Radiologic Assessment of Interbody Fusion

A Systematic Review on the Use, Reliability, and Accuracy of Current Fusion Criteria

Anneli A.A. Duits, MD
Paul R. van Urk, MD
A. Mechteld Lehr, PhD
Don Nutzinger, MD
Maarten R.L. Reijnders, MD
Harrie Weinans, PhD
Wouter Foppen, MD, PhD
F. Cuhmur Oner, MD, PhD
Steven M. van Gaalen, MD, PhD
Moyo C. Kruyt, MD, PhD

Investigation performed at the University Medical Center Utrecht, Utrecht, the Netherlands

Abstract

Background: Lumbar interbody fusion (IF) is a common procedure to fuse the anterior spine. However, a lack of consensus on image-based fusion assessment limits the validity and comparison of IF studies. This systematic review aims to (1) report on IF assessment strategies and definitions and (2) summarize available literature on the diagnostic reliability and accuracy of these assessments.

Methods: Two searches were performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines. Search 1 identified studies on adult lumbar IF that provided a detailed description of image-based fusion assessment. Search 2 analyzed studies on the reliability of specific fusion criteria/classifications and the accuracy assessed with surgical exploration.

Results: A total of 442 studies were included for search 1 and 8 studies for search 2. Fusion assessment throughout the literature was highly variable. Eighteen definitions and more than 250 unique fusion assessment methods were identified. The criteria that showed most consistent use were continuity of bony bridging, radiolucency around the cage, and angular motion $<5°$. However, reliability and accuracy studies were scarce.

Conclusion: This review highlights the challenges in reaching consensus on IF assessment. The variability in IF assessment is very high, which limits the translatability of studies. Accuracy studies are needed to guide innovations of assessment. Future IF assessment strategies should focus on the standardization of computed tomography–based continuity of bony bridging. Knowledge from preclinical and imaging studies can add valuable information to this ongoing discussion.

Level of Evidence: Diagnostic Level III. See Instructions for Authors for a complete description of levels of evidence.

Interbody fusion (IF) is a commonly performed surgical technique in the lumbar spine that has increased in popularity in recent years[1-3]. During an IF procedure, the intervertebral disk is removed and replaced with an IF cage or graft to promote bony fusion. Image-based fusion assessment is a commonly used primary outcome in IF studies[4,5] but is challenging for several reasons. Currently

Disclosure: The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJSREV/B53).

no widely accepted definition of fusion exists. Fusion is defined formally as "fusion of bones across a joint space by surgical means, which eliminates movement,"[6] but this definition lacks quantitative cutoffs, resulting in variable interpretations[7,8]. In addition, imaging limitations compounded by imaging artefacts from cage and graft materials complicate fusion assessment[9,10]. The lack of consensus for image-based fusion assessment may lead to methodological variations between studies, potentially limiting the relevance, and generalizability of their results. Although previous reviews have provided a general discussion of the existing imaging modalities[5,11,12], systematic reviews of image-based fusion criteria are lacking, adding another potential source of variation between studies.

This systematic review aims (1) to document what modalities, fusion definitions, and criteria/classifications are used for lumbar IF assessment and (2) to report the available evidence for diagnostic reliability (interobserver and intraobserver variation) and accuracy (compared with surgical exploration).

Methods

This systematic review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines[13] and involved 2 searches to cover both aims. The first search served to generate a database on current image-based fusion assessment methods of clinical studies after thoracolumbar and lumbar spine (T10-S1) IF. The second search identified diagnostic reliability and accuracy studies on fusion criteria and classifications.

Literature Search

Medline and Embase databases were searched from 1946 until November 2018 (search 1) and November 2021 (search 2). Additional articles were searched in the Cochrane databases for search 1 and by bibliography screening of relevant systematic reviews for search 2. Only articles written in English, German, or French were selected. Combinations of search terms related to the keywords listed in Figure 1 were used to create 2 databases (see Appendix A for the search string).

For both searches, 2 reviewers (A.A.A.D., A.M.L.) independently screened titles/abstracts using Rayyan, and full-texts for eligibility according to the inclusion and exclusion criteria reported in Figure 1 using Zotero (version 5; Corporation for Digital Scholarship). Disagreement was resolved through discussion, or a third review author was consulted (F.C.O.).

Risk of Bias Assessment

Two reviewers (A.A.A.D., A.M.L., and/or P.R.v.U.) independently assessed risk of bias for the studies included in search 2, using the Quality Appraisal of Reliability Studies (QAREL) and Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) checklists, depending on the study outcome. Risk of bias was deemed high for reliability studies if less than 60% of QAREL signaling questions were answered "yes" and for accuracy studies if more than 2 of the QUADAS-2 signaling questions were answered "no" or "unclear."

Data Extraction

Articles in search 1 were analyzed for (1) study characteristics including study

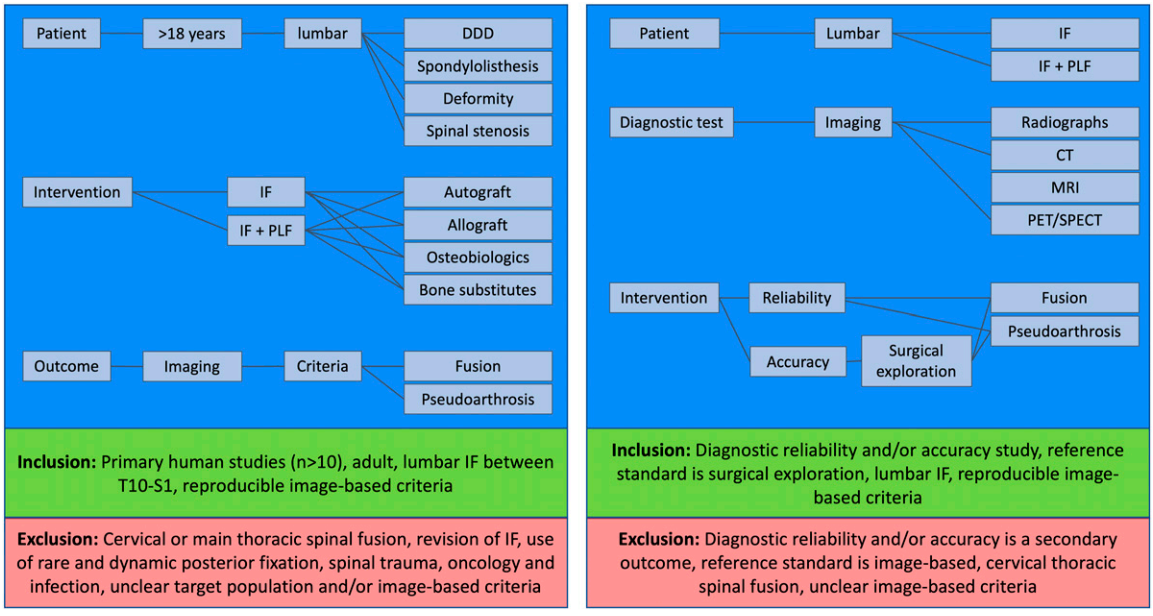

Fig. 1

Keywords, inclusion and exclusion criteria used for the search string of search 1 (left) and for search 2 (right). CT = computed tomography, DDD = degenerative disk disease, IF = interbody fusion, MRI = magnetic resonance imaging, PET = positron electron emission tomography, PLF = posterolateral fusion, and SPECT = single-photon emission computed tomography.

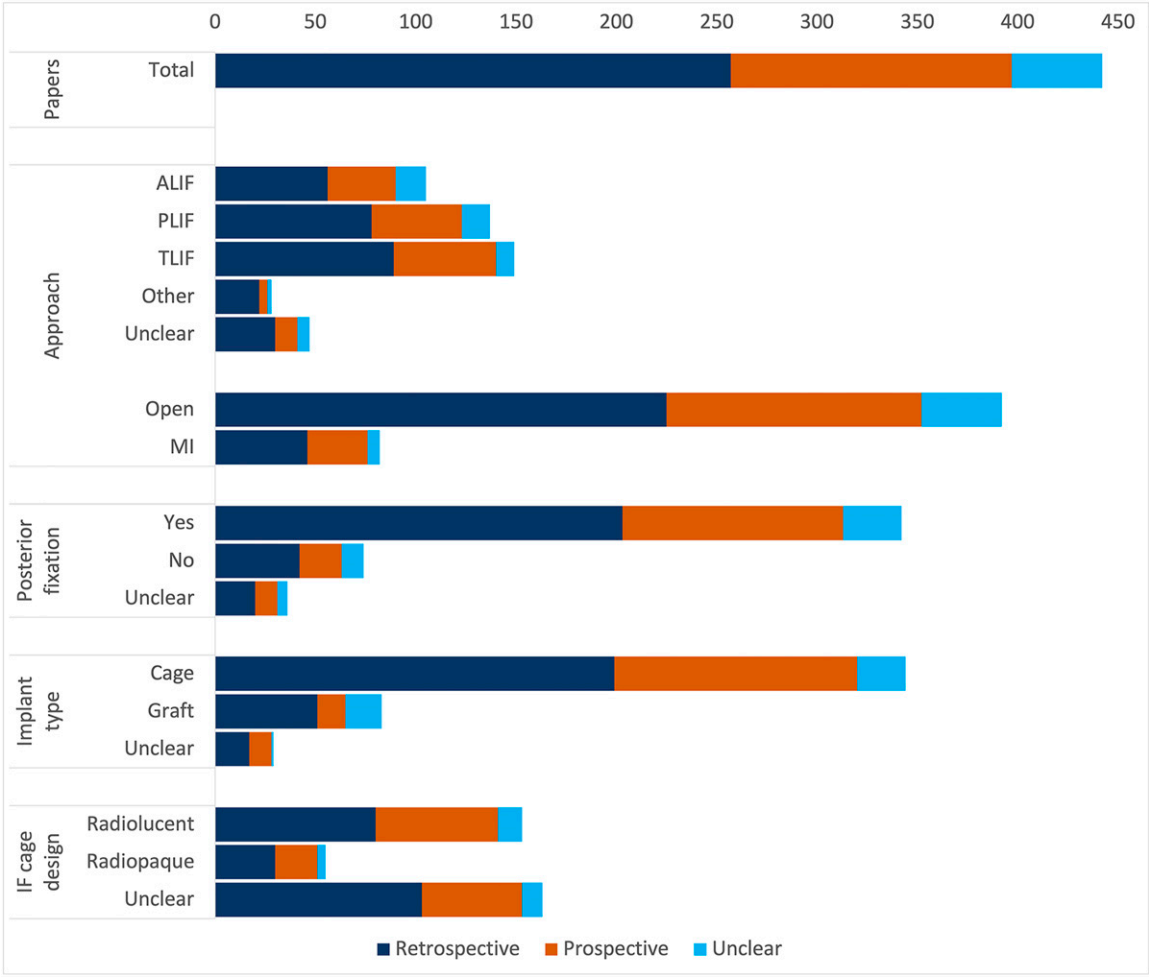

Fig. 2

Study characteristics of included studies from search 1. ALIF = anterior lumbar IF, IF = interbody fusion, MI = minimally invasive, PLIF = posterior lumbar IF, and TLIF = transformational lumbar IF.

type, year of publication, number of participants/segments, and surgical details; (2) imaging modality; (3) fusion criteria and cutoff; (4) follow-up time; and (5) fusion outcomes. For search 2, we analyzed (1) study type, (2) imaging modality, (3) fusion criteria and cutoff, (4) surgical exploration technique, (5) reliability parameters, and (6) accuracy compared with surgical exploration. All data were gathered and collected in an Excel file by a paired team of 2 of 4 authors (A.A.A.D., A.M.L., D.N., M.R.L.R.).

Data Analysis

The fusion criteria collected from search 1 were coded as "descriptive," in case of a pragmatic description of fusion not part of a classification, and "classification," in case of a grading or scoring system. Furthermore, all fusion criteria and cutoffs were recoded based on their use in the original article. Study characteristics and the frequency of use for the modalities, image-based fusion definitions, and criteria/classifications were calculated as absolute frequency (number of articles) and relative frequency (% of all articles).

For the diagnostic reability studies in search 2, the weighted Cohen κ-value for agreement between multiple observers and repeated measures by 1 observer was collected. The κ-values were interpreted as none (κ 0-0.20), minimal (κ 0.21-0.39), weak (κ 0.40-0.59), moderate (κ 0.60-0.79), strong (κ 0.80-0.90), and almost perfect (κ > 0.90) agreement[14]. In case other measures for reliability were used, the interpretation of the article was used. For the diagnostic accuracy studies, sensitivity, specificity, positive predictive value, negative predictive value, positive and negative likelihood ratios (LR+ and LR−), prevalence of pseudoarthrosis, and accuracy (percentage consistent with surgical exploration) were calculated from the contingency tables for fusion criteria/classifications compared with intraoperative findings. Pseudoarthrosis was defined as a positive test result. To allow for LR calculation in studies that reported 100% sensitivity or specificity, 0.5 was added to all cells of the contingency table to avoid division by 0[15].

Results

Literature Searches

Search 1 was performed in November 2018 and yielded 3,199 unique articles

TABLE I Criteria for Assessing IF*

Criteria Used to Define IF	No. of Articles (%)
1. Continuity of bony bridging	393 (89)
2. Signs of union†	142 (32)
3. Instability	313 (71)
a. Angular motion	135 (31)
b. Hardware failure	101 (23)
c. Loss of alignment	102 (23)
d. Translation	40 (9)
4. Radiolucency	266 (60)
a. Around the cage	229 (52)
b. Cleft in fusion mass	82 (19)
5. Signs of nonunion‡	17 (4)
Most common pragmatic definitions of IF	
1. Continuity of bony bridging/instability/radiolucency/ signs of union	95 (21)
2. Continuity of bony bridging/instability/radiolucency	87 (20)
3. Continuity of bony bridging/instability	81 (18)
4. Continuity of bony bridging	64 (14)
5. Instability/radiolucency	23 (6)

*IF = interbody fusion
†Signs of union: blurring of the endplates, bone maturation, increased bone density, mottling of the graft (sign of vascularization), resorption of the anterior traction spur, sentinel sign (continuous bone bridge anterior to the disk space), and trabecular structure in the graft.
‡Signs of nonunion: cystic lesions of the endplate, resorption of the bone graft, resorption of the endplate, sclerosis of the endplate, and vacuum phenomenon.

of which 830 were assessed for eligibility. A total of 442 reported the applied fusion method in detail and were included for full-text analysis. Sixty-eight articles were excluded specifically because the fusion assessment method was not described. Search 2 was performed in July 2021 and yielded 290 unique studies of which 229 were assessed for eligibility. A total of 8 studies were included for full-text analysis, 5 studies that assessed reliability and 3 studies that assessed both reliability and accuracy based on surgical exploration or histology. A detailed description of both searches is provided in the PRISMA flow diagram in Appendix B.

Part 1: What Is Used for Fusion Assessment

Study Characteristics

The first study of search 1 dated from 1968[16], and since the early 1990s, scientific interest increased. Studies had either a retrospective (58%), prospective (32%), or unclear design (10%). Initially, the proportion of prospective and retrospective studies was similar, but after 2008, relatively more retrospective studies were published reflecting the widespread clinical use of IF. Study characteristics are summarized in Figure 2. IF was mainly performed through a posterior approach (posterior lumbar IF or transformational lumber IF). Most studies used a cage that was predominantly radiolucent among the studies that reported the cage type. In 76% of studies, IF was supplemented with posterior fixation. Study characteristics were not always clearly described. The study design, approach, supplementary fixation, and/or implant/graft use were unclear in 6% to 10% of articles. However, especially the radiographic cage appearance (radiopaque/radiolucent) was commonly unclear (44% of articles).

Imaging Modality

All studies used conventional radiographs (CRs) (57%), dynamic radiographs (53%), and/or computed tomography (CT) scans (47%). The use of different modalities changed over time. Before 2013, almost half of the articles used CR for fusion assessment and less than 20% used CT scans. After 2013, CT scans were most often used (36% of articles), but both CR (32%) and dynamic radiographs (32%) remained popular. Dynamic CT (experimentally)[17] and magnetic resonance imaging (MRI) (as standard practice)[18] were used once for fusion assessment. None of the included articles used ultrasound, positron emission tomography or single-photon emission CT.

Radiographic Definitions of IF

Studies defined image-based fusion as positive signs of bony fusion and/or absence of negative signs. Positive signs consisted of continuity of bony bridging and other signs of union (such as sentinel sign and bone maturation) (Fig. 3). Negative signs consisted of dynamic measures of instability (angular motion or translation), static measures of instability (hardware failure and loss of correction), radiolucency, and others (such as cystic lesions of the endplate and sclerosis of the endplate) (Fig. 3). Studies most commonly assessed the continuity of bony bridging (89%), instability (71%), and radiolucency (60%) (Table I). Two-thirds (65%) of the articles described fusion with descriptive criteria, 32% with a classification and 3% with a combination of both. Eighteen radiographic IF definitions were identified of which the 5 most prevalent are listed in Table I.

Descriptive Criteria and Classifications

Fifty-seven descriptive fusion criteria were identified. These are presented in Figure 3. Criteria for continuity of bony bridging and radiolucency were usually qualitative. "Radiolucency <50% around the cage" was the only quantitative criterium that showed some consistent use (35 articles). The dynamic instability criteria were quantitative, but many different cutoffs existed. Acceptable translation ranged from 1.5 to 5 mm and angular motion from 0 to 11°. Of these criteria, angular motion of <5°

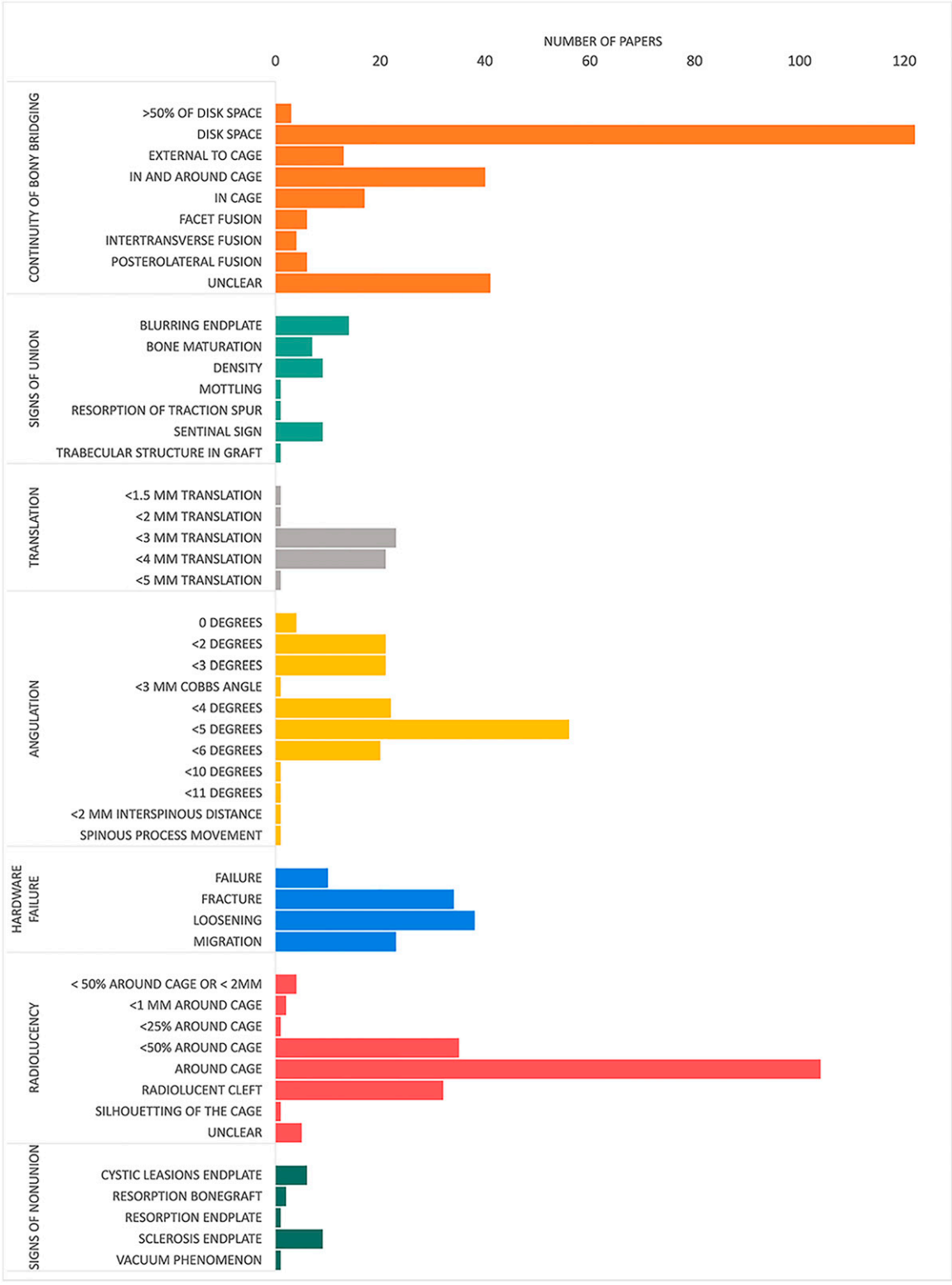

Fig. 3

Absolute frequency per used unique fusion criterion.

TABLE II Most Used IF Classifications*

Bridwell anterior fusion grades[19] (10% of articles)	
Grade I	Fused with remodeling and trabeculae
Grade II	Graft intact, not fully remodeled and incorporated though. No lucencies
Grade III	Graft intact, but a definite lucency at the top or bottom of the graft
Grade IV	Definitely not fused with resorption of bone graft and with collapse
Brantigan and Steffee[20] (5% of articles)	
Grade 1	Obvious radiographic pseudoarthrosis: based on collapse of the construct, loss of disk height, vertebral slip, broken screws, displacement of the carbon cage, or resorption of the bone graft
Grade 2	Probable radiographic pseudarthrosis: based on significant resorption of the bone graft, or a major lucency or gap visible in the fusion area (2 mm or more around the entire periphery of the graft or cage)
Grade 3	Radiographic status uncertain: Bone graft is visible in the fusion area at approximately the density originally achieved surgically. A small lucency or gap may be visible involving just a portion of the fusion area with at least half of the graft area showing no lucency between the graft bone and the vertebral bone
Grade 4	Probable radiographic fusion: Bone bridges the entire fusion area with at least the density originally achieved at surgery. No lucency between the donor bone and the vertebral bone
Grade 5	Radiographic fusion: The bone in the fusion area is more dense and mature than achieved at surgery. Optimally there is no interface between the donor bone and the vertebral bone; however, a sclerotic line between the graft and the vertebral bone indicates fusion. Other signs of solid fusion include mature bony trabeculae bridging the fusion area, resorption of anterior vertebral traction spur, anterior progression of the graft within the disk space, fusion of facet joints, the "ring" phenomenon on CT, or 3D imaging evidence
Lenke posterior fusion grades[21] (2% of articles)	
Grade A	Solid trabeculated transverse process and facet fusions bilaterally
Grade B	Thick fusion mass on one side. Difficult to visualize on the other side
Grade C	Suspected lucency or defect in the fusion mass
Grade D	Definite resorption of graft with fatigue of instrumentation

*3D = 3-dimensional, CT = computed tomography, IF = interbody fusion.

showed most consistent use (56 articles). In 51 articles, a combination of translation, angular motion, and hardware failure was used to define instability. Continuity of bony bridging (in the disk space) and radiolucency (around the cage) were the only criteria used in more than 75% of the studies.

Twenty-nine different classification systems were identified. These classifications typically consisted of 3- to 5-point scales or scorings. The anterior fusion grades (grades I-IV) of Bridwell et al.[19] and the 5-point scale (grades 1-5) developed by Brantigan and Steffee et al.[20] were most frequently used (Table II). The Bridwell classification and Brantigan and Steffee classification were based on CT in 16 and 8 articles, respectively, and on radiographs in 22 and 12 articles. For some classifications, various cutoff values were used. For instance, both Bridwell grade I (15 articles) and grades I and II (18 articles) could be considered fusion. In addition, in 10% of the studies, a grade describing an incomplete bony bridge was considered "fused" if the radiolucent gap around the cage was minimal. Classifications more often focused on positive signs of union, such as density and maturation, as compared with the pragmatic combinations of fusion criteria.

Overall, the descriptive criteria and classifications were used in a total of 256 unique combinations. Only 20% of these combinations were used in more than 5 studies and 45% was used by only a single article. Most frequently used were bridging bone in the disk space (23 articles), the Bridwell anterior fusion grading with fusion defined as grades I and II (20 articles), and the Brantigan and Steffee grading with fusion defined as grades 4 and 5 (13 articles).

Part 2: Reliability and Accuracy of Fusion Criteria and Classifications

Study Characteristics

Eight studies were included (7 human and 1 animal). The study characteristics are summarized in Table III. In total, the reliability of 11 criteria and 3 classifications was tested on 4 imaging modalities. Accuracy was measured for 4 criteria and 1 classification on 3 modalities.

Risk of Bias Analysis

The risk of bias for accuracy studies was low in 2 studies[22,23] and high in 1 (Appendix C)[24]. None of the reliability studies met the criteria for "high" quality.

Reliability

Seven studies assessed interobserver reliability between observers for descriptive criteria[22,23,25-29] and 2 studies for classifications[24,29] (Table IV). Most studies used κ values. Other measures of reliability were only reported incidentally. Zhou et al. reported an intraclass correlation coefficient for intraobserver reliability of 0.90 (CI 0.85-0.93) for radiolucency around the cage detected from signal intensity measurements on MRI, with an analysis interval of 6 months[23].

Soriano Sánchez et al. measured agreement between classification

TABLE III Study Characteristics Reliability and Accuracy Studies*

Author	Year	Participants (Segments)	Study Design	Cages	Imaging Modality
Reliability					
Fujibayashi et al.[22]	2012	76 (93)	Retrospective	Metallic	CT, dynamic radiographs
Kröner et al.[23]	2006	47 (49)	Prospective	Polymer	MRI
Shah et al.[24]	2003	53 (156)	Prospective	Metallic	CR, CT
Slosar et al.[25]	2015	33 (56)	Prospective	Polymer	CT
Soriano Sánchez et al.[26]	2020	50 (90)	RCT	Unclear	CT
Reliability and accuracy					
Carreon et al.[27]	2008	49 (69)	Unclear	Metallic	CT
Fogel et al.[28]	2008	90 (172)	Retrospective	Polymer	CR, CT
Zhou et al.[29]	2015	12 (36)	Prospective†	Tantalum	CR, MRI

*CR = conventional radiograph, CT = computed tomography, and MRI = magnetic resonance imaging.
†Animal study

systems for all observers. The intraobserver evaluation correlation of 3 observers was 0.602, 0.789, and 0.505 between the Lenke and Bridwell classifications; 0.639, 0.825, and 0.535 between the Lenke and Bratingan-Steffee-Fraser (BSF) classifications; and 0.685, 0.825, and 0.026 between the Bridwell and BSF classifications[29] (see Table II for a description of the classifications). Reported interobserver agreement for the descriptive criteria was usually strong or almost perfect[22,23,25-29]. Interobserver agreement for the classifications was reported poor by one study and good by another[29].

Accuracy

In both human accuracy studies, IF was combined with PLF[22,24]. In the study by Carreon et al., IF was tested separately during surgical exploration after removal of the posterior instrumentation[22]. In the animal study, a 3-level IF procedure was performed with supplemental anterior fixation in twelve 3-month-old female Danish Landrace pigs. Image-based findings were compared with histology instead of surgical exploration[23]. Quantitative accuracy measures are summarized in Table V. Agreement between the image-based analysis and surgical exploration ranged from 61% to 89%. Radiolucency and the BSF classification were the most accurate predictors of pseudoarthrosis.

Discussion

The first part of this systematic review analyzed circulating image-based IF definitions and assessment methods. An enormous variation in both the definition and assessment methodology of IF was found. There were 18 different definitions of IF and more than 250 combinations of criteria and classifications that showed very little repetition. These findings are similar to our previous review on posterolateral fusion assessment[30] and confirm the lack of consensus that was described earlier[5,8,31]. Somehow attempts at standardization of IF assessment have mainly focused on imaging modalities[5,7,11,32], but this review also shows substantial variation exists in fusion criteria/classifications that warrant attention. The influence of fusion criteria and classifications on reported fusion rates is demonstrated by various articles in this review[33-36]. For example, Isaacs et al. showed fusion rates ranging from 74% to 100%, depending on the criteria used[35]. Another study showed the same for different fusion classifications, where fusion rates ranged from 43% to 79%[29]. As a consequence, comparing fusion rates between IF studies is currently almost impossible, even when the same imaging modality is used. This means that the scientific impact of most studies is low, and comparison of techniques can only be performed in a series of comparative trials where consensus on assessment has been achieved.

Descriptive criteria were more used throughout the literature than classifications. The most common criteria were continuity of bony bridging, radiolucency around the cage, and angular motion less than 5°. Diagnostic reliability and accuracy studies were scarce and covered only a subset of criteria. None of the reliability or accuracy studies assessed criteria for dynamic instability. The other validated criteria demonstrated strong interobserver agreement for most imaging modalities. However, the accuracy of these criteria remains unknown because accuracy studies were too limited in number and too heterogenous for generalizable conclusions.

The lack of accuracy studies to confirm the real presence of a solid, bony fusion presents a significant challenge in the development of a relevant and accurate image-based fusion assessment. Many valuable insights are available from preclinical research and constraints of imaging techniques that can help standardize fusion assessment[9,10]. Unfortunately, these tend to be neglected in current literature[5,7,11]. For example, preclinical IF models show that the fusion mass forms at a variable rate and is irregular until it is matured[10]. These findings suggest pseudoarthrosis is more likely detected by 3-dimensional imaging, such as CT scans, than by

TABLE IV Reliability of Descriptive Fusion Criteria and Fusion Classifications*

Reliability	Modality	Fusion Rate (%)	Prevalence Criterium (%)†	Interobserver Variability (κ)
Continuity of bony bridging				
Disk space	CT[27]			0.25
In cage	CR[24]		4	0.74
	CT[24]		95/96	0.85
	MRI[23]	88	88/84	0.88
External to cage	CR[24]		88/92	0.86
	CT[24]		7/8	0.82
Between cages	MRI[23]	88	86	1.00
Lateral			57	0.88
Anterior			52/51	0.86
Posterior			85/80	0.84
Radiolucency				
Around the cage	CR[24,29]		20[29]	0.81
			3/1[24]	0.66
	CT[24,25]	80/91	2/4[25]	0.96
			6/4[24]	0.74
	MRI T1[29]		30	0.88
	MRI T2[29]		23	0.88
Signs of (non)union				
Trabecular bone	CT[25]	80/91	96/100	0.96
Anterior sentinel sign	CT[27]			0.34
	CT[25]	80/91	48/70	0.77
Posterior sentinel sign	CT[27]			0.23
Cystic lesions	CT[22,25]	80/91	4/5[25]	0.95
		75	17[22]	0.86
Other				
HU value	CT[26]			0.862-0.943‡
Combined criteria§	Dynamic x-ray/CT[22]			0.83
Classifications				
Lenke	CT[26]	79/67		0.248-0.315‡
Grade A			40/18	
Grade B			48/49	
Grade C			10/28	
Grade D			2/6	
Bridwell	CT[26]	43		0.246-0.346‡
Grade I			22/29/8	
Grade II			0/16/56	
Grade III			63/22/26	
Grade IV			14/33/11	
BSF	CR[28]	87		98.6%¶
	CT[26,28]	79[26], 77[28]		0.197-0.329[26]‡
Grade 3	CR[28]		87	
	CT[26,28]		77[28]	
			27/31/29[26]	
Grade 2	CR[28]		12	
	CT[26,28]		17[28]	
			59/43/49[26]	
Grade 1	CR[28]		2	
	CT[26,28]		2[28]	
			14/26/22[26]	

*BSF = Brantigan-Steffee-Fraser, CR = conventional radiograph, CT = computed tomography, HU = Hounsfield unit, MRI = magnetic resonance imaging.
†Percentage of the radiologic criterium was detected in the population. Results reported separate for multiple observers in case available.
‡Interobserver evaluation correlation.
§CT-based criteria: continuity of bony bridging, absence of radiolucency around the cage or screws. Dynamic radiograph criteria: <4° of motion.
¶Agreement between observers.

TABLE V Accuracy of IF Criteria and Classifications*

Criterium	Modality	Sens	Spec	PPV	NPV	LR+	LR−	Prev	Acc
Continuity of bony bridging									
Disk space	CT[27]	0.93	0.46	0.57	0.90	1.73	0.14	0.435	0.67
Radiolucency									
Around cage	CR[29]	0.79	0.92	0.92	0.79	22.4	0.34	0.364	0.82
	MRI T1[29]	0.85	0.80	0.80	0.85	7.5	0.46	0.364	0.83
	MRI T2[29]	0.57	0.87	0.87	0.80	12.1	0.49	0.364	0.82
Signs of union									
Anterior sentinel sign	CT[27]	0.20	0.92	0.66	0.60	2.5	0.9	0.435	0.61
Posterior sentinel sign	CT[27]	0.67	0.79	0.71	0.76	3.2	0.4	0.435	0.74
Classifications									
BSF	CR[28]	0.9	0.89	0.19	1.0	7.8	0.1	0.023	0.89
	CT[28]	0.9	0.85	0.23	1.0	6.2	0.1	0.023	0.86

*Acc = accuracy defined as rate consistent with surgical exploration, BSF = Brantigan-Steffee-Fraser, IF = interbody fusion, LR− = negative likelihood ratio, LR+ = positive likelihood ratio, NPV = negative predictive value, PPV = positive predictive value, Prev = prevalence defined as rate of pseudoarthrosis, Sens = sensitivity, and Specs = specificity.

2-dimensional (2D) imaging. This is supported by the typically higher 2D radiographic fusion rates compared with CT-based fusion rates for the same populations[34,37].

In our opinion, angular motion or other parameters of dynamic instability are currently the least suitable for IF assessment because IF is commonly combined with supplemental fixation that will influence mobility. Moreover, a cadaver study that compared controlled movement with radiographic measurements suggested false-positive and false-negative rates are high for dynamic instability measures[38]. We found no accuracy studies to prove otherwise. Currently, the U.S. Food and Drug Administration (FDA) still heavily relies on dynamic stability criteria based on their guidance documents for Investigational Device Exemption application for spinal systems[39]. Apart from continuity of bony bridging, the FDA defines fusion as the absence of translational motion (>3 mm) and angular motion (>5°)[39].

Recommendations

This systematic review shows the need for a straightforward, reproducible, and accurate method to assess IF. Although a multimodality approach as suggested by Choudhri et al. may still be useful for clinical practice, researchers should aim for a single–modality-based assessment, not dependent on additional fixation, to prevent unwanted variation between studies. Continuity of bony bridging in the disk space based on the CT scan seems most appropriate, although its superiority could not be inequivocally demonstrated in this systematic review. Continuity of bony bridging is the only criterion directly related to the fusion status. In most reliability studies, the interobserver reliability for CT-based continuity of bony bridging was substantial to almost perfect. Although the positive predictive value for fusion detection was suboptimal (57%), the detection of pseudoarthrosis was relatively accurate (90%)[22]. Among the commonly used imaging modalities, the CT scan seemed to overestimate the fusion rate the least[33,35,40]. Therefore, it is already commonly used as a last resort when pseudoarthrosis is suspected but could not be detected by other imaging techniques[37,41].

Some limitations of this systematic review should be noted. The full text of many non–English-language articles could not be retrieved, resulting in the inclusion of mostly English-language articles. Furthermore, only a very limited number of accuracy and reliability studies could be found in search 2. The included clinical accuracy studies were performed among a subpopulation in need of revision surgery. Therefore, the reported results may not be representative for the full IF population.

Future Research

Future research should focus on the development of reliable and accurate fusion classifications to guide CT-based assessment of the bony bridge. Potentially multiple classifications are needed depending on the cage design. For instance, the use of porous titanium implants especially without a lumen presents new challenges for assessing fusion because the classic bone bridge will not develop[42]. Currently, image-based determination of fusion status in human trials is imperfect. Surgical exploration of all cases in a prospective (randomized) trial is however highly unethical. Improved image quality and resolution of CT scans are the most promising strategies to replace surgical exploration as the gold standard. Photon-counting detector CT is an existing new development that can provide high-quality, ultra–high-resolution images[43,44]. However, this technique is still relatively new, and the accuracy of this technique to assess IF is yet to be determined. Thoroughly validating improved imaging techniques

to substitute surgical exploration are of utmost importance to improve and standardize image-based fusion assessment. Large animal IF models and postmortem studies should be used to test the diagnostic accuracy of fusion classifications using improved imaging techniques, for both innovative and standard implants, compared with manual palpation and histologic assessment[45].

Source of Funding

No financial support was received for this research.

Appendix

Supporting material provided by the authors is posted with the online version of this article as a data supplement at jbjs.org (http://links.lww.com/JBJSREV/B54). This content was not copyedited or verified by JBJS.

NOTE: The authors thank Wilco C.H. Jacobs, PhD (The Health Scientist, the Hague, the Netherlands) for his advice on the systematic review protocol, Paulien H. Wiersma (Utrecht University, Utrecht, the Netherlands) for her help on the search strings, Nicole Swart who used her knowledge as a chartered controller to help the excel analysis and the Department of Orthopedic surgery of the Diakonessenhuis Utrecht/Zeist, especially Arthur de Gast, for their help with the initiation of this systematic review.

Anneli A.A. Duits, MD[1,2,3],
Paul R. van Urk, MD[1],
A. Mechteld Lehr, PhD[1],
Don Nutzinger, MD[1],
Maarten R.L. Reijnders, MD[1],
Harrie Weinans, PhD[1,4],
Wouter Foppen, MD, PhD[1],
F. Cuhmur Oner, MD, PhD[1],
Steven M. van Gaalen, MD, PhD[1,5],
Moyo C. Kruyt, MD, PhD[1,6]

[1]Department of Orthopedic Surgery, University Medical Center Utrecht, Utrecht, the Netherlands

[2]Department of Orthopedic surgery, Diakonessenhuis, Utrecht, Zeist, the Netherlands

[3]Department of Orthopedics, Clinical Orthopedic Research Center (CORC-mN), Diakonessenhuis Utrecht/Zeist, Utrecht, the Netherlands

[4]Department of biomechanical Engineering, Delft University of Technology, Delft, the Netherlands

[5]Department of Orthopedic Surgery, Acibadem Internal Medical Center, Amsterdam, the Netherlands

[6]Department of Developmental BioEngineering, University of Twente, Enschede, the Netherlands

Email for corresponding author: a.a.a.duits@umcutrecht.nl

References

1. Cortesi PA, Assietti R, Cuzzocrea F, Prestamburgo D, Pluderi M, Cozzolino P, Tito P, Vanelli R, Cecconi D, Borsa S, Cesana G, Mantovani LG. Epidemiologic and economic burden attributable to first spinal fusion surgery: analysis from an Italian administrative database. Spine (Phila Pa 1976). 2017;42(18):1398-404.

2. Kobayashi K, Ando K, Nishida Y, Ishiguro N, Imagama S. Epidemiological trends in spine surgery over 10 years in a multicenter database. Eur Spine J. 2018;27(8):1698-703.

3. Pannell WC, Savin DD, Scott TP, Wang JC, Daubs MD. Trends in the surgical treatment of lumbar spine disease in the United States. Spine J. 2015;15(8):1719-27.

4. Finkelstein JA, Schwartz CE. Patient-reported outcomes in spine surgery: past, current, and future directions. J Neurosurg Spine. 2019;31(2):155-64.

5. Choudhri TF, Mummaneni PV, Dhall SS, Eck JC, Groff MW, Ghogawala Z, Watters WC, Dailey AT, Resnick DK, Sharan A, Wang JC, Kaiser MG. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 4: radiographic assessment of fusion status. J Neurosurg Spine. 2014;21(1):23-30.

6. Law J, Martin E, eds. Concise Medical Dictionary. Oxford, UK: Oxford University Press; 2020.

7. Gruskay JA, Webb ML, Grauer JN. Methods of evaluating lumbar and cervical fusion. Spine J. 2014;14(3):531-9.

8. McAfee PC, Boden SD, Brantigan JW, Fraser RD, Kuslich SD, Oxland TR, Panjabi MM, Ray CD, Zdeblick TA. Symposium: a critical discrepancy: a criteria of successful arthrodesis following interbody spinal fusions. Spine (Phila Pa 1976). 2001;26(3):320-34.

9. Bushberg JT, Seibert JA, Leidholdt EM, Boone JM, Abbey CK. The Essential Physics of Medical Imaging. 4th ed. Philadelphia, PA: Wolters Kluwer; 2021.

10. Walsh WR, Pelletier MH, Wang T, Lovric V, Morberg P, Mobbs RJ. Does implantation site influence bone ingrowth into 3D-printed porous implants? Spine J. 2019;19(11):1885-98.10.1016/j.spinee.2019.06.020

11. Lee Y-P, Farhan SA, Musa A, Bhatia N. Pseudarthrosis in spine surgery: diagnosis and treatment. Contemp Spine Surg. 2019;20(8):1-7.

12. Williams AL, Gornet MF, Burkus JK. CT evaluation of lumbar interbody fusion: current concepts. Am J Neuroradiol. 2005;26(8):2057-66.

13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

14. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276-82.

15. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2019.

16. Hodgson AR, Wong SK. A description of a technic and evaluation of results in anterior spinal fusion for deranged intervertebral disk and spondylolisthesis. Clin Orthop Relat Res. 1968;56:133-62.

17. Nakashima H, Yukawa Y, Ito K, Horie Y, Machino M, Kanbara S, Morita D, Imagama S, Ishiguro N, Kato F. Extension CT scan: its suitability for assessing fusion after posterior lumbar interbody fusion. Eur Spine J. 2011;20(9):1496-502.

18. Demirayak M, Sisman L, Turkmen F, Efe D, Pekince O, Goncu RG, Sever C. Clinical and radiological results of microsurgical posterior lumbar interbody fusion and decompression without posterior instrumentation for lateral recess stenosis. Asian Spine J. 2015;9(5):713-20.

19. Bridwell KH, Lenke LG, McEnery KW, Baldus C, Blanke K. Anterior fresh frozen structural allografts in the thoracic and lumbar spine. Do they work if combined with posterior fusion and instrumentation in adult patients with kyphosis or anterior column defects? Spine (Phila Pa 1976). 1995;20(12):1410-8.

20. Brantigan JW, Steffee AD. A carbon fiber implant to aid interbody lumbar fusion. Two-year clinical results in the first 26 patients. Spine (Phila Pa 1976). 1993;18(14):2106-7.

21. Lenke LG, Bridwell KH, Bullis D, Betz RR, Baldus C, Schoenecker PL. Results of in situ fusion for isthmic spondylolisthesis. J Spinal Disord. 1992;5(4):433-42.

22. Carreon L, Djurasovic M, Glassman S, Sailer P. 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 (Phila Pa 1976). 2007;32(8):892-5.

23. Zhou Z, Wei F, Huang S, Gao M, Li H, Stødkilde-Jørgensen H, Lind M, Bünger C, Zou X. In vivo magnetic resonance imaging evaluation of porous tantalum interbody fusion devices in a porcine spinal arthrodesis model. Spine (Phila Pa 1976). 2015;40(19):1471-8.

24. Fogel G, Toohey J, Neidre A, Brantigan J. Fusion assessment of posterior lumbar interbody fusion using radiolucent cages: x-ray films and helical computed tomography scans compared with surgical exploration of fusion. Spine J. 2008;8(4):570-7.

25. Fujibayashi S, Takemoto M, Izeki M, Takahashi Y, Nakayama T, Neo M. Does the formation of vertebral endplate cysts predict nonunion after lumbar interbody fusion? Spine (Phila Pa 1976). 2012;37(19):E1197-202.

26. Kröner AH, Eyb R, Lange A, Lomoschitz K, Mahdi T, Engel A. Magnetic resonance imaging evaluation of posterior lumbar interbody fusion. Spine (Phila Pa 1976). 2006;31(12): 1365-71.

27. Shah RR, Mohammed S, Saifuddin A, Taylor BA. Comparison of plain radiographs with CT scan to evaluate interbody fusion following the use of titanium interbody cages and transpedicular instrumentation. Eur Spine J. 2003;12(4):378-85.

28. Slosar PJ, Kaiser J, Marrero L, Sacco D. Interobserver agreement using computed tomography to assess radiographic fusion criteria with a unique titanium interbody device. Am J Orthop. 2015;44(2):86-9.

29. Soriano Sánchez JA, Soriano Solís S, Soto García ME, Soriano Solís HA, Torres BYA, Romero Rangel JAI. Radiological diagnostic accuracy study comparing Lenke, Bridwell, BSF, and CT-HU Fusion Grading Scales for minimally invasive lumbar interbody fusion spine surgery and its correlation to clinical outcome. Medicine (Baltimore). 2020;99(21): e19979.

30. Lehr AM, Duits AAA, Reijnders MRL, Nutzinger D, Castelein RM, Oner FC, Kruyt MC. Assessment of posterolateral lumbar fusion: a systematic review of imaging-based fusion criteria. JBJS Rev. 2022;10(10):e21.00129.

31. Bono CM, Lee CK. Critical analysis of trends in fusion for degenerative disc disease over the past 20 years: influence of technique on fusion rate and clinical outcome. Spine (Phila Pa 1976). 2004;29(4):455-63; discussion Z5.

32. Peters MJM, Bastiaenen CHG, Brans BT, Weijers RE, Willems PC. The diagnostic accuracy of imaging modalities to detect pseudarthrosis after spinal fusion: a systematic review and meta-analysis of the literature. Skeletal Radiol. 2019;48(10):1499-510.

33. Cho JH, Joo YS, Lim C, Hwang CJ, Lee DH, Lee CS. Effect of one- or two-level posterior lumbar interbody fusion on global sagittal balance. Spine J. 2017;17(12):1794-802.

34. Santos ER, Goss DG, Morcom RK, Fraser RD. Radiologic assessment of interbody fusion using carbon fiber cages. Spine (Phila Pa 1976). 2003;28(10):997-1001.

35. Isaacs RE, Sembrano JN, Tohmeh AG. Two-year comparative outcomes of MIS lateral and MIS transforaminal interbody fusion in the treatment of degenerative spondylolisthesis: part II: radiographic findings. Spine (Phila Pa 1976). 2016;41:S133-44.

36. Li J, Dumonski ML, Liu Q, Lipman A, Hong J, Yang N, Jin Z, Ren Y, Limthongkul W, Bessey JT, Thalgott J, Gebauer G, Albert TJ, Vaccaro AR. A multicenter study to evaluate the safety and efficacy of a stand-alone anterior carbon I/F cage for anterior lumbar interbody fusion: two-year results from a food and drug administration investigational device exemption clinical trial. Spine (Phila Pa 1976). 2010;35(26): E1564-70.

37. Deng QX, Ou YS, Zhu Y, Zhao ZH, Liu B, Huang Q, Du X, Jiang DM. Clinical outcomes of two types of cages used in transforaminal lumbar interbody fusion for the treatment of degenerative lumbar diseases: N-HA/PA66 cages versus PEEK cages. J Mater Sci Mater Med. 2016;27(6):102.

38. Shaffer WO, Spratt KF, Weinstein J, Lehmann TR, Goel V. 1990 Volvo Award in Clinical Sciences. The consistency and accuracy of roentgenograms for measuring sagittal translation in the lumbar vertebral motion segment. An experimental model. Spine (Phila Pa 1976). 1990;15(8):741-50.

39. Center for Devices and Radiological Health. Guidance Document for the Preparation of IDEs for Spinal Systems: Guidance for Industry and/or FDA Staff. Silver Spring, MD: U.S. Food and Drug Administration; 2000. Accessed January 27, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-document-preparation-ides-spinal-systems-guidance-industry-andor-fda-staff.

40. Bohinski RJ, Jain VV, Tobler WD. Presacral retroperitoneal approach to axial lumbar interbody fusion: a new, minimally invasive technique at L5-S1: clinical outcomes, complications, and fusion rates in 50 patients at 1-year follow-up. SAS J. 2010;4(2):54-62.

41. Govindasamy R, Solomon P, Sugumar D, Gnanadoss JJ, Murugan Y, Najimudeen S. Is the cage an additional hardware in lumbar interbody fusion for low grade spondylolisthesis? A prospective study. J Clin Diagn Res. 2017;11(5):RC05-8.

42. Malone H, Mundis GM, Collier M, Kidwell RL, Rios F, Jelousi M, Galli S, Shahidi B, Akbarnia BA, Eastlack RK. Can a bioactive interbody device reduce the cost burden of achieving lateral lumbar fusion? J Neurosurg Spine. 2022;2022: 1-8.

43. Kämmerling N, Sandstedt M, Farnebo S, Persson A, Tesselaar E. Assessment of image quality in photon-counting detector computed tomography of the wrist: an ex vivo study. Eur J Radiol. 2022;154:110442.

44. Booij R, Kämmerling NF, Oei EHG, Persson A, Tesselaar E. Assessment of visibility of bone structures in the wrist using normal and half of the radiation dose with photon-counting detector CT. Eur J Radiol. 2023:159:110662.

45. Duits A, Salvatori D, Schouten J, van Urk P, van Gaalen S, Ottink K, Öner C, Kruyt M. Preclinical model for lumbar interbody fusion in small ruminants: rationale and guideline. J Orthop Transl. 2023;38:167-74.