Eur Spine J (2012) 21:2085–2090
DOI 10.1007/s00586-012-2192-4

ORIGINAL ARTICLE

Interobserver agreement in fusion status assessment after instrumental desis of the lower lumbar spine using 64-slice multidetector computed tomography: impact of observer experience

Borislav Laoutliev · Inger Havsteen · Birthe Højlund Bech · Eva Narvestad · Hanne Christensen · Anders Christensen

Received: 14 August 2011 / Revised: 15 November 2011 / Accepted: 2 February 2012 / Published online: 19 February 2012

Abstract
Purpose Persistent lower back pain after instrumental posterolateral desis may arise from incomplete fusion. We investigate the impact of experience on interobserver agreement in fusion estimation.
Methods Four independent observers, two residents and two musculoskeletal radiologists, reviewed dedicated lumbar 64-MDCT scans and scored vertebral levels 1–5 after Glassman's grades, 1: solid bilateral fusion, 2: solid unilateral fusion, 3: partial bilateral fusion, 4: partial unilateral fusion, 5: non-fusion. We investigated two simplifying dichotomizations, solid bilateral fusion (Glassman 1) versus all others and uni- or bilateral fusion (Glassman 1–2) versus partial or non-fusion.
Results Thirty-six patients with 61 operated lumbar levels were included. Interobserver agreement rates for four observers using Glassman's system were fair (kappa 0.32), either dichotomization showed moderate agreement (kappa 0.53 and 0.59). Observer pairs had comparable prevalence adjusted interobserver agreement rates (residents: PABAK 0.67 and 0.54; consultants: PABAK 0.57 and 0.71).
Conclusions Difference in observer experience seems of minor impact.

B. Laoutliev · B. H. Bech · E. Narvestad
Department of Radiology, Diagnostic Centre, Copenhagen University Hospital Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

I. Havsteen (✉) · A. Christensen
Department of Radiology, Copenhagen University Hospital Bispebjerg, Bispebjerg Bakke 23, 2400 Copenhagen, Denmark
e-mail: seestein@gmail.com

H. Christensen
Department of Neurology, Copenhagen University Hospital Bispebjerg, Copenhagen, Denmark

Keywords Spine fusion · Interobserver variability · CT scan · Observer experience

Introduction

How to evaluate spinal fusion in patients with persistent or recurrent low back pain after posterolateral lumbar spine surgery remains controversial. In clinical context until recently, the most widely used non-invasive modalities were plain radiographs, static and bending images, now the use of computer tomography (CT) is widespread offering greater anatomical detail [1–3]. Visualized bony bridging as either continuous trabeculation or joint obliteration is suggested as best radiological evidence for solid fusion and several grading systems have been proposed [4–7]. Thin-sliced helical CT scans showed greater interobserver agreement and better illustration of bony bridging than static or bending radiographs [4, 8–12]. Comparison of either modality with surgical inspection documented not much agreement for plain radiographs [13] and posterolateral fusions on dedicated fine-cut CT were found moderately predictive of a solid fusion on surgical exploration [5]. The development of multidetector CT (MDCT) and dedicated protocols made it possible to minimize metallic artifacts [14, 15] and allowed for isotropic coronal and sagittal reconstruction. Dedicated fine-cut CT scans with sagittal and coronal reconstructions showed moderate inter- and intraobserver agreement with highest rates when the fusion was deemed solid [4].

The aim of the present study was to investigate interobserver agreement in estimating fusion after instrumental desis of the lower lumbar spine using 64 MDCT and the effects of observer experience. Simplified grading systems were included in the analysis.

Materials and methods

Patients

This study was based on a registry of patients referred for imaging at our hospital 2007–2009 due to persistent or recurrent lower back pain after instrumental desis/lumbar fusion of one or more levels among L3–S1. Surgical procedure included titanium pedicle screws and adjunctive spinal instrumentation in the form of vertical lateral rods and transversal pins. Patients with anterior fusion were excluded. 36 patients, 20 women and 16 men ($p = 0.6$), median age 54.4 years, were included. The study was approved by the National Data Protection Agency, File no. 2008-41-2004.

Image acquisition

CT scannings were performed with 64-slice MDCT (Toshiba Aquillion 64) using high X-ray kilovolt peak (135 kV) and high dose (275–500 mAs) to reduce metal artifacts and collimation 64 × 0.5 (isotropic voxel resolution). No additional metallic artifact reducing algorithms were employed. Coronal and sagittal reconstructions were performed using 2 mm slice thickness.

Observers and observed criteria

Four observers, 2 residents and 2 musculoskeletal subspecialized consultant radiologists, independently rated radiological bone fusion scoring MDCT scans of 61 operated vertebral fusion segments from 1 to 5 according to Glassman's classification system (Table 1; Figs. 1, 2, 3, 4). Observer background was for both residents 3 out of 4 years radiological training and subspecialists have held consultant positions for more than 10 years. Clinical information was limited to the presence of earlier lumbar desis and back pain. The postoperative status of the instrumented desis was observed including signs of absent bony fusion or changes to the osteosynthesis material such as fractures or surrounding lucencies. Bony fusion was defined as bone bridging. Causes of incomplete fusion were not considered separately in the evaluation.

We simplified the scoring system by dichotomizing the Glassman grades into 1–2 and 3–5. This corresponds to respectively uni- and bilateral solid fusion versus partial or non-fusion. We further analyzed Glassman grade 1, solid bilateral fusion, versus all others.

As a proxy for repeatability, we sought to determine if consensus grades per observed level from specialists and

Table 1 Glassman scoring system for posterolateral fusions [4]

1	Solid bilateral fusion
2	Solid unilateral fusion
3	Partial bilateral fusion
4	Partial unilateral fusion
5	No fusion

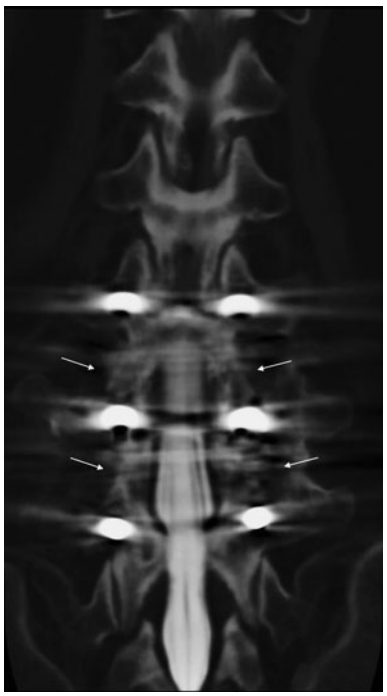

Fig. 1 64-MDCT of the lumbar spine, coronal reconstruction. Bilateral solid fusion (*arrows*), Glassman grade 1

Fig. 2 Solid fusion on the left (*arrow*), Glassman grade 2

Fig. 3 No fusion on both sides (*arrows*), Glassman grade 5

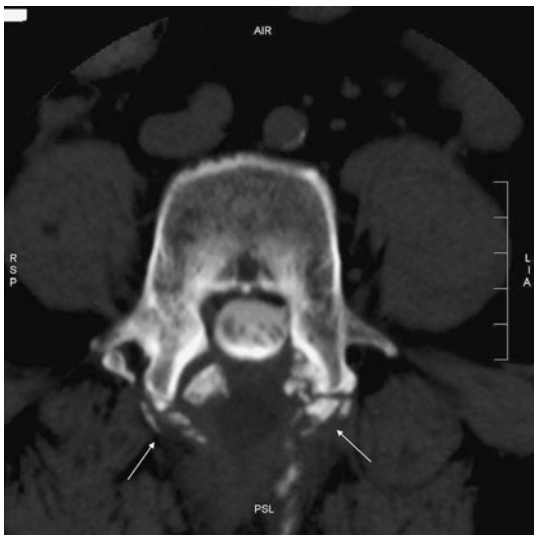

Fig. 4 Axial reconstruction. No fusion on both sides (*arrows*), Glassman grade 5

residents yielded the same rate of agreement as their respective subgroups. Interobserver variability was calculated for all operated levels and separately for each intervertebral level.

Statistics

We used R, version 2.10.0, for statistical analysis. Kappa was determined according to Fleiss for multiple observers. 95% confidence intervals, an estimate of the maximum attainable kappa and prevalence and bias indices were calculated for two observers where applicable [16–20]. Table 2 presents accepted guidelines for kappa values [21].

Table 2 Kappa values and guidelines for their interpretation [21]

κ	Interpretation
<0	No agreement
0.0–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

Results

We examined 61 operated levels in 36 patients (20 women and 16 men). Table 3 shows median Glassman scores per patient per operated segment and Table 4 shows the number of operated levels scored Glassman grades 1–5 per observer for all 61 operated levels and each intervertebral level separately. The highest interobserver agreement was obtained with Glassman grade 1, solid bilateral fusion, for all groups. This is most pronounced though not significant for consultants with kappa 0.71 corresponding to substantial agreement for all intervertebral levels (Table 5). Observer agreement rates decline successively with increasing Glassman score or degree of absent fusion (Table 4). The four observers agreed on uni- or bilateral fusion versus partial or non-fusion in 41 of 61 cases (67%) yielding kappa 0.53 corresponding to moderate agreement (Table 6). Consensus ratings among consultants versus residents showed agreement in 52 of 61 operated levels, yielding kappa 0.66 corresponding to substantial agreement.

Four observers with varying experience agreed on Glassman category in 22 of 61 operated levels (36%) yielding kappa 0.32, a fair agreement, as did the subgroups of residents and consultants. Consensus ratings among consultants versus residents showed agreement in 35 of 61 operated levels, yielding kappa 0.41 corresponding to moderate agreement. There was no significant difference between the groups in either dichotomized classification. Isolating Glassman grade 1 showed a prevalence index close to zero, thus with not much influence on the kappa value through chance agreement (Table 5). The pool of uni- or bilateral fusion showed prevalence indices close to 0.5; adjusting for prevalence their kappa (PABAK) values rose and resembled the bilateral fusion values. Disagreement was symmetrical among observers and the bias index in either simplified classification was near zero.

Table 3 Median Glassman scores per patient per operated segment

Patient	L3–L4	L4–L5	L5–S1
1	0	4.5	1
2	2	2	0
3	0	1.5	0
4	1	1	1
5	3	2.5	0
6	0	0	2
7	0	1	0
8	0	0	3.5
9	0	1	0
10	0	0	2.5
11	0	0	2.5
12	0	1	1
13	0	0	5
14	1	1	1
15	0	1	1
16	0	1.5	1
17	0	1.5	1
18	0	0	3
19	1	1	1.5
20	1	1	0
21	1	1	0
22	0	3.5	1
23	0	0	1.5
24	2.5	2	0
25	0	0	3.5
26	0	0	1
27	0	1	1
28	0	0	3
29	1.5	1.5	4
30	0	0	3
31	0	1	1
32	4.5	4	3
33	0	0	3
34	0	2	2
35	0	0	1
36	0	2	1.5

0 = no operative fusion attempted

Discussion

Non-invasive evaluation of arthrodesis in the postoperative lumbar spine in patients with persistent or recurrent low back pain is important and remains challenging. The gold standard for evaluation of fusion status remains surgical exploration [5, 8–10]; yet, this is not always possible due to patient concerns and for ethical reasons [11]. This study is limited to interobserver variability in non-invasive assessment with dedicated MDCT technique. Interobserver agreement for solid fusion among four observers of pairwise varying experience was moderate. Agreement rates were highest for bilateral solid fusion and declined successively with decreasing fusion grade, they were comparable in either dichotomous categorization. We found slightly higher agreement among consultants though without statistical significant difference.

Strengths are the homogeneous patient population in clinical assessment, age and gender and the standardized dedicated radiological approach. The sample size was comparable to earlier studies [9, 13]. Kappa values were stable in either dichotomy with relatively narrow confidence intervals.

This study has several limitations, it is retrospective and we investigated symptomatic patients only. The kappa value among four raters using five categories is slightly lower than earlier reported in a cross section of 1-year postoperative controls [4]. This may reflect different prevalence and bias indices in the two populations [22]. Difference in observer training was a less likely explanation, as agreement rates were not consistently higher for consultants and confidence intervals were robust in either simplified categorization. Clinical correlation was limited to patient selection as the study focused on interobserver agreement. Referral to radiology due to recurrent or persistent lower back pain and previous posterolateral fusion were the only inclusion criteria, pain or lack of recovery was not correlated to the study. Kamper and colleagues [23] find in a literature review on common definition for recovery from lower back pain no common definition, and Lamberg and colleagues [8] report that radiological and clinical outcomes do not appear to correlate very well.

In recent literature, solid fusion is the state with highest interobserver agreement [4] and best agreement with operative findings [5]. Our study population represented the clinically challenging patient segment; it reproduces the former and further establishes comparable interobserver agreement rates among subspecialists and radiological residents. On comparison to surgery, Carreon et al. [5] reported dedicated fine-cut CT moderately predictive of a solid fusion on surgical exploration in 93 patients over 163 levels and 3 experienced spine surgeons' interobserver agreement on the scans was moderate (facet joints) to substantial (posterolateral gutters). The limited agreement between experienced observers and with comparison to surgical exploration points to inherent difficulty in accurate prediction of fusion. The hierarchical structure of the 5-point scale describes different grades of stability well, but loses its usefulness when grading is uncertain. In this clinical context, we find CT assessment of fusion status as presence or absence of solid fusion useful and suggest interobserver variability is constant beyond a certain level

Table 4 61 operated levels. The number of levels with modified Glassman scores 1–5 [4] per observer, and the number of levels with agreement within groups of observers

Glassman score	1	2	3	4	5	Levels agreed	Kappa
All operated levels, $N = 61$							
Resident 1	33	11	9	6	2		
Resident 2	33	13	6	5	4		
2 Residents	26	4	1	1	0	32	$\kappa = 0.26$, $p = 6 \times 10^{-4}$
Consultant 1	31	13	6	6	5		
Consultant 2	30	15	5	7	4		
2 Consultants	26	5	1	0	1	33	$\kappa = 0.32$, $p = 2 \times 10^{-5}$
4 Observers	20	2	0	0	0	22	$\kappa = 0.32$, $p = 0$
L3/L4, $N = 10$							
Resident 1	6	1	2	0	1		
Resident 2	6	2	1	1	0		
2 Residents	5	1	1	0	0	7	$\kappa = 0.49$, $p = 0.01$
Consultant 1	5	2	0	3	0		
Consultant 2	5	2	1	1	1		
2 Consultants	5	0	0	0	0	5	$\kappa = 0.25$, $p = 0.20$
4 Observers	4	0	0	0	0	4	$\kappa = 0.35$, $p = 5 \times 10^{-6}$
L4/L5, $N = 23$							
Resident 1	13	5	2	3	0		
Resident 2	12	6	3	1	1		
2 Residents	9	2	0	1	0	12	$\kappa = 0.24$, $p = 0.07$
Consultant 1	13	5	2	2	1		
Consultant 2	12	8	0	2	1		
2 Consultants	11	4	0	0	0	15	$\kappa = 0.43$, $p = 0.002$
4 Observers	7	1	0	0	0	8	$\kappa = 0.31$, $p = 10^{-8}$
L5/S1, $N = 28$							
Resident 1	14	5	5	3	1		
Resident 2	15	5	2	3	3		
2 Residents	12	1	0	0	0	13	$\kappa = 0.20$, $p = 0.07$
Consultant 1	13	6	4	1	4		
Consultant 2	13	5	4	4	2		
2 Consultants	10	1	1	0	1	13	$\kappa = 0.24$, $p = 0.02$
4 Observers	9	1	0	0	0	10	$\kappa = 0.30$, $p = 10^{-12}$

Table 5 Bilateral fusion versus all others

	κ	95% CI	PI	BI	PABAK	κ_{max}
4 Observers	0.59					
2 Residents	0.54	0.33–0.75	0.08	0	0.54	1
2 Consultants	0.71	0.53–0.88	0	0.02	0.71	0.97
Residents versus consultants	0.60	0.40–0.79	0.15	0	0.61	1

61 operated levels. Prevalence index (PI). Bias index (BI). Prevalence adjusted bias adjusted kappa (PABAK) [20]

Table 6 Fusion versus partial/non-fusion

	κ	95% CI	PI	BI	PABAK	κ_{max}
4 Observers	0.53					
2 Residents	0.58	0.36–0.79	0.48	0.03	0.67	0.92
2 Consultants	0.46	0.23–0.69	0.46	0.02	0.57	0.96
Residents versus consultants	0.66	0.47–0.86	0.08	0.08	0.70	0.81

61 operated levels. *PI* prevalence index, *BI* bias index, *PABAK* prevalence adjusted bias adjusted kappa [20]

of experience. Larger studies are needed to investigate this further.

Conflict of interest None.

References

1. Christensen F (2004) Lumbar spinal fusion. Outcome in relation to surgical methods, choice of implant and postoperative rehabilitation. Acta Orthop Scand Suppl 75:2–43. doi:10.1080/03008820410002057
2. Resnick D, Choudhri T, Dailey A, Groff M, Khoo L, Matz P, Mummaneni P, Watters W, Wang J, Walters B, Hadley M (2005) American Association of Neurological Surgeons/Congress of Neurological Surgeons, Guidelines for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 4: radiographic assessment of fusion. J Neurosurg Spine 2:653–657. doi:10.3171/spi.2005.2.6.0653
3. Hayeri M, Tehranzadeh J (2009) Diagnostic imaging of spinal fusion and complications. Appl Radiol 38:14–28
4. Carreon L, Glassman S, Djurasovic M (2007) Reliability and agreement between fine-cut CT scans and plain radiography in the evaluation of posterolateral fusions. Spine J 7:39–43. doi:10.1016/j.spinee.2006.04.005
5. Carreon L, Djurasovic M, Glassman S, Sailer P (2007) Diagnostic accuracy and reliability of fine-cut CT scans with reconstructions to determine the status of an instrumented posterolateral fusion with surgical exploration as reference standard. Spine 32:892–895. doi:10.1097/01.brs.0000259808.47104.dd
6. Glassman S, Dimar J, Carreon L, Campbell M, Puno R, Johnson R (2005) Initial fusion rates with recombinant human bone morphogenetic protein-2/compression resistant matrix and a hydroxyapatite and tricalcium phosphate/collagen carrier in posterolateral spinal fusion. Spine 30:1694–1698. doi:10.1097/01.brs.0000172157.39513.80
7. Molinari R, Bridwell K, Klepps S, Baldus C (1999) Minimum 5-year follow-up of anterior column structural allografts in the thoracic and lumbar spine. Spine 24:967–972. doi:10.1097/00007632-199905150-00007
8. Lamberg T, Remes V, Helenius I, Schlenzka D, Yrjönen T, Österman K, Tervahartiala P, Seitsalo S, Poussa M (2005) Long-term clinical, functional and radiological outcome 21 years after posterior or posterolateral fusion in childhood and adolescence isthmic spondylolisthesis. Eur Spine J 14:639–644. doi:10.1007/s05586-004-0814-1
9. Laasonen E, Soini J (1989) Low-back pain after lumbal fusion. Surgical and computed tomographic analysis. Spine 14:210–213. doi:10.1097/00007632-198902000-00011
10. Brodsky A, Kovalsky E, Khalil M (1991) Correlation of radiologic assessment of lumbar spine fusions with surgical exploration. Spine 16:S261–S265. doi:10.1097/00007632-199106001-00017
11. Shah R, Mohammed S, Saifuddin A, Taylor B (2003) Comparison of plain radiographs and CT scan to evaluate interbody fusion following the use of titatiuminterbody cages and transpedicular instrumentation. Eur Spine J 12:378–385. doi:10.1007/s00586-002-0517-4
12. Siambanes D, Mather S (1998) Comparison of plain radiographs and CT scans in instrumented posterior lumbar interbody fusion. Orthopedics 21:165–167
13. Larsen JM, Rimoldi RL, Capen DA, Nelson RW, Nagelberg S, Thomas JC (1996) Assessment of pseudoarthrosis in pedicle screw fusion: a prospective study comparing plain radiographs, flexion/extension radiographs, CT scanning, and bone scintigraphy with operative findings. J Spinal Disord 9:117–120. doi:10.1097/00002517-199604000-00005
14. Robertson D, Weiss P, Fishman E, Magid D, Walker P (1988) Evaluation of CT techniques for reducing artifacts in the presence of metallic orthopedic implants. J Comput Assist Tomogr 12:236–241. doi:10.1097/00004728-198803000-00012
15. Douglas-Akinwande A, Buckwalter K, Rydberg J, Rankin J, Choplin R (2006) Multichannel CT: evaluating the spine in postoperative patients with orthopedic hardware. Radiographics 26:S97–S110. doi:10.1148/rg.26si065512
16. Hanley J (1987) Standard error of the kappa statistic. Psychol Bull 102:315–321. doi:10.1037//0033-2909.102.2.315
17. Dunn G (1989) Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold, London
18. Brennan P, Silman A (1992) Statistical methods for assessing observer variability in clinical measures. BMJ 304:1491–1494. doi:10.1136/bmj.304.6840.1491
19. Feinstein A, Cicchetti D (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43:543–549. doi:10.1016/0895-4356(90)90158-L
20. Byrt T, Bishop J, Carlin J (1993) Bias, prevalence and kappa. J Clin Epidemiol 46:423–429. doi:10.1016/0895-4356(93)90018-V
21. Landis J, Koch G (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. doi:10.2307/2529310
22. Randolph JJ (2005) Free-marginal multirater kappa: an alternative to Fleiss' fixed-marginal multirater kappa. In: Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland
23. Kamper SJ, Stanton TR, Williams CM, Maher CG, Hush JM (2010) How is recovery from low back pain measured? A systematic review of the literature. Eur Spine J 20:9–18. doi:10.1007/s00586-010-1477-8