J (2001) Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 286:821–830
29. Kim SW, Limson MA, Kim SB, Arbatin JJ, Chang KY, Park MS, Shin JH, Ju YS (2009) Comparison of radiographic changes after ACDF versus Bryan disc arthroplasty in single and bi-level cases. Eur Spine J 18:218–231
30. Landewe R, van der Heijde D (2007) Primer: challenges in randomized and observational studies. Nat Clin Pract Rheumatol 3:661–666
31. 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
32. 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
33. Mannion AF, Porchet F, Kleinstück F, Lattig F, Jeszenszky D, Bartanusz V, Dvorak J, Grob D (2009) The quality of spine surgery from the patient's perspective: part 1. The Core Outcome Measures Index (COMI) in clinical practice. Eur Spine J 18(Suppl 3):367–373
34. Mannion AF, Porchet F, Kleinstuck FS, Lattig F, Jeszenszky D, Bartanusz V, Dvorak J, Grob D (2009) The quality of spine surgery from the patient's perspective: part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. Eur Spine J 18(Suppl 3):374–379
35. Miller JN, Colditz GA, Mosteller F (1989) How study design affects outcomes in comparisons of therapy. II: surgical. Stat Med 8:455–466
36. Morshed S, Tornetta P III, Bhandari M (2009) Analysis of observational studies: a guide to understanding statistical methods. J Bone Joint Surg Am 91(Suppl 3):50–60
37. Mummaneni PV, Burkus JK, Haid RW, Traynelis VC, Zdeblick TA (2007) Clinical and radiographic analysis of cervical disc arthroplasty compared with allograft fusion: a randomized controlled clinical trial. J Neurosurg Spine 6:198–209
38. Murrey D, Janssen M, Delamarter R, Goldstein J, Zigler J, Tay B, Darden B (2009) Results of the prospective, randomized, controlled multicenter Food and Drug Administration investigational device exemption study of the ProDisc-C total disc replacement versus anterior discectomy and fusion for the treatment of 1-level symptomatic cervical disc disease. Spine J 9:275–286
39. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, Bouter LM, de Vet HC (2008) Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine 33:90–94
40. Papanikolaou PN, Christidi GD, Ioannidis JP (2006) Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ 174:635–641
41. Pauza KJ, Howell S, Dreyfuss P, Peloza JH, Dawson K, Bogduk N (2004) A randomized, placebo-controlled trial of intradiscal electrothermal therapy for the treatment of discogenic low back pain. Spine J 4:27–35
42. Porchet F, Metcalf NH (2004) Clinical outcomes with the Prestige II cervical disc: preliminary results from a prospective randomized clinical trial. Neurosurg Focus 17:E6
43. Sackett DL (1998) Evidence-based medicine. Spine 23:1085–1086

44. Sasso RC, Best NM, Metcalf NH, Anderson PA (2008) Motion analysis of Bryan cervical disc arthroplasty versus anterior discectomy and fusion: results from a prospective, randomized, multicenter, clinical trial. J Spinal Disord Tech 21:393–399
45. Sasso RC, Smucker JD, Hacker RJ, Heller JG (2007) Artificial disc versus fusion: a prospective, randomized study with 2-year follow-up on 99 patients. Spine 32:2933–2940 discussion 2941–2932
46. Sasso RC, Smucker JD, Hacker RJ, Heller JG (2007) Clinical outcomes of BRYAN cervical disc arthroplasty: a prospective, randomized, controlled, multicenter trial with 24-month follow-up. J Spinal Disord Tech 20:481–491
47. Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273:408–412
48. Szpalski M, Gunzburg R, Mayer M (2002) Spine arthroplasty: a historical review. Eur Spine J 11(Suppl 2):S65–S84
49. White P, Lewith G, Prescott P (2004) The core outcomes for neck pain: validation of a new outcome measure. Spine 29:1923–1930