ORIGINAL ARTICLE

Development of appropriateness criteria for the surgical treatment of symptomatic lumbar degenerative spondylolisthesis (LDS)

A. F. Mannion · V. Pittet · F. Steiger · J.-P. Vader · H.-J. Becker · F. Porchet · The Zürich Appropriateness of Spine Surgery (ZASS) Group

Received: 11 July 2013 / Revised: 22 December 2013 / Accepted: 22 March 2014 / Published online: 24 April 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Purpose Spine surgery rates are increasing worldwide. Treatment failures are often attributed to poor patient selection and inappropriate treatment, but for many spinal disorders there is little consensus on the precise indications for surgery. With an aging population, more patients with lumbar degenerative spondylolisthesis (LDS) will present for surgery. The aim of this study was to develop criteria for the appropriateness of surgery in symptomatic LDS.

Methods A systematic review was carried out to summarize the current level of evidence for the treatment of LDS. Clinical scenarios were generated comprising combinations of signs and symptoms in LDS and other relevant variables. Based on the systematic review and their own clinical experience, twelve multidisciplinary international experts rated each scenario on a 9-point scale (1 highly inappropriate, 9 highly appropriate) with respect to performing decompression only, fusion, and instrumented fusion. Surgery for each theoretical scenario was classified as appropriate, inappropriate, or uncertain based on the median ratings and disagreement in the ratings.

Results 744 hypothetical scenarios were generated; overall, surgery (of some type) was rated appropriate in 27 %, uncertain in 41 % and inappropriate in 31 %. Frank panel disagreement was low (7 % scenarios). Face validity was shown by the logical relationship between each variable's subcategories and the appropriateness ratings, e.g., no/mild disability had a mean appropriateness rating of 2.3 ± 1.5, whereas the rating for moderate disability was 5.0 ± 1.6 and for severe disability, 6.6 ± 1.6. Similarly, the average rating for no/minimal neurological abnormality was 2.3 ± 1.5, increasing to 4.3 ± 2.4 for moderate and 5.9 ± 1.7 for severe abnormality. The three variables most likely ($p < 0.0001$) to be components of scenarios rated "appropriate" were: severe disability, no yellow flags, and severe neurological deficit.

Conclusion This is the first study to report criteria for determining candidacy for surgery in LDS developed by a multidisciplinary international panel using a validated method (RAM). The panel ratings followed logical clinical rationale, indicating good face validity. The work refines clinical classification and the phenotype of degenerative spondylolisthesis. The predictive validity of the criteria should be evaluated prospectively to examine whether patients treated "appropriately" have better clinical outcomes.

Keywords Lumbar degenerative spondylolisthesis · Appropriateness of surgery · RAND appropriateness method (RAM)

Introduction

Symptomatic lumbar degenerative spondylolisthesis (LDS) is a spinal pathology that presents a common problem in daily spinal practice. The combination of osteoarthritic and degenerative changes in the disc and facet joints causes

The Zürich Appropriateness of Spine Surgery (ZASS) Group:
F. Balagué, J. I. Brox, C. Cedraschi, J. Fairbank, T. F. Fekete, P. Fritzell, D. Jeszenszky, J. Lurie, F. Pellisé, C. Reitmann, V. Sonntag, C. Standaert, M. Szpalski.

A. F. Mannion (✉) · F. Steiger · H.-J. Becker · F. Porchet
Spine Center, Schulthess Klinik Zürich, Lengghalde 2,
8008 Zürich, Switzerland
e-mail: anne.mannion@yahoo.com

V. Pittet · J.-P. Vader
Institute of Social and Preventive Medicine,
University of Lausanne, Lausanne, Switzerland

anterior vertebral displacement of one vertebral body over another and ensuing spinal stenosis [1]. The resulting symptoms are usually a combination of stenotic-type radiating buttock and leg pain and mechanical low back pain. Conservative management is usually applied in the first instance, but if unsuccessful, surgery is often advocated [2–4]. In earlier years, decompression was the most common type of surgical procedure used [5], followed by decompression and uninstrumented fusion [6]. However, LDS is now considered by many to represent an inherent instability of the lumbar spine, with the commonly recommended treatment being a combination of decompression and instrumented lumbar fusion [4, 7–10]. Both a systematic review of comparative non-randomized studies [1] and the subsequent guidelines of NASS [4] suggest that a satisfactory clinical outcome is significantly more likely with fusion than with decompression alone. However, the studies included in the systematic review were of low methodological quality, with only short- to mid-term follow-up, and outcome was not always assessed using validated methods. The use of adjunctive instrumentation was reported to increase the probability of attaining solid fusion, but it did not result in significant improvements in clinical outcomes [1]. A recent high-quality study with 4-year follow-up also confirmed that there were no consistent differences in clinical outcome dependent on the type of fusion (instrumented or not) [11].

LDS occurs mainly in elderly patients, in whom comorbidities are common, and the use of fusion increases the perioperative risk [12]. To avoid these risks, less invasive surgical treatment such as decompression alone is sometimes advocated, especially in the face of predominantly stenotic or radiating pain symptoms [13, 14]. However, it is unclear whether such an approach compromises outcome. Many questions remain concerning the appropriate management of LDS and the extent of surgery needed in any individual case. In clinical practice, the decision is often influenced by the age of the patient, their comorbidity, duration of symptoms, degree of slip, stability of the slip during flexion–extension imaging, and presenting symptoms—though such decisions are rarely based on hard evidence. It is often believed that patients with mainly symptoms of neural compression due to stenosis may benefit from simple decompression and forego more extensive fusion surgery, despite the underlying slippage. However, there is little evidence to substantiate this view. Similarly, fusion is often advocated in the face of significant “instability”, yet there is little consensus in the literature as to how this should be defined or measured.

In spine surgery, many treatment failures are attributable to poor patient selection and the application of inappropriate treatment [15, 16]. A procedure is considered appropriate when the expected health benefits (quality of life or life span, reduced pain, improved functioning) exceed the potential risks (mortality, morbidity, pain, impairment, anxiety caused by the procedure, time lost from work) by a sufficiently wide margin that the procedure is worth performing for the patient, exclusive of cost [17]. Many clinical practice guidelines have been developed to help define what care is appropriate in different fields of medicine [18], but often they do not rest on solid scientific evidence or explicit, validated methods. The lack of availability of good quality RCTs in the field of spine surgery leaves us with notable gaps in our knowledge, and in clinical practice many decisions have to be taken without the benefit of high-quality evidence. Determining the appropriateness of surgery must, therefore, depend on other methods. One such alternative is the RAND appropriateness method (RAM) [19], which combines a detailed review of the literature with a modified Delphi panel approach to gage collective expert opinion. It is one of the most respected methods for defining appropriate medical care [20]. A major strength of the RAM is the level of clinical detail that can be attained in the subsequent recommendations, increasing their acceptability for clinicians. The RAM is considered most useful for procedures that are used frequently, associated with a substantial amount of morbidity and/or mortality, consume significant resources, have wide variations among geographic areas in rates of use and whose use is controversial [17]. All these criteria apply to the procedures commonly used in the surgical treatment of LDS [21]. The aim of the present study was to use the RAM with a multispecialty panel to develop explicit criteria for the appropriate management of individual patients with LDS.

Methods

The panel followed the standardized procedure for the RAM [19] (Fig. 1).

Systematic literature review

First, a systematic literature search was carried out to identify studies evaluating the efficacy/effectiveness, outcome, safety, side effects, and complications of surgery for LDS. This was written up as a research article [22] for later use by the expert panel (see below). Searches were carried out for articles in the English or German language in Medline, Cochrane Library, Embase, and Cinahl from 1990 to 2012. Medical sub-headings were used as search terms, including spondylolisthesis, lumbar vertebra OR back, degenerative, lumbar, decompression surgery, spine fusion, laminectomy, laminotomy, arthrodesis, surgical technique, lumb* OR uninstrumented fusion. The bibliographies of

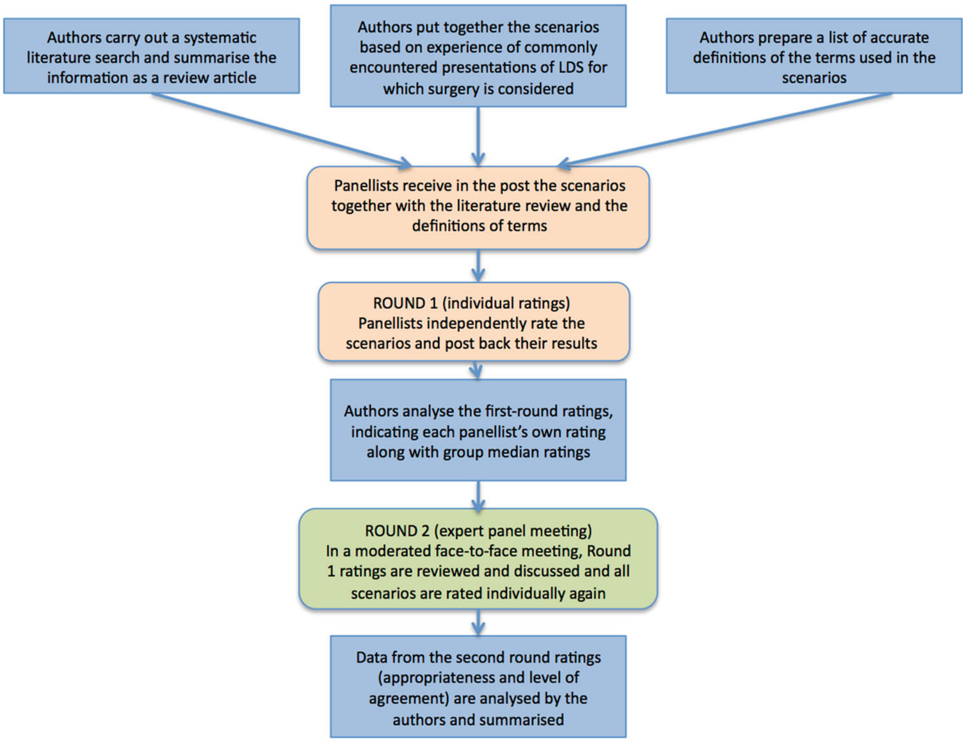
Authors carry out a systematic literature search and summarise the information as a review article
Authors put together the scenarios based on experience of commonly encountered presentations of LDS for which surgery is considered
Authors prepare a list of accurate definitions of the terms used in the scenarios
Panellists receive in the post the scenarios together with the literature review and the definitions of terms
ROUND 1 (individual ratings) Panellists independently rate the scenarios and post back their results
Authors analyse the first-round ratings, indicating each panellist's own rating along with group median ratings
ROUND 2 (expert panel meeting) In a moderated face-to-face meeting, Round 1 ratings are reviewed and discussed and all scenarios are rated individually again
Data from the second round ratings (appropriateness and level of agreement) are analysed by the authors and summarised

Fig. 1 Summary of the RAM procedure used in the study

retrieved articles and relevant conference proceedings were also reviewed.

Clinical scenarios

Based on the literature review and in consultation with the authors' own Spine Center team, a detailed catalog of potential scenarios (i.e., signs-and-symptoms profiles) for which elective surgery might be used/proposed in connection with LDS was prepared (cauda equina syndrome was excluded, as this was considered as an emergency procedure unquestionably requiring surgery). The theoretical scenarios were to be viewed in consideration of LDS as the most distinct radiographic finding associated with the symptoms, i.e., for the case where symptomatic LDS was the only or the predominant diagnosis. Ten major groups of clinical presentations ("chapters") were created based on the main signs and symptoms: (1) back pain only without significant instability; (2) back pain only with significant instability; (3) radicular pain without back pain, without significant instability; (4) radicular pain without back pain, with significant instability; (5) radicular pain with back pain, without significant instability; (6) radicular pain with back pain, with significant instability; (7) neurogenic claudication without back pain, without significant instability; (8) neurogenic claudication without back pain, with significant instability; (9) neurogenic claudication with back pain, without significant instability; (10) neurogenic claudication with back pain, with significant instability.

A list of variables was then created that would allow further classification of patients in terms of the factors that surgeons take into account in deciding whether to recommend a particular procedure. This included the severity of any neurological abnormality, radiological signs of significant stenosis (central or foraminal), comorbidity status, and level of disability. Definitions of all terms employed in the formulation of indications were agreed upon and documented (Table 1). The variables were structured into ordinal-scaled levels, and a matrix of indications was generated for all possible permutations of these variables, with the intention of covering virtually all conceivable clinical patterns of LDS for which surgical treatment might be considered. Each unique combination of variables was considered an "indication" or "patient clinical scenario". In total, 372 such scenarios were prepared, to be assessed in relation to the appropriateness of each of three different surgical procedures: decompression only (including unilateral or bilateral fenestration, hemilaminectomy, laminectomy, laminarthrectomy, laminotomy, foraminotomy, discectomy, flavectomy, sequestrectomy); fusion with/without decompression (including anterior interbody fusion between adjacent vertebrae, posterolateral or posterior fusion with autologous bone graft or other fusion materials); instrumented fusion with/without decompression (including intervertebral stabilization such as transforaminal interbody fusion (TLIF), anterior lumbar interbody fusion (ALIF) or posterior lumbar interbody fusion (PLIF), extreme lateral lumbar interbody fusion (XLIF); pedicular

Table 1 Definition of variables used in the clinical presentations and in forming the clinical scenarios

Variable	Categories	Definition
Back pain	Yes/no	Presence/absence of pain on the posterior aspect of the body from the lower margin of the twelfth ribs to the lower gluteal folds; intensity ≥3 on a 0–10 pain scale
Radicular pain	Yes/no	Presence/absence of unilateral pain in the lower extremity, compatible with the involvement of specific nerve roots
Neurogenic claudication	Yes/no	Presence/absence of unilateral or bilateral pain, discomfort, perceived weakness or sensory disturbances in the lower limb or buttock, precipitated by walking or standing and relieved by flexion of the trunk, and limiting walking capacity. Vascular claudication and other conditions/joint problems limiting walking capacity have been ruled out
Instability considered to be clinically relevant	Yes/no	Although the term instability is in common use and is felt to be important by some, no commonly agreed definition of the concept could be found by the expert panel (in the study, this will be assessed in relation to the examining physician's statement in the clinical report, which should include his/her criteria used to determine the presence of instability. It is understood that these criteria may vary from surgeon to surgeon)
Severity of neurological abnormality	1	"No or minimal neurological abnormality", which includes any of the following: positive straight-leg raising test (pain distal to the knee at 60 degrees or less of passive hip flexion; includes crossed and/or straight-leg raising); reflex asymmetry at knee or ankle; mild dermatomal sensory loss in lower extremity
	2	"Moderate neurological abnormality", which includes any of the following: moderate unilateral dermatomal sensory disturbances in a lower extremity; nonprogressive unilateral muscle weakness which does not meet criteria for major neurologic abnormality
	3	"Major neurologic abnormality", which includes any of the following: unilateral weakness in foot dorsiflexion, plantar flexion or knee extension, with strength rated 3 out of 5 or less, where 3/5 = ability to maintain position against gravity, but not against resistance; includes foot drop; documented progressive motor weakness; neurogenic bladder and/or bowel dysfunction
Radiological signs	1	Presence/absence of significant foraminal stenosis (a bony or soft tissue lesion that compresses or impinges on a lumbar nerve root as it exits the spinal canal that is compatible with the signs and symptoms)
	2	Presence/absence of significant central stenosis {reduced dimension of the portion of the spinal canal that surrounds the dural sac that is compatible with the signs and symptoms. Includes lateral recessal stenosis)
Comorbidity status	1	≤Moderate systemic disturbance caused either by the condition that is to be treated on surgical intervention or which is caused by other existing pathological processes (ASA < 3)
	2	Severe systemic disturbance from any cause or causes (ASA ≥ 3)
Disability	1	Mild disability—able to work full time and/or perform other activities of daily living (including housework or child care) without inordinate difficulty; symptoms may limit participation in sports or other recreational activities
	2	Moderate disability—symptoms prevent patient from working full time at usual job, or prevent full participation in other activities of daily living (including housework or child care)
	3	Severe disability—unable to work or to perform other activities of daily living (including housework or child care) due to symptoms
Yellow flags	Yes/no	Given by the presence/absence of psychosocial/behavioral factors, psychological distress, unhelpful coping strategies, etc

instrumentation; transarticular instrumentation; and facet screws).

The panel

A multispecialty, multinational expert panel was assembled comprising 14 experts from USA, Belgium, Sweden, Norway, UK, Spain, and Switzerland/Hungary. All were members of the largest international spine societies, and they represented the fields of orthopedic surgery, neurosurgery, rheumatology, physical medicine and rehabilitation, and clinical psychology. The panel was designed to best reflect the variety of specialties that are involved in patient treatment decisions and to obtain a mix of surgeons who perform the procedures ("doers") and specialists who treat LDS and refer them for surgery ("referrers"). As previously recommended [17, 23], the main selection criteria included acknowledged leadership in the field, absence of conflicts of interest, geographic diversity and diversity of practice setting, some evidence of ability to deconstruct and analyze decision making, and a reputation for being reliable and dependable. The members of the panel were all comfortable with and adept at speaking the chosen working language (English). The panel was moderated by a research clinician who was highly experienced in the RAM process (J-PV), assisted by an experienced neurosurgeon (FP) and clinical researcher (AFM).

Rating process

The panelists rated the appropriateness of surgery for each of the scenarios in two rounds (Fig. 1). In the first round, done remotely and on an individual basis, panelists were sent (by registered post) a package containing the systematic review, rating sheets and definition of terms. They were instructed to consider the conclusions of the systematic review and use their clinical expertise to judge the appropriateness of each of the three operative interventions for each of the clinical scenarios. The scenarios were rated on a 9-point scale (1 = extremely inappropriate, 5 = equivocal/uncertain, 9 = extremely appropriate). "Appropriate" was considered to mean that the expected health benefit (e.g., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeded the expected negative consequences (mortality, morbidity, anxiety of anticipating the procedure, pain produced by the procedure, time lost from work) by a margin wide enough to make the procedure worth doing, regardless of cost. Having accomplished this task, each expert mailed his rating sheets back to the main center responsible for entering and analyzing the data. One panelist did not feel confident completing the ratings since her specialty (Clinical Psychology) was too far removed from spine surgery to actually rate the indications; she subsequently acted as an advisor only, contributing to the discussion at the Expert Panel Meeting, during round 2. The ratings were entered into a customized computer program and the first-round results were then prepared ready for the Expert Panel Meeting (see later; Data analysis).

Approximately 5 days prior to the meeting, one of the three main project leaders contacted each expert to discuss any general questions they may have had about the panel process, the definitions, the rating scale, the rating process, the review document, issues arising from the first-round ratings, general and specific points, or indications that needed to be clarified before the panel meeting. This was intended to allow the time during the meeting itself to be used as efficiently as possible for the task in hand.

The morning of the first day of the Expert Panel Meeting was spent modifying the original list of indications and/or definitions until consensus among all panelists was reached. It was decided to add the presence/absence of psychosocial "yellow flags" (i.e., possible psychosocial barriers to recovery from a back problem, such as psychological distress, inappropriate beliefs about back pain, unhelpful coping strategies, etc.) [24] as an additional dichotomous variable to be considered along with the others in the matrix. This then doubled the total number of scenarios to be rated in the second round to 744 for each of the three surgical procedures. In round 2, the panel members discussed the first-round ratings, under the leadership of the moderator, focusing on areas of disagreement. The panelists were provided with reports showing their own initial ratings and the anonymous distribution of all the other panelists' ratings for each indication (this was done one chapter at a time). Panelists compared their own responses with those of the others and were given the opportunity to explain the reasons that led them to consider individual indications as appropriate or inappropriate. Indications and disagreements were discussed, after which panelists again individually re-rated all the indications. The results of these second-round ratings (from $N = 12$ experts; one was unable to attend the meeting) were used to ascertain the degree of panel agreement on indications and hence to determine appropriateness or inappropriateness of surgery for each indication (see later; Data analysis).

A feedback questionnaire was sent to the participants after the meeting to gain further information regarding their impressions of the systematic review, the first-round ratings, the panel meeting itself, and their overall experience in relation to their participation.

Data analysis

To facilitate comparison and following the standard procedure for the RAND appropriateness method, the 9-point scale was condensed into three categories based on the median value of the panel's ratings (1–3 inappropriate, 4–6 uncertain, and 7–9 appropriate) together with the degree of intra-panel disagreement. All indications for which there was disagreement were classified as uncertain, irrespective of the panel's median score. In summary, the following definitions were used:

- Appropriate: panel median of 7–9, without disagreement
- Uncertain: panel median of 4–6 OR any median with disagreement
- Inappropriate: panel median of 1–3, without disagreement

Intra-panel disagreement was defined as occurring when (in the panel of 12) at least three panelists rated an indication in the range of 1–3 and at least three others rated it in the range of 7–9. Intermediate situations, where there was neither agreement nor disagreement, were classified as uncertain.

Statistical analysis

Proportions of appropriate, inappropriate and uncertain ratings were calculated. Descriptive statistics for each variable were stratified by appropriateness status to examine possible predictors of appropriateness. Chi-squared tests were used to

Table 2 Proportion of clinical scenarios in each type of clinical presentation (chapter) that were considered appropriate (A), inappropriate (I) or uncertain (U) for the different types of surgery

Main chapter	Any surgery			Decompression only			Fusion (±decomp)			Instrumented fusion (±decomp)			Fusion or Instrumented fusion (±decomp)		
	% A	% I	% U	% A	% I	% U	% A	% I	% U	% A	% I	% U	% A	% I	% U
Back pain only, no instability	7.1	61.9	31.0	1.2	83.3	15.5	0.0	84.5	15.5	0.0	78.6	21.4	1.2	72.6	26.2
Back pain, instability	11.9	44.0	44.0	0.0	98.8	1.2	4.8	61.9	33.3	11.9	44.1	44.1	11.9	44.1	44.1
Radicular pain, no back pain, no instability	27.8	27.8	44.4	27.8	27.8	44.4	0.0	87.5	12.5	0.0	83.3	16.7	0.0	68.1	31.9
Radicular pain, back pain, no instability	29.2	26.4	44.4	18.1	29.2	52.8	0.0	56.9	43.1	5.6	51.4	43.1	12.5	34.7	52.8
Radicular pain, no back pain, instability	30.6	27.8	41.7	0.0	65.3	34.7	8.3	48.6	43.1	19.4	29.2	51.4	25.0	27.8	47.2
Radicular pain, back pain, instability	33.3	25.0	41.7	0.0	70.8	29.2	9.7	43.1	47.2	30.6	27.8	41.7	30.6	25.0	44.4
Neurogenic claudication, no back pain, no instability	38.9	19.4	41.7	34.7	22.2	43.1	0.0	52.8	47.2	0.0	58.3	41.7	2.8	45.8	51.4
Neurogenic claudication, back pain, no instability	33.3	22.2	44.4	25.0	27.8	47.2	0.0	47.2	52.8	9.7	38.9	51.4	18.1	33.3	48.6
Neurogenic claudication, no back pain, instability	31.9	27.8	40.3	0.0	56.9	43.1	0.0	44.4	55.6	20.8	29.2	50.0	29.2	27.8	43.1
Neurogenic claudication, back pain, instability	36.1	22.2	41.7	0.0	51.4	48.6	11.1	30.6	58.3	31.9	25.0	43.1	33.3	22.2	44.4

compare distributions of categorical variables, and *t* tests were used to compare between-group (e.g., doers and referrers) differences in the average median ratings for indications. Logistic regression analysis was used to determine the relevance of each of the variables (individual indications) in determining appropriateness [dichotomized as yes (appropriate) versus no (uncertain and inappropriate)]. Statistical analyses were carried out using Statview 5.0 (SAS Institute Inc, San Francisco, CA, USA) and IBM SPSS v21.0 for Apple Macintosh (Chicago, IL, USA). Statistical significance was accepted at the $p < 0.05$ level.

Results

Main findings

Of the 744 hypothetical scenarios generated for evaluation, surgery of one type or other was considered appropriate in 27 %, uncertain/equivocal in 41 % and inappropriate in 31 %. Frank disagreement (in considering surgery of any type) was low (7 % of scenarios).

The results for the appropriateness of surgery for each of the main chapters (clinical presentations) are shown in Table 2. It was noted from this data that fusion was considered appropriate far more often when instrumentation was used than when it was omitted. In discussion, there appeared to be differences of opinion between experts regarding the optimal role of instrumentation that could not be resolved with reference to available data or other consensus building approaches. This led to the situation where, for some scenarios, one expert might rate simple fusion as inappropriate because they felt that instrumentation should be used, while another might rate instrumented fusion as inappropriate or uncertain because they thought simple fusion would suffice, yet both felt that fusion of some kind was appropriate. Therefore, to more easily compare the appropriateness of decompression alone vs. the addition of fusion, the data were further analyzed by combining the results of all fusions by taking, for each scenario, the highest rating given for either fusion or instrumented fusion and using this to generate the median appropriateness scores for the "new" category of treatment (fusion ± instrumentation (inst)). These comparisons along with "surgery of some type" (highest rating out of decompression, fusion or instrumented fusion, followed by determination of the median values for this new category) are displayed in Fig. 2. In summary, Table 2 is useful for providing comparisons of the appropriateness of instrumented vs. non-instrumented fusion, while Fig. 2 combines these data and more easily compares the appropriateness of fusion (± instrumentation) with decompression alone.

Generally, the experts were extremely reticent to consider surgery of any type for scenarios describing back pain only and no instability (in just 7 % of the detailed scenarios it was appropriate, and in 62 % it was decidedly inappropriate). When back pain was present with instability, the appropriateness of surgery increased slightly to approx

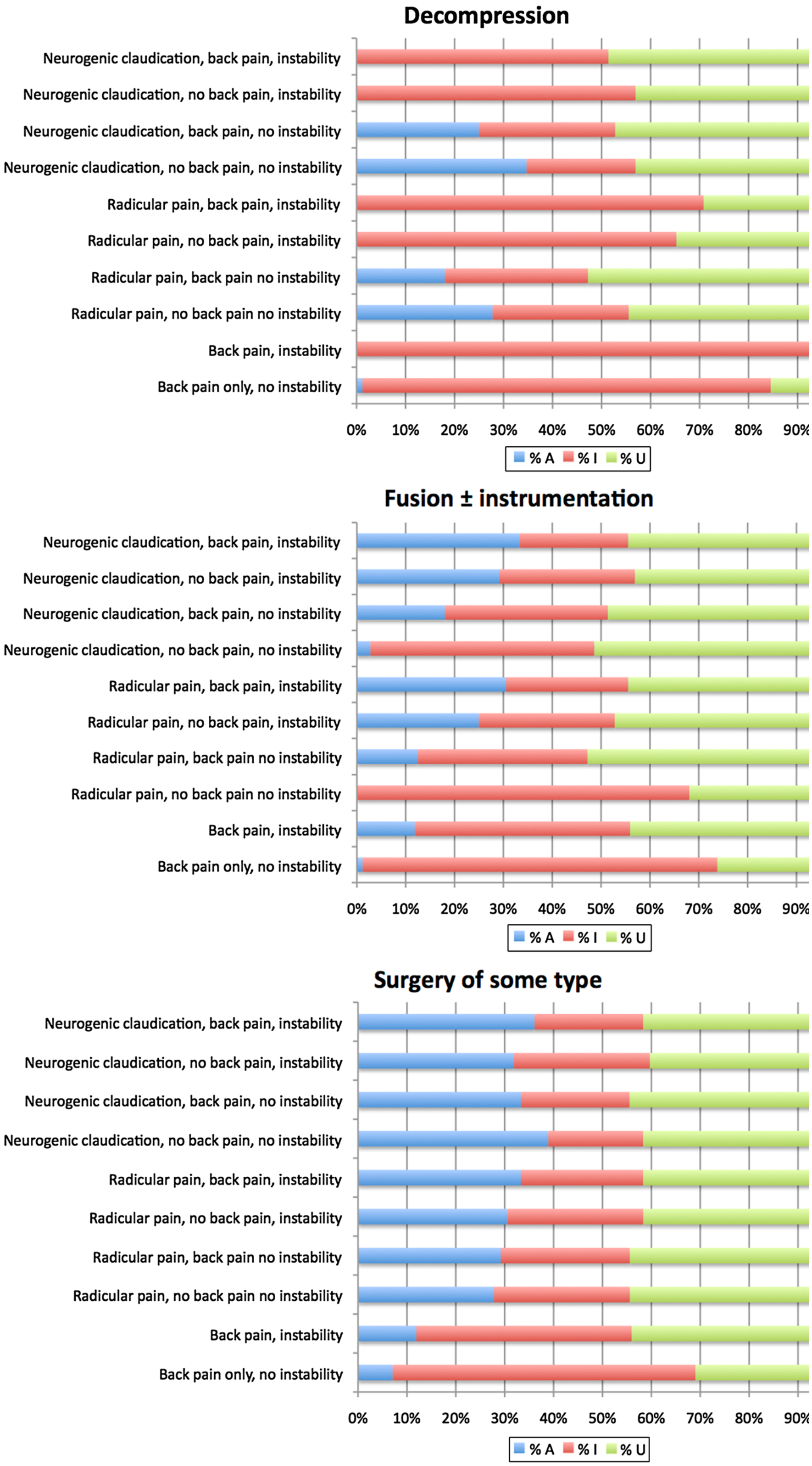

Fig. 2 Appropriateness ratings for the different chapters, in relation to decompression, fusion ± instrumentation, and surgery "of some type" (*A* appropriate, *I* inappropriate, *U* uncertain)

12 % scenarios (Fig. 2) and with fusion ± inst being the preferred surgery (Table 2).

Surgery of some type was considered appropriate for radicular pain in approximately 30 % scenarios and for neurogenic claudication in approximately 35 % scenarios (Table 2). For these two symptoms, when there was no accompanying instability, decompression was more often considered appropriate (18–35 % scenarios) than was fusion ± inst (0–10 % scenarios); however, the addition of back pain (even with no instability) increased the number of scenarios for which fusion ± inst was considered appropriate (from 0 to 6 % for radicular pain and from 0 to 10 % for neurogenic claudication; Table 2). In the presence of instability, no scenarios at all were considered appropriate for decompression alone, and a high proportion (51–99 %, depending on the clinical presentation) were considered decidedly inappropriate (Table 2).

Features of clinical scenarios considered appropriate

Table 3 summarizes for each clinical presentation (chapter) the general trends regarding the characteristics of scenarios considered appropriate for some type of surgery. For all chapters, the absence of yellow flags was a key feature, although flags were less important when disability was severe (except for back pain ± instability, where even with severe disability no scenarios with yellow flags were appropriate). Overall, the presence of yellow flags served to markedly reduce the number of scenarios for which surgery of some type was considered appropriate (Fig. 3). Severe neurological abnormality (i.e., category 3 in Table 1) was a feature of many of the "appropriate" scenarios, although some scenarios were nonetheless appropriate with less severe abnormalities, as long as disability was severe. All scenarios that were rated appropriate for some type of surgery included the radiological sign of significant central or foraminal stenosis. The majority of the scenarios rated appropriate for some type of surgery included "moderate comorbidity, at most", but severe comorbidity featured in some appropriate scenarios when there was also severe disability. For almost all chapters, severe disability (category 3) was a characteristic of the "appropriate scenarios", although some chapters (most notably those with instability) featured only moderate disability (Table 3). None of the scenarios with minimal disability (category 1) were considered appropriate for any type of surgery.

Face validity of the criteria

There was a logical relationship between each variable's subcategories and the appropriateness ratings for some type of surgery, e.g., no/mild disability had a mean appropriateness rating of 2.3 ± 1.5, whereas the rating for moderate disability was 5.0 ± 1.6 and for severe disability, 6.6 ± 1.6. Similarly, the mean rating for no/minimal neurological abnormality was 2.3 ± 1.5, increasing to 4.3 ± 2.4 for moderate and 5.9 ± 1.7 for severe neurological abnormality. In multiple regression, these variables that served to provide detail to the clinical presentations (i.e., degree of disability, severity of neurological abnormality, presence of yellow flags, degree of comorbidity, presence of radiological signs) accounted for 60 % of the variance in median ratings of appropriateness of some type of surgery ($p < 0.0001$).

In agreement with the results summarized above for the individual chapters, multiple logistic regression revealed that the three variables most likely ($p < 0.0001$) to be components of scenarios considered "appropriate" (as opposed to inappropriate or uncertain) for some type of surgery were: severe disability, no yellow flags, and severe neurological abnormality; the presence of neurogenic claudication, radicular pain and low/moderate comorbidity were additional significant predictors.

Value of the multidisciplinary discussion

The value of the interaction between the specialists in the multidisciplinary discussion was shown by comparison between the first-round and second-round expert ratings: there was a decrease in frank disagreement in considering surgery of any type, from 22 to 7 % of all rated scenarios. For decompression only, the proportion of scenarios with disagreement decreased from 19 % during the first round to 4 % during the second; for fusion ± inst it reduced from 46 to 11 %.

The proportion of scenarios considered appropriate for any surgery was lower in the second round than in the first (27 versus 37 %, respectively) but was similar for fusion ± inst (16 versus 14 %, respectively).

Difference between doers and referrers

Overall, surgeons gave significantly ($p < 0.0001$) higher ratings than non-surgeons (4.9 ± 2.2 versus 4.3 ± 2.2 points respectively), especially for the scenarios in the category "back pain only" (Fig. 4). The least discrepancy was found for the scenarios in the category "radicular pain with no back pain and no instability".

Panelists' evaluation of the process

The results of the feedback questionnaire (from 11 of the 12 experts) are shown in Table 4. Participation in the project was clearly a positive experience for the majority of the panelists. When asked how informative the expert panel discussion was, the experts gave a median value of 5 (range 3–5) on the 1–5 numeric scale. The experts' own satisfaction in participating as a member of the panel was also rated 5 (range 3–5). The extent to which they thought the ratings of their panel would reflect the appropriateness of specific surgical treatments for LDS was rated 4 (range 2–5), although their belief that the panel process could lead

Table 3 Overview of characteristics generally associated with "surgery of some type" being appropriate for each clinical presentation

Clinical presentation (chapter)										
Variable	Back pain, no instability	Back pain, instability	Radicular pain, no back pain, no instability	Radicular pain, back pain, no instability	Radicular pain, no back pain, instability	Radicular pain, back pain, instability	Neurogenic claud, no back pain, no instability	Neurogenic claud, back pain, no instability	Neurogenic claud, no back pain, instability	Neurogenic claud, back pain, instability
Yellow flags	No yellow flags	No yellow flags	No yellow flags, except with severe disability	No yellow flags, except with severe disability	No yellow flags, except with severe disability	No yellow flags, except with severe disability	No yellow flags, except with severe disability	No yellow flags, except with severe disability	No yellow flags, except with severe disability	No yellow flags, except with severe disability
Neurologic abnormality	Major neurologic abnormality (but moderate abnormality acceptable if severe disability and only moderate comorbidity)	Major neurologic abnormality (but moderate/no abnormality acceptable if severe disability and only moderate comorbidity)	Major neurologic abnormality (but moderate/no abnormality acceptable if severe disability)	Major neurologic abnormality (but moderate/no abnormality acceptable if severe disability)	Major neurologic abnormality (but moderate/no abnormality acceptable if severe disability)	Major neurologic abnormality (but moderate/no abnormality acceptable if severe disability)	Extent of neurologic abnormality less important (though no or only moderate abnormality only acceptable with severe disability)	Extent of neurologic abnormality less important (though no or only moderate abnormality only acceptable with severe disability)	Extent of neurologic abnormality less important (though no or only moderate abnormality only acceptable with severe disability)	Extent of neurologic abnormality less important (though no or only moderate abnormality only acceptable with severe disability)
Radiological signs	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis	Significant central or foraminal stenosis
Comorbidity	At most, moderate comorbidity (but severe comorbidity acceptable if severe disability)	At most, moderate comorbidity (but severe comorbidity acceptable if severe disability)	At most, moderate comorbidity (but severe comorbidity acceptable if severe disability)	At most, moderate comorbidity (but severe comorbidity acceptable if severe disability)	At most, moderate comorbidity (but severe comorbidity acceptable if severe disability)	Comorbidity not relevant in any consistent way	Comorbidity not relevant in any consistent way	Comorbidity not relevant in any consistent way	Comorbidity not relevant in any consistent way	Comorbidity not relevant in any consistent way
Disability	Severe disability	Severe disability	Severe disability	Severe disability	Severe disability (occasionally moderate)	Severe disability (occasionally moderate)	Severe disability	Severe disability	Severe disability (occasionally moderate)	Severe disability (occasionally moderate)

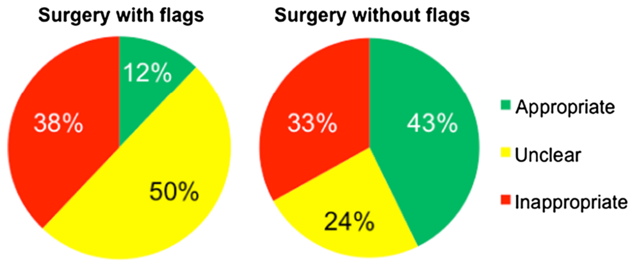

Fig. 3 Influence of the presence of *yellow flags* on the appropriateness of surgery (of some type)

to guidelines for physicians to assist with decision making was only 3 (range 3–5).

Discussion

General summary

In medicine, in general, there is increasing awareness of the need to deliver evidence-based treatment, with the evidence ideally being derived from the results of randomized controlled trials (RCTs). However, in spine surgery, the problems associated with performing RCTs are considerable [25–27]. RCTs demand precise clinical classification or phenotyping, and eligibility criteria are often kept very rigid, such that the results may not translate to the patients seen in everyday clinical practice [28–31]. Often, only a minority of eligible patients declare themselves willing to be randomized to treatment. Blinding of patients and surgeons is difficult, and this opens up an obvious potential for bias. Further, many patients withdraw their participation or cross over into the other treatment arm, representing another potential source of bias [32, 33]. The lack of good quality RCTs in LDS leaves us with notable gaps in our knowledge, and determining the appropriateness of surgery must therefore depend on other methods. The present study used the RAND Appropriateness Method (RAM) to develop criteria for assessing the appropriateness of surgery for LDS in theoretical cases presenting with a variety of signs and symptoms. The study was successful in its aim, and criteria were developed which appeared to show good face validity (see later).

The detailed output of the RAM is an appropriateness rating for each scenario (combination of variables) or "patient type" that is considered. Since there were over seven hundred different patient types considered here, the results are not easily tabulated or interpreted in simple terms. While the clinical specificity of appropriateness criteria allows them to perform better than widely acclaimed guidelines developed by specialty societies [34], it also makes their practical implementation more difficult. As pointed out by Porchet et al. [35], it is not feasible for clinicians to wade through hundreds of clinical scenarios to find the one that most closely resembles the specific patient they have in front of them. Instead, the main value of the current work will come in future developments where, following further validation studies, the derived information will be compiled into a computerized/web-based algorithm or "decision tool" for determining candidacy for surgery. The work also refines clinical classification of LDS and contributes to the definition of its phenotype for future, e.g., genetic studies.

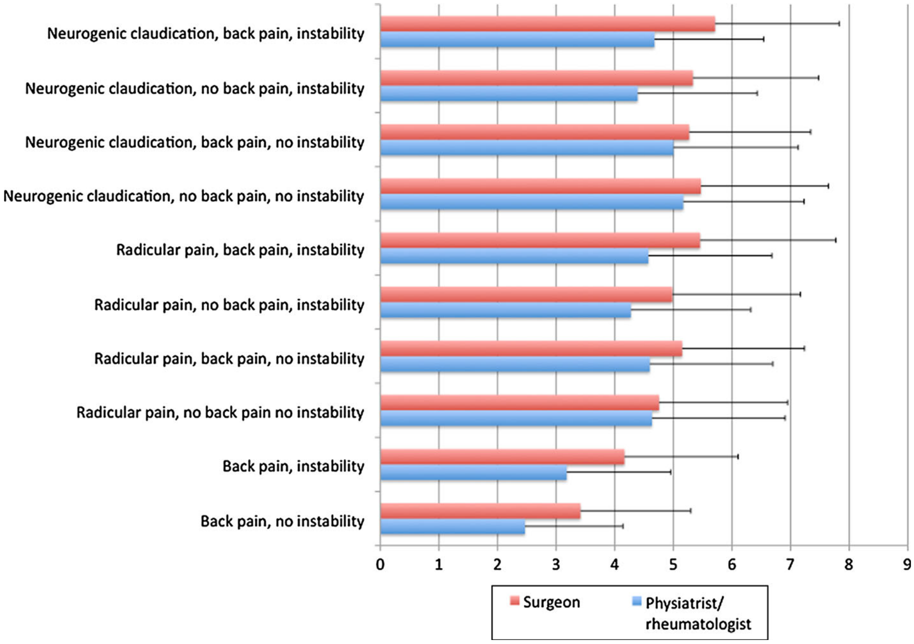

Fig. 4 Surgeon and physiatrist/rheumatologist mean (SD) ratings of appropriateness for all the scenarios in each of the chapters

Table 4 Results of the expert panel feedback questionnaire concerning the RAM process

Item	Responses (from $N =$ ll experts)			
	Response categories			
	1 not at all, 2 a little, 3 somewhat, 4 pretty much, 5 very much			
Review of the scientific literature	Mean	Median	Min	Max
How completely did you read it?	4.6	5	4	5
How many hours did you spend reading it? _______hours	4.3	2.5	1	15
How informative was it?	3.7	4	3	5
How much did it influence your first-round ratings?	3.2	3	2	5
First-round ratings (done before the meeting)				
How easy did you find the task?	2.7	3	1	4
How onerous did you find the task?	3.3	3	2	4
How clear were the instructions?	4.1	4	2	5
How consistent do you believe you were? (the effects of fatigue, memory, different times to rate, format of instrument, etc.)	3.5	3	3	4
How many hours did it take you? _______hours	4.1	4	2	8
Panel meeting				
How knowledgeable about the subject matter were the moderators?	4.5	4	4	5
How well did the moderators function as group leaders?	4.7	5	4	5
How informative was the discussion?	4.5	5	3	5
How argumentative was the discussion?	3.6	4	1	5
How much were you influenced in your second ratings by the feedback from the first-round ratings?	3.5	4	3	4
How much were you influenced in your second ratings by the discussion?	3.6	4	3	5
Overall impressions of your experience				
How satisfying did you find your participation on this panel?	4.5	5	3	5
How well do you believe your own ratings reflect the appropriateness of specific surgical treatments for LDS?	3.9	4	3	5
How well do you estimate that your panel's ratings will reflect the appropriateness of specific surgical treatments for LDS?	3.8	4	2	5
How much do you believe that this panel process can lead to guidelines to assist physician decision making for surgery for LDS?	3.5	3	3	5
How did your participation on this panel compare with your expectations? (1 much worse to 5 much better)	4.0	4	3	5

Appropriate indications

The main findings of the present report were able to highlight the impact of specific variables on the appropriateness of surgery, and, in doing so, serve to confirm the face validity of the criteria developed. Significant predictors of a scenario being considered appropriate for surgery were severe disability, severe neurological abnormalities and no yellow flags. That a certain threshold for surgery is required in terms of severe disability and neurological abnormality would appear to make sense, intuitively. At first sight, the importance of yellow flags may also appear obvious. Yellow flags are psychosocial factors that increase the risk of developing or perpetuating pain and long-term disability, and may therefore act as barriers to recovery. However, while yellow flags have been shown to be a predictor of poor outcome in the treatment of low back pain (LBP) (especially non-specific LBP) [36], their role in relation to the outcome of surgery for specific spinal disorders such as LDS is less clear. There are some lines of argument that suggest that the more convincing the medical indication for surgery, the less relevant is the presence of psychosocial factors [15, 37]. This was also partly seen in the present study, in the sense that, of all the appropriate scenarios, a greater proportion had yellow flags in the more typically "clear-cut" clinical presentations (e.g., neurogenic claudication with no back pain and no instability) than in the more controversial presentations (e.g., back pain with no instability). In fact, there was a linear relationship between the percentage of scenarios considered appropriate within a chapter and the percentage of these scenarios containing yellow flags ($r = 0.91$, $p < 0.0001$; actual data not shown). Back pain is a common symptom accompanying LDS. However, surgery was rarely indicated for back pain alone, especially in the absence of instability: it was considered appropriate in only 7 % scenarios in the chapter “back pain with no instability”, typically those with high levels of disability, major neurological abnormality and low comorbidity. In the presence of yellow flags, even with this severe presentation, surgery was rated uncertain for back pain alone.

A lower comorbidity level was also a feature of appropriate scenarios, although this seemed to have less influence on determining appropriateness than did neurological abnormalities or disability. While comorbidity may, to an extent, be expected to influence whether surgery should be considered and if so what type, it may be the case that as long as the clinical threshold for surgery is met (by the key signs and symptoms), then comorbidity remains of lesser importance.

Expert panel process, and differences between doers and referrers

Interdisciplinary discussion is a key element of the RAM, as is the non-directive involvement of a trained moderator and non-dominant participation by the panel members [35]. The benefit of the discussion between the two rounds of ratings was shown in the present study by the reduced disagreement in the second-round ratings. Rating independently, without any knowledge of the other experts' opinions or ratings, there was disagreement in 22 % of scenarios; this reduced to 7 % following the discussion of discordant ratings and exchange of opinions among the different specialists. This low level of disagreement is especially impressive because the RAM does not force consensus; instead, the panelists give their final ratings independently and privately, and the statistics subsequently determine the level of consensus or otherwise. The panelists' feedback on their involvement in the RAM process was also testament to the value of the interdisciplinary discussion, with median ratings between 4 and 5 (out of a maximum 5) for each of the 6 questions concerned with the quality of the panel meeting.

Analyses of the mean ratings of those who perform the procedure (“doers”, i.e., neurosurgeons and orthopedic surgeons) and those who do not (“referrers”) revealed significant differences, especially in relation to the scenarios describing back pain only. In essence, the study “quantified” the more conservative stance of non-surgeons in relation to the appropriateness of surgery for LDS. This has been observed before, in previous studies using the RAM to evaluate the appropriateness of spine surgery, and has been suggested to reflect the different case mixes typically seen by surgeons and non-surgeons and their respective appreciation of the risks and benefits of surgery compared with alternative treatments [38]. It may also be that different specialists favor their own approach to treatment, or see a biased selection of patients: the failures of back surgery usually return to non-operative specialists, while patients with good outcomes are not seen again [35]. The results nonetheless emphasize the importance of having a mix of “doers” and “referrers” when developing treatment appropriateness criteria. They also reveal important information regarding a possible need for improved cross-specialty communication and decision making within the different specialties. The development of the aforementioned computerized “decision tool”, based on the appropriateness criteria, may help referring physicians to assess whether referral to the spine specialist is actually appropriate. It is sometimes difficult to persuade a patient that surgery is not appropriate once he/she has been referred to the specialist for surgical assessment, and access to decision aids should serve to minimize this problem. In a recent study, it was shown that of the 303 lumbar spine referrals to a group of 10 neurosurgeons, 80 (26 %) were appropriate, 92 (30 %) were uncertain and 131 (44 %) were inappropriate for surgical assessment [39]. The authors concluded that physicians seeking specialist consultations for patients with lumbar spine complaints need to be better informed of the criteria that indicate an appropriate referral for surgical treatment. Avoiding inappropriate referrals could reduce waiting-times for both consultation and surgery for patients who actually require it [39]. This would allow resources to be more focused, improving the overall level of care.

Implications of the findings

Most quality improvement efforts in spine surgery focus on reducing risk by improving the technical quality of care provided [40]. However, since risk is inherent in any procedure, reducing the number of unnecessary operations is an important issue in patient safety and quality improvement. Recent years have seen increasing criticism of the excessive and potentially inappropriate—or at least unsubstantiated—use of lumbar spine surgery. It is hypothesized that reducing inappropriate surgery may have a greater impact on complication rates than does improving the technical quality of surgery performed [40]. The issue of appropriateness is not only relevant in relation to the unnecessary risks accompanying unnecessary surgery, but also in relation to poor outcomes. Many studies have sought to identify personal risk factors such as age, smoking habit, psychological status, etc. that influence the outcome of spine surgery. In spine surgery, however, treatment failures are also commonly attributable to poor patient selection. Unclear indications for a given procedure are a strong risk factor for a poor outcome [16].

It is incumbent on the surgeon to perform an accurate diagnostic work-up and to critically assess the indications for surgery; any shortcomings in this respect will naturally increase the potential for an unsatisfactory result. However, in the absence of clear guidelines, this relies on the "best judgment" of the surgeon and in such circumstances it might be easy to be distracted by irrelevant signs that can increase the pressure for surgery. Waddell et al. [15] have, for example, reported that psychological factors manifest as inappropriate symptoms and signs may obscure the physical assessment, leading to a mistaken diagnosis of a surgically treatable lesion. In this instance, inappropriate illness behavior leads to inappropriate surgery and, consequently, to a poor outcome [15]. The availability of clear guidelines for the appropriateness of surgery should allow attention to be concentrated on the relevant indications only and allow for more focused decision making, and hence better outcomes.

There is great variability in the rates of surgery across the world and even between different spine service areas of any given State in the USA [41, 42]. These differing rates are also associated with differing proportions of surgical success. The variability may be related to differences in physicians' preferences or thresholds for surgery and their criteria for the selection of patients. The availability and implementation of standardized appropriateness criteria for surgical indications, once validated and accepted on a national and international basis, may help to create greater consistency in this respect.

The recent study of Danon-Hersch et al. [43] was the first in the field of lumbar spine surgery to document an association between the appropriate use of treatment and clinical outcome. This is extremely encouraging as regards the development of such criteria for the more contentious of our spine surgical procedures. Similar prospective investigations using the appropriateness criteria developed in the present study should verify whether their use does indeed improve outcome in patients with LDS.

Finally, approximately 40 % of the scenarios left the panel undecided, i.e., were rated "uncertain". This is indicative of a strong need for further research; these particular scenarios are the ones that should be targeted in future clinical trials.

Limitations of the study

Our study has a number of limitations. First, even though there are many articles describing and/or comparing different surgical options for LDS, the systematic review conducted in connection with the RAM process revealed that there is insufficient high-quality evidence to draw firm conclusions concerning indications for surgical treatment or predictors of outcome in LDS [22]. Indeed, the RAM was developed specifically for use in such cases and uses the best evidence available, but it nonetheless means that the evidence base upon which the ratings were made was not the most robust. Second, the appropriateness criteria developed are pertinent to the current literature and current expert opinion on the treatment of LDS; the evidence base should be re-evaluated regularly to examine whether new knowledge has any impact on the existing criteria. Third, the expert panel comprised a balanced group of renowned spine clinicians from different disciplines and around the world. However, it cannot be ruled out that a panel comprising different members, or from different countries, would have arrived at different conclusions. Previous reliability studies comparing the results of panels in different countries [38, 44, 45] have shown substantial agreement, although it is still generally recommended that the need for any local adaptation of recommendations or guidelines be examined prior to their use in different settings. Fourth, given the measurement error inherent in flexion–extension radiographic measures, the ability to accurately establish the presence of grade 1 spondylolisthesis may be questionable. Similarly, although many consider it an important factor and something they can readily recognize, there is no consensus regarding the precise definition or method of measurement of "instability" [46] (see Table 1), meaning the interpretation of these signs may remain somewhat subjective. Nonetheless, even though there is no clear definition of instability, our results indicate that the panel considered it relevant, rating otherwise identical indications differently depending on the presence or not of instability considered to be clinically relevant (see Table 2). Fifth, LDS is often present in conjunction with other significant degenerative changes and the appropriateness criteria are to be considered only in relation to the case where LDS is the most distinct radiographic finding associated with the given symptoms. The same recommendations may not hold in the setting of concomitant spinal pathology such as scoliosis or previous spine surgery. Finally, a relatively low proportion of scenarios was considered appropriate for surgery (27 %). However, this is not to suggest that of all patients who are surgically treated for LDS only 27 % are done so appropriately; many of the scenarios would scarcely arise in clinical practice and emerge only through the need to create all possible permutations of the relevant variables. Follow-on studies are required to examine both the prevalence of the examined scenarios in clinical practice and the validity of the criteria on a prospective basis before they are implemented in clinical practice.

In conclusion, using the best available evidence together with collective expert opinion, we employed a systematic, transparent, and validated method to develop criteria for determining candidacy for surgery for LDS. The appropriateness ratings of the international, multidisciplinary panel followed logical clinical rationale, indicating good face validity of the criteria developed. The criteria should be evaluated for their predictive validity on a prospective basis to examine whether patients treated “appropriately” do indeed have better clinical outcomes.

Acknowledgments This study was supported by grants from “Eurospine, the Spine Society of Europe”, and the “Swiss National Science Foundation” (IZ32Z0_141535).

Conflict of interest None.

References

1. Martin CR, Gruszczynski AT, Braunsfurth HA, Fallatah SM, O’Neil J, Wai EK (2007) The surgical management of degenerative lumbar spondylolisthesis: a systematic review. Spine (Phila Pa 1976) 32:1791–1798
2. Weinstein JN, Lurie JD, Tosteson TD, Zhao W, Blood EA, Tosteson AN, Birkmeyer N, Herkowitz H, Longley M, Lenke L, Emery S, Hu SS (2009) Surgical compared with nonoperative treatment for lumbar degenerative spondylolisthesis. Four-year results in the Spine Patient Outcomes Research Trial (SPORT) randomized and observational cohorts. J Bone Joint Surg Am 91:1295–1304
3. Pearson AM, Lurie JD, Blood EA, Frymoyer JW, Braeutigam H, An H, Girardi FP, Weinstein JN (2008) Spine patient outcomes research trial: radiographic predictors of clinical outcomes after operative or nonoperative treatment of degenerative spondylolisthesis. Spine (Phila Pa 1976) 33:2759–2766
4. Watters WC 3rd, Bono CM, Gilbert TJ, Kreiner DS, Mazanec DJ, Shaffer WO, Baisden J, Easa JE, Fernand R, Ghiselli G, Heggeness MH, Mendel RC, O’Neill C, Reitman CA, Resnick DK, Summers JT, Timmons RB, Toton JF (2009) An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spondylolisthesis. Spine J 9:609–614. doi:10.1016/j.spinee.2009.03.016
5. Epstein NE, Epstein JA, Carras R, Lavine LS (1983) Degenerative spondylolisthesis with an intact neural arch: a review of 60 cases with an analysis of clinical findings and the development of surgical management. Neurosurgery 13:555–561
6. Katz JN, Lipson SJ, Lew RA, Grobler LJ, Weinstein JN, Brick GW, Fossel AH, Liang MH (1997) Lumbar laminectomy alone or with instrumented or noninstrumented arthrodesis in degenerative lumbar spinal stenosis. Patient selection, costs, and surgical outcomes. Spine (Phila Pa 1976) 22:1123–1131
7. Kornblum MB, Fischgrund JS, Herkowitz HN, Abraham DA, Berkower DL, Ditkoff JS (2004) Degenerative lumbar spondylolisthesis with spinal stenosis: a prospective long-term study comparing fusion and pseudarthrosis. Spine (Phila Pa 1976) 29:726–733 discussion 733–724
8. Ghogawala Z, Benzel EC, Amin-Hanjani S, Barker FG 2nd, Harrington JF, Magge SN, Strugar J, Coumans JV, Borges LF (2004) Prospective outcomes evaluation after decompression with or without instrumented fusion for lumbar stenosis and degenerative Grade I spondylolisthesis. J Neurosurg Spine 1:267–272
9. Okuda S, Oda T, Miyauchi A, Haku T, Yamamoto T, Iwasaki M (2007) Surgical outcomes of posterior lumbar interbody fusion in elderly patients. Surgical technique. J Bone Joint Surg Am 89(Suppl 2 Pt.2):310–320
10. Fischgrund JS, Mackay M, Herkowitz HN, Brower R, Montgomery DM, Kurz LT (1997) 1997 Volvo Award winner in clinical studies. Degenerative lumbar spondylolisthesis with spinal stenosis: a prospective, randomized study comparing decompressive laminectomy and arthrodesis with and without spinal instrumentation. Spine (Phila Pa 1976) 22:2807–2812
11. Abdu WA, Lurie JD, Spratt KF, Tosteson AN, Zhao W, Tosteson TD, Herkowitz H, Longely M, Boden SD, Emery S, Weinstein JN (2009) Degenerative spondylolisthesis: does fusion method influence outcome? Four-year results of the spine patient outcomes research trial. Spine (Phila Pa 1976) 34:2351–2360
12. Schaeren S, Broger I, Jeanneret B (2008) Minimum four-year follow-up of spinal stenosis with degenerative spondylolisthesis treated with decompression and dynamic stabilization. Spine (Phila Pa 1976) 33:E636–E642
13. Kelleher MO, Timlin M, Persaud O, Rampersaud YR (2010) Success and failure of minimally invasive decompression for focal lumbar spinal stenosis in patients with and without deformity. Spine (Phila Pa 1976) 35:E981–E987
14. Toyoda H, Nakamura H, Konishi S, Dohzono S, Kato M, Matsuda H (2010) Clinical outcome of microsurgical bilateral decompression via unilateral approach for lumbar canal stenosis: minimum five-year follow-up. Spine (Phila Pa 1976) 36(5):410–415
15. Waddell G, Morris EW, Di Paola MP, Bircher M, Finlayson D (1986) A concept of illness tested as an improved basis for surgical decisions in low-back disorders. Spine (Phila Pa 1976) 11:712–719
16. Mannion AF, Elfering A (2006) Predictors of surgical outcome and their assessment. Eur Spine J 15(Suppl 1):S93–S108
17. Fitch K, Bernstein SJ, Aguilar MD, Burnand B, LaCalle JR, Lazaro P, van het Loo M, McDonell J, Vader J, Kahan JP (2001) The RAND/UCLA appropriateness method user’s manual. RAND Corporation, Santa Monica
18. Fletcher SW, Fletcher RH (1998) Development of clinical guidelines. Lancet 352:1876
19. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE (1986) A method for the detailed assessment of the appropriateness of medical technologies. Int J Technol Assess Health Care 2:53–63
20. Naylor CD (1998) What is appropriate care? N Engl J Med 338:1918–1920
21. Deyo RA, Nachemson A, Mirza SK (2004) Spinal-fusion surgery—the case for restraint. N Engl J Med 350:722–726
22. Steiger F, Becker H-J, Standaert CJ, Balague F, Vader J-P, Porchet F, Mannion AF (2014) Surgery in lumbar degenerative spondylolisthesis—indications, outcomes and complications: a systematic review. Eur Spine J [Epub ahead of print]
23. Tan C, Treasure T, Browne J, Utley M, Davies CW, Hemingway H (2007) Seeking consensus by formal methods: a health warning. J R Soc Med 100:10–14
24. Kendall NA, Linton SJ (1997) Guide to assessing psychosocial Yellow Flags in acute low back pain: risk factors for long-term disability and work loss. In: Accident Rehabilitation and Compensation Insurance Corporation of New Zealand and the National Advisory Committee on Health and Disability Wellington, New Zealand
25. Boutron I, Ravaud P, Nizard R (2007) The design and assessment of prospective randomised, controlled trials in orthopaedic surgery. J Bone Joint Surg Br 89:858–863
26. Fairbank J (1999) Randomized controlled trials in the surgical management of spinal problems. Spine (Phila Pa 1976) 24:2556–2561 discussion 2562–2553
27. Winter RB (1999) The prospective, randomized, controlled clinical trial in spine surgery: fact or fiction? Spine (Phila Pa 1976) 24:2550–2552

28. 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
29. 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
30. Hoppe DJ, Schemitsch EH, Morshed S, Tornetta P 3rd, 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
31. Horwitz RI (1987) Complexity and contradiction in clinical trial research. Am J Med 82:498–510
32. 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
33. 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
34. Kahn KL, Park RE, Vennes J, Brook RH (1992) Assigning appropriateness ratings for diagnostic upper gastrointestinal endoscopy using two different approaches. Med Care 30:1016–1028
35. Porchet F, Vader JP, Larequi-Lauber T, Costanza MC, Burnand B, Dubois RW (1999) The assessment of appropriate indications for laminectomy. J Bone Joint Surg Br 81:234–239
36. Nicholas MK, Linton SJ, Watson PJ, Main CJ (2011) Early identification and management of psychological risk factors (“yellow flags”) in patients with low back pain: a reappraisal. Phys Ther 91:737–753. doi:10.2522/ptj.20100224
37. Carragee EJ (2001) Psychological screening in the surgical treatment of lumbar disc herniation. Clin J Pain 17:215–219
38. Vader JP, Porchet F, Larequi-Lauber T, Dubois RW, Burnand B (2000) Appropriateness of surgery for sciatica: reliability of guidelines from expert panels. Spine (Phila Pa 1976) 25:1831–1836
39. Findlay JM, Deis N (2010) Appropriateness of lumbar spine referrals to a neurosurgical service. Can J Neurol Sci 37:843–848
40. Deyo RA, Mirza SK (2009) The case for restraint in spinal surgery: does quality management have a role to play? Eur Spine J 18(Suppl 3):331–337
41. Keller RB, Atlas SJ, Soule DN, Singer DE, Deyo RA (1999) Relationship between rates and outcomes of operative treatment for lumbar disc herniation and spinal stenosis. J Bone Joint Surg Am 81:752–762
42. Volinn E, Mayer J, Diehr P, Van Koevering D, Connell FA, Loeser JD (1992) Small area analysis of surgery for low-back pain. Spine (Phila Pa 1976) 17:575–581
43. Danon-Hersch N, Samartzis D, Wietlisbach V, Porchet F, Vader JP (2010) Appropriateness criteria for surgery improve clinical outcomes in patients with low back pain and/or sciatica. Spine 35(6):672–683
44. Vader JP, Burnand B, Froehlich F, Dupriez K, Larequi-Lauber T, Pache I, Dubois RW, Gonvers JJ, Brook RH (1997) Appropriateness of upper gastrointestinal endoscopy: comparison of American and Swiss criteria. Int J Qual Health Care 9:87–92
45. Burnand B, Vader JP, Froehlich F, Dupriez K, Larequi-Lauber T, Pache I, Dubois RW, Brook RH, Gonvers JJ (1998) Reliability of panel-based guidelines for colonoscopy: an international comparison. Gastrointest Endosc 47:162–166
46. Szpalski M (1996) The mysteries of segmental instability. Bull Hosp Jt Dis 55:147–148