Surgical Treatment for Discogenic Low-Back Pain: Lumbar Arthroplasty Results in Superior Pain Reduction and Disability Level Improvement Compared With Lumbar Fusion

Fred H. Geisler

doi: [https://doi.org/10.1016/SASJ-2006-0002-RR](https://doi.org/10.1016/SASJ-2006-0002-RR)
http://ijssurgery.com/content/1/1/12

This information is current as of April 26, 2019.

**Email Alerts**  Receive free email-alerts when new articles cite this article. Sign up at: [http://ijssurgery.com/alerts](http://ijssurgery.com/alerts)
ABSTRACT

Background
The US Food and Drug Administration approved the Charité artificial disc on October 26, 2004. This approval was based on an extensive analysis and review process; 20 years of disc usage worldwide; and the results of a prospective, randomized, controlled clinical trial that compared lumbar artificial disc replacement to fusion. The results of the investigational device exemption (IDE) study led to a conclusion that clinical outcomes following lumbar arthroplasty were at least as good as outcomes from fusion.

Methods
The author performed a new analysis of the Visual Analog Scale pain scores and the Oswestry Disability Index scores from the Charité artificial disc IDE study and used a nonparametric statistical test, because observed data distributions were not normal. The analysis included all of the enrolled subjects in both the nonrandomized and randomized phases of the study.

Results
Subjects from both the treatment and control groups improved from the baseline situation ($P < .001$) at all follow-up times (6 weeks to 24 months). Additionally, these pain and disability levels with artificial disc replacement were superior ($P < .05$) to the fusion treatment at all follow-up times including 2 years.

Conclusions
The a priori statistical plan for an IDE study may not adequately address the final distribution of the data. Therefore, statistical analyses more appropriate to the distribution may be necessary to develop meaningful statistical conclusions from the study. A nonparametric statistical analysis of the Charité artificial disc IDE outcomes scores demonstrates superiority for lumbar arthroplasty versus fusion at all follow-up time points to 24 months.

Key Words: lumbar arthroplasty, total disc replacement, clinical trial, statistical analysis.
rehabilitation of the patient may require a similar time to be completed to achieve the maximum successful clinical benefit of the procedure.

Though lumbar artificial disc technology has been used commercially in Europe since 1987, the first disc implanted in the United States at the Texas Back Institute (Plano, Tex) by Scott Blumenthal was in March 2000 at the start of the US Food and Drug Administration (FDA) investigational device exemption (IDE) trial. Therefore, lumbar arthroplasty is a relatively new procedure to treat lumbar DDD in the United States. The hypothesized benefits of arthroplasty over fusion procedures with significant clinical benefit include (1) reduction or elimination of disc-derived (also known as discogenic) pain; (2) restoration and maintenance of normal segmental range of motion and sagittal balance; and (3) potential reduction or retardation of progressive adjacent-level DDD, which necessitates further surgical intervention involving not only reducing the forces and angulation of the adjacent level compared with a fusion but also normalizing the adjacent level biomechanics.9

On October 26, 2004, the FDA approved the world’s first lumbar artificial disc, the Charité artificial disc (DePuy Spine, Raynham, Mass), for use in the United States.10 In doing so, the FDA followed the recommendation of its expert Orthopaedic and Rehabilitation Devices Panel, which on June 2, 2004, unanimously recommended approval.11 The FDA decision was based on the results of a prospective, randomized, controlled clinical trial that compared lumbar artificial disc replacement to anterior lumbar interbody fusion (ALIF) and on 20 years of worldwide surgical experience. FDA approval meant that the manufacturer could market the device as safe and effective for the treatment of single-level lumbar DDD in indicated patients at either the L4-5 or L5-S1 level.

This clinical trial was the first in the history of spine surgery to compare two different surgical treatments for lumbar DDD according to a multicenter, prospective, randomized, controlled study design. The results of the study were published in peer-reviewed journals, including Journal of Neurosurgery in September 20049 and Spine in July 2005.12,13 Because of the prospectively specified non-inferiority design of the study and the complex FDA-required success/failure criteria, the primary conclusion of the study for FDA labeling purposes was that treatment with artificial disc replacement was clinically at least as good as a fusion procedure.

I present here new level I medical evidence that surgical treatment with single-level arthroplasty is not only “at least as good as” a fusion procedure, as the FDA label states, but that reduction in pain and disability improvement are highly statistically superior in patients receiving treatment with lumbar arthroplasty compared to both baseline and an anterior fusion procedure. A more rapid decrease in pain and disability (postoperative healing/recovery) was also noted in the Charité group compared to the fusion group.

MATERIALS AND METHODS

A multicenter, prospective, randomized, controlled trial was performed under an FDA-approved protocol (IDE# G990303). Local institutional review board approval was obtained at all 14 study sites, and all subjects enrolled in the study gave written informed consent. The trial incorporated a non-inferiority design with a 2:1 randomization: treatment with artificial disc replacement versus the control, a fusion procedure. Enrollment in the study constituted 71 subjects in an initial nonrandomized treatment phase (approximately 5 subjects per site) and then a total of 304 subjects in the randomized phase: 205 in the treatment group and 99 in the control group. Subjects in the control group underwent an ALIF procedure with BAK threaded fusion cages (Zimmer Spine, Minneapolis, Minn) and bone graft. Subjects were assessed clinically and radiographically before surgery, 6 weeks after surgery, and then at 3, 6, 12, and 24 months after surgery. Demographics, inclusion/exclusion criteria, subject accountability, clinical outcomes, and all other detailed study information conforming to the CONSORT checklist were previously described by Blumenthal et al.12

The Oswestry Disability Index (ODI) is a 10-question validated measure (score 0–100) of disability and pain among the population with low-back pain.14 Subjects were required to complete an ODI questionnaire and a 0–100 Visual Analog Scale (VAS) pain questionnaire preoperatively and at each follow-up visit.

STATISTICAL METHODS

The statistical analysis of the ODI and VAS scores performed for the FDA and reported by Blumenthal et al., required the use of the Student’s t test. This methodology was prespecified in the statistical plan of the protocol (1) before FDA approval of the protocol, (2) before subsequent subject enrollment, and (3) before the results/distributions of the data were known. In the FDA-approved protocol the methodology could not be altered post hoc for FDA labeling claims. Using Student’s t-test, mean ODI and VAS scores were significantly better in the treatment group compared with the control group at all follow-up time points except for the 2-year follow-up.12 These results, combined with the non-inferiority study design, resulted in the primary conclusion of the study: that treatment with artificial disc replacement is at least as good as a fusion procedure in properly indicated patients.

However, Student’s t test, by definition, assumes a normal distribution of data. Therefore, Student’s t test is simply not the appropriate test with which to analyze non-normally distributed data. The ODI and VAS scores reasonably approximated a
normal distribution at baseline. At the 2-year follow-up—the endpoint of the study—the distributions were heavily nonsymmetric, skewed, and clearly not normally distributed (Figure 1). Using Student’s $t$ test might be compared to driving a car on tires designed to be inflated to 35 psi, but only inflating them to 10 psi.

A more appropriate statistical test for analysis of non-normally distributed data would be a nonparametric test such as the Wilcoxon rank sum test, which is appropriate for a non-normal data distribution. I performed a separate analysis of the ODI and VAS scores using the Wilcoxon rank sum test in which all subjects enrolled in both the nonrandomized and randomized phases of the treatment group ($n = 276$) were compared to the study control group (the complete FDA IDE dataset). The data utilized for this analysis were the same data submitted to the FDA as part of the postmarketing application submission—this was not a subset analysis. The results were verified by an independent third party (Stat Tech Services, Chapel Hill, NC) using the SAS version 8.2 statistical software package (SAS Institute, Cary, NC).

Figure 1

Distribution of VAS pain scores and ODI disability scores for all subjects (i.e., both treatment and control groups). A reasonable approximation of a normal distribution of scores is evident at baseline. However, at 2-year follow-up the distributions are not normal but skew towards the low cut-off end of both scales, indicating a favorable clinical response.
RESULTS
Both the ODI and VAS scores had almost identical initial mean values for each of the 2 groups (treatment and control). Thus, the randomization worked in this study and no baseline corrections were necessary or used in the analysis of this data.

The nonparametric analysis of ODI and VAS scores demonstrated that subjects enrolled in both the treatment and the control groups had highly significantly lower scores at all time points compared to baseline, including the 2-year follow-up ($P < .001$). The improved scores in both groups were (1) sustained over the 2-year period, (2) monotonically decreasing, and (3) more than twice the difference considered to be of minimum clinical significance. This triad makes a placebo effect of surgery an unlikely explanation for the observed improvement. Furthermore, inspection of the recovery curves reveals significant improvement at 6 weeks with maintenance out to 2 years (Figures 2, 3).

Patients in the treatment group attained a greater proportion of the total 2-year recovery in this early phase of the postoperative period in both clinical indexes. Additionally, significantly lower scores occurred in the Charité artificial disc group compared to the control fusion group at all postoperative time points, including the 2-year follow-up ($P < .05$). These results demonstrate superiority of arthroplasty over fusion in indicated patients, according to these key clinical measures, and a major improvement from baseline in both treatment groups. The control group scores closely followed the results of the BAK cage IDE study described by Kuslich et al.16

DISCUSSION
Given the distribution of the ODI and VAS scores at 2 years as shown in Figure 1, it is clear that a nonparametric test is the appropriate statistical test for analyzing the ODI and VAS data from the Charité artificial disc clinical trial. This new analysis does not and cannot change the primary FDA study conclusion: that arthroplasty is “at least as good as” a fusion procedure. This limitation of labeling claims occurred because the FDA study was conducted with a non-inferiority design with prespecified criteria for clinical success. Furthermore, FDA claims of clinical superiority cannot emanate from non-inferiority studies that are not a priori sufficiently powered to demonstrate superiority of one treatment over another.

However, improvement in ODI score was a primary clinical endpoint of this study and improvement in VAS score was a secondary clinical endpoint. These 2 clinical outcome measures are the most relied-upon measures of clinical outcome following low-back surgery. There is no doubt, following this analysis, that subjects receiving arthroplasty attained superiority in improved pain and disability levels compared with baseline levels and at

Figure 2

Mean ODI disability scores at baseline and at each protocol-specified follow-up time point through 2 years. The dotted line indicates the ODI score necessary for minimum clinical improvement, 10 points.14 (a) Mean scores for 276 subjects in the Charité artificial disc treatment group. There was a significant difference in level of disability at all time points compared to baseline ($P < .001$), Wilcoxon rank sum test. (b) Mean scores for subjects in the treatment group compared to 99 enrolled subjects in the control group. There was a significant difference in the treatment group with respect to disability, compared to the control group at all postoperative time points ($P < .05$), Wilcoxon rank sum test.
Deyo et al. have pointed out in a number of publications that, in their view, many treatments for low-back pain are ineffective, including osteopathic manipulation,17 chiropractic care,18 physical therapy,18 transcutaneous electrical nerve stimulation therapy,19 and fusion.20,21 Critics of lumbar arthroplasty cite mixed short-term22–25 and long-term26 results with the Charité artificial disc as well as review articles 27,28 written before publication of the US trial (level IV medical evidence) as the primary reason why arthroplasty is not a reasonable treatment for discogenic low-back pain. Yet other more favorable long-term results in large patient cohorts (level IV medical evidence) are often only offhandedly considered.29–32

Lumbar arthroplasty with the Charité artificial disc has been performed outside the United States for more than 20 years and, as McAfee33 eloquently pointed out, the overwhelming majority of the early disc arthroplasty cases were performed with widely variable indications; basic, rudimentary instrumentation; different sizing options; nonexistent diagnostic testing; and a lack of fundamental understanding of lumbar spine biomechanics. These early issues and failures were well known to the clinical trial investigators at the time the FDA IDE study protocol was developed. In fact, it was this knowledge of the previous successes and failures that led to the inclusion/exclusion criteria and the surgical technique used in the FDA IDE study. Thus, the FDA IDE study was designed, based upon the earlier experience, to provide level I medical evidence of the efficacy and safety of treatment with lumbar arthroplasty learned from the earlier experience, and to better define the appropriate patient selection, surgical technique, and implant sizing. The previous lumbar artificial disc clinical information has been analyzed and clinical and surgical techniques have been refined through the 20-year history of the device. If this historical review of the previous series had not occurred, then the clinically superior results presented here would not have been possible. Using the historical data and clinical experience to criticize the FDA IDE results in a vacuum, without considering the advancements described above, seems unwarranted and unscientific.

As for lack of safety, another continuing cause for criticism of arthroplasty, the same historical literature is often cited without proper context and without any relevance to today’s indications, implants, instrumentation, and knowledge. Contemporary information about complications in today’s arthroplasty patients has been presented at dozens of medical meetings over the past 3 years and has been published.12,34–36

Mean VAS pain scores at baseline and at each protocol-specified follow-up time point through 2 years. The dotted line indicates the VAS score necessary for minimum clinical improvement, 19 points.14 (a) Mean scores for 276 subjects in the Charité artificial disc treatment group. There was a significant difference in pain at all time points compared to baseline (P < .001), Wilcoxon rank sum test. (b) Mean scores for subjects in the treatment group compared to the 99 enrolled subjects in the control group. There was a significant difference in the treatment group with respect to pain, compared to the control group at all postoperative points (P < .05), Wilcoxon rank sum test.
Revision rates for fusion are just as high or higher than for disc replacement. No evidence exists that the incidence of complications with or without revision is higher in arthroplasty patients compared with fusion patients. However, differences exist in the types of potential complications.

Before the 1970s–1980s, fusion was the standard of care for surgical treatment of degenerative conditions of the hip and knee. This fusion standard was replaced in the ensuing years with artificial joint arthroplasty, which today is the standard of care for surgical treatment of these conditions in indicated patients. Often cited as the grandfather of modern hip arthroplasty, the early ideas of John Charnley for avoiding fusion of the hip were not readily accepted. Though Charnley’s work began in the late 1950s, the first hip replacement in the United States was not performed until 1969, with the first appearance in the US literature in 1970. Modern disc replacement has taken a similar track, with the third-generation Charité artificial disc having been developed and used in Europe as early as 1987 and used in the first such procedure performed in the United States only in March 2000.

As noted earlier, Deyo et al. have denounced both non-operative and operative treatment for low-back pain, including fusion and now disc replacement. What else is there for the patient with chronic low-back pain? Despite Deyo’s criticism, there are in fact multiple prospective, multicenter low-back fusion studies that describe good results and that have been published in the peer-reviewed literature. All of these studies were performed under FDA-approved protocols with narrow indications, as was the Charité clinical trial.

Highly significant improvements in pain and level of disability compared to baseline in indicated disc replacement patients is not in question. Superior pain reduction and reduced disability level compared to the current standard of care are not in question. Critics of surgical intervention for patients with low-back pain decry the lack of level I data, yet when level I data demonstrating superior outcomes are produced, the surgical intervention is still criticized as “too new,” “ineffective,” or “unsafe” despite extraordinary evidence to the contrary. In the face of level I data, such critics use lower levels of evidence in an attempt to essentially cancel out the results of a level I study. But level II, III, IV, and V data, by definition, do not trump level I data, and lower levels of published data exist for every treatment in medicine. If lower levels of data are allowed to trump Level I data, then nothing in medicine would be proven safe and effective, and if that is the case, why perform Level I studies at all?

Disc replacement in the low back is not for every patient with chronic low-back pain. Currently in the United States, disc replacement is narrowly indicated to patients with painful DDD at 1 level (L4-5 or L5-S1); who have no contraindications such as multilevel disease, scoliosis, or instability; and who fail at least 6 months of non-operative treatment. In my practice area, approximately 1 in 5000 patients with low-back pain are indicated for disc replacement after applying the indications and contraindications for the procedure. Thus, disc replacement is not being “sold” by physicians or by industry as a cure for low-back pain, though the lay press and Wall Street investors often jump to that conclusion.

As of this writing, a majority of private payers are not covering lumbar arthroplasty, a treatment backed by level I clinical data that demonstrates superior pain and disability improvement compared to baseline and to a fusion procedure. This leaves the patient with a choice between having an inferior surgical procedure covered by insurance or waiting in pain and disability for an indefinite period, hoping for a reversal in a payer’s coverage decision. All of this takes the decision of treatment out of the hands of physicians and patients and places it in the capricious hands of government and third-party payers. As a result, only a small number of financially well-off patients will be able to receive this FDA-approved treatment. But the middle class—and more importantly the working and nonworking poor—will not have access to an FDA-approved treatment in the United States. Further, though a CPT (current procedural terminology) code (22857) now exists for the procedure, the payment ($1382) is decidedly inadequate compared with that for a less technically demanding ALIF procedure ($1975), a disparity that serves as a disincentive to surgeons to perform lumbar arthroplasty in indicated patients.

Fred H. Geisler, MD, PhD
From the Illinois Neuro-Spine Center at Rush-Copley Medical Center, Aurora.

The author acknowledges consulting and research relationships with the manufacturer of the Charité artificial disc (DePuy Spine, Raynham, Mass), and has received funding in excess of $500.

Address correspondence and reprint requests to Fred H. Geisler, MD, PhD, Illinois Neuro-Spine Center at Rush-Copley Medical Center, 2020 Ogden Ave, Suite 335, Aurora, IL 60504 (e-mail: fgeisler@concentric.net)

This submission was received December 7, 2006, and accepted for publication February 26, 2007.


Institutional review board approval was obtained at the 14 hospital sites that participated in the original IDE study, in accordance with FDA regulations.
REFERENCES


48. bb350cc260751a2bb065792ca35d33103c0cc2f8a1b74b419f80e 8e57b0ee4bfbe2eb74552cd48915e951016e6950fa7611d78f20d 752c0387c3de10d7b35de.pdf. Accessed March 9, 2007.