Dear Editor:

Please find enclosed our manuscript entitled, *Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-Analysis*. We are submitting this manuscript for consideration by *Annals of Internal Medicine*. No authors have conflicts of interest and this work has not been submitted elsewhere. This review is based on research conducted by the Oregon Evidence-based Practice Center (EPC) under contract to the Yale University Open Data Access Project.

I will serve as the corresponding author. Please contact me if you have any questions. Thank you for the opportunity to submit our work.

Sincerely,

Rongwei Fu, PhD
Oregon Evidence-based Practice Center
Department of Public Health & Preventive Medicine
Oregon Health & Science University
Mailcode CB669
3181 SW Sam Jackson Park Rd.
Portland, OR 97239

Phone: 503-494-6069
Email: fur@ohsu.edu
2. **Authors: Original Submitted Manuscript (1 November 2012)**
Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) In Spine Fusion: A Systematic Review and Meta-Analysis

Investigators:
Rongwei Fu, PhD
Shelley Selph, MD
Mark Helfand, MD, MS, MPH
Kimberly Peterson, MS
Arpita Tiwari, MHS
Roger Chou, MD
Marian McDonagh, PharmD

Author affiliations:
Rongwei Fu: The Oregon Evidence-based Practice Center, the Department of Public Health and Preventive Medicine, and the Department of Emergency Medicine; Oregon Health & Science University, Portland, OR
Shelley Selph: The Oregon Evidence-based Practice Center, the Department of Medical Informatics and Clinical Epidemiology; Oregon Health & Science University, Portland, OR
Mark Helfand: Portland VA Medical Center, Portland VA Medical Center and the Oregon Evidence-based Practice Center; Oregon Health & Science University, Portland, OR
Kimberly Peterson: The Oregon Evidence-based Practice Center; Oregon Health & Science University, Portland, OR and the VA Evidence-based Synthesis Program – Portland VA Medical Center.
Arpita Tiwari: The Oregon Evidence-based Practice Center; Oregon Health & Science University, Portland, OR
Roger Chou: The Oregon Evidence-based Practice Center, the Department of Medicine, and the Department of Medical Informatics & Clinical Epidemiology; Oregon Health & Science University, Portland, OR
Marian McDonagh: The Oregon Evidence-based Practice Center, the Department of Medical Informatics and Clinical Epidemiology; Oregon Health & Science University, Portland, OR

Running title: Effectiveness and harms of rhBMP-2 in spine fusion

Key words: rhBMP-2; reporting bias; meta-analysis of individual patient data; spinal fusion; systematic review

Corresponding author: Rongwei Fu, PhD; 3181 SW Sam Jackson Park Road, Mail Code: CSB669, Portland, OR 97239; Email: fur@ohsu.edu; Phone 503-494-6069

Text word count: 4,479; Abstract word count: 315 (excluding subheadings); Figures: 2; Tables: 4.
Acknowledgments: The authors acknowledge Robin Paynter MLIS, for conducting literature searches; Howard Balshem MS, Susan Carson MPH, Allison Lowe BA, Basmah Rahman MPH, Katie Reitel MPH, MSW, Sujata Thakurta MA, and Leah Williams BS, for their contributions to this article; all are affiliated with Oregon Health & Science University.

Grant support: The Yale University Open Data Access Project.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Rongwei Fu, PhD, Oregon Health & Science University, Mailcode BICC, 3181 SW Sam Jackson Park Rd, Mailcode CB669, Portland, OR 97239.

Current Author Addresses:
Rongwei Fu: 3181 SW Sam Jackson Park Rd, Mailcode CSB669, Portland, OR 97239

Shelley Selph, Mark Helfand, Kimberly Peterson, Arpita Tiwari, Roger Chou, Marian McDonagh: 3181 SW Sam Jackson Park Rd, Mailcode BICC, Portland, OR 97239
Abstract

**Background:** Under-reporting of adverse events in industry-sponsored trials of rhBMP-2 may limit the accuracy of synthesis based on published results.

**Purpose:** To independently assess the benefits and harms of rhBMP-2 in spinal fusion and the quality of reporting in Medtronic’s published articles.

**Data Sources:** Individual patient data (IPD) of 17 industry-sponsored trials, related proprietary documents, MEDLINE, EMBASE, Cochrane Database, and reference lists.

**Study Selection:** Randomized controlled trials (RCTs) and cohort studies comparing rhBMP-2 with any control group, and uncontrolled studies of rhBMP-2.

**Data Extraction:** Details about population, study design, and results were extracted by one investigator and confirmed by another. Two investigators independently assessed quality using predefined criteria.

**Data Synthesis:** Thirteen RCTs and 30 cohort studies were included. RhBMP-2 and controls were generally associated with similar benefits. The occurrence of any adverse event in RCTs was high with 78% to 88% of patients having at least one adverse event at 24 months, but also generally similar between rhBMP-2 versus controls in lumbar fusion. There were more wound complications and dysphagia with rhBMP-2 in anterior cervical spine fusion. In anterior lumbar interbody fusion, rhBMP-2 was associated with increased risk of retrograde ejaculation and urogenital problems, but the difference was not statistically significant and the confidence intervals were wide. At 24 months the overall cancer risk was increased with rhBMP-2 (RR 3.04; 95% CI 1.23 to 7.48), but event rates were low and cancer types were heterogenous.

**Limitations:** Selective outcome reporting limited the usefulness of the published trials. Trials had unblinded outcome assessment and poor ascertainment of some harms.

**Conclusions:** In spinal fusion, rhBMP-2 has no proven clinical benefit over iliac bone graft and may be associated with important harms. Although IPD can reduce the problem of publication and reporting biases and allow a more thorough evaluation than published results, study design limitations and sparse data still leave uncertainty about specific harms.

**Primary Funding Source:** The Yale University Open Data Access Project.
Spinal fusion is the most commonly performed surgery for chronic non-specific back pain caused by degenerative conditions (1). Traditionally, spinal fusions are performed by using graft material harvested from the patient’s iliac crest. In 2002, the U.S. Food and Drug Administration (FDA) approved a genetically engineered protein, recombinant human bone morphogenic protein-2 (rhBMP-2), as a bone graft substitute for single-level anterior lumbar interbody fusion (ALIF) in conjunction with an implant. However, rhBMP-2 has been used primarily “off label” in posterolateral spine fusion (PLF) and transforaminal lumbar interbody fusion (TLIF) (2).

Publications of industry-sponsored trials reported beneficial effects of rhBMP-2 in spinal fusion with no device-related adverse events (3-6). However, observational studies reported serious complications associated with rhBMP-2 (7-10) and in 2008 the FDA issued a public health notification of life-threatening complications associated with use of rhBMP-2 in cervical spine fusion (11).

Underreporting, particularly of adverse events, has made it difficult to draw reliable conclusions about the balance of benefits and harms of rhBMP-2 compared with alternative osteoinductive materials (12, 13). The primary aims of our review are to assess 1) the benefits and harms of rhBMP-2 in spinal fusion using standard systematic review methods supplemented with individual patient data (IPD) and relevant documents provided by Medtronic Inc. and 2) the quality of reporting of the Medtronic trials in published articles.
Methods

Detailed methods, additional analyses, and complete evidence tables for this review are available in the full report (14).

Data Sources and Searches

The Yale University Open Data Access Project (YODAP) provided datasets, trial protocols, and reports submitted to the FDA of 17 completed Medtronic-funded studies (13 randomized controlled trials and four uncontrolled trials) and 1113 MedWatch reports of adverse events concerning rhBMP-2. We also searched MEDLINE® (1996 to August 2012), Embase®, the Cochrane Database of Systematic Reviews®, the Cochrane Central Register of Controlled Trials® (3rd Quarter 2012), Scopus, Clinicaltrials.gov, and the FDA web site, and manually searched reference lists of relevant papers.

Study Selection

Two reviewers independently assessed titles and abstracts of citations and then full-text articles for inclusion. Controlled clinical trials and cohort studies of rhBMP-2 that evaluated patient-centered outcomes (e.g., measures of pain, disability or functional health), spinal fusion, or any adverse events (e.g., inflammation, heterotopic bone formation, osteolysis and instability, leg pain, retrograde ejaculation, and cancer) in humans with symptomatic spinal disease were eligible for inclusion. For fusion and harms we also included uncontrolled intervention series that followed patients who received rhBMP-2. We excluded trials reported only in abstracts and studies that combined results of rhBMP-2 with other bone morphogenetic proteins.
Data Abstraction and Quality Rating

For publicly available articles and reports, one investigator abstracted patient and study characteristics and results and a second investigator reviewed data abstraction for accuracy. We used (proprietary) study protocols, reports, and data dictionaries from the manufacturer to assess ascertainment of outcomes, publication bias, and selective reporting (15). Two investigators independently rated the quality of controlled trials and cohort studies as good, fair, or poor using criteria adapted from the Cochrane Back Review Group (16) and the US Preventive Services Task Force (17). Discrepancies were resolved through consensus. We rated the strength of evidence for outcome measures by assessing the aggregate risk of bias, consistency, directness, and precision of the evidence (18).

Quantitative Synthesis

For studies with IPD, we used a two-step approach. In the first step, for each study, we recalculated all benefit outcomes using consistent definitions and calculated study-level estimates for each outcome. We defined overall success as radiographic fusion; improvement of Oswestry Disability Index [ODI] score by at least 15 points, maintenance or improvement in neurological status, and no serious adverse event or additional surgery classified as implanted related. For outcomes based on multiple criteria (overall success and fusion), patients had to satisfy all criteria; patients with data for some but not all criteria were categorized as failures. In sensitivity analyses, patients with data for some but not all criteria were categorized as missing and not included in the analysis. Overall and specific adverse events were obtained directly from
IPD (no recalculation) except for urinary retention, wound infection, wound dehiscence and possible lumbar radiculitis, which were obtained by reviewing case histories in proprietary reports. We defined “possible radiculitis” as 1) back pain plus leg, thigh, or buttock pain or weakness (unilateral or bilateral); or 2) adverse events described as "sciatica" or "radiculopathy"; or 3) back and/or leg pain with use of epidural steroids or surgery for radiculopathy (e.g., discectomy, foraminotomy). We excluded cervical/arm symptoms, numbness/paresthesias without weakness or pain, just back pain, just leg pain and pain attributed to trauma. For study-level estimates, we used risk ratios (RR) for binary outcomes and analysis of covariance (ANCOVA) to estimate mean differences for continuous outcomes.

In the second step, we combined the estimates from different studies using a random effects model (19) except for rare events, where a fixed effects model was used.(20) We conducted meta-analysis only if studies were similar enough to produce a meaningful combined estimate (20). Based on input from clinical experts, for all outcomes (except for cancer and death) we stratified data synthesis by spinal area (lumbar, cervical) and surgical approach (e.g., ALIF, PLF, TLIF). For benefits, the analyses focused on time periods up to 24 months. For harms, we aggregated data into two periods: 1) operative and up to four weeks post-operative, 2) up to 24 months post-operative. For cancer and death we analyzed the cumulative number of events up to 24 and 48 months and combined all surgical approaches because these outcomes were rare and not believed to necessarily be affected by surgical techniques. We excluded patients who had a preexisting cancer from all cancer analyses.

Only the ALIF trials and PLF trials provided adequate data for meta-analyses. We assessed statistical heterogeneity using standard $\chi^2$ tests and the $I^2$ statistic (21). We performed sensitivity analyses by excluding poor quality studies, studies that utilized a lower rhBMP-2
concentration (posterolateral fusion), and graft site related adverse events in the analysis of harms. Except for cancer, results of sensitivity analyses were similar and not separately reported. All meta-analyses were performed using Stata/IC 12.1 (StataCorp LP, College Station, TX).

Role of the Funder

YODAP proposed the aims for the review, served as the intermediary for our requests for additional information or data from Medtronic, and provided comments on our initial draft report. Medtronic had no influence over the conduct of our analyses or the publication of our findings.

Results

Overview

Thirteen randomized trials, 12 sponsored by Medtronic (n=1,879) and one by Norton Healthcare (n=106) (22), met the inclusion criteria (Figure 1). We excluded one Medtronic trial (n=3) and nine additional controlled trials described in abstracts only (23-31). Ten Medtronic trials (3-5, 22, 32-37) were published in journals. Information on four trials (all published) was available from the FDA site (38-40).

The Medtronic studies applied similar exclusion criteria and generally enrolled a homogenous population (Table 1). Six enrolled fewer than 50 subjects and 8 enrolled fewer than 100 (sample size range 14 to 85). Twelve trials compared spinal fusion with rhBMP-2 versus iliac crest bone grafts (ICBG); one compared rhBMP-2 with implantation of the MAVERICK™ artificial disk (41).
Missing data varied widely between outcomes and time points. In all but one study (42), there was no pre-specified algorithm on how to handle missing data. While there were important baseline differences in some of the trials, we did not detect a consistent pattern favoring the intervention groups. The main risks for bias were lack of blinding of surgeons, patients, and (except for radiologic endpoints) outcome assessors to treatment group. The quality of ascertainment varied for different endpoints. Benefit endpoints (e.g., pain, function, and fusion) seemed to be ascertained reliably using well-designed questionnaires or scales. One outcome—pain in the bone graft harvest site—was assessed only in the control group and only on the side of graft harvest. For harms, a general classification was used for many adverse events (e.g., cardiovascular and urogenital), without specific symptom questionnaires or objective tests. For example, for retrograde ejaculation, a condition in which patients do not always volunteer information, it was unclear whether investigators asked about symptoms and how an event was defined. No trial defined radiculitis and adverse events consistent with possible radiculitis were variously classified within the same trial as back and leg, neurological, or spinal events. Very little information was available in the Medtronic datasets about local effects such as inflammation, heterotopic bone formation, or osteolysis. For cancer, it was not considered in the Medtronic clinical protocols as a pre-specified endpoint. Cancer events were captured only by voluntary reporting through a generic adverse event text field and could be under-reported.

Most trials were phrased as non-inferiority or equivalence studies in study objectives and analysis of endpoints, but sample size calculations in most trials were not based on a non-inferiority or equivalence design and most outcomes in the published trials were analyzed as endpoints from superiority trials. In reports to the FDA, Medtronic emphasized prespecified primary and secondary effectiveness endpoints. In contrast, there was less complete reporting in
journal publications. Rates of overall success were reported in the literature for two of the 14 studies in which it was measured (Table 1) and only two of five studies where it was the primary objective (36, 41).

As previous reviews have noted (12, 13), adverse events were underreported in the journal articles for both rhBMP-2 and autograft groups (Table 2). In the pivotal trial of ALIF, for example, Burkus, et al. (4) reported only 25 adverse events at 24 months, but IPD indicated 315 adverse events in the rhBMP-2 group and 274 adverse events in the autograft group two years after surgery. Infection rates were reported in only two publications (37, 41). Until 2009, specific adverse events were rarely reported in the published studies but were reported thoroughly in documents Medtronic submitted to the FDA (43-48).

We found 30 cohort studies comparing rhBMP-2 with autograft and/or allograft, 47 intervention series, and 34 case series or case reports of patients who received rhBMP-2 in spine surgery, including four Medtronic prospective intervention series (none fully published) (14). Most observational studies were retrospective, small, off-label use and provided little information on patient characteristics. The main risks for bias were unclear comparability of groups at baseline, unclear blinding of outcome assessors, and failure to adjust for potential confounding variables and baseline differences. Some observational studies that were designed specifically to assess adverse events had more reliable or complete ascertainment (10, 49-53). One large cohort study used ICD-9 codes from large administrative datasets to ascertain serious complications (7).
Benefits and Harms

**Anterior Lumbar Interbody Fusion (ALIF)**

*Randomized Trials.* In July, 2002, the FDA gave premarket approval for the use of the InFUSE™ Bone Graft with the LT-Cage® for ALIF procedures in patients with degenerative disk disease at one level from L4-S1. The approval was based on results from the pilot study (Study 1) and two “pivotal” studies: a randomized trial of open spinal fusion using a retroperitoneal or transperitoneal approach with either the rhBMP-2 (InFUSE™) graft or ICBG (Study 2), and a separate series of patients who underwent laparoscopic implantation of the InFUSE® Bone Graft without a control group (Study 3). Later, Medtronic evaluated InFUSE™ with the INTER FIX™ threaded fusion device (Study 9), which was approved by the FDA in June 2004, and a bone dowel (Studies 4, 5).

In our IPD analysis, the five RCTs (n=465) provided moderately strong evidence that there were no consistent differences between rhBMP-2 and ICBG in overall success, fusion rates, or most other effectiveness measures from the immediate postoperative period through 24 months (Table 3). The one exception was a 3-point higher SF-36 physical component summary score for patients in the rh-BMP-2 group at 6, 12, and 24 months. At 24 months, fusion rates ranged from 60% to 100% and the average overall success rate was about 50% for the rhBMP-2 group and 60% for the ICBG group.

Another RCT compared rhBMP-2 with an artificial disc (n=577). Our primary analysis of IPD data from this trial showed that fusion rates for rhBMP-2 at 12 and 24 months were 81% and 79%, lower than the rates reported in the publication of the trial (100%) (41) and lower than the rates in the published pivotal trials of rhBMP-2 (97% and 95%) (4).
Adverse events were common. Meta-analysis of the five trials (n=465) showed that at 4 weeks, 38% of patients in the rhBMP-2 group and 45% of patients in the ICBG group had experienced at least one adverse event; at 24 months, about 80% in both groups had experienced an adverse event (Table 4). We found no significant difference between rhBMP-2 and ICBG for any specific adverse events. For retrograde ejaculation, subsidence, and urogenital problems, rhBMP-2 was associated with increased risk, but the association was not significant and confidence intervals were wide. The trials provided sparse, inconclusive results on these and other events (Table 4).

Observational Studies. Three cohort studies (50, 54, 55) and four intervention series (49, 56-58) evaluated rhBMP-2 in ALIF. One small cohort study reported a fusion rate similar to those of the trials (54).

A cohort study comparing rhBMP-2 to ICBG reported an increase in retrograde ejaculation associated with rhBMP-2 use (7% of 69 rhBMP-2 patients vs. 1% of 174 control patients, \( P = 0.0025 \)) (55). The rhBMP-2 arm of a retrospective cohort study had a similar rate (7.4%) (50). Some observational studies reported higher rates of subsidence, but there was variation in how subsidence was defined and when it was measured (57, 59).

Selective Reporting in Other Publications. In addition to the primary publications of some of the trials, Medtronic sponsored several selective analyses and reviews of the ALIF studies. In 2002 and 2003, the journal Spine published supplements sponsored by Medtronic Sofamor Danek Inc. that reported selective effectiveness endpoints but no adverse event endpoints for the Medtronic trials that had been completed by that time (60-62).

Although the FDA concluded that, in the pivotal trials, rhBMP-2 and ICBG were equivalent at 24 months (39), some publications implied that they had found rhBMP-2 to be superior. In 2001, for example, Kleeman and colleagues published results for 22 of the 137 patients...
laparoscopic ALIF study (Study 3) (63). The publication highlights that 100% of these subjects had a successful fusion and “improvement in back pain, leg pain, and function.” These results were not representative of the overall results in the laparoscopic ALIF study. Duplicative publications featured a post hoc “integrated analysis” designed to demonstrate a statistically significant increase in fusion rates with rhBMP-2 (64-66). This analysis added a laparoscopic series of 266 patients using ICBG from an older, unpublished trial of the LT-CAGE as the control group of the laparoscopic ALIF study and combined data from the pivotal ALIF trial using open surgical approach, resulting in a statistically significant difference between rhBMP-2 and ICBG for several endpoints. In internal documents, but not in publications, Medtronic noted that the surgeons in the earlier study were likely less skilled at laparoscopic ALIF than the surgeons in the later study (67).

**Posterolateral Fusion (PLF) in Lumbar spine**

We included five RCTs (22, 32, 36, 37, 68) (including one of patients over 60 year old) (22), eight cohort studies (69-76) and seven intervention series (67, 77-82). Three of the five trials (32, 37, 68) used a higher dose and concentration of rhBMP-2 than InFUSE™. One trial did not report dose (22).

In our meta-analysis, the four Medtronic RCTs (n=733) provided moderate strength evidence of no difference in overall success, fusion rates, and other effectiveness outcomes between rhBMP-2 and ICBG (Table 3). The only exception was that rhBMP-2 was associated with slightly better SF-36 physical health scores at 6 and 12 months (WMD 1.79, 95% CI 0.26 to 3.31, $I^2 = 0$; WMD 1.89, 95% CI 0.26 to 3.53, $I^2 = 0$, respectively). The fusion rate at 24 months ranged from 70% to 90% in the ICBG group and 86% to 93% in the rhBMP-2 group. The rate of
overall success was in the range of 40% to 60% in both groups. A subgroup analysis in patients over 60 years old from the 4 Medtronic trials combined with another trial (22) had similar fusion results (RR 1.12, 95% CI 0.98 to 1.29, $I^2 = 52\%$). Fusion results from cohort studies and intervention series were consistent with RCTs.

In the trials, there was no difference between rhBMP-2 versus ICBG in the risk of experiencing at least one adverse event, one serious adverse event or one device-related adverse event at four weeks and 24 months (Table 4). In both groups, about 50% of patients experienced at least an adverse event at four weeks and 88% at 24 months. Risk of back and leg pain events was significantly higher with rhBMP-2 versus ICBG, but the types of adverse events classified as back and leg pain were very heterogeneous (e.g., radiculopathy, Baker’s cyst, sacroiliac joint pain, arthritic knee pain or ankle pain). There was no difference in the risk of possible radiculitis at 4 weeks and 24 months. We also found similar rates for other specific harms at 4 weeks and 24 months (Table 4). Findings from two cohort studies (74, 75) and one intervention series (79) that reported specific adverse events were consistent with the RCT findings.

In our IPD analysis of the largest trial assessing the benefits and harms of rhBMP-2 using the PLF approach (AMPLIFY™) (37), rhBMP-2 and ICBG did not differ in rates of overall success (61% vs. 56%) and fusion (90.0% vs. 89.5%). In contrast, the journal publication and FDA summary reported that use of rhBMP-2 resulted in a higher fusion rate (96% vs. 89%, $P=0.014$) (37, 40). The reason for this difference may be that we classified patients with partial data on fusion criteria as failure, though it is not clear why this would only affect the rhBMP-2 group.

In the journal articles, counts of additional surgeries routinely excluded reoperations and sometimes, elective removal. For AMPLIFY™, Dimar, et al. reported 20 secondary surgery
events in the rhBMP-2 group, significantly lower than the 36 events in the ICBG group
($P=0.015$) (37). However, when elective removal and reoperation were included in the IPD
analysis, the difference was not significant (rhBMP-2: 36 events in 34 patients; ICBG, 57 events
in 43 patients; $P=0.15$).

**Other Fusion Techniques in Lumbar Spine**

We were not able to reach any definitive conclusion on the comparative effectiveness or
harms of rhBMP-2 in other lumbar fusion techniques. Except for one small Medtronic trial
(n=67) (33), only low-quality observational studies about effectiveness and harms were available
for these approaches (see full report (14).)

**Cervical Spine Fusion**

For anterior cervical spine fusion, a small, fair-quality (n=33) Medtronic trial (3) found
no difference between rhBMP-2 and ICBG in likelihood of fusion, overall success, and other
effectiveness endpoints. These results were consistent with those of three pertinent cohort studies
(9, 10, 59).

In the randomized trial, at 24 months, rhBMP-2 was associated with more adverse events
than ICBG (42 AEs in 18 patients vs. 13 AEs in 15 patients; RR 2.69, 95% CI 1.25 to 5.80).
Results from a large fair-quality cohort study (n=27,067) provided moderate strength of evidence
that use of rhBMP-2 in anterior cervical spine fusion was associated with increased risk of
complications (OR 1.43, 95% CI 1.12 to 1.70), increased risk of dysphagia/dysphonia (OR 1.63;
95% CI 1.30 to 2.05), and wound complications (OR 1.67, 95% CI 1.10 to 2.53) (7). These
results were consistent with results from smaller cohort studies (total n=346) (8-10). Intervention
series reported 5% to 60% of patients developed dysphagia, depending on how dysphagia was defined (83-87).

There were no prospective, controlled trials of rhBMP-2 in posterior cervical spine fusion. We included four retrospective cohort studies (n=3,233) (one fair-quality (7) and three poor-quality (88-90)) and one intervention series (n=53) (91). A single cohort study (n=204) (90) provided low strength of evidence that rhBMP-2 was associated with a higher fusion rate (100% vs. 88%, P=0.01) but more recurrent neck pain (48% vs. 29%, P=0.003) at 24 months. Moderate strength of evidence indicated that rhBMP-2 use was associated with similar risk of overall complications (7, 88-90) and wound complications (7, 90), compared with no rhBMP-2 use.

Cancer and Death

We pooled five Medtronic trials (Studies 2, 4, 5, 10, 14) that reported at least one cancer event at 24 months. Compared with ICBG, the use of rhBMP-2 was associated with a 1.9% increase in the absolute risk of cancer (95% CI 0.4 to 3.3; NNH=53, 95% CI 31 to 250) with an RR of 3.04 (95% CI 1.23 to 7.48) (Figure 2). The effect of dosage was unclear: 10 of 17 cancers in the rhBMP-2 occurred in the AMPLIFY™ trial, but another high-dose study (Study 13) had no cancers in the rhBMP-2 group (n=98). It was also unclear whether under-reporting played a role. To assess the potential impact of the seven Medtronic trials with zero incident cancer in both treatment groups (sample sizes 14 to 197), we performed a sensitivity analysis by considering these trials as a combined “pseudo-trial” (n= 429) and included it in the meta-analysis by conservatively assuming that no cancer occurred in the rhBMP-2 group and one cancer occurred in the control group. The sensitivity analysis showed a 1.3% (95% CI 0.2 to 2.4; NNH=77, 95% CI 42 to 500) absolute increase in cancer risk associated with rhBMP-2 (RR
2.45; 95% CI 1.08 to 5.58). At 48 months, the association was attenuated due to fewer studies and no-longer significant (four studies; RR 1.75; 95% CI 0.82 to 3.74). The above meta-analyses included all types of malignancies. When we excluded cancers not included in the Surveillance, Epidemiology and End Results Program (http://seer.cancer.gov/), the increased risk associated with rhBMP-2 at 24 months was no longer statistically significant (RR 2.32; 95% CI 0.90 to 5.98 at 24 months; RR 1.97; 95% CI 0.88 to 4.42 at 48 months). One cohort study (92) of 125 patients (24 rhBMP-2, 101 ICBG) provided results consistent with the RCT (RR 2.10; 95% CI 0.69 to 6.41). The total number of cancers included in the IPD meta-analysis is low (Figure 2), and the strength of evidence is low.

There was no increase in risk of deaths at 24 months (nine trials, RR 0.71; 95% CI 0.32 to 1.55) (Studies 2, 4, 6-10, 13-14) or 48 months (four trials, RR 0.66; 95% CI 0.28 to 1.56) (Studies 4, 10, 13-14).

Discussion

While rhBMP-2 appears to be similarly effective as ICBG in terms of beneficial outcomes for ALIF and PLF, the current evidence does not provide definite answers on beneficial outcomes for other surgical approaches, or the comparative risk for harms with rhBMP-2. Although SF-36 PCS scores were better for patients undergoing ALIF and PLF, the difference was only 2 to 3 points on a 0-100 point scale, failing to meet typical criteria for a clinically meaningful difference (93).

Adverse events were common. In lumbar spine fusion, we generally found similar results between rhBMP-2 versus autograft/allograft group. In contrast, use of rhBMP-2 in anterior cervical spine fusion was associated with increased overall adverse events, wound complications
and dysphagia or dysphonia. We found an increased risk of cancer associated with the use of rhBMP-2 with any fusion surgery at 24 months. However, the included cancers in the meta-analysis were heterogeneous, cancers could have been under-reported, and results were no longer statistically significant when non-SEER cancers were excluded.

Meta-analysis of IPD has been considered the gold standard of meta-analysis (94). In this review, the availability of IPD allowed a more thorough evaluation of both benefits and harms that is not possible with only the published studies, and reduced the problem of publication and reporting biases. In addition to providing data on unpublished Medtronic studies, IPD allowed us to examine all outcomes from all time points for the manufacturer sponsored trials, whereas the published studies were likely to provide complete information only on statistically significant results at selected time points. Due to serious under-reporting in the early published RCTs, available information for adverse events from these studies was minimal. However, the availability of IPD could not overcome the inadequacy of adverse events ascertainment.

The availability of IPD data improved the quality of the meta-analyses by allowing recalculation or re-categorization of outcome measures based on a consistent definition, calculating a mean difference adjusted for potential baseline imbalances for continuous outcomes, and allowing better handling of missing data. For example, as coded by Medtronic, back and leg pain adverse events included many events unlikely related to back surgery or use of rhBMP-2 while radiculitis was not consistently assessed. Possible radiculitis events were variably categorized as back and leg, neurological or spinal events, even within the same trial. By re-classifying events that appeared consistent with radiculitis, we were able to create a “possible radiculitis” variable using adverse event case histories.
On the other hand, meta-analysis based on IPD requires substantially more time and resources than meta-analysis based on study-level data, and cannot compensate for flawed data collection or sparse data. For example, recoding could not overcome the failure to adequately ascertain radiculitis, resulting in low confidence in the finding of no difference in risk. Even with IPD on 1879 patients, from 12 trials, and many observational studies, the evidence base is small within each surgical approach. Since rhBMP-2 was evaluated as part of a device, only one pivotal study was needed for each indication. Only two pivotal trials each were available for meta-analyses of ALIF and PLF. Although abstracts suggested that additional trials have been performed, we found no published RCTs with a funding source truly independent of the manufacturer.

Additionally, there have been no prospective, well-designed, adequately-powered studies specifically aimed to systematically assess harms using adequate ascertainment methods. Concerns over increased risk of retrograde ejaculation, urinary retention, osteolysis/instability, and subsidence in ALIF and heterotopic bone formation in PLIF were raised in the FDA review process and in a thorough systematic review (95), but we found insufficient data to reach definitive conclusions, with wide confidence intervals for effect measures.

There was also insufficient information to adequately evaluate the effects of dose on risk of harms. Eleven Medtronic studies, including all six ALIF RCTs, used rhBMP-2 at a concentration of 1.5mg/mL (INFUSE Bone Graft), with an overall rhBMP-2 dose of 0.6-16.8 mg depending on spinal location and levels fused. Higher and unapproved concentrations of rhBMP-2 (2.0-3.0 mg/mL) were used in five of the six PLF studies, with total doses ranging from 15.0-63.0 mg. Determining the effects of higher versus lower concentrations and doses of
rhBMP-2 was not possible due to differences in surgical approach, rhBMP-2 carrier, fusion hardware, as well as the small number of studies and small sample sizes.

Although we had unusual access to protocols and documents submitted by the manufacturer to the FDA, other information, such as operative notes and internal correspondence, might have been helpful in assessing the extent of design and reporting bias. For example, Carragee noted that the control intervention—iliac crest bone graft—might have been less effective in the PLF trials because the protocols required different graft volume and site preparation than would be used in actual practice (12). Internal correspondence could indicate whether the manufacturer chose the control conditions for valid scientific reasons. Internal correspondence is also essential to evaluate selective analysis reporting, ghostwriting, time lag bias, and misrepresentation of facts (96). Finally, case report forms would have enabled us to evaluate the integrity of adverse event adjudication.
References


34. Gornet MF, Dryer RF, Peloza JH, Sehranck FW. Lumbar disc arthroplasty vs. Anterior lumbar interbody fusion: Five-year outcomes for patients in the Maverick(degrees) disc IDE study. Spine Journal. 2010;10(9):64S.
40. U.S. Food and Drug Administration. Executive Summary for P050036 Medtronic's AMPLIFY™ rhBMP-2 Matrix. Orthopaedic and Rehabilitation Devices Advisory Panel. FDA [serial on the Internet]. 2010; (Study 14): Available from:


with the CD HORIZON spinal system for posterolateral lumbar fusion in patients with symptomatic degenerative disc disease at two adjacent vertebral levels- final progress report. (Study 15, Final Report/Final Antibody Report). 2010. 1-277


<table>
<thead>
<tr>
<th>IDE Clinical Trial Name</th>
<th>Sample size, n</th>
<th>rhBMP-2 Conc. (mg/cc)</th>
<th>Baseline Characteristics</th>
<th>IPD Results 24 Months</th>
<th>Published Results, 24 months</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean age, year</td>
<td>Male, %</td>
<td>Diabetes, %</td>
<td>Smoking, %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I: 42.5</td>
<td>I: 46</td>
<td>I: 9</td>
<td>I: 46</td>
</tr>
<tr>
<td>LT-CAGE® Pilot (1) Boden, 2000</td>
<td>RCT</td>
<td>11</td>
<td>1</td>
<td>1.5</td>
<td>3.9-7.8 ACS</td>
<td>I: 43.3</td>
</tr>
<tr>
<td>INFUSE®/ INTER (2) Burkus, 2002</td>
<td>RCT</td>
<td>142</td>
<td>136</td>
<td>1.5</td>
<td>4.2-8.4 ACS</td>
<td>I: 43.2</td>
</tr>
<tr>
<td>INFUSE®/ LT-CAGE® Lap Pivotal (3) Unpublished</td>
<td>IS</td>
<td>134</td>
<td>1</td>
<td>1.5</td>
<td>4.2-8.4 ACS</td>
<td>I: 42.4</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pilot (4) Burkus, 2002</td>
<td>RCT</td>
<td>24</td>
<td>22</td>
<td>1.5</td>
<td>8.1-11.7 ACS</td>
<td>I: 41.5</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pivotal (5) Unpublished</td>
<td>RCT</td>
<td>55</td>
<td>30</td>
<td>1.5</td>
<td>8.1-11.7 ACS</td>
<td>I: 39.7</td>
</tr>
<tr>
<td>INFUSE®/ INTER FIX™ ALIF Pilot (9) Unpublished</td>
<td>RCT</td>
<td>25</td>
<td>20</td>
<td>1.5</td>
<td>8.4-16.8 ACS</td>
<td>I: 45.9</td>
</tr>
<tr>
<td>MAVERICK™ Disc Pivotal (10) Gout, 2011</td>
<td>RCT</td>
<td>172</td>
<td>405</td>
<td>1.5</td>
<td>4.2-12.0 ACS</td>
<td>I: 40.2</td>
</tr>
<tr>
<td>n/N</td>
<td>RR (95% CI)</td>
<td>n/N</td>
<td>RR (95% CI)</td>
<td>Fusion</td>
<td>Overall Success</td>
<td>Fusion, n/N</td>
</tr>
</tbody>
</table>

**Table 1. Included Medtronic Studies of rhBMP-2 with comparison of IPD and published data**
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name (Study #)</th>
<th>Publication</th>
<th>Design</th>
<th>Sample size, n</th>
<th>rhBMP-2 Conc. (mg/cc)</th>
<th>Dose (mg)</th>
<th>Carrier</th>
<th>Baseline Characteristics</th>
<th>IPD Results 24 Months</th>
<th>Published Results, 24 months (Blank cells=unpublished study)</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFUSE®/ MASTER GRAFT® Pilot (8)</td>
<td>RCT</td>
<td>25 21</td>
<td>1.5 12.0 ACS</td>
<td>I: 55.9/ C: 56.9</td>
<td>I: 0/ C: 14</td>
<td>I: 24/ C: 24</td>
<td>I: 24/ C: 28</td>
<td>I: 28/ C: 43</td>
<td>I: 19/22/ C: 14/20</td>
<td>I: 15/24/ C: 10/20</td>
</tr>
<tr>
<td>AMPLIFIX™/ rhBMP-2/ CRM Pilot (14)</td>
<td>RCT</td>
<td>239 224</td>
<td>2.0 40.0 CRM</td>
<td>I: 53.2/ C: 52.3</td>
<td>I: 45.2/ C: 42.4</td>
<td>I: 26.4/ C: 26.3</td>
<td>I: 30.5/ C: 27.7</td>
<td>I: 34.7/ C: 41.1</td>
<td>I: 169/210/ C: 162/161</td>
<td>I: 118/211/ C: 105/166</td>
</tr>
<tr>
<td>rhBMP-2/ CRM 2-level Pilot (15)</td>
<td>IS</td>
<td>29 29</td>
<td>2.0 40.0 CRM</td>
<td>I: 53.9/ C: 52.0</td>
<td>I: 10/ C: 24</td>
<td>I: 41/ C: 45</td>
<td>I: 18/26</td>
<td>--</td>
<td>Not measured</td>
<td>--</td>
</tr>
<tr>
<td>PLIF</td>
<td>RCT</td>
<td>34 33</td>
<td>1.5 4.2-8.4 ACS</td>
<td>I: 46.3/ C: 46.1</td>
<td>I: 50/ C: 46</td>
<td>I: 3/ C: 3</td>
<td>I: 53/ C: 46</td>
<td>I: 35/ C: 40</td>
<td>I: 26/32/ C: 21/30</td>
<td>I: 15/33/ C: 10/31</td>
</tr>
<tr>
<td>Cervical/PLIF</td>
<td>IS</td>
<td>30 30</td>
<td>1.5 8.4 ACS</td>
<td>I: 51.0/ C: 40.0</td>
<td>I: 7/ C: 27</td>
<td>I: 47/ C: 31</td>
<td>24/25</td>
<td>--</td>
<td>Not measured</td>
<td>--</td>
</tr>
<tr>
<td>ACDF</td>
<td>RCT</td>
<td>18 15</td>
<td>1.5 0.6-1.2 ACS</td>
<td>I: 51.3/ C: 47.1</td>
<td>I: 44/ C: 47</td>
<td>I: 0/ C: 0</td>
<td>I: 28/ C: 47</td>
<td>I: 6/ C: 0†</td>
<td>I: 11/12/ C: 12/12</td>
<td>I: 10/12/ C: 10/12</td>
</tr>
</tbody>
</table>

ACDF = anterior cervical disectomy and fusion, ACS = absorbable collagen sponge, ALIF = anterior lumbar interbody fusion, C = comparator group (ICBG group), BCP = biphasic calcium phosphate, CRM = compression resistant matrix, I = Investigational group (rhBMP-2 group), IDE = investigational device exemption, IS = intervention series, PEEK = polyetheretherketone, PLF = posterior lumbar fusion, PLIF = posterior lumbar interbody fusion, RCT = randomized controlled trial, rhBMP-2 = recombinant human bone, US = United States

* Study 3 data not published independently. Burkus, 2003 contains pooled data from Study 3 and 2.
† n not reported, results reported only as percentages.
‡ Study 5 data not published independently. Burkus, 2005 contains pooled data from Study 4 and 5.
§ Comparator is Maverick, not ICBG
∥ The Mexico pilot was an Intervention Series with two cohorts.
¶ I1 = rhBMP-2 without internal fixation, I2 = rhBMP-2 + TSRH (Texas Scottish Rite Hospital) pedicle screw instrumentation, C: autograft + TSRH
**Rounded down from percentage reported in text (92.3). The table in this publication reports a slightly higher percentage (97.3%).
†† Prior neck surgery
Table 2. Comparison of reported adverse events in the published trials versus adverse events in the IPD at 24 months

<table>
<thead>
<tr>
<th>Author/Trial Study number</th>
<th>Patients, n</th>
<th>Number of AE reported by published study</th>
<th>Number of AE based on IPD*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rhBMP-2</td>
<td>Control</td>
<td>rhBMP-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rhBMP-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Control</td>
</tr>
<tr>
<td>ALIF</td>
<td></td>
<td></td>
<td>rhBMP-2</td>
</tr>
<tr>
<td>Boden, 2000</td>
<td>11</td>
<td>3</td>
<td>6 (1 ileus and delay in gait training, 1 wound dehiscence, 1 low back pain and 3 trauma)</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pilot (Study 1)</td>
<td>11</td>
<td>3</td>
<td>2 (1 ileus and delay in gait training, 1 urinary retention)</td>
</tr>
<tr>
<td>Burkus, 2002</td>
<td>143</td>
<td>136</td>
<td>6 (6 intraoperative vascular)†</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Open Pivotal (Study 2)</td>
<td>143</td>
<td>136</td>
<td>13 (5 intraoperative vascular, 8 graft side related)</td>
</tr>
<tr>
<td>Burkus, 2002</td>
<td>24</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pilot (Study 4)</td>
<td>24</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Burkus, 2005‡</td>
<td>55</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pivotal (Study 5)</td>
<td>55</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>Gornet, 2011</td>
<td>172</td>
<td>405</td>
<td>407, complete reporting of AE in a supplemental table</td>
</tr>
<tr>
<td>MAVERICK™ Disc Pivotal§ (Study 10)</td>
<td>172</td>
<td>405</td>
<td>982, complete reporting of AE in a supplemental table</td>
</tr>
<tr>
<td>PLF</td>
<td></td>
<td></td>
<td>449</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1139</td>
</tr>
<tr>
<td>Dawson, 2009</td>
<td>25</td>
<td>21</td>
<td>2 (1 durotomy, 1 wound infection)</td>
</tr>
<tr>
<td>INFUSE®/ MASTERGRAFT® Pilot (Study 8)</td>
<td>25</td>
<td>21</td>
<td>3 (1 durotomy, 1 wound infection, 1 graft side related)</td>
</tr>
<tr>
<td>Boden, 2002</td>
<td>11 + 11</td>
<td>5</td>
<td>4 (1 leg pain, 1 back pain, 2 hematoma), all led to second surgery</td>
</tr>
<tr>
<td>rhBMP-2/BCP US Pilot RCT (Study 12)</td>
<td>11</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Dimar II, 2009</td>
<td>239</td>
<td>224</td>
<td>603, complete reporting of AE in a table</td>
</tr>
<tr>
<td>AMPLIFY™ (rhBMP-2/ CRM) Pivotal (Study 14)</td>
<td>239</td>
<td>224</td>
<td>579, complete reporting of AE in a table</td>
</tr>
</tbody>
</table>

Downloaded From: http://annals.org/pdfaccess.ashx?url=/data/journals/aim/927097/ on 09/13/2018
<table>
<thead>
<tr>
<th>Author/Trial Study number</th>
<th>Patients, n</th>
<th>Number of AE reported by published study</th>
<th>Number of AE based on IPD*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rhBMP-2</td>
<td>rhBMP-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLIF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haid, 2004</td>
<td>34</td>
<td>29 (19 Neurological, 10 bone formation outside the disc space with leg pain increase)</td>
<td>35 (1 cardiovascular, 20 neurological, 2 graft side related, 12 bone formation outside the disc space with leg pain increase)</td>
</tr>
<tr>
<td>INFUSE®/ INTER FIX™ PLIF</td>
<td>33</td>
<td>112</td>
<td>122</td>
</tr>
<tr>
<td>(Study 6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACDF</td>
<td>18</td>
<td>2 ectopic bone formation, as part of radiographic outcomes</td>
<td>45</td>
</tr>
<tr>
<td>Baskin 2003</td>
<td>15</td>
<td>1 ectopic bone formation, as part of radiographic outcomes</td>
<td>13</td>
</tr>
<tr>
<td>INFUSE®/ CORNERSTONE® ACDF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pilot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Study 7)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AE = adverse events; IPD = individual patient data
* The number and type specific AEs can be found in Appendix F of the full report. There are too many specific AEs to be included in the table.
† Six cases of retrograde ejaculation were reported, but not by intervention groups.
‡ Burkus, 2005 contains pooled data from Study 4 and 5.
§ The comparison group in this study is artificial disk, not ICBG. Discrepancy in numbers between published trial and IPD was partially due to an updated date set sent by Medtronic.
Table 3. Effectiveness Endpoints for ALIF and PLF with rhBMP-2 vs. ICBG

<table>
<thead>
<tr>
<th>Outcome Scale</th>
<th>6 weeks</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALIF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Relative risk (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>----</td>
<td>1.14 (0.94 to 1.37)</td>
<td>1.13 (0.87 to 1.47)</td>
<td>1.27 (0.89 to 1.82)</td>
</tr>
<tr>
<td>Fusion</td>
<td>----</td>
<td>1.10 (1.02 to 1.19)</td>
<td>1.09 (0.95 to 1.24)</td>
<td>1.05 (0.88 to 1.24)</td>
</tr>
<tr>
<td>Neuro success</td>
<td>1.01 (0.94 to 1.09)</td>
<td>0.98 (0.90 to 1.06)</td>
<td>1.02 (0.95 to 1.10)</td>
<td>1.04 (0.95 to 1.13)</td>
</tr>
<tr>
<td>ODI success</td>
<td>1.19 (0.79 to 1.78)</td>
<td>1.06 (0.94 to 1.20)</td>
<td>1.03 (0.82 to 1.30)</td>
<td>1.08 (0.94 to 1.24)</td>
</tr>
<tr>
<td>Return to work</td>
<td>1.21 (0.73 to 1.98)</td>
<td>1.01 (0.89 to 1.15)</td>
<td>1.01 (0.90 to 1.13)</td>
<td>1.04 (0.93 to 1.16)</td>
</tr>
<tr>
<td><strong>Weighted mean difference (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>-3.33 (-6.59 to 2.93)</td>
<td>-3.62 (-8.02 to 0.78)</td>
<td>-3.24 (-8.30 to 1.81)</td>
<td>-6.94 (-13.90 to 0.02)</td>
</tr>
<tr>
<td>Fusion</td>
<td>-0.57 (-1.21 to 0.02)</td>
<td>2.81 (0.85 to 4.76)</td>
<td>2.95 (0.86 to 5.04)</td>
<td>3.34 (0.92 to 5.75)</td>
</tr>
<tr>
<td>Neuro success</td>
<td>-0.36 (-2.45 to 1.72)</td>
<td>-0.31 (-2.22 to 1.60)</td>
<td>-1.06 (-2.60 to 1.47)</td>
<td>2.86 (-0.20 to 5.92)</td>
</tr>
<tr>
<td>ODI success</td>
<td>0.55 (0.26 to 3.31)</td>
<td>1.79 (0.26 to 3.31)</td>
<td>1.89 (0.26 to 3.53)</td>
<td>1.10 (-0.66 to 2.86)</td>
</tr>
<tr>
<td>Return to work</td>
<td>0.74 (-1.68 to 3.16)</td>
<td>0.96 (0.84 to 1.08)</td>
<td>1.07 (0.89 to 1.29)</td>
<td>1.02 (0.91 to 1.15)</td>
</tr>
<tr>
<td><strong>PLF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Relative risk (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>----</td>
<td>1.55 (0.90 to 2.67)</td>
<td>1.17 (0.84 to 1.63)</td>
<td>1.04 (0.90 to 1.20)</td>
</tr>
<tr>
<td>Fusion</td>
<td>----</td>
<td>1.44 (0.95 to 2.19)</td>
<td>1.29 (0.94 to 1.78)</td>
<td>1.16 (0.96 to 1.41)</td>
</tr>
<tr>
<td>Neuro success</td>
<td>1.04 (0.99 to 1.10)</td>
<td>1.02 (0.97 to 1.09)</td>
<td>0.99 (0.93 to 1.04)</td>
<td>1.02 (0.96 to 1.09)</td>
</tr>
<tr>
<td>ODI success</td>
<td>1.01 (0.85 to 1.20)</td>
<td>1.07 (0.98 to 1.17)</td>
<td>1.01 (0.85 to 1.21)</td>
<td>1.03 (0.94 to 1.13)</td>
</tr>
<tr>
<td>Return to work</td>
<td>1.28 (0.73 to 2.25)</td>
<td>0.96 (0.84 to 1.08)</td>
<td>1.07 (0.89 to 1.29)</td>
<td>1.02 (0.91 to 1.15)</td>
</tr>
<tr>
<td><strong>Weighted mean difference (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>0.74 (-1.68 to 3.16)</td>
<td>-2.41 (-4.86 to 0.04)</td>
<td>-2.23 (-4.95, 0.49)</td>
<td>-1.92 (-5.03 to 1.18)</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.10 (-0.27 to 0.48)</td>
<td>-0.45 (-1.07 to 0.17)</td>
<td>-0.41 (-1.34 to 0.52)</td>
<td>-0.31 (-0.76 to 0.15)</td>
</tr>
<tr>
<td>Neuro success</td>
<td>-0.23 (-0.21 to 0.66)</td>
<td>-0.27 (-0.71 to 0.17)</td>
<td>-0.29 (-0.74 to 0.17)</td>
<td>-0.35 (-0.82 to 0.13)</td>
</tr>
<tr>
<td>ODI success</td>
<td>0.50 (-0.95 to 1.96)</td>
<td>0.06 (-1.47 to 1.60)</td>
<td>-0.48 (-2.21 to 1.25)</td>
<td>0.54 (-2.74 to 3.83)</td>
</tr>
</tbody>
</table>
Table 4. Overall and specific adverse events for ALIF and PLF with rhBMP-2 vs. ICBG

<table>
<thead>
<tr>
<th>ALIF</th>
<th>4 weeks</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event type</td>
<td>% Patients with BMP vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
</tr>
<tr>
<td>≥ 1 Adverse event, any type</td>
<td>38 vs. 45</td>
<td>0.86 (0.70 to 1.06)</td>
</tr>
<tr>
<td>≥ 1 Serious adverse event</td>
<td>9 vs. 8</td>
<td>1.12 (0.61 to 2.07)</td>
</tr>
<tr>
<td>≥ 1 device-related adverse event</td>
<td>---</td>
<td>----</td>
</tr>
</tbody>
</table>

Individual adverse events

| Anatomical/technical difficulty | 0.9 vs. 3 | 0.29 (0.08 to 1.09) | 0.0 | Same as four weeks |
| Back and/or Leg Pain | 4 vs. 3 | 1.03 (0.42 to 2.52) | 0.0 | 26 vs. 24 | 1.07 (0.47 to 2.46) | 68.7 |
| Cardiovascular | 2 vs. 4 | 0.56 (0.18 to 1.74) | 0.0 | 6 vs. 7 | 0.85 (0.43 to 1.69) | 0.0 |
| Gastrointestinal | 13 vs. 15 | 0.84 (0.53 to 1.33) | 0.0 | 17 vs. 19 | 0.89 (0.60 to 1.30) | 0.0 |
| Implant Problems | 2 vs. 1 | 1.00 (0.28 to 3.59) | 0.0 | 3 vs. 0.9 | 1.77 (0.55 to 5.64) | 0.0 |
| Infection (all types) | 6 vs. 5 | 1.09 (0.49 to 2.43) | 0.0 | 10 vs. 10 | 1.07 (0.62 to 1.83) | 0.0 |
| Neurological | 3 vs. 4 | 0.80 (0.29 to 2.22) | 0.0 | 16 vs. 14 | 1.09 (0.69 to 1.72) | 0.0 |
| Possible Radiculitis | 3 vs. 3 | 0.97 (0.35 to 2.71) | 0.0 | 23 vs. 24 | 0.99 (0.71 to 1.38) | 0.0 |
| Respiratory | 2 vs. 3 | 0.57 (0.16 to 2.00) | 0.0 | 3 vs. 5 | 0.47 (0.16 to 1.36) | 0.0 |
| Retrograde Ejaculation | 4 vs. 1 | 2.62 (0.28 to 24.56) | --- | 6 vs. 1 | 4.36 (0.52 to 36.40) | --- |
| Spinal Event | 0 vs. 2 | 0.23 (0.03 to 2.07) | 0.0 | 12 vs. 11 | 1.13 (0.68 to 1.89) | 0.0 |
| Subsidence | 2 vs. 1 | 1.43 (0.24 to 8.41) | --- | 4 vs. 1 | 2.57 (0.63 to 10.50) | 0.0 |
| Urogenital | 7 vs. 4 | 1.77 (0.81 to 3.87) | 15.8 | 13 vs. 8 | 1.56 (0.88 to 2.77) | 0.0 |
| Vertebral Fracture | 1 vs. 0 | 2.63 (0.28 to 24.71) | 0.0 | Same as four weeks |
| Urinary Retention* | --- | ---- | --- | 5 vs. 2 | 0.90 (0.08 to 9.74) | 0.0 |
| Wound Infection* | --- | ---- | --- | 4 vs. 6 | 0.75 (0.33 to 1.68) | 0.0 |
| Wound Dehiscence* | --- | ---- | --- | 1 vs. 0 | 2.16 (0.26 to 18.17) | 0.0 |
| Relevant additional surgeries | --- | ---- | --- | 11 vs. 13 | 0.79 (0.40 to 1.54) | 23.3 |
**PLF**

### Event type

<table>
<thead>
<tr>
<th>Event type</th>
<th>Events 51</th>
<th>Events 49</th>
<th>Events 88</th>
<th>Events 87</th>
<th>Ratio (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1 Adverse event, any type</td>
<td>1.02 (0.83 to 1.27)</td>
<td>32.0</td>
<td>1.01 (0.96 to 1.06)</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 1 Serious adverse event</td>
<td>0.89 (0.68 to 1.18)</td>
<td>0.0</td>
<td>0.97 (0.84 to 1.11)</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 1 device-related adverse event</td>
<td>---</td>
<td>----</td>
<td>6 vs. 5</td>
<td>1.35 (0.73 to 2.49)</td>
<td>49.1</td>
<td></td>
</tr>
</tbody>
</table>

### Individual adverse events

| Anatomical/technical difficulty                 | 4.90 (0.58 to 41.31) | Same as four weeks |
| Back and/or leg pain                           | 1.84 (1.01 to 3.37) | 1.18 (1.01 to 1.38) |
| Cardiovascular                                 | 0.97 (0.68 to 1.39) | 0.93 (0.70 to 1.24) |
| Dural injury                                   | 0.75 (0.43 to 1.31) | 0.78 (0.45 to 1.33) |
| Gastrointestinal                               | 0.72 (0.44 to 1.17) | 0.81 (0.51 to 1.29) |
| Implant problems                               | 2.43 (0.57 to 10.44) | 1.52 (0.57 to 4.01) |
| Infection (all types)                          | 1.04 (0.55 to 1.98) | 1.00 (0.66 to 1.50) |
| Possible Radiculitis                           | 1.29 (0.52 to 3.22) | 0.95 (0.74 to 1.22) |
| Neurological                                   | 1.48 (0.70 to 3.14) | 1.13 (0.88 to 1.46) |
| Respiratory                                    | 1.34 (0.60 to 3.02) | 1.45 (0.80 to 2.61) |
| Spinal event                                   | 0.97 (0.29 to 3.26) | 0.88 (0.56 to 1.39) |
| Urogenital                                     | 1.03 (0.61 to 1.74) | 1.10 (0.76 to 1.60) |
| Vertebral fracture                             | 1.23 (0.31 to 4.90) | 0.95 (0.26 to 3.47) |
| Relevant additional surgeries                  | ---       | ----      | 12 vs. 14 | 0.81 (0.55 to 1.17) | 0.0 |

Categories of adverse events are based on Medtronic’s protocols.
* based on individual case histories in the reports provided by Medtronic.
Figure 1. Literature flow chart\(^a\)

14697 records identified from database searches after removal of duplicates

14742 records screened

13969 records excluded at abstract level

45 additional records identified through other sources

773 full-text articles assessed for eligibility

636 full-text articles excluded
- 44 non-English language
- 75 ineligible outcome
- 66 ineligible intervention
- 16 ineligible population
- 351 ineligible publication type
  - 338 general
  - 9 RCTs, abstract only
  - 3 pending trials – currently not completed
- 4 ineligible study design, including 1 Medtronic RCT with only three subjects
- 80 ineligible area of the body

Included Articles:
125 studies reported in 136 publications included in qualitative synthesis, plus Individual Patient Data provided by Medtronic

By study design:
- 13 randomized controlled trials reported in 18 publications, including data from 2 unpublished trials
- 30 cohort studies reported in 31 publications
- 47 intervention series in 50 publications, including data from 3 unpublished trials
- 34 case report/series in 35 publications
- 1 combination studies reported in 2 publications

Included articles by approach and study design:

<table>
<thead>
<tr>
<th></th>
<th>ALIF</th>
<th>PLF</th>
<th>Other</th>
<th>Cervical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1 (Anterior)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Medtronic)</td>
<td>(PLIF)</td>
<td>0 (Posterior)</td>
</tr>
<tr>
<td>Cohorts</td>
<td>3</td>
<td>8</td>
<td>16</td>
<td>5 (Anterior)</td>
</tr>
<tr>
<td>Intervention</td>
<td>1 (Medtronic)</td>
<td>2 (Medtronic)</td>
<td>1 (Medtronic)</td>
<td>4 (Posterior)</td>
</tr>
<tr>
<td>Series</td>
<td>4 (Other)</td>
<td>5 (Other)</td>
<td>26 (Other)</td>
<td>7 (Anterior)</td>
</tr>
</tbody>
</table>

\(^a\) Modified from the PRISMA flow diagram.(92)
Figure 2. Comparison of risk of cancer between rhBMP-2 versus control

<table>
<thead>
<tr>
<th>Study</th>
<th>Control Rate</th>
<th>Risk Ratio (95% CI)</th>
<th>Events, rhBMP-2</th>
<th>Events, Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>24 month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infuse LT Cage Pivotal, 1998</td>
<td>0.74%</td>
<td>1.90 (0.17, 20.74)</td>
<td>2/143</td>
<td>1/136</td>
</tr>
<tr>
<td>Infuse Bone Dowel Pilot, 1998</td>
<td>0%</td>
<td>2.76 (0.12, 64.41)</td>
<td>1/24</td>
<td>0/22</td>
</tr>
<tr>
<td>Infuse Bone Dowel Pivotal, 2000</td>
<td>0%</td>
<td>1.66 (0.07, 39.55)</td>
<td>1/55</td>
<td>0/30</td>
</tr>
<tr>
<td>Maverick Disc Pivotal, 2003</td>
<td>0.74%</td>
<td>2.35 (0.48, 11.55)</td>
<td>3/172</td>
<td>3/405</td>
</tr>
<tr>
<td>Amplify Pivotal, 2002</td>
<td>0.89%</td>
<td>4.69 (1.04, 21.15)</td>
<td>10/239</td>
<td>2/224</td>
</tr>
<tr>
<td>Subtotal (I² = 0.0%, p = 0.951)</td>
<td></td>
<td>3.04 (1.23, 7.48)</td>
<td>17/633</td>
<td>6/817</td>
</tr>
</tbody>
</table>

| **48 month**                               |              |                     |                 |                 |
| Infuse Bone Dowel Pilot, 1998              | 0            | 2.76 (0.12, 64.41)  | 1/24            | 0/22            |
| Maverick Disc Pivotal, 2003                | 1.24%        | 1.41 (0.34, 5.85)   | 3/172           | 5/405           |
| BCP Canada, 1999                           | 2.04%        | 0.34 (0.01, 8.15)   | 0/48            | 1/49            |
| Amplify Pivotal, 2002                      | 2.23%        | 2.25 (0.81, 6.28)   | 12/239          | 5/224           |
| Subtotal (I² = 0.0%, p = 0.701)            |              | 1.75 (0.82, 3.74)   | 16/483          | 11/700          |

Downloaded From: http://annals.org/pdfaccess.ashx?url=/data/journals/aim/927097/ on 09/13/2018
3. *Annals of Internal Medicine*: External Peer Reviewer Comments (December 2012)

Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-Analysis

Fu R et al.

Comments to Editors

Although this paper is of great interest to me as a Spine specialist, I do not think that the internal medicine audience of the Annals of Internal Medicine will find it compelling reading. I suspect that few internists are asked whether bone morphogenetic protein is a better choice for a fusion than bone taken from a patient’s pelvis. The more important question is whether a fusion operation is needed at all.

The authors of the paper should not be faulted but they are working with a limited number of randomized trials and cohort studies. They are trying to be fair in demonstrating some benefit from BMP compared to the standard operation but cannot find any. On the other hand, harm can be shown when the material is used “off-label” in the cervical spine.

One of the larger studies used in the analysis involves an artificial disc. I am not sure that information from this experimental device study using 2 experimental devices (the artificial disc and BMP) is appropriate for inclusion. It is hard to separate out factors related to the disc versus BMP.

I think the paper has a better home in one of the spine journals.
Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-Analysis

Fu R et al.

Comments for Authors:

Major Comments

The authors are to be commended for addressing a difficult topic with a modest amount of information. Although the awareness of internists to the benefits and risks of new devices is important, the topic of bone morphogenetic protein for fusion operations seems to be too specialized a topic for an internal medicine journal. Most internists do not believe that the deliberation about a specific operation for the spine is in their purview let alone the decision about autologous bone graft versus BMP.

The major aim of the paper was to highlight discrepancies between individual patient data and reports of randomized clinical trials using the same patients.

In regard to study selection, the authors excluded trials with other BMPs but did not exclude a study with a large population that included another experimental device – artificial disc. The artificial disc has its own series of adverse events. The inclusion of this study makes it difficult to separate relative harms.

In Quantitative Synthesis – It is confusing regarding the inclusion and exclusion criteria. Success was defined with fusion, ODI improvement of 15 and no adverse events. Back pain is not specially stated. However, possible radiculitis as an adverse event includes back and leg pain, but excludes back pain alone or leg pain alone. These criteria seem arbitrary. Is a patient excluded from success if they have back pain? Is a patient excluded from adverse events if they have back pain alone?

Results – Page 9 The authors point out the inconsistencies of data gathering in the studies. For example, only those who had bone graft were asked about buttock pain versus all individuals in the study. As you pointed out, many studies did not ascertain adverse events in an established manner. This suggests that the number of adverse events are potentially underreported. Also on 2 of the 14 studies measured no overall success rate.
Page 14 - The lack of data available to determine adverse events related to BMP versus alternative causes
is highlighted by the increase in back and leg events but with causes very unlikely to be related to the
device like “arthritic knee or ankle pain” It makes data related to adverse events difficult to interpret.
Page 18 – The possible radiculitis category remains a problem. The authors are trying to reclassify
patients more stringently to increase the likelihood of relationship of BMP use and radicular pain. Some
might suggest that leg pain alone is adequate for radiculitis and should be included.
In the final analysis, the authors can not make up for the absence of data to determine the relative risk and
benefit of BMP versus bone graft It is clear that the call for more randomized controlled trials for
evaluation of efficacy and safety of BMP should be the final sentence of the paper.
4. **Annals of Internal Medicine: Statistical Editor Comments (December 2012)**

Manuscript: M12-2731

Title: Effectiveness and harms of recombinant human bone morphogenetic protein-2 (rhBMP-2) in spine fusion: A systematic review and meta-analysis

Author(s): Fu R, Selph S, Helfand M, Peterson K, Tiwari A, Chou R, McDonagh M

John E. Cornell is the primary statistical editor for your manuscript and prepared this review. However, the issues raised in this review represent the composite advice of Annals of Internal Medicine’s statistical team after discussion of your manuscript at one of our regular statistical meetings. Our aim is for the revision process to be as efficient as possible and to result in a published report that is methodologically robust and reported clearly. The primary statistical and physician editors for your manuscript will review the revised manuscript and may discuss it with the full team. If you have questions about any of the issues raised in our Annals’ review, please direct them to Darren Taichman at DTaichman@mail.acponline.org. Most often we can resolve these issues via e-mail communication, but in the event that initial e-mail communication does not satisfactorily address your concerns, we will arrange for members of our team to discuss the issues with you in a telephone conference.

The relative effectiveness and harms associated the use of rhBMP-2 is spinal fusion is a timely, but controversial topic. A prior review by Carragee, et al. (Spine J 2011; 11: 471-91) revealed a number of potential study design biases and inconsistencies in reporting of adverse events between published studies and data submitted to FDA. A strength of the current review is that the authors have direct access to individual-patient level data from the 17 industry (Medtronic) sponsored trials (13 randomized and 4 non-randomized trials). Access to this level of data provides the opportunity to conduct a more consistent, complete, and rigorous analysis of the effectiveness and harms of rhBMP-2 in spinal fusion. The authors also have access to 1113 MedWatch reports of adverse events associated with rhBMP-2, and they conducted a traditional literature search of MEDLINE, EMBASE, the Cochrane Database of Systematic
Reviews, the Cochrane Central Register of Controlled Trials, Scopus, Clinicaltrials.gov, the FDA website, and manual searching of reference lists.

The authors have two primary objectives for their review: 1. Conduct an independent analysis of the effectiveness and harms of rhBMP-2, and 2. Compare their results against those presented in Medtronic’s published studies.

1. The first objective is more than a simple reanalysis based on the individual patient-level data from the Medtronic trials. It appears to utilize both the industry IDP data plus the eligible trials, both independent and industry sponsored studies, published since the original trials. The authors need to be clearer about these aspects of their review and the purpose that each serves in the manuscript in their articulation of the study objectives.

2. The second is comparable to the analyses provided in the Carragee, et al. review, though they don’t appear to utilize some of the additional data sources Carragee, et al. used: Scoliosis Research Society, the CDC, or the National Inpatient Database, all of which could provide updated data that is informative with respect to determining the adverse events associated with rhBMP-2 and their respective rates. We wonder why the authors choose to exclude these additional government and institutional databases from these analyses. Since these databases continue to collect adverse event data relevant to the safety of rhBMP-2, we recommend that the authors review these additional databases, as well.

3. This is clearly a complex undertaking. The mixing of the different data sources and analyses in the manuscript is often confusing, making it difficult for the reader to digest. We need the authors to provide a more careful description and presentation of the data organized by data sources: IPD from Medtronic trials, additional trials and observational studies captured through the extensive literature search, and the FDA sources. Beginning with the Methods section, the authors need to make clear to the reader which sections refer to IPD analyses, which report results based on the traditional literature review and study-level meta-analyses, and which report/summarize additional information based on review of the FDA data and documents.
3.1. The method section reads like a traditional study-level systematic review, rather than reflecting the more complex sources of information utilized in this review. This may not be the best model for this review. I suggest organizing the methods around the data sources. The subsections within each are tailored to the data sources

3.1.1. IPD data source: Data Sources, Quality Assessment, Definitions for Outcomes, Rules for Attributing of Adverse Events to rhBMP-2, Selection and Coding of Covariates, Managing Missing Data, and Data Synthesis and Analysis

3.1.2. Study-level sources: Search Strategy, Study Section, Data Abstraction and Quality Assessment, Data Synthesis and Analysis

3.1.3. FDA data (and other government or institutional) sources: Data Source, Dates, etc.

3.2. Be clear about the role of each data source in your review. We need a Table that presents your definitions for the outcomes and contrast your definitions with those used in the Medtronic publications. This Table may appear either in the Methods section under IPD data sources or early in the Results section where the IPD analyses are described.

3.3. Although the authors have IPD available, the current analytic approach aggregates this rich source of data to form a series of study-level statistics that the authors combine using traditional meta-analytic methods. This approach is disappointing and inefficient. If you have the patient-level data, you need to make maximum use of this rich data source.

3.3.1. Use more sophisticated mixed-effect models to combine the outcomes and adverse events across trials, like you would in a multi-center clinical trial. Doing so will clearly strengthen the review beyond the information already published by Carragee, et al.

3.3.2. While you note that few of these trials described how they handled missing data, this approach provides an opportunity to apply appropriate methods for handling missing data and provide more appropriate estimates of the treatment effects and adverse event rates from these trials.

3.3.3. Cumulative adverse event rates are presented at 4 and 24 months. Since the actual rates are proportional to length of time, one would think that a Poisson (or negative binomial) model would be the
best approach to estimating event rates and incidence rate ratios. Of course this depends on the availability of person-time information. If this level of data is unavailable, it would be interesting and informative to look at (perhaps graphically) the cumulative event rates for the groups reported across all the time-points represented in the data, rather than the static 4 and 24 months. While it may not be possible to do these analyses for all adverse events, it would be important to do this for general categories of adverse events and the more serious adverse events if possible.

3.4. Use traditional meta-analytic methods to combine the actual study-level estimates obtained through your literature search, but exercise more care in summaries with sparse data and zero events. The method used in the current analyses uses a fixed-effects Mantel-Haenszel method with 0.5 continuity correction for these cases (Stata metaan default). I suggest that the authors use a fixed-effect logistic model based on the grouped counts within study. This approach does not require a correction for continuity, and it is easily done with the Stata glogit command, using the margins command to generate study-level estimates and 95% Confidence Intervals. This approach is more in line with recommendations made by the first author in their nicely written J Clin Epidemiol 2011; 64:1187-97 article.

3.5. You note that data reported only in abstracts was excluded from this review. It is common for reviewers to include such data in a systematic review. With all the various data sources included in this analysis, conducting a formal sensitivity analysis that includes the data from these abstracts simply adds to the complexity of the presentation. You should, however, in the Discussion section, describe the pattern of results reported in these abstracts relative to your findings.

3.6. Comparing and combining the results from the IPD with the additional study-level data should be based on the pooled estimates for each. This is analogous to the mixed-treatment effect meta-analysis problem. A hierarchical Bayesian model is the optimal approach, though a simpler frequentist method is acceptable and would reduce the complexity of the analyses (See Jansen Res Syn Meth 2012; 3: 177-190).

3.7. We also suggest that you follow Carragee, et al. and use the CONSORT recommendations and calculate 90% confidence intervals for serious adverse events.
4. The Results section is appropriately organized around type of procedure, but within the sections the presentation mixes information from the various sources both within and across the primary study objectives. It creates a confusing picture of the data and its meaning.

4.1. Separate out the analyses for the two-objectives: 1. Effectiveness and Harms and 2. Comparisons with Published Reports.

4.2. Move all the comparative material that contrasts the findings from your independent analyses with those presented in the various industry trials into a separate section at the end of the results section.

4.3. Within the sections for each procedure, organize the presentation of the data so that the results from each data source are clearly identified.

4.3.1. Start with the IPD analyses, followed by the study-level data, followed by the FDA (and other government and institutional data).

4.3.2. A brief opening paragraph within the section that describes each procedure that provides a description of the trials and their methodological quality, along with a succinct summary of the findings would help the reader get oriented to the data.

4.3.3. Separate effectiveness analyses from harms analyses more clearly to minimize confusion. This is done to some extent, but the text wonders off into the comparison and contrast between the IPD and FDA data against the published studies. Separating these objectives will help clarify the presentation of the data.

4.4. Tables need to be reorganized to better delineate the data sources and suggested structure for the analyses.

4.4.1. Table 1 needs to include study duration or duration of follow-up for each of the studies.

4.4.2. As you rework your tables, be sure to include numerator and denominator information whenever counts of events or percentages are reported.

4.4.3. Report the number of studies and total N whenever you report summary estimates or counts.

4.4.4. Report absolute risk differences and confidence intervals for adverse event rates. Use CONSORT guidelines when computing the confidence intervals for these events: 90% CIs for serious adverse events.
5.  *Annals of Internal Medicine*: Comments on Format Requirements

*Annals of Internal Medicine*

Revision Requirements

---

**Manuscript No.:**  M12-2731  
**First Author:**  Fu

---

**Text**

- Add section heading "Data Synthesis and Analysis."
- Please provide full corporate names, city, and state of the manufacturers of brand-name materials: Medtronic, Inc.; MAVERICK; inFUSE; Medtronic Sofamor Danek; AMPLIFY.
- There are too many tables and figures for this manuscript. Please reduce to no more than 4 tables and figures for print publication.

**References**

- Add the following to these references: reference 15-city and state of publication; reference 14-is the report number known yet?; reference 27-page numbers; references 42 through 28-city and state of publication; references 50, 71-page numbers.
- Please follow these principles when citing URLs in the reference list: 1) Please cite print references rather than Web-based references (i.e., URLs) if a print version is available. 2) When citing URLs, please be sure to always include the date on which the URL was accessed. 3) If a reference is available only on the Web, please make sure that the URL is valid. References to invalid links are not acceptable. 4) If a reference is published both in print and electronically, please cite the print version. However, if you consulted an updated Web version of a print reference (e.g., a government report that was published in print in June 2005 but updated online in December 2005), please cite both the print and Web versions, along with the date each was published and the date on which the URL was accessed.

**Tables**

- Define all abbreviations in footnotes to the tables.
- Remove all vertical lines and all internal rule lines from your tables. Retain only horizontal lines beneath the title, beneath the column headings, and at the end of the tables.
- If you have not already done so, please prepare tables by using the Tables feature in Word. Data should appear in individual cells. Please do not use tabs.

**Figures**

- In the figure legends, define all abbreviations used in the figures.

**Permissions**

- Provide letters of permission from everyone listed in the acknowledgment section.

**Electronic Manuscript Submission**
• Please submit an electronic file for the manuscript, a separate file containing all tables, and individual JPEG files for each figure. For further instructions on how to convert files created in Microsoft Office into JPEGs, please e-mail the contact person below.
Dear Dr. Fu:

Thank you for submitting to Annals of Internal Medicine your manuscript, "Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-Analysis." The editors, an external reviewer and a statistician read the manuscript. In addition, the editorial team discussed it at our weekly manuscript and statistical meetings.

We are potentially interested in publishing your manuscript as a Review article. However, before we can consider it further, you must prepare a revision that satisfactorily addresses the issues outlined in this letter and in the accompanying external review and statistical review. A technical/format review from one of our production editors will be forthcoming.

General Comments:

1. Your writing requires improvement to make the subject matter more accessible to the clinically experienced reader who is not an expert in spinal surgery. Bear in mind when revising your paper that most readers will not be well versed in the abbreviations or trial acronyms you are currently using. Minimize their use, and be sure they are well defined when it is necessary to use them.

2. Your results section requires reorganization so that the stated goals (to review the benefits and harms of rhBMP-2, and assess the reporting by Medtronic) are each clearly addressed. Currently, your second goal (to assess the quality of and possibility of biased/selective reporting) is buried in a rather haphazard manner within the other results presented, and so it is difficult for the reader to evaluate your findings regarding this aspect of your study. Please see the advice of our Statistical Editor regarding how to better organize your results so as to overcome this problem.
3. We are concerned that additional data are available that might inform your review. These are noted in the review you reference (Carragee et al, Spine J, 2011). Please see the comments of our Statistical Editor to this issue as well.

Specific Comments:

4. Abstract: Your abstract also requires revision to make the topic accessible to non-experts in this area. For example, many readers will not know what rhBMP-2 is. They may not know what “similar benefits” are referred to. Currently, the abstract’s data synthesis contains no information regarding your assessment of reporting quality. The conclusion also does not provide readers with your assessment of your results with regard to this stated objective (to assess reporting).

5. Introduction: Most readers of Annals will not be familiar with ALIF, PLF and TLIF. You should provide a little more information so that the non-expert will understand the goals of these procedures and their differences. In addition, it is important to better orient the non-expert reader to what each rhBMP-2 and bone grafts are meant to do. I realize this might seem foolish, but many readers might not know the purpose of the material and how they are thought to work. Finally, knowing just a little more about what rhBMP-2 does in a very general way (1 or 2 sentences) might help readers understand the kinds of complications worried about (a point that will require some elaboration in your discussion).

6. Introduction: Although obvious after reading your full paper, or knowing a bit about this subject, make it clear early on if Medtronic is the only maker of rhBMP-2 and to whom you refer when noting, “…industry-sponsored trials…” Also, help orient the reader a little better to the concerns raised by the FDA. What life-threatening (or other) complications were of concern? Also, did the FDA or others raise the concern that there had been purposeful selective reporting by the manufacturer?

7. Methods: Data Sources and Searches: Tell readers specifically whether the 17 studies submitted by Medtronic to the FDA represent all studies known to have been completed by the company.

8. Methods: In stating the goals of your study, if an assessment of biased or selective reporting was a specific goal in your evaluation of the “quality of reporting” by Medtronic, say so.

9. Page 6: Tell non-expert readers whether a change in ODI of 15 has been established as a clinically significant change. Also, how was, “maintenance of improvement” defined?

10. Page 7: Why did you exclude each cervical/arm symptoms, numbness/parasthesias without weakness or pain, just back pain and just leg pain?

11. Page 7: Why are harms (other than death or cancer) limited to 24 months?

12. Page 7: How often were patients with pre-existing cancer excluded? Further, as these outcomes were rare we would think these patients should be included (except where it might be
clear that the pre-existing cancer caused an oncological issue occurring after spinal fusion, such as the appearance of metastatic lesions from a previously known breast cancer).

13. Results: It is currently difficult for the reader to gain a clear message regarding the quality of the available data sources, how they differed, and your results addressing the question of selective / biased reporting. Please see the advice from our Statistical Editor regarding more informative means of organizing your results.

14. Page 9: Re: pain at the bone graft harvest site being assessed only in the control group. Please confirm that patients in the rhBMP-2 groups in these studies did not undergo sham harvests.

15. Page 10: Although it is clear that a greater number of adverse events were reported to the FDA than reported in published manuscripts, can we be sure that they have truly been, “reported thoroughly in documents Medtronic submitted to the FDA?”

16. Page 11, first paragraph: this is an example of where the non-expert reader will be lost in many abbreviations and acronyms with which s/he will not be familiar, nor likely understand. Work to minimize their use, and be sure of clarity to the non-expert throughout. Please do not misinterpret this advice as suggesting you “dumb down” your presentation. Rather, you need to be sure it is understandable to the clinically informed reader without expertise in this area, while still providing detailed information that will be of interest to experts in the area. I will not point out each time such problematic presentation occurs. Please bear this issue in mind throughout your revisions.

17. Page 12: “…in both groups…” Please clarify whether you mean about 80% of patients from both groups combined, or about 80% in each of the two groups.

18. Page 12 and elsewhere: I suspect that many non-orthopedic / neurosurgical readers will not know what subsidence is with regard to implants. Be sure to define.

19. Please provide greater detail regarding the actual tumors reported (e.g., histology, type of surgery, timing in relation to surgery, dose of rhBMP used (for those in that group); consider doing so in tabular form).

20. Explain to readers the importance / rationale for an analysis in which you excluded cancers not included in SEER.

21. “Since rhBMP-2 was evaluated as part of a device…” – this might not be understood by readers unfamiliar with the procedures you are assessing.

22. Tables: these will change according to the advice our Statistical Editor. In current Table 1, for the INFUSE /L.T. CAGE Pilot study (Boden 2000) – is there a typographical error or were there really 46 rh-BMP-2 patients with prior back surgery and 0 among those receiving ICBG? Table 3: it is unclear why you are showing these data at 6 weeks, 6, 12 and 24 months. Figure 1- in a footnote provide information regarding the 3 pending trials (e.g., the trial registration numbers).
23. Discussion: It would be very helpful to add a brief discussion of rhBMP’s known biologic properties and whether there are proposed (or even established) mechanisms by which it might cause the harmful effects suggested.

24. As your revise your discussion, consider structuring along the following lines: Provide a brief synopsis of key findings. Discuss possible mechanisms and explanations for the findings. Discuss relevant previous reports, and explain how your work adds to or differs from them. Then discuss limitations of your study. Mention important future research directions in this area. Conclude with a brief section that summarizes in a straightforward and circumspect manner the clinical implications of the work.

Please send your revised manuscript and cover letter to us within four weeks of receiving this letter. Number the lines of your revised manuscript consecutively throughout. In the cover letter, group and number your responses to correspond to the comments from the Editor, Reviewer, Statistician, and the production editor who asked for technical revisions. Please restate each comment and follow it with your response indicating what you did, why you did it, and the line number of the revised manuscript where the changes may be found. Submit your revised manuscript and cover letter at https://www.acponline.org/authors/ by clicking the "Revise Paper" link listed below your "Tasks."

Annals now requires authors to complete the International Committee of Medical Journal Editors (ICMJE) conflict of interest disclosure form. This form, which is discussed further at www.icmje.org/format.pdf, is intended to facilitate detailed reporting of conflicts of interest and standardize the format of reporting across ICMJE member journals. Upon final acceptance, each author will receive an email with a link to access, and upload, the ICMJE form. Individual forms for an article are published online on the day of publication for readers to access disclosure information.

Please keep all editorial correspondence confidential, and refrain from sharing either the correspondence itself or the essence of its content with individuals who are not your collaborators. Doing so helps ensure we can offer you advice that is in the best interests of your paper, without concern for how it might be considered or used by others.

We look forward to receiving your revised manuscript.

Sincerely,

Darren Taichman, MD, PhD

Executive Deputy Editor
7. Authors: E-mail Inquiry (11 January 2013)

January 11, 2013

Darren Taichman
Executive Deputy Editor, Annals of Internal Medicine

REF: M12-2731

Dear Dr. Taichman,

I would like to thank you and the other editors and reviewers for the careful comments on our manuscript (M12-2731 Effectiveness and harms of recombinant human morphogenetic protein-2 (rhBMP-2) in spine fusion: A systematic review and meta-analysis). Our team has met and discussed each of the comments and we have conducted some additional analyses in response to the comments.

Prior to finalizing the revisions to the manuscript, we would like to discuss four of the comments from the primary statistical editor, Dr. John E. Cornell, specifically:

1. Use of mixed-effects model
2. Analysis method and effect measure for rare adverse events
3. Method for combining IPD and additional study-level summary data
4. Use of 90% Confidence Interval

We would be happy to discuss any additional changes you may suggest. Please see the attached file for more detail.

Sincerely,

Rongwei Fu, Ph.D
3181 SW Sam Jackson Park Rd.
MC: CB669
Portland, OR 97239
USA

Statistical Editor Comments:

1. Use of the mixed-effects model (comment 3.3)

“Although the authors have IPD available, the current analytic approach aggregates this rich source of data to form a series of study-level statistics that the authors combine using traditional meta-analytic methods. This approach is disappointing and inefficient. If you have the patient-level data, you need to make maximum use of this rich data source…Use more sophisticated mixed-effect models to combine the outcomes and adverse events across trials, like you would in
a multi-center clinical trial. Doing so will clearly strengthen the review beyond the information already published by Carragee, et al.”

Response:

We agree that, in general, the mixed-effect model offers advantages in exploring the association between treatment differences and participant-level characteristics through meta-regression. However, we would like to propose keeping our two-step study level approach as the primary analysis for the following reasons:

a) When the primary interest of an IPD analysis, as in this case, is to assess overall treatment differences, rather than effects based on participant-level characteristics, the mixed effects model and study level analysis produce very similar results (see, for example, Higgins et al. 2001; Olkin and Sampson, 1998) (1, 2). To confirm this, we conducted the mixed-effect model for continuous outcomes (assuming random treatment effects and heterogeneous residual variance across included studies) and got very similar results compared to the two-step approach (see attached table). We propose adding the mixed effects model as a sensitivity analysis and including the main results of the mixed effects model analysis as an appendix. We do plan to investigate the effects of patient level characteristics in more detail in future analyses, and would use the mixed-effects model for those evaluations.

b) Although the mixed effects model is also more flexible than traditional methods in handling missing data (e.g., the mixed effects model can include all available data even if there is missing data for some subjects, as long as the data are missing at random), our two-step analysis (calculating summary statistics in the first step and then using traditional meta-analysis methods) allowed us to use all available data from post-operative time points, as in the mixed model approach. In addition, an advantage of IPD is to “redefine multiple version of variables based on missing patterns”, and we were able to conduct analyses incorporating more subjects with missing data by redefining outcome variables, partially addressing this problem. Use of different definitions for variables generally did not make a difference in the results and overall conclusions.

c) We believe our two-step approach also has important advantages. The two-step approach makes it easier for readers to contrast our results with published results from individual studies and provides estimates for measures of heterogeneity (for example, I^2) that are more familiar than the variance components from the mixed model. It also provides more easily interpretable measures for some outcomes, such as harms, since in the mixed model approach the best measure for a dichotomous outcome is an odds ratio(OR). Converting an OR to a more clinically interpretable relative risk or risk difference requires additional calculations based on assumptions about the baseline risk.

d) While we did not use the IPD to explore the association between treatment differences and participant-level characteristics, our approach highlights other advantages of the availability of IPD. The IPD allowed us to evaluate both benefits and harms much more thoroughly than Carragee could. We had much more complete data on adverse events than Carragee had, and were able to 1) check the quality of data and define the outcome variables in a more consistent
manner; 2) define new variables such as radiculitis, and redefine multiple version of variables based on missing patterns as mentioned above; and 3) generate more appropriate summary statistics, with complete data from all time points.

e) Keeping the two-step approach as the primary analysis is consistent with the methods in our grant proposal and in our protocol submitted to the PROSPERO systematic review registry (available online at: http://www.crd.york.ac.uk/PROSPERO/ ). (3) We also believe we can be clearer in our methods and discussion about why the two-step approach was used.

2. Analysis method and effect measure for rare adverse events (comment 3.4)

“Use traditional meta-analytic methods to combine the actual study-level estimates obtained through your literature search, but exercise more care in summaries with sparse data and zero events. The method used in the current analyses uses a fixed-effects Mantel-Haenszel method with 0.5 continuity correction for these cases (Stata metaan default). I suggest that the authors use a fixed-effect logistic model based on the grouped counts within study. This approach does not require a correction for continuity, and it is easily done with the Stata glogit command, using the margins command to generate study-level estimates and 95% Confidence Intervals. This approach is more in line with recommendations made by the first author in their nicely written J Clin Epidemiol 2011; 64:1187-97 article.”

Response:

For analyzing rare adverse events, we agree (with ourselves!) that using the fixed-effects Mantel-Haenszel method with 0.5 continuity correction is not the best approach. However, the Stata glogit procedure produces an odds ratio, which is not ideal either. As an alternative, we propose using the Mantel-Haenszel method with no continuity correction, and conducting sensitivity analyses based on additional correction values (in addition to 0.5).

3. Method for combining IPD and additional study-level summary data (comment 3.6)

Response:

We wanted to be clear that because only one study provided additional study-level data beyond IPD, we could not use more complex methods to combine data; as discussed previously we used the two-step approach to combine IPD and study level data.

4. Use of 90% Confidence Interval (CI). (comment 3.7)

“Report absolute risk differences and confidence intervals for adverse event rates. Use CONSORT guidelines when computing the confidence intervals for these events: 90% CIs for serious adverse events.”

Response:
Although we calculated absolute risk differences (RDs), we did not report them for all harms because only two of the adverse events had statistically significant differences: cancer and “back and leg pain”. For cancer, we reported the risk difference and number need to harm (NNH) in the Results section. The latter consisted of a set of poorly defined, heterogeneous outcomes, and was therefore not a very meaningful category, thus RD was not reported. In addition, for rare adverse events, we focused on the risk ratio (RR) rather than the risk difference, since studies suggest that combined estimates based on the RD are often biased and have conservative CI coverage and low statistical power (4). We did report the absolute risk of each adverse event for each group in Table 4, which we believe helps readers understand the frequency of adverse events.

We have reviewed the CONSORT statement and were unable to identify any explicit recommendation on using a 90% CI. Rather, it states, “A 95% confidence interval is conventional, but occasionally other levels are used.” We also reviewed the CONSORT extension on harms, which does not discuss which CI’s to use. It is unclear to us why a 90% CI interval was selected by Carragee et al, as we believe a 95% CI is considered the standard.
<table>
<thead>
<tr>
<th>Outcome Scale</th>
<th>6 weeks</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALIF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ODI (0-50)</td>
<td>-2.52 (-7.10 to 2.05)</td>
<td>-3.79 (-8.69 to 1.11)</td>
<td>-3.74 (-9.09 to 1.60)</td>
<td>-7.35 (14.00 to 0.70)</td>
</tr>
<tr>
<td>Back pain (0-10)</td>
<td>0.21 (-0.28 to 0.71)</td>
<td>-0.36 (-0.94 to 0.22)</td>
<td>-0.51 (-1.18 to 0.16)</td>
<td>-0.74 (-1.49 to 0.00)</td>
</tr>
<tr>
<td>Leg pain (0-10)</td>
<td>-0.57 (-1.12 to -0.02)</td>
<td>-0.20 (-0.72 to 0.32)</td>
<td>-0.49 (-1.07 to 0.08)</td>
<td>-0.60 (-1.28 to 0.08)</td>
</tr>
<tr>
<td>SF-36 PCS (0-100)</td>
<td>0.58 (-0.99 to 2.14)</td>
<td><strong>2.99 (0.69 to 5.31)</strong></td>
<td><strong>2.94 (0.85 to 5.03)</strong></td>
<td><strong>3.68 (0.86 to 6.49)</strong></td>
</tr>
<tr>
<td>SF-36 MCS (0-100)</td>
<td>-0.39 (-2.48 to 1.69)</td>
<td>-0.33 (-2.24 to 1.59)</td>
<td>-0.56 (-2.60 to 1.48)</td>
<td>2.90 (-0.29 to 6.08)</td>
</tr>
<tr>
<td><strong>Weighted mean difference (95% CI) – traditional meta-analysis methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ODI (0-50)</td>
<td>-2.33 (-6.59 to 1.93)</td>
<td>-3.62 (-8.02 to 1.81)</td>
<td>-3.24 (-8.30 to 1.81)</td>
<td>-6.94 (-13.90 to 0.02)</td>
</tr>
<tr>
<td>Back pain (0-10)</td>
<td>0.22 (-0.38 to 0.82)</td>
<td>-0.31 (-0.82 to 0.20)</td>
<td>-0.51 (-1.19 to 0.16)</td>
<td><strong>-0.62 (-1.23 to 0.02)</strong></td>
</tr>
<tr>
<td>Leg pain (0-10)</td>
<td>-0.57 (-1.12 to 0.02)</td>
<td>-0.20 (-0.72 to 0.31)</td>
<td>-0.51 (-1.13 to 0.12)</td>
<td>-0.55 (-1.15 to 0.05)</td>
</tr>
<tr>
<td>SF-36 PCS (0-100)</td>
<td>0.55 (-1.02 to 2.11)</td>
<td><strong>2.81 (0.85 to 4.76)</strong></td>
<td><strong>2.95 (0.86 to 5.04)</strong></td>
<td><strong>3.34 (0.92 to 5.75)</strong></td>
</tr>
<tr>
<td>SF-36 MCS (0-100)</td>
<td>-0.36 (-2.45 to 1.72)</td>
<td>-0.31 (-2.22 to 1.60)</td>
<td>-0.56 (-2.60 to 1.47)</td>
<td>2.86 (-0.20 to 5.92)</td>
</tr>
<tr>
<td><strong>PLF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ODI (0-50)</td>
<td>0.74 (-1.68 to 3.17)</td>
<td>-2.40 (-4.85 to 0.04)</td>
<td>-2.09 (-5.28, 1.10)</td>
<td>-1.98 (-4.86 to 0.90)</td>
</tr>
<tr>
<td>Back pain (0-10)</td>
<td>0.10 (-0.27 to 0.48)</td>
<td>-0.46 (-1.14 to 0.23)</td>
<td>-0.42 (-1.34 to 0.50)</td>
<td>-0.31 (-0.76 to 0.15)</td>
</tr>
<tr>
<td>Leg pain (0-10)</td>
<td>0.23 (-0.21 to 0.66)</td>
<td>-0.27 (-0.71 to 0.17)</td>
<td>-0.29 (-0.75 to 0.16)</td>
<td>-0.34 (-0.82 to 0.13)</td>
</tr>
<tr>
<td>SF-36 PCS (0-100)</td>
<td>-0.10 (-1.15 to 0.96)</td>
<td><strong>1.79 (0.27 to 3.31)</strong></td>
<td>1.83 (-0.19 to 3.85)</td>
<td>1.10 (-0.65 to 2.86)</td>
</tr>
</tbody>
</table>
### References

January 29, 2013

Rongwei Fu, Ph.D
3181 SW Sam Jackson Park Rd.
#CB669
Portland, OR 97239
USA

REF: M12-2731

Dear Dr. Fu:

Below I am pasting the response from our statistical editorial team to your questions:

Response to Dr. Fu’s inquiry regarding our statistical review

We greatly appreciate your allowing us to review your response to our statistical review prior to resubmission of your manuscript. Hopefully, this will help speed up the review and revision cycle. We apologize that we failed to respond sooner to your comments and suggestions, but we thought it best to discuss your comments among the statisticians involved in the initial review.

1. With respect to the use of a mixed-effects model, we are aware of the Higgins et al. 2001 and Olkin and Sampson, 1998 article. While it is true that study-level effects are often similar, the loss of information due to aggregating the data tends to underestimate the variances. It is also clear from a number of simulation studies, including one by DerSimonian and Kacker, 2007, that the one-step DL estimator provides poorer coverage than other iterative methods for estimating random-effects. The problems with the DL estimator are greater when the data is sparse, such as is the case with adverse events (see Shuster, et al. Research Synthesis Methods 2012; DOI: 10.1002/jrsm.1039). So, while the estimates you provided are similar in most cases, we strongly prefer that you present the full-information mixed-effects model results as your primary analysis for publication in Annals of Internal Medicine. It is the stronger and more reliable analysis. We are far less concerned about comparability of “methods” across publications, than we are with providing the most optimum analysis for a given problem. As you know, when event rates are low, OR and RR are nearly equivalent; but, if you are concerned about consistency, you can easily generate RR estimates from any generalized mixed-effects software. You don’t need to make unnecessary assumptions when you have patient-level data. Patient-level analyses are analogous to multi-site randomized controlled trials. You simply fit a model for a binomial response variable with a log link, rather than a logit link. It is easily done with GLLAMM in Stata or Proc NLMIXED in SAS. Theo Stijnen, et al.’s Tutorial in Biostatistics (Stat Med 2010; 29: 3046-3067) shows how to fit these models to study-level data with sparse outcomes with Proc NLMIXED. The models are easily adapted to accommodate IPD, as well.
2. We agree that you can use the MH method without continuity correction to pool the data, rather than the glogit procedures. The latter was simply a suggestion. However, we still prefer that you use the full-information in the data rather than pool study-level aggregate data.

3. You are right that it makes little sense to invoke complicated Bayesian models to combine IPD and study-level estimates when we only have one published study. We prefer that you take a simpler approach and simply compare and contrast the estimates from the mixed-effect model with the study-level estimates from this single study. There is no need to quantitatively combine the results from these two levels of analysis.

4. You are also correct that the CONSORT statement is silent regarding the coverage for confidence intervals. A more careful reading of Carragee et al. suggests that they are simply referring to the principle that confidence need to be calculated and presented. So, how did they come up with the principle of using 90% confidence intervals for severe or catastrophic adverse events? The FDA often uses 90% confidence intervals for such events in their review and postings of adverse events. We have not reviewed the FDA adverse event reports for rhBMP-2, but we expect that 90% CI is used in these reports. If this is the case, then we suggest that your presentation be consistent with the type of information the FDA uses. Though we don’t always agree with the FDA’s rules, particularly with handling of missing data in clinical trials, we believe that it is prudent to be consistent with their approach to reporting serious adverse events in this case. If you prefer, you can present both the 90% and 95% in your main table, and comment on the use of different confidence coverage in your discussion. It would be interesting to know if the two representations lead to differing conclusions about the safety of rhBMP-2.

We look forward to seeing your responses to these items as well as the others raised in our prior communication with comments from our editorial team and reviewers, as well of-course your revised manuscript. Please refer to my prior letter with instructions regarding how to prepare and submit these responses / revision.

Thanks again!

Sincerely,

Darren Taichman, MD, PhD
Executive Deputy Editor
Dear Dr Taichman,

We are submitting a revised version of our manuscript, "Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-analysis." Our responses to the comments from the Editor, the Statistical Editor, the Reviewers, and the Production/Technical Editor are detailed below. We discussed some comments with the Editor and the Statistical Editor before this submission. The additional comments and our responses are attached at the end of this document. The manuscript has undergone substantial revisions and reorganization based on the comments, including expansion of the Methods section and an expanded presentation of results for Aim #2 (Reporting Bias).

We would be happy to discuss any additional changes you may suggest. Thank you for the opportunity to resubmit our manuscript.

Sincerely,

Rongwei Fu, Ph.D.
3181 SW Sam Jackson Park Rd.
MC: CB669
Portland, OR 97239
USA

**Editor’s comments:**

**General**

1. Your writing requires improvement to make the subject matter more accessible to the clinically experienced reader who is not an expert in spinal surgery. Bear in mind when revising your paper that most readers will not be well versed in the abbreviations or trial acronyms you are currently using. Minimize their use, and be sure they are well defined when it is necessary to use them.

Response: We have revised the text to make it more generally understandable. The Introduction more clearly describes spinal fusion and the different surgical techniques evaluated. We
eliminated many abbreviations and acronyms names (AMPLIFY and MAVERICK) in the text and minimized the use of abbreviations overall. For the main surgical techniques (ALIF, PLIF, TLIF, PLF), we used abbreviations after defining them on first use. We have spelled out the names of surgical techniques in the headings in the Results section and in tables and figures.

2. Your results section requires reorganization so that the stated goals (to review the benefits and harms of rhBMP-2, and assess the reporting by Medtronic) are each clearly addressed. Currently, your second goal (to assess the quality of and possibility of biased /selective reporting) is buried in a rather haphazard manner within the other results presented, and so it is difficult for the reader to evaluate your findings regarding this aspect of your study. Please see the advice of our Statistical Editor regarding how to better organize your results so as to overcome this problem.

Response: We reorganized the Results so that our analysis of Reporting Bias (Aim #2) is reported separately, after presenting the analysis of effectiveness and harms (Aim #1). Please see below our specific responses to the Statistical Editor comments.

3. We are concerned that additional data are available that might inform your review. These are noted in the review you reference (Carragee et al, Spine J, 2011). Please see the comments of our Statistical Editor to this issue as well.

Response: We reviewed the additional data sources mentioned in the Carragee review and the Statistical Editor comments. Data cited by Carragee from the Scoliosis Research Society did not meet our inclusion criteria because it combined results for rhBMP-2 and BMP-7 (OP-1 putty); it was also reported in a journal publication that we excluded (Williams BJ, Smith JS, Fu KM, Hamilton DK, Polly DW Jr, Ames CP, et al. Does BMP increase the incidence of perioperative complications in spinal fusion? A comparison of 55,862 cases of spinal fusion with and without BMP. Spine (Phila Pa 1976). 2011 Mar 9. [Epub ahead of print] PMID: 21394069.) Carragee incorrectly cited the Centers for Disease Control and Prevention as a data source, but this was in fact data from the Center for Devices and Radiological Health within the FDA, available at the FDA website, which we included. We included the study from the National Inpatient Database (Cahill KS, Chi JH, Day A, Claus EB. Prevalence, complications, and hospital charges associated with use of bone-morphogenetic proteins in spinal fusion procedures. JAMA. 2009 Jul 1;302(1):58-66. PMID: 19567440. )

Specific Comments:
4. Abstract: Your abstract also requires revision to make the topic accessible to non-experts in this area. For example, many readers will not know what rhBMP-2 is. They may not know what “similar benefits” are referred to. Currently, the abstract’s data synthesis contains no information regarding your assessment of reporting quality. The conclusion also does not provide readers with your assessment of your results with regard to this stated objective (to assess reporting).

Response: We revised the Abstract/Background to better describe vertebral fusion and the purpose of rhBMP-2 “Recombinant human bone morphogenetic protein-2 (rhBMP-2) is used as a bone graft substitute in spinal fusion, a procedure that unites (fuses) bones in the spine. The
accuracy and completeness of journal publications of industry-sponsored trials on the effectiveness and harms of rhBMP-2 has been called into question.” To clarify what is “similar benefits”, we revised the original sentence in the Abstract/Data Synthesis section and it now says “Based on meta-analysis of IPD from RCTs, rhBMP-2 and bone graft had similar effects on overall success, fusion, and other measures of effectiveness in lumbar and cervical spine fusion.”

Also to describe the outcomes measures more accurately, we replaced the term “benefits” with the term “effectiveness” throughout the Abstract and the main text.

We revised the Abstract/Data Synthesis section to include a summary of reporting bias: “The journal publications misrepresented the effectiveness and harms through selective reporting, duplicate publication, and underreporting.”

We also revised the Conclusion to describe our assessment with regard to reporting bias: “Earlier disclosure of all relevant data would have better informed clinicians and the public than the journal publications did.”

5. Introduction: Most readers of Annals will not be familiar with ALIF, PLF and TLIF. You should provide a little more information so that the non-expert will understand the goals of these procedures and their differences. In addition, it is important to better orient the non-expert reader to what each rhBMP-2 and bone grafts are meant to do. I realize this might seem foolish, but many readers might not know the purpose of the material and how they are thought to work. Finally, knowing just a little more about what rhBMP-2 does in a very general way (1 or 2 sentences) might help readers understand the kinds of complications worried about (a point that will require some elaboration in your discussion).

Response: We revised the Introduction to better describe vertebral fusion and various techniques (Page 4 starting from line 106) “The most common surgery for chronic low back pain with lumbar disc degenerative conditions is vertebral fusion (1) to restrict spinal motion and remove the presumed cause of pain. An interbody fusion, involving removal of a degenerated intervertebral disc and fusion of the adjacent vertebral bodies, can be performed via an anterior (anterior lumbar interbody fusion, ALIF), posterior (posterior lumbar interbody fusion, PLIF), or transforaminal (transforaminal lumbar interbody fusion, TLIF) approach. Fusion involving adjacent transverse processes is referred to as posterolateral lumbar fusion (PLF).”

We also revised the Introduction to describe more clearly the use of bone graft and rhBMP-2 (Page 4 starting from line 113): “Traditionally, spinal fusions are performed using graft material harvested from the patient’s iliac crest to promote fusion. In 2002, the U.S. Food and Drug Administration (FDA) approved recombinant human bone morphogenetic protein-2 (rhBMP-2), a genetically engineered protein with bone-growth stimulating properties, as a bone graft substitute in conjunction with a device implant...”

6. Introduction: Although obvious after reading your full paper, or knowing a bit about this subject, make it clear early on if Medtronic is the only maker of rhBMP-2 and to whom you refer when noting, “…industry-sponsored trials...”
Response: We added the sentence (Page 4 line 120) “Medtronic, Inc. (Medtronic) is the sole manufacturer of devices involving rhBMP-2 for spinal fusion.”

Also, help orient the reader a little better to the concerns raised by the FDA. What life-threatening (or other) complications were of concern? Also, did the FDA or others raise the concern that there had been purposeful selective reporting by the manufacturer?

Response: We revised to state (Page 4 starting from line 126): “In 2008, the FDA issued a public health notification of life-threatening complications associated with off-label use of rhBMP-2 in cervical spine fusion—swelling of the neck and throat resulting in compression of the airway and other structures.”

The FDA should have received full results for all specified outcomes for the trials for device approval before manuscripts were submitted for journal publication, and the FDA did not raise concern about under-reporting or selective reporting. However, under-reporting or selective reporting was discussed in publications that we referred to in the Introduction (Page 5 line 130): “Selective reporting or underreporting of outcomes in journal publications could have led to misleading conclusions about the balance of benefits and harms of rhBMP-2 (13, 15).”

Methods: Data Sources and Searches: Tell readers specifically whether the 17 studies submitted by Medtronic to the FDA represent all studies known to have been completed by the company.

Response: We revised the Methods/Data Source section to state (Page 6 line 151): “…the YODA Project provided de-identified patient-level data, protocols, data dictionaries, and internal reports submitted to the FDA for all 17 Medtronic-funded studies of rhBMP-2 in spinal fusion that were completed or terminated by December, 2011.”

8. Methods: in stating the goals of your study, if an assessment of biased or selective reporting was a specific goal in your evaluation of the “quality of reporting” by Medtronic, say so.

Response: We revised the Introduction (Page 5 line 133) to be clearer that one of the goals was to assess “…2) assess reporting biases in published articles of industry-sponsored studies”

9. Page 6: Tell non-expert readers whether a change in ODI of 15 has been established as a clinically significant change. Also, how was, “maintenance of improvement” defined?

Response: We added a Table defining various outcomes (Appendix Table 1). For ODI success, we now says in the Appendix table that “At least a 15-point improvement in ODI score for back pain at each visit postoperatively compared with pre-operative index score (FDA’s recommendation, a 15-point improvement is clinically meaningful based on Copay, 2008)” and “At least a 15-point improvement in NDI score for neck pain at each visit postoperatively compared with pre-operative index score (FDA’s recommendation, a 15-point improvement is clinically meaningful based on Copay, 2008).”
The term “maintenance of improvement” was used to defined “Neurological Success” and now in Appendix Table 1, we have provided detailed information about how “Neurological Success” was defined in different studies. For example, “Four neurologic tests evaluated motor function, sensory function, deep tendon reflexes, and sciatic tension signs (straight-leg raise). A score was developed for each test.” For studies 4, 8, 10, 14, “neurologic success was defined as having the same or better score in all four tests compared to pre-operative score”.

10. Page 7: Why did you exclude each cervical/arm symptoms, numbness/parasthesias without weakness or pain, just back pain and just leg pain?

Response: We excluded cervical and arm symptoms from our primary definition of lumbar radiculitis, since cervical radiculopathy should not occur following a lumbar surgery. We excluded numbness/parasthesias without weakness or pain, as well as back pain or leg pain without weakness, since these are nonspecific symptoms that do not necessarily indicate radiculitis. However, we conducted sensitivity analysis using three alternative definitions for radiculitis, some broader and some more restrictive. For example, one definition classified patients with leg symptoms with or without back pain that followed a radicular distribution as having radiculitis, and another definition classified patients with any back and leg pain as having radiculitis. Results were similar when using alternative definitions (Table 2). Appendix Table 1 shows alternative definitions we used for radiculitis.

11. Page 7: Why are harms (other than death or cancer) limited to 24 months?

Response: We limited analyses of harms other than death or cancer to 24 months in the manuscript since all trials provided data up to 24 months, and few trials reported harms after 24 months. However, we reported longer-term results in the full report of our systematic review and meta-analysis, as stated in the Methods (Page 9 starting line 217): “Data beyond 24 months were sparse and are reported elsewhere (17), except for cancer and death, and for these outcomes, we also analyzed the cumulative number of events up to 48 months.”

12. Page 7: How often were patients with pre-existing cancer excluded? Further, as these outcomes were rare we would think these patients should be included (except where it might be clear that the pre-existing cancer caused an oncological issue occurring after spinal fusion, such as the appearance of metastatic lesions from a previously known breast cancer).

Response: To clarify, we did not exclude patients with preexisting cancers. We only excluded 2 cancers that were identified in the immediate postoperative period but determined to exist before surgery, as shown in the Footnote of the newly added Appendix Table 3. We apologize for this confusion and have removed the sentence from the Methods section.

13. Results: It is currently difficult for the reader to gain a clear message regarding the quality of the available data sources, how they differed, and your results addressing the question of selective / biased reporting. Please see the advice from our Statistical Editor regarding more informative means of organizing your results.
Response: We revised the beginning part of the Methods section to more clearly describe the different sources of data and how they were used to address our two aims (Page 5 starting from line 143): “To achieve the two aims of our review, we used four sources of data: 1) Medtronic IPD, related protocols, and data dictionaries; 2) Medtronic internal reports; 3) summary documents from the FDA Web site, and 4) a broad-based literature search to identify a) additional studies on rhBMP-2 and b) publications related to Medtronic-sponsored studies. For aim 1, we used data from sources 1), 2) and 4 a), and for aim 2, we compared the results from journal publications with data from the other sources.”

For trials with individual patient data, we based our assessments of quality on the information provided in study protocols and internal documents provided by Medtronic, as the journal publications did not provide additional information. We revised the Methods section to make this clearer (Page 7 line 172): “For Medtronic-funded studies, quality assessment was based on information from trial protocols and internal reports.”

As discussed in the response to comment 2 above, we have re-organized the results section so that our findings with regard to selective reporting are now presented separately (Page 16 starting from line 392).

14. Page 9: Re: pain at the bone graft harvest site being assessed only in the control group. Please confirm that patients in the rhBMP-2 groups in these studies did not undergo sham harvests.

Response: Correct, patients in the rhBMP-2 group did not undergo sham iliac bone graft harvest.

15. Page 10: Although it is clear that a greater number of adverse events were reported to the FDA than reported in published manuscripts, can we be sure that they have truly been, “reported thoroughly in documents Medtronic submitted to the FDA?”

Response: Yes, we found no differences in the rates of adverse events reported to the FDA and the rates of adverse events based on individual patient data.

16. Page 11, first paragraph: this is an example of where the non-expert reader will be lost in many abbreviations and acronyms with which s/he will not be familiar, nor likely understand. Work to minimize their use, and be sure of clarity to the non-expert throughout. Please do not misinterpret this advice as suggesting you “dumb down” your presentation. Rather, you need to be sure it is understandable to the clinically informed reader without expertise in this area, while still providing detailed information that will be of interest to experts in the area. I will not point out each time such problematic presentation occurs. Please bear this issue in mind throughout your revisions.

Response: We have removed this paragraph as we agree it provides unnecessary detail about the trials. Instead, we replaced it with a more general overview describing the number and size of trials, quality, and trial characteristics (Page 12 starting from line 296) “Five Medtronic-sponsored trials with individual patient data (four fair quality—Studies 2, 4, 5, and 9—and one...
poor quality—Study 1) evaluated rhBMP-2 vs. ICBG in ALIF. Studies 1, 2, and 9 used rhBMP-2 with either the LT-Cage or the INTER FIX device, and Studies 4 and 5 used rhBMP-2 off-label with bone dowels. Study 5 was terminated early when less than half of the projected sample (n = 180) were enrolled.”

We have made similar revisions at other places when necessary.

17. Page 12: “…in both groups…” Please clarify whether you mean about 80% of patients from both groups combined, or about 80% in each of the two groups.

Response: We clarified that the 80% was from each of the two groups (Page 13 line 308).

18. Page 12 and elsewhere: I suspect that many non-orthopedic / neurosurgical readers will not know what subsidence is with regard to implants. Be sure to define.

Response: We revised to include a definition of subsidence (Page 13 line 312): “…defined as sinking or settling of the device into bone.”

19. Please provide greater detail regarding the actual tumors reported (e.g., histology, type of surgery, timing in relation to surgery, dose of rhBMP used (for those in that group); consider doing so in tabular form).

Response: We added an Appendix Table 3 with additional information for each tumor as requested.

20. Explain to readers the importance / rationale for an analysis in which you excluded cancers not included in SEER.

Response: We provided a rationale for an analysis excluding cancers not reportable by SEER (Page 10 line 239): “skin cancers with low propensity to metastasize”. Cancers not reportable by SEER (Squamous and basal cell carcinoma) were footnoted in Appendix Table 3.

21. “Since rhBMP-2 was evaluated as part of a device…” – this might not be understood by readers unfamiliar with the procedures you are assessing.

Response: These sentences have been removed to avoid confusion.

22. Tables: these will change according to the advice our Statistical Editor. In current Table 1, for the INFUSE /LT_CAGE Pilot study (Boden 2000) – is there a typographical error or were there really 46 rh-BMP-2 patients with prior back surgery and 0 among those receiving ICBG? Table 3: it is unclear why you are showing these data at 6 weeks, 6, 12 and 24 months. Figure 1- in a footnote provide information regarding the 3 pending trials (e.g., the trial registration numbers).
Response: In Table 1, Boden 2000, the rates of 46% vs. 0% for prior back surgery are correct, but there were only 3 patients in the control group. We have added the numerators in the table so this is clearer.

In Table 3, we focused on data at 6 weeks, 6, 12 and 24 months, since these were typical time points for follow-up in the Medtronic trials. We also added the results from 3 months.

In Figure 1 (now Appendix Figure 1), we added the trial registration numbers for the pending trials.

23. Discussion: It would be very helpful to add a brief discussion of rhBMP’s known biologic properties and whether there are proposed (or even established) mechanisms by which it might cause the harmful effects suggested.

Response: We revised to state (Page 21 line 502): “Potential mechanisms for these adverse events may be related to pro-inflammatory or carcinogenic effects of rhBMP-2 (63).”

24. As your revise your discussion, consider structuring along the following lines: Provide a brief synopsis of key findings. Discuss possible mechanisms and explanations for the findings. Discuss relevant previous reports, and explain how your work adds to or differs from them. Then discuss limitations of your study. Mention important future research directions in this area. Conclude with a brief section that summarizes in a straightforward and circumspect manner the clinical implications of the work.

Response: We revised the Discussion section that includes all of the components discussed above. For example, we added a paragraph comparing our results with the reviews by Carragee and others (Page 21 starting from line 504): “In their review, Carragee et al. demonstrated that adverse events were underreported in publications of five trials for which the FDA had made full results public (13). Our study confirms this finding and also demonstrates that under-reporting was also pervasive in trials of off-label uses for which results previously had not been available to the public. Other reviews generally found rhBMP-2 and ICBG associated with similar benefits for ALIF and PLIF based on low or moderate strength evidence, but lacked data to evaluate harms. Ratko et al. noted that the absence of reported harms could be due to non-reporting by investigators (15).”

We also added a sentence describing possible pro-inflammatory and carcinogenic effects of rhBMP-2 (see response to comment #23). We mentioned important future research direction as (Page 23 starting line 542) “More research is needed to provide more reliable estimates of risk of cancer and other adverse events and to identify patient populations in which use of rhBMP-2 may be associated with greater effectiveness, such as cases where use of bone graft alone is associated with a high risk of pseudoarthrosis.” and we concluded with the clinical implication of this work as (Page 23 line 545) “Based on the currently available evidence, it is difficult to identify clear indications for rhBMP-2 in spinal fusion.”
Reviewer comments:

Major Comments:
The authors are to be commended for addressing a difficult topic with a modest amount of information. Although the awareness of internists to the benefits and risks of new devices is important, the topic of bone morphogenetic protein for fusion operations seems to be too specialized a topic for an internal medicine journal. Most internists do not believe that the deliberation about a specific operation for the spine is in their purview let alone the decision about autologous bone graft versus BMP.
The major aim of the paper was to highlight discrepancies between individual patient data and reports of randomized clinical trials using the same patients.

Response: Thanks for your comment. To clarify, our aims were to assess the benefits and harms of rhBMP-2 as well as to evaluate for selective reporting and related biases.

In regard to study selection, the authors excluded trials with other BMPs but did not exclude a study with a large population that included another experimental device – artificial disc. The artificial disc has its own series of adverse events. The inclusion of this study makes it difficult to separate relative harms

Response: The MAVERICK trial met inclusion criteria because rhBMP-2 was used in the control arm. We agree that it should not be pooled with trials of rhBMP-2 vs. bone graft to evaluate effectiveness and harms given the differences in comparisons, and we did not do that in our analysis. It was excluded from meta-analysis except for cancer and death, for which the surgical comparison may be less important than the fact that rhBMP-2 was used in one group and not the other.

In Quantitative Synthesis – It is confusing regarding the inclusion and exclusion criteria. Success was defined with fusion, ODI improvement of 15 and no adverse events. Back pain is not specially stated. However, possible radiculitis as an adverse event includes back and leg pain, but excludes back pain alone or leg pain alone. These criteria seem arbitrary. Is a patient excluded from success if they have back pain? Is a patient excluded from adverse events if they have back pain alone?

Response: We now provide a table (Appendix table 1) to provide detailed information about how overall success (and other outcomes) is defined. Because studies varied slightly in how they defined overall success, as described in the Methods (Page 7 line 183) we used the raw IPD to re-calculate overall success using a standardized definition as radiographic fusion; improvement of Oswestry Disability Index [ODI] score, maintenance or improvement in neurological status, and no serious adverse event classified as implant- or implant/surgical-associated and no additional surgical procedure classified as “failure”.

That is, in the definition of overall success, back pain was measured by ODI, but not by the back and leg pain (or radiculitis) as an adverse event. Patients with these symptoms were not excluded from “overall success” unless they met the criteria above for “serious” adverse events classified as implant- or implant/surgical-associated. We agree that the presence of possible radiculitis or back or leg pain alone is difficult to interpret.
Results – Page 9  The authors point out the inconsistencies of data gathering in the studies. For example, only those who had bone graft were asked about buttock pain versus all individuals in the study. As you pointed out, many studies did not ascertain adverse events in an established manner. This suggests that the number of adverse events are potentially underreported, Also on 2 of the 14 studies measured no overall success rate.

Response:  Noted. Thank you for your comment.

Page 14 - The lack of data available to determine adverse events related to BMP versus alternative causes is highlighted by the increase in back and leg events but with causes very unlikely to be related to the device like “arthritic knee or ankle pain” It makes data related to adverse events difficult to interpret.

Response:  We agree with the Reviewer, and therefore applied standardized definitions for possible radiculitis that we believe are more specific (see response to Editor comment #10). We also agree with the reviewer that failure to define adverse events and inconsistency in categorization makes it difficult to interpret results, as stated in the Discussion (Page 22 line 519): “…it cannot compensate for flawed data collection or sparse data.”

Page 18 – The possible radiculitis category remains a problem. The authors are trying to reclassify patients more stringently to increase the likelihood of relationship of BMP use and radicular pain. Some might suggest that leg pain alone is adequate for radiculitis and should be included.

Response:  As noted by this reviewer in the previous comment, leg pain alone is non-specific for radiculitis and includes many other unrelated conditions (such as knee pain from osteoarthritis, Baker’s cyst, and other musculoskeletal conditions). Therefore, we did not include it in our primary definition for possible radiculitis. However, as described in the response to Editor comment #10, we evaluated alternative definitions that classified leg pain with a radicular distribution as “possible radiculitis” (Table 2). Use of different definitions for “possible radiculitis” did not change the results of no association.

In the final analysis, the authors cannot make up for the absence of data to determine the relative risk and benefit of BMP versus bone graft. It is clear that the call for more randomized controlled trials for evaluation of efficacy and safety of BMP should be the final sentence of the paper.

Response:  Given the failure to demonstrate significant benefits, we do not believe that additional randomized trials in previously evaluated populations are necessarily indicated. However, we agree that additional research may be indicated for harms, and in higher risk populations in whom fusion is typically less successful. We revised the last part of the Discussion to be clearer on this point (Page 23 starting from line 542): “More research is needed to provide more reliable estimates of risk of cancer and other adverse events and to identify patient populations in which use of rhBMP-2 may be associated with greater effectiveness, such as cases where use of bone graft alone is associated with a high risk of pseudoarthrosis.
Statistical Editor Comments:

John E. Cornell is the primary statistical editor for your manuscript and prepared this review. However, the issues raised in this review represent the composite advice of Annals of Internal Medicine’s statistical team after discussion of your manuscript at one of our regular statistical meetings. Our aim is for the revision process to be as efficient as possible and to result in a published report that is methodologically robust and reported clearly. The primary statistical and physician editors for your manuscript will review the revised manuscript and may discuss it with the full team. If you have questions about any of the issues raised in our Annals’ review, please direct them to Darren Taichman at DTaichman@mail.acponline.org. Most often we can resolve these issues via e-mail communication, but in the event that initial e-mail communication does not satisfactorily address your concerns, we will arrange for members of our team to discuss the issues with you in a telephone conference.

The relative effectiveness and harms associated the use of rhBMP-2 is spinal fusion is a timely, but controversial topic. A prior review by Carragee, et al. (Spine J 2011; 11: 471-91) revealed a number of potential study design biases and inconsistencies in reporting of adverse events between published studies and data submitted to FDA. A strength of the current review is that the authors have direct access to individual-patient level data from the 17 industry (Medtronic) sponsored trials (13 randomized and 4 non-randomized trials). Access to this level of data provides the opportunity to conduct a more consistent, complete, and rigorous analysis of the effectiveness and harms of rhBMP-2 in spinal fusion. The authors also have access to 1113 MedWatch reports of adverse events associated with rhBMP-2, and they conducted a traditional literature search of MEDLINE, Embase, the Cochrane Database of Systematic Reviews, the Cochrane Central Register of Controlled Trials, Scopus, Clinicaltrials.gov, the FDA website, and manual searching of reference lists.

The authors have two primary objectives for their review: 1. Conduct an independent analysis of the effectiveness and harms of rhBMP-2, and 2. Compare their results against those presented in Medtronic’s published studies.

The first objective is more than a simple reanalysis based on the individual patient-level data from the Medtronic trials. It appears to utilize both the industry IPD data plus the eligible trials, both independent and industry sponsored studies, published since the original trials. The authors need to be clearer about these aspects of their review and the purpose that each serves in the manuscript in their articulation of the study objectives.

Response: We revised the Methods (Page 5 starting from line 143) to be clearer about the data sources and how they were used to assess each of our aims: “To achieve the two aims of our review, we used four sources of data: 1) Medtronic IPD, related protocols, and data dictionaries; 2) Medtronic internal reports; 3) summary documents from the FDA Web site, and 4) a broad-based literature search to identify a) additional studies on rhBMP-2 and b) publications related to Medtronic-sponsored studies. For aim 1, we used data from sources 1), 2) and 4 a), and for aim 2, we compared the results from journal publications with data from the other sources.”
We also revised the Methods to be clearer that we used the IPD to re-code outcomes in a consistent way (Page 7 starting from line 183): “To standardize the outcome measures, we applied consistent definitions (Appendix Table 1, available at www.annals.org) across studies and used the IPD to recode and recalculate effectiveness outcomes.”

We also revised to clarify that we searched for additional trials and other studies (Page 6 starting line 160): “For data sources 3) and 4), we searched MEDLINE® (1996 to August 2012), Embase®, the Cochrane Database of Systematic Reviews®, the Cochrane Central Register of Controlled Trials® (3rd Quarter 2012), Scopus, Clinicaltrials.gov, and the FDA website, and manually searched reference lists of relevant papers.”

Only one additional trial was found.

The second is comparable to the analyses provided in the Carragee, et al. review, though they don’t appear to utilize some of the additional data sources Carragee, et al. used: Scoliosis Research Society, the CDC, or the National Inpatient Database, all of which could provide updated data that is informative with respect to determining the adverse events associated with rhBMP-2 and their respective rates. We wonder why the authors choose to exclude these additional government and institutional databases from these analyses. Since these databases continue to collect adverse event data relevant to the safety of rhBMP-2, we recommend that the authors review these additional databases, as well.

Response: As noted in our response above to Editor comment #3: We reviewed the additional data sources mentioned in the Carragee review and the Statistical Editor comments. Data cited by Carragee from the Scoliosis Research Society did not meet our inclusion criteria because it combined results for rhBMP-2 and BMP-7 (OP-1 putty); it was also reported in a journal publication that we excluded (Williams BJ, Smith JS, Fu KM, Hamilton DK, Polly DW Jr, Ames CP, et al. Does BMP increase the incidence of perioperative complications in spinal fusion? A comparison of 55,862 cases of spinal fusion with and without BMP. Spine (Phila Pa 1976). 2011 Mar 9. [Epub ahead of print] PMID: 21394069.) Carragee incorrectly cited the Centers for Disease Control and Prevention as a data source, but this was in fact data from the Center for Devices and Radiological Health within the FDA, which we included. We included the study from the National Inpatient Database (Cahill KS, Chi JH, Day A, Claus EB. Prevalence, complications, and hospital charges associated with use of bone-morphogenetic proteins in spinal fusion procedures. JAMA. 2009 Jul 1;302(1):58-66. PMID: 19567440. ) in our review.

This is clearly a complex undertaking. The mixing of the different data sources and analyses in the manuscript is often confusing, making it difficult for the reader to digest. We need the authors to provide a more careful description and presentation of the data organized by data sources: IPD from Metronic trials, additional trials and observational studies captured through the extensive literature search, and the FDA sources. Beginning with the Methods section, the authors need to make clear to the reader which sections refer to IPD analyses, which report results based on the traditional literature review and study-level meta-analyses, and which report/summarize additional information based on review of the FDA data and documents.
The method section reads like a traditional study-level systematic review, rather than reflecting the more complex sources of information utilized in this review. This may not be the best model for this review. I suggest organizing the methods around the data sources. The subsections within each are tailored to the data sources.

IPD data source: Data Sources, Quality Assessment, Definitions for Outcomes, Rules for Attributing of Adverse Events to rhBMP-2, Selection and Coding of Covariates, Managing Missing Data, and Data Synthesis and Analysis:

Study-level sources: Search Strategy, Study Section, Data Abstraction and Quality Assessment, Data Synthesis and Analysis

FDA data (and other government or institutional) sources: Data Source, Dates, etc.

Response: We re-organized the text for the methods section as suggested by the Statistical Reviewer to the extent possible given the method used to conduct the review. To clarify, for Medtronic sponsored trials, we used protocols and internal reports provided by Medtronic to assess quality and the corresponding journal publications for these trials did not provide additional information about these trials. For data synthesis and analysis, we identified only one additional trial and used all data sources in one process.

In addition, we did not make attributions of adverse events to rhBMP-2 ourselves and used attribution made by Medtronic as designated in the Medtronic datasets. This is described in Appendix Table 2 on outcome variable definitions.

For the subsection of “Selection and Coding of Covariates”, we listed variables used to explore heterogeneity in the section of data synthesis and quantitative analysis (Page 9 line 233): “We evaluated baseline age, sex, smoking status, diabetes status, previous back surgery, and whether the patient worked before surgery as potential sources of heterogeneity.”

We have also added a section on managing missing data (Page 8 starting from line 194): “In our primary analysis of overall success and fusion success, patients meeting some of the multiple defined criteria but missing data for other criteria were classified as failures, and patients without data for any of the criteria were excluded from the analysis. We also performed two sensitivity analyses: in one, patients with missing data for some or all criteria were excluded from the analysis; in the other, patients missing data for some or all criteria were included as failures. For other binary effectiveness outcomes, patients with missing data were excluded in the primary analysis but included as failures in the sensitivity analysis. For adverse events, all patients were included since we analyzed cumulative adverse events from the time of surgery.”.

Data from documents on FDA websites did not provide additional information on effectiveness and harms beyond what we obtained from IPD and were not included to evaluate effectiveness and harms. This is clarified at the beginning of the Methods section (Page 5 line 146): “For aim 1, we used data from sources 1), 2) and 4 a)”, where FDA data represents data source 3).
Be clear about the role of each data source in your review. We need a Table that presents your definitions for the outcomes and contrast your definitions with those used in the Metronic publications. This Table may appear either in the Methods section under IPD data sources or early in the Results section where the IPD analyses are described.

Response: As discussed above, the role of each data source was clarified in the beginning part of Methods. We also added a Table (Appendix Table 1) showing definitions for the outcomes used in our IPD analysis, published trials, and Medtronic protocols.

Although the authors have IPD available, the current analytic approach aggregates this rich source of data to form a series of study-level statistics that the authors combine using traditional meta-analytic methods. This approach is disappointing and inefficient. If you have the patient-level data, you need to make maximum use of this rich data source. Use more sophisticated mixed-effect models to combine the outcomes and adverse events across trials, like you would in a multi-center clinical trial. Doing so will clearly strengthen the review beyond the information already published by Carragee, et al. While you note that few of these trials described how they handled missing data, this approach provides an opportunity to apply appropriate methods for handling missing data and provide more appropriate estimates of the treatment effects and adverse event rates from these trials.

Response: We discussed the use of mixed-effects models with the editors and the details of that discussion and our responses are attached at the end of this response letter.

We have re-analyzed the data based on the recommendations above and from the subsequent discussion. We revised the Methods section to reflect this (Starting page 9 line 220). For continuous outcomes, we used a linear mixed effects model and for common binary outcomes, we used a generalized linear mixed effects model with binomial distribution and a log link. For rare binary outcomes, we used a generalized linear fixed effects model with binomial distribution and a log link. Such an approach still makes use of all the data, and current literature supports the use of a fixed effect approach when analyzing rare events. In addition, analyses of rare adverse events did not exhibit heterogeneity among studies so a random effects model would approximate a fixed effects model.

All results on effectiveness and harms were based on the above methods except for two common binary outcomes (fusion, neurological success) at some time points, where the generalized linear mixed effects model with binomial distribution and a log link could not produce a combined estimate. In such cases, we replaced the results with those from a two-stage approach. This is explained in the Methods (Page 9 starting from line 229) “When the generalized linear model with log link could not produce a combined estimate due to ill-fitting data, we provided combined estimates from a two-step approach described elsewhere (17).” and such results were footnoted in Table 2.

Since the log link does not constrain the expected probability to be less than or equal to 1, it seemed that when the observed probability was 100% for one group, the generalized linear mixed effects model with binomial distribution and a log link could not provide a good fit of the
data. Alternative integration methods and optimization techniques, or using a different SAS procedure (PROC NLMIXED vs. PROC GLIMMIX) did not help solve the problem. Therefore, in such case for fusion and neurological success, we reported estimates based on the two-stage approach.

Cumulative adverse event rates are presented at 4 and 24 months. Since the actual rates are proportional to length of time, one would think that a Poisson (or negative binomial) model would be the best approach to estimating event rates and incidence rate ratios. Of course this depends on the availability of person-time information. If this level of data is unavailable, it would be interesting and informative to look at (perhaps graphically) the cumulative event rates for the groups reported across all the time-points represented in the data, rather than the static 4 and 24 months. While it may not be possible to do these analyses for all adverse events, it would be important to do this for general categories of adverse events and the more serious adverse events if possible.

Response: We agree that a Poisson or negative binomial model would be better when data on person-time are available. Unfortunately, we don’t have such data. We followed the editor’s suggestion and made graphs to show cumulative event rates for the two groups across all the time-points for two general categories of adverse events: having at least one adverse event and having at least one serious adverse event. They are presented as Appendix Figure 2 in the revised manuscript.

Use traditional meta-analytic methods to combine the actual study-level estimates obtained through your literature search, but exercise more care in summaries with sparse data and zero events. The method used in the current analyses uses a fixed-effects Mantel-Haenszel method with 0.5 continuity correction for these cases (Stata metaan default). I suggest that the authors use a fixed-effect logistic model based on the grouped counts within study. This approach does not require a correction for continuity, and it is easily done with the Stata glogit command, using the margins command to generate study-level estimates and 95% Confidence Intervals. This approach is more in line with recommendations made by the first author in their nicely written J Clin Epidemiol 2011; 64:1187-97 article.

Response: Thank you. We agree that combining sparse data and zero events requires more care. However, since we only found one additional trial that provided study-level data, there is no need to combine them.

You note that data reported only in abstracts was excluded from this review. It is common for reviewers to include such data in a systematic review. With all the various data sources included in this analysis, conducting a formal sensitivity analysis that includes the data from these abstracts simply adds to the complexity of the presentation. You should, however, in the Discussion section, describe the pattern of results reported in these abstracts relative to your findings.

Response: We did review abstracts to determine whether they provided information on unpublished trials and identified no additional trials of rhBMP-2 versus bone graft reported only
as abstracts or in Clinicaltrials.gov. Therefore we did not add information about abstracts in the manuscript.

Comparing and combining the results from the IPD with the additional study-level data should be based on the pooled estimates for each. This is analogous to the mixed-treatment effect meta-analysis problem. A hierarchical Bayesian model is the optimal approach, though a simpler frequentist method is acceptable and would reduce the complexity of the analyses (See Jansen Res Syn Meth 2012; 3: 177-190).

Response: We only found one additional trial that provided study-level data. Based on the recommendations from the statistical editor, we did not combine the IPD with study-level data from that one study. We only contrasted the results qualitatively (Page 14 line 333): “The additional trial (20) also found no difference in fusion rates at 24 months (ICBG 71% vs. rhBMP-2 86%; RR 1.12, 95% CI 0.98 to 1.29).”

We also suggest that you follow Carragee, et al. and use the CONSORT recommendations and calculate 90% confidence intervals for serious adverse events.

Response: We reviewed the CONSORT recommendations, which do not recommend using a 90% confidence interval for serious adverse events. We discussed this with Dr. Taichman, who suggested that we determine whether the FDA used a 90% confidence interval, which we could use to guide our confidence interval. The FDA documents report confidence intervals based on 95% confidence intervals or only report event rates without confidence intervals, so we retained with the standard 95% confidence interval.

The Results section is appropriately organized around type of procedure, but within the sections the presentation mixes information from the various sources both within and across the primary study objectives. It creates a confusing picture of the data and its meaning.

Separate out the analyses for the two-objectives: 1. Effectiveness and Harms and 2. Comparisons with Published Reports.
Move all the comparative material that contrasts the findings from your independent analyses with those presented in the various industry trials into a separate section at the end of the results section.

Response: We re-organized the Results so that Aim #1 (effectiveness and harms) is now presented separately from Aim #2 (selective reporting and related biases) - see also response to Editor comment #2.

Within the sections for each procedure, organize the presentation of the data so that the results from each data source are clearly identified. Start with the IPD analyses, followed by the study-level data, followed by the FDA (and other government and institutional data).
Response: To clarify, because there was only one additional trial without IPD data, results of effectiveness and harms were mostly based on IPD. As discussed above, results from the one additional trial were reported separately from the IPD results (Page 14 line 333). Results from the study-level data from cohort studies and intervention series were presented after the IPD results as suggested.

Data from documents on FDA websites did not provide additional information on effectiveness and harms beyond what we obtained from IPD and were not included to evaluate effectiveness and harms.

A brief opening paragraph within the section that describes each procedure that provides a description of the trials and their methodological quality, along with a succinct summary of the findings would help the reader get oriented to the data.

Response: We revised the beginning of each section to provide an overview of the trials (number and type of studies, quality or doses), and provide a summary of findings before going into more detail regarding Results (e.g., for PLF, page 14 starting from line 329): “Meta-analysis based on the IPD (n=722) provided moderate strength evidence showing no consistent difference between rhBMP-2 and ICBG in effectiveness outcomes through 24 months…”

Separate effectiveness analyses from harms analyses more clearly to minimize confusion. This is done to some extent, but the text wonders off into the comparison and contrast between the IPD and FDA data against the published studies. Separating these objectives will help clarify the presentation of the data.

Response: We revised the Results to more clearly separate analyses of effectiveness from harms (e.g., for ALIF, see Page 12 starting line 295). As discussed above, the comparison and contrast between the IPD against the published studies is presented in the section for reporting bias now.

Tables need to be reorganized to better delineate the data sources and suggested structure for the analyses.

Response: We have reorganized Table 1 and Table 2, such that one table only has the study characteristics of Medtronic sponsored studies (currently Appendix table 1), and the other table has information for overall success, fusion and Adverse events as contrasted between IPD and published studies. This latter table is used to support aim 2 of assessing reporting bias.

Table 1 needs to include study duration or duration of follow-up for each of the studies.

Response: We revised Table 1 (currently Appendix table 1) showing longest duration of follow-up for each study.

As you rework your tables, be sure to include numerator and denominator information whenever counts of events or percentages are reported.
Response: We revised throughout the Tables to include numerators and denominators (current Appendix Table 1 and table 3)

Report the number of studies and total N whenever you report summary estimates or counts.

Response: We revised the Tables to include the total number of studies and total N when reporting pooled estimates and counts (Current table 2 and table 3).

Report absolute risk differences and confidence intervals for adverse event rates. Use CONSORT guidelines when computing the confidence intervals for these events: 90% CIs for serious adverse events.

Response: For adverse events (Table 2), to follow Dr. Taichman’s suggestion, we used a generalized linear model with binomial distribution and a log link. Such model produced a RR so we reported RR in the table. For cancer, when there is a significant difference between the rhBMP-2 and the control group, we calculated both RR and absolute risk difference and reported them in the text. We retained 95% CI’s (see responses above).

Production Editor (Technical) Comments: Revision requirements
Text:

Add section heading "Data Synthesis and Analysis."

Response: We added “Data Synthesis and Analysis” to the Methods.

Please provide full corporate names, city, and state of the manufacturers of brand-name materials: Medtronic, Inc.; MAVERICK; inFUSE; Medtronic Sofamor Danek; AMPLIFY.

Response: The brand names no longer appear in the text.

There are too many tables and figures for this manuscript. Please reduce to no more than 4 tables and figures for print publication

Response: We have reduced the number of in-text figures to 1 and the number of in-text tables to 3. All of the other figures and tables are provided as Appendices available at www.annals.org

References:
Add the following to these references: reference 15-city and state of publication; reference 14-is the report number known yet?; reference 27-page numbers; references 42 through 28-city and state of publication; references 50, 71-page numbers.

Response: We have updated all the references; the report number for previous reference 14 (now reference 17) is not yet known.
Please follow these principles when citing URLs in the reference list: 1) Please cite print references rather than Web-based references (i.e., URLs) if a print version is available. 2) When citing URLs, please be sure to always include the date on which the URL was accessed. 3) If a reference is available only on the Web, please make sure that the URL is valid. References to invalid links are not acceptable. 4) If a reference is published both in print and electronically, please cite the print version. However, if you consulted an updated Web version of a print reference (e.g., a government report that was published in print in June 2005 but updated online in December 2005), please cite both the print and Web versions, along with the date each was published and the date on which the URL was accessed.

Response: Thank you. We have revised references based on these requirements.

Tables:
Define all abbreviations in footnotes to the tables.
Remove all vertical lines and all internal rule lines from your tables. Retain only horizontal lines beneath the title, beneath the column headings, and at the end of the tables.
If you have not already done so, please prepare tables by using the Tables feature in Word. Data should appear in individual cells. Please do not use tabs.

Response: All of the tables have been reformatted as requested.

Figures:
In the figure legends, define all abbreviations used in the figures.

Response: We have added figure legend text to the end of the manuscript MS word file and have defined all the abbreviations used in the figures.

Permissions:
Provide letters of permission from everyone listed in the acknowledgment section.

Response: We have provided letters of permission for everyone listed in the acknowledgements section.

Electronic Manuscript Submission:
Please submit an electronic file for the manuscript, a separate file containing all tables, and individual JPEG files for each figure. For further instructions on how to convert files created in Microsoft Office into JPEGs, please e-mail the contact person below.

Response: The files have been prepared as requested.
10. **Authors: Revised Manuscript 1 (20 March 2013)**
Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-analysis

Investigators:
Rongwei Fu, PhD
Shelley Selph, MD
Marian McDonagh, PharmD
Kimberly Peterson, MS
Arpita Tiwari, MHS
Roger Chou, MD
Mark Helfand, MD, MS, MPH

Author affiliations:
Rongwei Fu: Oregon Evidence-based Practice Center, Department of Public Health & Preventive Medicine, and Department of Emergency Medicine, Oregon Health & Science University, Portland, OR
Shelley Selph: Oregon Evidence-based Practice Center, Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University; Portland, OR
Marian McDonagh: Oregon Evidence-based Practice Center, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; Portland, OR
Kimberly Peterson: Oregon Evidence-based Practice Center, Oregon Health & Science University, Portland, OR
Arpita Tiwari: Oregon Evidence-based Practice Center, Oregon Health & Science University; Portland, OR
Roger Chou: Oregon Evidence-based Practice Center, Department of Medicine, and Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University; Portland, OR
Mark Helfand: Oregon Evidence-based Practice Center, Oregon Health & Science University; Portland VA Medical Center; Portland, OR

Running title: Effectiveness and harms of rhBMP-2 in spine fusion

Key words: rhBMP-2; reporting bias; meta-analysis of individual patient data; spinal fusion; systematic review

Corresponding author: Rongwei Fu, PhD; 3181 SW Sam Jackson Park Road, Mail Code: CSB669, Portland, OR 97239; Email: fur@ohsu.edu; Phone 503-494-6069

Text word count: 5,168 (excluding headings); Abstract word count: 395 (excluding subheadings); Figures: 1; Tables: 3; Appendix Tables: 3. Appendix figures: 2.
Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Yale University Open Access Data Project.

Contract support: Research Subcontract to Oregon Health & Science University under Sponsored Research Agreement between Yale University and Medtronic, Inc.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Rongwei Fu, PhD, Oregon Health & Science University, 3181 S.W. Sam Jackson Park Rd., Mail code CSB669, Portland, OR 97239.

Current Author Addresses:
Rongwei Fu: 3181 S.W. Sam Jackson Park Rd, Mail code CSB669, Portland, OR 97239
Shelley Selph, Marian McDonagh, Kimberly Peterson, Arpita Tiwari, Roger Chou, Mark Helfand: 3181 S.W. Sam Jackson Park Rd, Mail code BICC, Portland, OR 97239
Abstract

Background: Recombinant human bone morphogenetic protein-2 (rhBMP-2) is used as a bone graft substitute in spinal fusion, a procedure that unites (fuses) bones in the spine. The accuracy and completeness of journal publications of industry-sponsored trials on the effectiveness and harms of rhBMP-2 has been called into question.

Purpose: To independently assess the effectiveness and harms of rhBMP-2 in spinal fusion and reporting bias in industry-sponsored journal publications.

Data Sources: Individual patient data (IPD) of 17 industry-sponsored studies, related internal documents, and searches of MEDLINE (1996 to August 2012), other databases, and reference lists.

Study Selection: Randomized controlled trials (RCTs) and cohort studies of rhBMP-2 versus any control, and uncontrolled studies of harms.

Data Extraction: Effectiveness outcomes in IPD were recalcuted using consistent definitions. Details about population, study design, and results were abstracted by one investigator and confirmed by another. Two investigators independently assessed quality using predefined criteria.

Data Synthesis: Thirteen RCTs and 30 cohort studies were included. Based on meta-analysis of IPD from RCTs, rhBMP-2 and bone graft had similar effects on overall success, fusion, and other measures of effectiveness in lumbar and cervical spine fusion. The risk of any adverse events was similar between rhBMP-2 versus bone graft in lumbar fusion, though rates were high across interventions (78% to 88% at 24 months from surgery). For anterior lumbar interbody fusion, rhBMP-2 was associated with increased risk of retrograde ejaculation and urogenital problems, but the differences were not statistically significant and the confidence intervals were wide. RhBMP-2 was associated with increased risk of wound complications and dysphagia in anterior cervical spine fusion. At 24 months the overall cancer risk was increased with use of rhBMP-2 (RR 3.04; 95% CI 1.23 to 7.48), but event rates were low and cancer types were heterogeneous. The journal publications misrepresented the effectiveness and harms through selective reporting, duplicate publication, and underreporting.

Limitations: Outcome assessment was not blinded, with poor ascertainment of some harms in trials. Confounding variables were often not adjusted in cohort studies. No trials were truly independent of industry sponsorship.

Conclusion: In spinal fusion, rhBMP-2 has no proven clinical advantage over iliac crest bone graft and may be associated with important harms, making it difficult to identify clear indications for rhBMP-2. Earlier disclosure of all relevant data would have better informed clinicians and the public than the journal publications did.

Primary Funding Source: Research Subcontract to Oregon Health & Science University under Sponsored Research Agreement between Yale University and Medtronic, Inc.
The most common surgery for chronic low back pain with lumbar disc degenerative conditions is vertebral fusion (1) to restrict spinal motion and remove the presumed cause of pain. An interbody fusion, involving removal of a degenerated intervertebral disc and fusion of the adjacent vertebral bodies, can be performed via an anterior (anterior lumbar interbody fusion, ALIF), posterior (posterior lumbar interbody fusion, PLIF), or transforaminal (transforaminal lumbar interbody fusion, TLIF) approach. Fusion involving adjacent transverse processes is referred to as posterolateral lumbar fusion (PLF).

Traditionally, spinal fusions are performed using graft material harvested from the patient’s iliac crest to promote fusion. In 2002, the U.S. Food and Drug Administration (FDA) approved recombinant human bone morphogenetic protein-2 (rhBMP-2), a genetically engineered protein with bone-growth stimulating properties, as a bone graft substitute in conjunction with a device implant (LT-CAGE™) for single-level ALIF (2). In December 2003, FDA approved the use of rhBMP-2 with another implant (INTER FIX™) for similar indications (3). In clinical practice, rhBMP-2 has primarily been used “off label” in PLF and TLIF (4).

Medtronic, Inc. (Medtronic) is the sole manufacturer of devices involving rhBMP-2 for spinal fusion. Publications of Medtronic-sponsored trials before 2009 reported beneficial effects of rhBMP-2 in spinal fusion with either no or few adverse events and “no unanticipated device-related adverse events” (5-8). However, observational studies subsequently reported serious complications associated with rhBMP-2 in cervical spine fusion, (9-12) and FDA documents summarizing Medtronic-sponsored trials appeared to indicate substantially more adverse events than reported in the journal publications (13). In 2008, the FDA issued a public health notification of life-threatening complications associated with off-label use of rhBMP-2 in
cervical spine fusion—swelling of the neck and throat resulting in compression of the airway and other structures (14).

Selective reporting or underreporting of outcomes in journal publications could have led to misleading conclusions about the balance of benefits and harms of rhBMP-2 (13, 15). Our study aimed to 1) estimate effectiveness and harms of rhBMP-2 in spinal fusion in a systematic review, using individual patient data (IPD) when available and 2) assess reporting biases in published articles of industry-sponsored studies.

Methods
A short version of the review protocol was registered at the PROSPERO international prospective register of systematic reviews (16) on February 23, 2012, and the full protocol was deposited with the Yale University Open Data Access (YODA) Project at the same time. Detailed methods and additional analyses are available elsewhere (17).

Data Sources
To achieve the two aims of our review, we used four sources of data: 1) Medtronic IPD, related protocols, and data dictionaries; 2) Medtronic internal reports; 3) summary documents from the FDA Web site, and 4) a broad-based literature search to identify a) additional studies on rhBMP-2 and b) publications related to Medtronic-sponsored studies. For aim 1, we used data from sources 1), 2) and 4 a), and for aim 2, we compared the results from journal publications with data from the other sources.
**Patient-level data and Medtronic internal reports**

For data sources 1) and 2), the YODA Project provided de-identified patient-level data, protocols, data dictionaries, and internal reports submitted to the FDA for all 17 Medtronic-funded studies of rhBMP-2 in spinal fusion that were completed or terminated by December, 2011. The Medtronic internal reports for the studies included summaries of results on study outcomes as well as brief adverse event case histories. We also received 1,229 MedWatch reports of adverse events concerning rhBMP-2 and other documents summarizing features of the trials.

**Additional data sources**

For data sources 3) and 4), we searched MEDLINE® (1996 to August 2012), Embase®, the Cochrane Database of Systematic Reviews®, the Cochrane Central Register of Controlled Trials® (3rd Quarter 2012), Scopus, Clinicaltrials.gov, and the FDA website, and manually searched reference lists of relevant papers.

Two reviewers independently assessed each study to determine inclusion eligibility. We included controlled clinical trials and cohort studies that evaluated effectiveness or harms of rhBMP-2 in spinal fusion. For harms, we also included uncontrolled intervention series. We excluded studies that combined results of rhBMP-2 with other bone morphogenetic proteins, unless rhBMP-2 was the predominant bone morphogenetic protein used.
Data Abstraction and Quality Rating

For all data sources, one investigator abstracted patient and study characteristics and results, and a second investigator reviewed data abstraction for accuracy. For Medtronic-funded studies, quality assessment was based on information from trial protocols and internal reports. Two investigators independently rated the quality of controlled trials and cohort studies as good, fair, or poor, using criteria adapted from the Cochrane Back Review Group (18) and the US Preventive Services Task Force (19). Discrepancies were resolved through consensus.

Definitions and calculations of endpoints for individual patient data

We used the study protocols and ClinicalTrials.gov entries to determine pre-specified primary outcomes. In nine studies, the primary effectiveness measure was “overall success” (at 24 months) and fusion was the primary endpoint in the remaining studies. Other effectiveness outcomes included pain, disability, neurologic status, function, and return to work. Studies differed slightly in how they specifically defined effectiveness outcomes. To standardize the outcome measures, we applied consistent definitions (Appendix Table 1, available at www.annals.org) across studies and used the IPD to recode and recalculate effectiveness outcomes.

We obtained overall and specific adverse events directly from IPD (no recalculation) except for urinary retention, wound infection, wound dehiscence, and possible lumbar radiculitis, which we identified by reviewing case histories in internal reports. Lumbar radiculitis was not a pre-specified outcome in the trials or the case histories, and we applied four alternative definitions of radiculitis (Appendix Table 1, available at www.annals.org) in primary and sensitivity analyses.
Management of missing data

In our primary analysis of overall success and fusion success, patients meeting some of the multiple defined criteria but missing data for other criteria were classified as failures, and patients without data for any of the criteria were excluded from the analysis. We also performed two sensitivity analyses: in one, patients with missing data for some or all criteria were excluded from the analysis; in the other, patients missing data for some or all criteria were included as failures. For other binary effectiveness outcomes, patients with missing data were excluded in the primary analysis but included as failures in the sensitivity analysis. For adverse events, all patients were included since we analyzed cumulative adverse events from the time of surgery.

Data Synthesis and Analysis

Aim #1: Effectiveness and harms of rhBMP-2

Based on input from clinical experts and similarity among studies, we stratified our analyses by spinal area (lumbar, cervical) and surgical approach (e.g., ALIF, PLF) for all outcomes except cancer and death, for which we combined trials of all surgical approaches because these outcomes were rare and not necessarily affected by the specific surgical technique. Only the ALIF and PLF trials provided sufficient data for meta-analyses. Meta-analyses were based on IPD from Medtronic-sponsored trials. From the broad search we identified only one additional trial without corresponding IPD (20) and qualitatively compared its results with IPD results.

For effectiveness endpoints, we calculated outcomes at 6 weeks and at 3, 6, 12, and 24 months after surgery, based on the time points typically evaluated in the trials. For harms, we
aggregated data into two periods: 1) operative and up to four weeks post-operative, and 2) up to 24 months post-operative. Data beyond 24 months were sparse and are reported elsewhere (17), except for cancer and death, and for these outcomes, we also analyzed the cumulative number of events up to 48 months.

We used mixed effects models to combine IPD. For continuous outcomes, a linear mixed effects model was used to obtain a combined mean difference between rhBMP-2 and control groups after adjusting for baseline values and individual study effects (21). We assumed random treatment effects and heterogeneous residual variance across included studies. For common binary outcomes, a generalized linear mixed effects model assuming random treatment effects and binomial distribution with log link was used to obtain a combined risk ratio (RR). For rare binary outcomes, we used a generalized linear fixed effects model assuming binomial distribution with log link. We also performed a sensitivity analysis for cancer by including all zero event trials in the meta-analysis as a combined “pseudo-trial” with an assumption of no cancers in the rhBMP-2 group and one cancer in the control group. When the generalized linear model with log link could not produce a combined estimate due to ill-fitting data, we provided combined estimates from a two-step approach described elsewhere (17).

We assessed statistical heterogeneity based on the estimated between-study variance from the mixed effects model (21). We evaluated baseline age, sex, smoking status, diabetes status, previous back surgery, and whether the patient worked before surgery as potential sources of heterogeneity. We also performed sensitivity analyses by excluding poor quality studies and studies that utilized a lower rhBMP-2 concentration, or by excluding graft-site-related adverse events in analyses of harms. For cancer, we performed sensitivity analyses by excluding events not reportable to the National Cancer Institute Surveillance Epidemiology and End Results Program.
(SEER) Program (skin cancers with low propensity to metastasize). Results of sensitivity analyses were generally similar and, except for cancer and possible lumbar radiculitis, are not reported separately. IPD meta-analyses were performed using SAS® software 9.2 (SAS Institute Inc., Cary, NC, USA).

We rated the strength of evidence by outcomes based on the aggregate risk of bias, consistency, directness, and precision of the evidence (22).

**Aim #2: Assessment of reporting and related biases**

We assessed publication and outcome reporting biases and quality of reporting (23) by comparing journal publications with corresponding study protocols, reports, and data dictionaries provided by Medtronic. We used a previously published protocol for identifying and categorizing potential sources of reporting bias (24, 25).

**Role of the Funder**

The YODA Project proposed the aims for the review, served as the intermediary for our requests for additional information or data from Medtronic, and provided comments on our initial draft report. Neither the YODA Project nor Medtronic influenced the conduct of our analyses or the publication of our findings.
Results

Aim #1: Benefits and Harms of rhBMP-2

Overview of included studies

We included 13 randomized controlled trials (RCTs), 12 sponsored by Medtronic (n=1,879) and one by Norton Healthcare (n=106) (20) (Appendix Figure 1, available at www.annals.org). All RCTs compared rhBMP-2 with iliac crest bone graft (ICBG) except for Study 10, which compared rhBMP-2 versus artificial disc replacement. We excluded one very small (n=3) Medtronic trial.

The Medtronic studies applied similar eligibility criteria and enrolled similar populations across studies within each surgical approach (Appendix Table 2 available at www.annals.org). Eight studies enrolled fewer than 100 subjects (sample sizes ranged 14 to 85). At 24 months, 9 of the 12 randomized trials had follow-up rates over 90% in both groups.

While there were some important baseline differences, we did not detect a consistent pattern favoring rhBMP-2. The main risks for bias were lack of blinding of surgeons, patients, and outcome assessors (except for radiologic endpoints). The quality of ascertainment varied for different outcomes. Outcomes related to potential effectiveness (e.g., pain, function, fusion) were generally ascertained with well-designed questionnaires or scales. For harms, the studies used broad classifications for many adverse events, and events were generally not actively elicited using specific symptom questionnaires or objective tests. For example, for retrograde ejaculation, a condition which may not be volunteered, it was unclear whether investigators asked about symptoms or how the outcome was defined. No trial defined radiculitis, and adverse events consistent with possible radiculitis were variously classified even within the same trial as back and leg pain, neurological events, or spinal events. Cancer was not considered in the protocols as
a pre-specified endpoint; it was only captured by voluntary reporting and was potentially under-ascertained. Very little information was available about local effects, such as inflammation, heterotopic bone formation, or osteolysis.

We also identified 30 cohort studies comparing rhBMP-2 with autograft and/or allograft and 47 intervention series, including four Medtronic prospective intervention series (none fully published) (17) and 34 case series or case reports of patients who received rhBMP-2 in spinal surgery. Most of these studies were retrospective, small, provided little information on patient characteristics, and evaluated off-label indications. Most of the cohort studies had baseline differences between groups or did not report baseline characteristics, had unclear blinding of outcome assessors, and failed to adjust for potential confounding variables and baseline differences.

Anterior lumbar interbody fusion (ALIF)

Five Medtronic-sponsored trials with individual patient data (four fair quality—Studies 2, 4, 5, and 9—and one poor quality—Study 1) evaluated rhBMP-2 vs. ICBG in ALIF. Studies 1, 2, and 9 used rhBMP-2 with either the LT-Cage or the INTER FIX device, and Studies 4 and 5 used rhBMP-2 off-label with bone dowels. Study 5 was terminated early when less than half of the projected sample (n = 180) were enrolled.

The five RCTs (n=465) provided moderately strong evidence showing no consistent differences between rhBMP-2 and ICBG in overall success, fusion rates, or other effectiveness measures from the immediate postoperative period through 24 months (Table 1). An exception was a 3-point higher SF-36 physical component summary score in the rh-BMP-2 group at 3, 6,
12, and 24 months. At 24 months, fusion rates ranged from 60% to 100% and the average overall success rate was about 60% for the rhBMP-2 group and 50% for the ICBG group.

Adverse events were common, with 38% of rhBMP-2 patients and 45% of ICBG experiencing at least one adverse event through 4 weeks postoperatively, and 80% in each group through 24 months (Table 2 and Appendix Figure 2, available at www.annals.org). Meta-analysis showed no significant differences between rhBMP-2 and ICBG groups for any specific adverse event, though estimates were frequently imprecise, precluding strong conclusions (Table 2). For retrograde ejaculation, subsidence (defined as sinking or settling of the device into bone), and urogenital problems, risk estimates favored ICBG but the differences were not statistically significant and confidence intervals were wide. There was no difference in the risk of possible lumbar radiculitis between the rhBMP-2 and ICBG groups, based on several alternative definitions.

Three cohort studies (26-28) and four intervention series (29-32) also evaluated rhBMP-2 in ALIF. One cohort study found rhBMP-2 associated with increased risk of retrograde ejaculation (7% of 69 patients vs. 1% of 174, \( P = 0.0025 \)) (27) and two cohort studies found rhBMP-2 associated with increased risk of subsidence (30, 33), but there was variation in subsidence definition and measurement, and studies did not adjust for confounders.

**Posterolateral fusion (PLF) in the lumbar spine**

Four Medtronic-sponsored randomized trials with IPD (three fair quality—Studies 8, 13, and 14, and one poor quality—Study 12) evaluated rhBMP-2 for PLF. Studies 12, 13, and 14 used a higher dose and concentration of rhBMP-2 than used in ALIF trials. Study-level data was
available for one trial without IPD (20); it restricted enrollment to patients over 60 years old, and did not report the rhBMP-2 dose used.

Meta-analysis based on the IPD (n=722) provided moderate strength evidence showing no consistent difference between rhBMP-2 and ICBG in effectiveness outcomes through 24 months (Table 1). The fusion rate at 24 months ranged from 70% to 90% in the ICBG group and 86% to 93% in the rhBMP-2 group; the rate of overall success ranged from 40% to 60% in both groups. The additional trial (20) also found no difference in fusion rates at 24 months (ICBG 71% vs. rhBMP-2 86%; RR 1.12, 95% CI 0.98 to 1.29).

For harms, there was no difference between rhBMP-2 versus ICBG in the risk of experiencing at least one adverse event, one serious adverse event, or one device-related adverse event, at four weeks and 24 months in the trials (Table 2 and Appendix Figure 2, available at www.annals.org). There was no difference in the risk of possible radiculitis up to 4 weeks or 24 months using our primary or alternative definitions. We also found similar rates for other specific harms, except for back and leg pain up to 4 weeks (Table 2); however, back and leg pain events were very heterogeneous (e.g., radiculopathy, Baker’s cyst, sacroiliac joint pain, arthritic knee pain, or ankle pain) and included events unlikely to be related to fusion surgery.

Results from cohort studies (34-41) and intervention series (42-48) appeared consistent with the randomized trials, though few studies (34, 39, 40, 44) reported specific adverse events.

Other lumbar spine fusion techniques

We were not able to reach any definitive conclusion regarding the comparative effectiveness or harms of rhBMP-2 for other lumbar fusion techniques. Except for one small
Cervical spine fusion

For anterior cervical spine fusion, Study 7, a small (n=33), fair-quality Medtronic trial with IPD, provided low strength evidence showing no difference between rhBMP-2 and ICBG in effectiveness endpoints. Three cohort studies also found no clear differences in effectiveness (11, 12, 33).

In Study 7, rhBMP-2 was associated with greater risk of adverse events than ICBG at 24 months (45 adverse events in 18 patients vs. 13 adverse events in 15 patients; RR 2.88, 95% CI 1.30 to 6.41). A large fair quality cohort study (n=27,067) using the National Inpatient Sample provided moderate strength evidence that rhBMP-2 was associated with increased risk of complications (OR 1.43, 95% CI 1.12 to 1.70), dysphagia/dysphonia (OR 1.63; 95% CI 1.30 to 2.05), and wound complications (OR 1.67, 95% CI 1.10 to 2.53) (9). Smaller cohort studies (total n=346) were consistent with these results (10-12). Intervention series reported 5% to 60% of patients developed dysphagia, with some variability attributable to how dysphagia was defined (49-53).

There were no controlled trials of rhBMP-2 in posterior cervical spine fusion. In the National Inpatient Sample study, posterior cervical spine fusion with or without rhBMP-2 (n = 2,869) was associated with similar risks of overall complications and wound complications (9).

Cancer and death

Five Medtronic-sponsored trials with IPD (Studies 2, 4, 5, 10, 14) reported at least one cancer through 24 months and were included in our meta-analysis. Detailed information about cancer events is shown in Appendix Table 3, available at www.annals.org). Compared with
ICBG, rhBMP-2 was associated with increased risk of cancer (RR 3.45, 95% CI 1.34 to 8.85 and absolute difference 1.9 percentage points, 95% CI 0.5 to 3.2), with a number needed to harm of 53 (95% CI 31 to 200) (Figure 1). Data were insufficient to determine the effect of rhBMP-2 dose on estimates of cancer risk. Although 10 of 17 cancers with rhBMP-2 occurred in the largest high-dose trial (n=239, Study 14), another high-dose study (Study 13) reported no cancers with rhBMP-2 (n=98). The increased risk remained significant when we performed a sensitivity analysis to account for the seven (n=14 to 197, total sample size 429) zero-event Medtronic-sponsored trials (RR 2.90, 95% CI 1.19 to 7.08; absolute difference 1.3 percentage points, 95% CI 0.2 to 2.4, for number needed to harm of 77, 95% CI 42 to 500). At 48 months, the increased risk was no longer statistically significant (4 studies; RR 1.82; 95% CI 0.84 to 3.95). Overall, the event rates and the strength of evidence are low.

Estimates were similar at 24 months (RR 2.92; 95% CI 1.10 to 7.72) and through 48 months (RR 1.92; 95% CI 0.86 to 4.32) when excluding non-SEER cancers. One cohort study (54) of 125 patients (24 rhBMP-2, 101 ICBG) reported a similar increased risk of cancer, though the difference was not statistically significant (RR 2.10; 95% CI 0.69 to 6.41).

There was no difference between rhBMP-2 and ICBG in risk of death through 24 months (9 trials, RR 0.71; 95% CI 0.32 to 1.55—Studies 2, 4, 6-10, 13-14) or 48 months (4 trials, RR 0.66; 95% CI 0.28 to 1.56—Studies 4, 10, 13-14).

Aim #2: Quality of Reporting in Published Trials

On-label use in anterior lumbar interbody fusion (ALIF)

In 2002, the FDA approved the use of rhBMP-2 with the LT-Cage in ALIF on the basis of three premarketing studies (2). By 2004, at least 11 journal articles reporting results from
these three studies had been published in major orthopedic journals (6, 55-63). Comparison with internal documents and our meta-analysis indicated that, through selective outcome reporting, selective analysis, duplicate publication, and reframing, the publications implied inaccurately that the device was superior to the autograft control.

The primary publications of the pivotal trials did not report the primary endpoint, overall success at 24 months, which was in the range of 50% to 60% for both groups (Table 3, Studies 2 and 3 (6, 56, 59). Instead, publications emphasized other outcomes that appeared to favor rhBMP-2. For example, the internal summary for Study 2 [Medtronic, #10] concluded “[s]tatistical superiority cannot be claimed”, but the corresponding publication reiterated the fusion rates (94.5% vs. 88.7%) in the abstract, results, and conclusion sections while failing to mention that the difference was not statistically significant. (6).

In 2001, Kleeman and colleagues described the results for one site in Study 3 (22 of the 137 subjects), stating that 100% of these subjects had a successful fusion and “improvement in back pain, leg pain, and function” (58), which did not represent the overall results for this study (Table 3, Study 3). Other Medtronic-supported articles (5, 8, 57, 59, 62, 65, 69) cited Kleeman’s article instead of the full results when referring to Study 3.

In 2003, Burkus and colleagues published a post hoc “integrated analysis” of Studies 2 and 3 that promoted the idea that rhBMP-2 would have superior fusion rates relative to ICBG with sufficient sample size (59). For “controls,” the authors combined the control arm of Medtronic Study 2 (open surgery with the LT-CAGE) with a group of 266 patients from an older, completely unrelated, unpublished series of patients who underwent laparoscopic surgery with the LT-CAGE (59). While we did not receive data from this earlier study, available information makes clear that it was inappropriate to construct the control group in this way. In internal documents, but not in publications, Medtronic noted that surgeons in the earlier study
were likely less skilled, evidenced by longer operative times, higher blood loss, and longer hospital stays (70). Because it was already known that the older series had worse fusion outcomes than would be expected in more recent studies, the combined analysis appeared to “stack the deck” in favor of rhBMP-2 over ICBG. In 2004, the investigators republished this integrated analysis in two other journals (57, 62), misrepresenting the results as “proof” of “statistically superior fusion rates and statistically lower pain scores” (62). The IPD analysis also confirmed the findings of previous reviews (13, 15) that adverse events for on-label use were dramatically underreported in both rhBMP-2 and ICBG groups (Table 3, Studies 1-3).

Off-label use in anterior lumbar interbody fusion (ALIF)

Between 2002 and 2006, the use of rhBMP-2 in the U.S. increased from 0.7% to 25% in all spinal fusions (9). In over 80% of these procedures, rhBMP-2 was used in posterolateral fusions; posterior or transforaminal interbody fusions, or other off-label applications (4). Three pilot studies of rhBMP-2 in off-label cervical (5) and lumbar applications (7, 66) were published in 2002. These articles reported selective effectiveness endpoints and no adverse event endpoints for the Medtronic trials (Table 3, Studies 4, 7, 12). In one of these, Burkus and colleagues reported that 24 out of 24 patients (100%) receiving rhBMP-2 achieved fusion at 24 months compared with 13 out of 19 in the control group (68%) (Table 3, Study 4) (7, 57). However, a larger, pivotal bone dowel trial was terminated early. Instead of publishing these results separately, the investigators combined the results of Studies 4 and 5, misrepresented the two separate trials as “a two-part, prospective, randomized, multicenter study” with “two sequential phases,” and reported that “[f]usion rates were significantly better in the study group (p < 0.001)” without mentioning early termination (65, 71).
**Off-label use in posterior lumbar interbody fusion (PLIF)**

At the time of approval of rhBMP-2 for use in ALIF with the LT-CAGE, Medtronic and the FDA knew that the use of rhBMP-2 in other procedures could present safety challenges (63, 72). In December, 1999, Medtronic suspended enrollment in Study 6, a randomized trial of rhBMP-2 in a posterior interbody fusion, because of concerns about ectopic bone formation in two patients (8), which could lead to radiculopathy due to nerve root impingement.

Medtronic followed the 67 patients who had already been enrolled out to 24 months. In March 2002 Medtronic requested FDA permission to terminate the study. The same year, Medtronic Sofamor Danek Inc. sponsored a supplement in the journal *Spine*, in which review articles were published along with conclusions from an “international panel of experts” that included outside experts, investigators associated with Medtronic, and Medtronic employees. Two articles in the supplement, one authored by the Medtronic vice president for research, discussed the concern about ectopic bone formation in Study 6. While noting that large randomized trials were needed to establish the safety of rhBMP-2 in the off-label procedures, the supplement argued that ectopic bone formation, and complications it might cause, were due to poor technique (61, 63). No data from Study 6 were presented that would support this argument. The international panel finding was that “[w]hen used properly, BMPs currently appear to be extremely safe for spine fusion” (73).

After Study 6 was terminated, an article was published in 2004 (Table 3, Study 6) (8) stating, for the first time, the numbers of patients that had ectopic bone formation (24/34 rhBMP-2 patients vs. 4/33 ICBG patients, p<0.001). However, despite the small size of the study, the authors emphasized the importance of small, insignificant differences in some effectiveness
measures and the lack of association between ectopic bone formation and leg pain, and they
gave an incomplete account of the reasons the study was terminated (13, 74).

Off-label use in posterolateral lumbar fusion (PLF)

There were six off-label studies of rhBMP-2 in PLF—three published (Studies 8, 12, 14),
two unpublished (Studies 13, 15), and one study with partial data presented (Study 16). As with
other studies, adverse events were under-reported in both groups, except in Study 14, the most
recently published and largest trial (Table 3). Only Study 8 reported overall success, though it
was the primary effectiveness endpoint in Studies 8 and 14.

Based on IPD from Study 14 (68), rhBMP-2 and ICBG did not differ in rates of overall
success (56% vs. 56%) and fusion (90% vs. 90%). In contrast, the journal publication and FDA
summary reported that use of rhBMP-2 resulted in a higher fusion rate (96% vs. 89%, \( P=0.014 \))
(68, 75). The reason for this difference may be that we classified patients with partial data as
failure for fusion, though it is not clear why this would differentially affect the rhBMP-2 group.

Discussion

In spinal fusion, rhBMP-2 and ICBG appear to be similarly effective when used in ALIF
and PLF; however, the current evidence does not allow definitive conclusions regarding the
effectiveness for other surgical approaches. For just one outcome measure, SF-36 physical
component summary, the scores were slightly better with rhBMP-2 than with ICBG in patients
undergoing ALIF through 24 months, but the difference was only 2 to 3 points on a 0-100 point
scale, failing to meet typical criteria for a clinically meaningful difference (76).
The use of rhBMP-2 in anterior cervical spine fusion was associated with statistically significant increases in overall adverse events, wound complications, and dysphagia or dysphonia. For lumbar fusion—both on-label and off-label—adverse events were common in both rhBMP-2 and ICBG patients and the risk of any adverse events was similar. The full data raise concerns over increased risk of retrograde ejaculation, urinary retention, osteolysis/instability, subsidence, and heterotopic bone formation. However, with the exceptions of cancer risk and back and leg pain, we could not draw definitive conclusions regarding specific harms because the studies were small and the quality of harm ascertainment was often poor. Our analysis makes it evident that more definitive evidence about harms was needed before rhBMP-2 became widely used.

We found an increased risk of cancer associated with the use of rhBMP-2 with lumbar or cervical fusion surgery at 24 months, with or without inclusion of non-SEER-reportable cancers. However, this finding should be interpreted with caution; cancers in the meta-analysis were heterogeneous and cancer may have been under-reported. Potential mechanisms for these adverse events may be related to pro-inflammatory or carcinogenic effects of rhBMP-2 (63).

In their review, Carragee et al. demonstrated that adverse events were underreported in publications of five trials for which the FDA had made full results public (13). Our study confirms this finding and also demonstrates that under-reporting was also pervasive in trials of off-label uses for which results previously had not been available to the public. Other reviews generally found rhBMP-2 and ICBG associated with similar benefits for ALIF and PLIF based on low or moderate strength evidence, but lacked data to evaluate harms. Ratko et al. noted that the absence of reported harms could be due to non-reporting by investigators (15).
Meta-analysis of individual patient data offers several important advantages over traditional, study-level meta-analysis (77). Compared with other reviews, (15, 78) we had a more complete and standardized evaluation of outcomes, with data from unpublished studies as well as data unreported or incompletely reported by published studies, thus reducing potential effects of publication and reporting bias. Additionally, outcome measures could be recalculated or recategorized using consistent definitions, potential baseline imbalances adjusted for using patient-level information, and multiple sensitivity analyses performed to handle missing data.

On the other hand, meta-analysis based on IPD requires substantially more time and resources than meta-analysis based on study-level data, and it cannot compensate for flawed data collection or sparse data. Even with IPD on 1,879 patients from 12 trials, one additional trial, and many observational studies, the evidence base remains relatively small within each surgical approach. We found no published trials truly independent of the manufacturer. Additionally, there have been no prospective, well-designed, adequately-powered studies specifically aimed to systematically assess a number of important harms using adequate ascertainment methods.

There was also insufficient information to adequately evaluate the effects of dose on risk of harms. Eleven Medtronic studies (Studies 1-11) used rhBMP-2 at a concentration of 1.5mg/mL, with total doses ranging from 0.6-16.8 mg. Higher and unapproved concentrations of rhBMP-2 (2.0-3.0 mg/mL) were used in five of the six PLF studies, with total doses ranging from 15.0-63.0 mg. Determining the effects of higher versus lower concentrations and doses of rhBMP-2 was not possible due to differences in surgical approach, rhBMP-2 carrier, and fusion hardware, as well as the small number of studies and small sample sizes.

Journal practices, as well as those of the manufacturer, contributed to an incomplete and sometimes misleading evidence base. Although we had unusual access to protocols and
documents submitted by the manufacturer to the FDA, other information, such as operative notes
and internal correspondence, might have been helpful in assessing the extent of design and
reporting bias. Internal correspondence is essential to evaluate selective analysis reporting,
ghostwriting, time lag bias, and misrepresentation of facts (25). Finally, availability of detailed
case report forms would have enabled us to better evaluate the integrity of adverse event
adjudication.

In conclusion, we found substantial evidence of reporting bias and no evidence that rhBMP-2
is more effective than ICBG in spinal fusion, with some evidence of an association with important
harms. More research is needed to provide more reliable estimates of risk of cancer and other adverse
events and to identify patient populations in which use of rhBMP-2 may be associated with greater
effectiveness, such as cases where use of bone graft alone is associated with a high risk of
pseudoarthrosis. Based on the currently available evidence, it is difficult to identify clear indications
for rhBMP-2 in spinal fusion.
References


49. Lanman TH, Hopkins TJ. Early findings in a pilot study of anterior cervical interbody fusion in which recombinant human bone morphogenetic protein-2 was used with poly(L-lactide-co-D,L-lactide) bioabsorbable implants. Neurosurg Focus. 2004;16(3):E6. [PMID: 15198494]


70. **Medtronic, Inc. II B: Report of pivotal clinical trial results (G960065) laparoscopic use of InFUSE bone graft/LT-CAGE lumbar tapered fusion device. Internal document:2072-423.


72. **Medtronic, Inc. Summary Information on Medtronic Clinical Trials. [2011].


Acknowledgments: The authors acknowledge Robin Paynter, MLIS, for conducting literature searches; Howard Balshem, MS, Susan Carson, MPH, Elaine Graham, MLS, Allison Lowe, BA, Edwin Reig, MS, MAT, Katie Reitel, MPH, MSW, Sujata Thakurta, MA, Ngoc Wasson, MPH, and Leah Williams, BS, for their contributions to this article; all are affiliated with Oregon Health & Science University.

Figure Legends

Figure 1. Comparison of Cancer Risk Between rhBMP-2 Versus Control
Shows a forest plot of the comparison of cancer risk in 5 studies at 24 months and 4 studies at 48 months.

Figure footnote:
*The combined risk ratio (RR) was obtained using a generalized fixed effects model with binomial distribution and log link without correction for zero events. The RR from each study was estimated, when there is zero event, by adding a continuity correction of 0.5, for illustrative purposes.

Appendix Figure 1. Literature Flow Chart.
Shows the literature retrieved and the numbers of abstract records and full-text articles excluded and included.

Figure footnote:
ALIF = anterior lumbar interbody fusion, PLF = posterolateral lumbar fusion, RCT = randomized controlled trial

Appendix Figure 2. Cumulative Proportion of Patients With At Least One Adverse Event: (a) ALIF and (b) PLF.
Shows the cumulative proportion of patients with at least one adverse event for ALIF and PLF at time points from operation.
Figure footnote:
AE = adverse event; ALIF = anterior lumbar interbody fusion; ICBG = iliac crest bone graft; PLF = posterolateral lumbar fusion; rhBMP-2 = recombinant human bone morphogenetic protein-2; SAE = serious adverse event.
*There was no significant difference between the rhBMP-2 versus ICBG groups at any time point for either outcome and surgery approach.
Table 1. Effectiveness Endpoints for ALIF and PLF With rhBMP-2 Versus ICBG

**Anterior lumbar interbody fusion**

<table>
<thead>
<tr>
<th>Endpoint (Scale)</th>
<th>6 weeks</th>
<th>3 months</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk ratio (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>----</td>
<td>----</td>
<td>1.18 (0.93 to 1.50)</td>
<td>1.12 (0.95 to 1.33)</td>
<td>1.19 (0.99 to 1.42)</td>
</tr>
<tr>
<td>Fusion</td>
<td>----</td>
<td>----</td>
<td><strong>1.10 (1.02 to 1.19)</strong> †</td>
<td>1.09 (0.95 to 1.24) †</td>
<td>1.05 (0.88 to 1.24) †</td>
</tr>
<tr>
<td>Neurological success</td>
<td>1.02 (0.93 to 1.13) 434 (4)</td>
<td>1.06 (0.97 to 1.16) 442 (4)</td>
<td>1.01 (0.91 to 1.12) 433 (4)</td>
<td>1.04 (0.94 to 1.14) 420 (4)</td>
<td>1.08 (0.98 to 1.19) 400 (4)</td>
</tr>
<tr>
<td>ODI success</td>
<td>1.04 (0.83 to 1.29) 442 (4)</td>
<td>1.03 (0.87 to 1.23) 455 (5)</td>
<td>1.09 (0.95 to 1.25) 450 (5)</td>
<td>1.03 (0.92 to 1.15) 436 (5)</td>
<td>1.10 (0.97 to 1.24) 417 (5)</td>
</tr>
<tr>
<td>Return to work‡</td>
<td>1.21 (0.71 to 2.05) 211 (4)</td>
<td>0.97 (0.70 to 1.32) 210 (4)</td>
<td>1.02 (0.89 to 1.17) 207 (4)</td>
<td>1.01 (0.90 to 1.14) 201 (4)</td>
<td>1.06 (0.94 to 1.19) 196 (4)</td>
</tr>
</tbody>
</table>

<p>| Weighted mean difference (95% CI) | Sample Size, n (Studies) | |
| ODI (0-50)§ | -2.36 (-6.91 to 2.19) 444 (4) | -5.05 (-10.21, 0.10) 461 (5) | -3.79 (-8.69 to 1.11) 456 (5) | -3.74 (-9.09 to 1.60) 441 (5) | <strong>-7.35 (14.00 to -0.70)</strong> 423 (5) |
| Back pain (0-10)§ | 0.21 (-0.28 to 0.71) 443 (4) | <strong>-0.57 (-1.06 to -0.09)</strong> 446 (4) | -0.36 (-0.94 to 0.22) 442 (4) | -0.51 (-1.18 to 0.16) 426 (4) | <strong>-0.74 (-1.49 to 0.00)</strong> 409 (4) |
| Leg pain (0-10)§ | <strong>-0.57 (-1.12 to -0.02)</strong> 443 (4) | -0.37 (-1.02 to 0.27) 446 (4) | -0.20 (-0.72 to 0.32) 442 (4) | -0.49 (-1.07 to 0.08) 426 (4) | -0.60 (-1.28 to 0.08) 409 (4) |
| SF-36® PCS (0-100)‖ | 0.55 (-1.02 to 2.11) 356 (3) | <strong>2.91 (0.28 to 5.53)</strong> 374 (4) | <strong>3.00 (0.69 to 5.31)</strong> 449 (5) | <strong>2.94 (0.85 to 5.03)</strong> 440 (5) | <strong>3.68 (0.86 to 6.49)</strong> 421 (5) |
| SF-36® MCS (0-100)‖ | -0.36 (-2.45 to 1.73) 356 (3) | 0.74 (-1.34 to 2.83) 374 (4) | -0.33 (-2.24 to 1.59) 449 (5) | -0.56 (-2.60 to 1.48) 440 (5) | 2.90 (-0.29 to 6.08) 421 (5) |</p>
<table>
<thead>
<tr>
<th>Endpoint (Scale)</th>
<th>6 weeks</th>
<th>3 months</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk Ratio (95% CI)</strong>&lt;br&gt;Sample Size, n (Studies)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>----</td>
<td>----</td>
<td>1.34 (1.10 to 1.64)</td>
<td>1.07 (0.93 to 1.25)</td>
<td>1.05 (0.91 to 1.21)</td>
</tr>
<tr>
<td>Fusion</td>
<td>----</td>
<td>----</td>
<td>1.37 (1.19 to 1.59)</td>
<td>1.29 (0.94 to 1.78)‡</td>
<td>1.16 (0.96 to 1.41)‡</td>
</tr>
<tr>
<td>Neurological success</td>
<td>1.03 (0.94 to 1.13)</td>
<td>1.0 (0.93 to 1.08)</td>
<td>1.02 (0.96 to 1.09) †</td>
<td>1.01 (0.95 to 1.07)</td>
<td>1.01 (0.92 to 1.10)</td>
</tr>
<tr>
<td>ODI success</td>
<td>1.00 (0.81 to 1.23)</td>
<td>1.03 (0.91 to 1.17)</td>
<td>1.07 (0.98 to 1.17)</td>
<td>1.01 (0.91 to 1.11)</td>
<td>1.01 (0.91 to 1.12)</td>
</tr>
<tr>
<td>Return to work‡</td>
<td>1.26 (0.71 to 2.21)</td>
<td>1.09 (0.85 to 1.40)</td>
<td>0.87 (0.67 to 1.14)</td>
<td>1.07 (0.96 to 1.19)</td>
<td>1.03 (0.94 to 1.14)</td>
</tr>
</tbody>
</table>

| Weighted mean difference (95% CI)**Sample Size, n (Studies) |
|-----------------|---------|----------|----------|-----------|-----------|
| ODI (0-50)§     | 0.74 (-1.68 to 3.17) | -1.97 (-4.36 to 0.42) | -2.40 (-4.85 to 0.04) | -2.09 (-5.28, 1.10) | -1.98 (-4.86 to 0.90) |
| Back pain (0-10)§ | 0.10 (-0.27 to 0.48) | -0.25 (-0.62 to 0.12) | -0.46 (-1.14 to 0.23) | -0.42 (-1.34 to 0.50) | -0.31 (-0.76 to 0.15) |
| Leg pain (0-10)§ | 0.23 (-0.21 to 0.66) | **-0.44 (-0.87 to -0.01)*** | -0.27 (-0.71 to 0.17) | -0.29 (-0.75 to 0.16) | -0.34 (-0.82 to 0.13) |
| SF-36® PCS (0-100)║ | -0.10 (-1.15 to 0.96) | 0.64 (-0.68 to 1.96) | **1.79 (0.27 to 3.31)*** | 1.83 (-0.19 to 3.85) | 1.10 (-0.65 to 2.86) |
| SF-36® MCS (0-100)║ | 0.52 (-0.94 to 1.98) | -0.05 (-1.59 to 1.50) | 0.06 (-1.48 to 1.60) | -0.50 (-2.56 to 1.57) | 0.54 (-3.16 to 4.25) |

ALIF = anterior lumbar interbody fusion, rhBMP-2 = recombinant human bone morphogenetic protein-2, ODI = Oswestry Disability Index, PCS = physical component summary, MCS = mental component summary, PLF = posterolateral lumbar fusion
Values in bold font are significant at 0.05 level.

* For ALIF, a total n = 465 was included in the analysis, excluding 4 patients who underwent open surgery in study 1; for PLF, a total n = 722 was included in the analysis, excluding 11 patients randomized to rhBMP-2 without instrumentation in study 12.
† These combined estimates were obtained using a two-stage approach.
‡ Includes only patients who worked before surgery. For ALIF, 221 patients worked before surgery; for PLF, the number was 241.
§ For ODI, back pain, and leg pain, high values represent worse outcomes and a negative difference favors rhBMP-2.
║ For SF-36® PCS and MCS, low values represent worse outcomes and a positive difference favors rhBMP-2.
### Table 2. Overall and Specific Adverse Events in Randomized Controlled Trials of ALIF and PLF with rhBMP-2 Versus ICBG

<table>
<thead>
<tr>
<th>Event type†</th>
<th>≤ 4 weeks</th>
<th>≤ 24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
</tr>
<tr>
<td>Anterior lumbar interbody fusion*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall adverse events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 1 Adverse event, any type</td>
<td>38% vs. 45%</td>
<td>0.84 (0.61 to 1.17)</td>
</tr>
<tr>
<td>≥ 1 Serious adverse event</td>
<td>9% vs. 8%</td>
<td>1.09 (0.59 to 2.00)</td>
</tr>
<tr>
<td>≥ 1 device-related adverse event</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Specific adverse events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomical/technical difficulty</td>
<td>0.9% vs. 3%</td>
<td>0.27 (0.04 to 1.73)</td>
</tr>
<tr>
<td>Back and/or leg pain</td>
<td>4% vs. 3%</td>
<td>1.08 (0.41 to 2.86)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>2% vs. 4%</td>
<td>0.60 (0.19 to 1.85)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>13% vs. 15%</td>
<td>0.86 (0.54 to 1.36)</td>
</tr>
<tr>
<td>Implant problems</td>
<td>2% vs. 1%</td>
<td>1.36 (0.23 to 8.07)</td>
</tr>
<tr>
<td>Infection (all types)</td>
<td>6% vs. 5%</td>
<td>1.10 (0.49 to 2.45)</td>
</tr>
<tr>
<td>Neurological</td>
<td>3% vs. 4%</td>
<td>0.83 (0.30 to 2.33)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (primary) ‡</td>
<td>3% vs. 3%</td>
<td>0.98 (0.34 to 2.88)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 2) ‡</td>
<td>2% vs. 3%</td>
<td>0.48 (0.14 to 1.62)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 3) ‡</td>
<td>3% vs. 3%</td>
<td>0.84 (0.30 to 2.36)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 4) ‡</td>
<td>0.8% vs. 2%</td>
<td>0.34 (0.07 to 1.72)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>2% vs. 3%</td>
<td>0.67 (0.18 to 2.46)</td>
</tr>
<tr>
<td>Retrograde ejaculation</td>
<td>4% vs. 1%</td>
<td>2.62 (0.28 to 24.56)</td>
</tr>
<tr>
<td>Spinal event</td>
<td>0% vs. 2%</td>
<td>0/167 vs. 3/158</td>
</tr>
<tr>
<td>Subsidence</td>
<td>2% vs. 1%</td>
<td>1.43 (0.24 to 8.41)</td>
</tr>
<tr>
<td>Urogenital</td>
<td>7% vs. 4%</td>
<td>2.03 (0.86 to 4.78)</td>
</tr>
<tr>
<td>Event type†</td>
<td>≤ 4 weeks</td>
<td>≤ 24 months</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
</tr>
<tr>
<td>Vertebral fracture</td>
<td>1% vs. 0%</td>
<td>2/168 vs. 0/156</td>
</tr>
<tr>
<td>Urinary retention‡</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Wound infection‡</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Wound dehiscence‡</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Relevant additional surgeries</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

**Posterolateral lumbar fusion§**

**Overall adverse events**

| ≥ 1 Adverse event, any type | 51% vs. 49% | 0.99 (0.74 to 1.34) | 722 (4) | 88% vs. 87% | 1.02 (0.96 to 1.07) | 722 (4) |
| ≥ 1 Serious adverse event | 20% vs. 23% | 0.89 (0.67 to 1.18) | 722 (4) | 50% vs. 52% | 0.96 (0.83 to 1.11) | 722 (4) |
| ≥ 1 device-related adverse event | --- | --- | | 6% vs. 5% | 1.35 (0.72 to 2.50) | 722 (4) |

**Specific adverse events**

<p>| Anatomical/technical difficulty | 1% vs. 0% | 4/337 vs. 0/323 | 660 (2) | Same as four weeks |
| Back and/or leg pain | 8% vs. 4% | 1.84 (1.00 to 3.37) | 706 (3) | 49% vs. 42% | 1.18 (0.91 to 1.52) | 722 (4) |
| Cardiovascular | 14% vs. 14% | 0.85 (0.40 to 1.81) | 706 (3) | 19% vs. 21% | 0.90 (0.57 to 1.40) | 722 (4) |
| Dural injury | 6% vs. 7% | 0.76 (0.43 to 1.32) | 722 (4) | 6% vs. 8% | 0.80 (0.47 to 1.36) | 722 (4) |
| Gastrointestinal | 7% vs. 10% | 0.72 (0.44 to 1.17) | 722 (4) | 16% vs. 18% | 0.88 (0.64 to 1.21) | 722 (4) |
| Implant problems | 2% vs. 0.6% | 2.85 (0.58 to 14.03) | 706 (3) | 3% vs. 2% | 1.58 (0.58 to 4.31) | 706 (3) |
| Infection (all types) | 9% vs. 10% | 0.99 (0.60 to 1.62) | 706 (3) | 18% vs. 19% | 0.96 (0.71 to 1.31) | 706 (3) |
| Neurological | 5% vs. 3% | 1.59 (0.74 to 3.43) | 722 (4) | 26% vs. 23% | 0.97 (0.62 to 1.51) | 722 (4) |
| Possible lumbar radiculitis (Primary) ‡ | 3% vs. 2% | 1.34 (0.51 to 3.47) | 722 (4) | 24% vs. 26% | 0.95 (0.73 to 1.22) | 722 (4) |
| Possible lumbar radiculitis (Definition 2) ‡ | 3% vs. 2% | 1.57 (0.46 to 5.36) | 722 (4) | 14% vs. 15% | 0.90 (0.54 to 1.51) | 722 (4) |
| Possible lumbar radiculitis (Definition 3) ‡ | 3% vs. 3% | 1.35 (0.59 to 3.12) | 722 (4) | 24% vs. 26% | 0.91 (0.71 to 1.18) | 722 (4) |
| Possible lumbar radiculitis (Definition 4) ‡ | 2% vs. 1% | 1.64 (0.48 to 5.54) | 455 (4) | 10% vs. 11% | 0.89 (0.42 to 1.87) | 455 (4) |
| Respiratory | 4% vs. 3% | 1.37 (0.59 to 3.17) | 706 (3) | 7% vs. 5% | 1.45 (0.80 to 2.63) | 706 (3) |
| Spinal event | 1% vs. 1% | 1.27 (0.13 to 12.71) | 9% vs. 10% | 0.89 (0.46 to 1.73) |</p>
<table>
<thead>
<tr>
<th>Event type†</th>
<th>≤ 4 weeks</th>
<th></th>
<th>≤ 24 months</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
<td>Sample Size, n (Studies)</td>
<td>Patients with rhBMP-2 vs. ICBG</td>
</tr>
<tr>
<td>Urogenital</td>
<td>7% vs. 7%</td>
<td>0.93 (0.38 to 2.26)</td>
<td>722 (4)</td>
<td>13% vs. 12%</td>
</tr>
<tr>
<td>Vertebral fracture</td>
<td>2% vs. 0.9%</td>
<td>1.28 (0.29 to 5.67)</td>
<td>660 (2)</td>
<td>1% vs. 1%</td>
</tr>
<tr>
<td>Relevant additional surgeries</td>
<td>---</td>
<td>----</td>
<td>---</td>
<td>12% vs. 14%</td>
</tr>
</tbody>
</table>

ALIF = anterior lumbar interbody fusion; ICBG = iliac crest bone graft; rhBMP-2 = recombinant human bone morphogenetic protein-2; PLF = posterolateral lumbar fusion.

Values in bold font are significant at 0.05 level.

* For ALIF, a total n = 465 was included in the analysis, excluding 4 patients who underwent open surgery in study 1.
† Categories of adverse events are based on Medtronic datasets, except for those indicated otherwise.
‡ Based on individual adverse event case histories in the proprietary reports provided by Medtronic.
§ For PLF, a total n = 722 was included in the analysis, excluding 11 patients randomized to rhBMP-2 without instrumentation in study 12.
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name (Study #)</th>
<th>Sample Size, n</th>
<th>Overall Success, 24 Months</th>
<th>Fusion, 24 Months</th>
<th>Cumulative Number of Adverse Events up to 24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IPD Results</td>
<td>Published Results*</td>
<td>IPD Results†</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rhBMP-2 (%)</td>
<td>ICBG (%)</td>
<td>Published Results‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICBG (95% CI)</td>
<td>rhBMP-2 (%)</td>
<td>ICBG (%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RR</td>
<td>ICBG (95% CI)</td>
<td>rhBMP-2 (%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>ICBG (%)</td>
</tr>
<tr>
<td>Anterior lumbar interbody fusion – on-label use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pilot (1)</td>
<td>11</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Boden, 2000 (55) RCT/Poor</td>
<td></td>
<td>NA</td>
<td>ICBG (%)</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pivotal (2)</td>
<td>143</td>
<td>136</td>
<td>77/133 (58%)</td>
<td>127/132 (96%)</td>
</tr>
<tr>
<td>Burkus, 2002 (6) RCT/Fair</td>
<td></td>
<td>68/123 (55%)</td>
<td>68/123 (55%)</td>
<td>108/121 (89%)</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Lap Pivotal (3)</td>
<td>134</td>
<td>70/115 (61%)</td>
<td>70/115 (61%)</td>
<td>93/114 (81.6%)</td>
</tr>
<tr>
<td>Burkus, 2003 (59) § IS/Fair</td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/INTER FIX™ ALIF Pilot (9)</td>
<td>25</td>
<td>20</td>
<td>11/23 (48%)</td>
<td>15/22 (68%)</td>
</tr>
<tr>
<td>Unpublished RCT/Fair</td>
<td></td>
<td>7/17 (41%)</td>
<td>7/17 (41%)</td>
<td>13/15 (87%)</td>
</tr>
<tr>
<td>MAVORICK™ Disc Pivotal (10)</td>
<td>172</td>
<td>405</td>
<td>58/139 (42%)</td>
<td>NA</td>
</tr>
<tr>
<td>Gornet, 2011 (64) RCT/Fair¶</td>
<td></td>
<td>233/371 (63%)</td>
<td>57/103 (55.3%)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64/0.64 (0.53, 0.77)</td>
<td>230/313 (73.5%)</td>
<td>107/136 (79%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.64</td>
<td>230/313 (73.5%)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>57/103 (55.3%)</td>
<td>107/136 (79%)</td>
<td>100%**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>107/136 (79%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%**</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>449</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>1,139</td>
</tr>
<tr>
<td>Anterior lumbar interbody fusion – off-label use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pilot (1)</td>
<td>11</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Boden, 2000 (55) RCT/Poor</td>
<td></td>
<td>NA</td>
<td>ICBG (%)</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pivotal (2)</td>
<td>143</td>
<td>136</td>
<td>77/133 (58%)</td>
<td>127/132 (96%)</td>
</tr>
<tr>
<td>Burkus, 2002 (6) RCT/Fair</td>
<td></td>
<td>68/123 (55%)</td>
<td>68/123 (55%)</td>
<td>108/121 (89%)</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Lap Pivotal (3)</td>
<td>134</td>
<td>70/115 (61%)</td>
<td>70/115 (61%)</td>
<td>93/114 (81.6%)</td>
</tr>
<tr>
<td>Burkus, 2003 (59) § IS/Fair</td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/INTER FIX™ ALIF Pilot (9)</td>
<td>25</td>
<td>20</td>
<td>11/23 (48%)</td>
<td>15/22 (68%)</td>
</tr>
<tr>
<td>Unpublished RCT/Fair</td>
<td></td>
<td>7/17 (41%)</td>
<td>7/17 (41%)</td>
<td>13/15 (87%)</td>
</tr>
<tr>
<td>MAVORICK™ Disc Pivotal (10)</td>
<td>172</td>
<td>405</td>
<td>58/139 (42%)</td>
<td>NA</td>
</tr>
<tr>
<td>Gornet, 2011 (64) RCT/Fair¶</td>
<td></td>
<td>233/371 (63%)</td>
<td>57/103 (55.3%)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64/0.64 (0.53, 0.77)</td>
<td>230/313 (73.5%)</td>
<td>107/136 (79%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.64</td>
<td>230/313 (73.5%)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>57/103 (55.3%)</td>
<td>107/136 (79%)</td>
<td>100%**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>107/136 (79%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%**</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>449</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>1,139</td>
</tr>
</tbody>
</table>

**p<0.001
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name (Study #)</th>
<th>Study Design/Quality</th>
<th>Sample Size, n</th>
<th>Overall Success, 24 Months</th>
<th>Fusion, 24 Months</th>
<th>Cumulative Number of Adverse Events up to 24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>IPD Results</td>
<td>Published Results*</td>
<td>IPD Results†</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rhBMP-2 (%)</td>
<td>ICBG (%)</td>
<td>RR (95% CI)</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pilot (4)</td>
<td>RCT/Fair</td>
<td>24</td>
<td>17/24 (71%)</td>
<td>4/20 (20%)</td>
<td>3.54 (1.42, 8.83)</td>
</tr>
<tr>
<td>Burkus, 2002 (7)</td>
<td></td>
<td>55</td>
<td>33/50 (66%)</td>
<td>15/27 (56%)</td>
<td>1.19 (0.80, 1.76)</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pivotal (5)</td>
<td>RCT/Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burkus, 2005 (65)</td>
<td></td>
<td>34</td>
<td>15/31 (48%)</td>
<td>10/31 (32%)</td>
<td>1.50 (0.80, 2.81)</td>
</tr>
<tr>
<td>Posterior lumbar interbody fusion – off-label use</td>
<td>RCT/Fair</td>
<td>30</td>
<td>13/25 (52%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/ INTER FIX™ PLIF (6)</td>
<td>RCT/Fair</td>
<td>34</td>
<td>15/31 (48%)</td>
<td>10/31 (32%)</td>
<td>1.50 (0.80, 2.81)</td>
</tr>
<tr>
<td>Haid, 2004 (8)</td>
<td></td>
<td>30</td>
<td>13/25 (52%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/ TELAMON PEEK PLIF Pilot (11) Unpublished IS/Poor</td>
<td></td>
<td>30</td>
<td>13/25 (52%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Posterior lumbar fusion – off-label use</td>
<td>RCT/Fair</td>
<td>30</td>
<td>13/25 (52%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rhBMP-2/BCP Mexico Pilot (16)</td>
<td>RCT/Poor</td>
<td>11</td>
<td>4/11 (36%)</td>
<td>4/10 (40%)</td>
<td>0.73 (0.21, 2.55)</td>
</tr>
<tr>
<td>Boden, 2002 (66)</td>
<td></td>
<td>11</td>
<td>4/11 (36%)</td>
<td>4/10 (40%)</td>
<td>0.73 (0.21, 2.55)</td>
</tr>
<tr>
<td>IDE Clinical Trial Name (Study #)</td>
<td>Sample Size, n</td>
<td>Overall Success, 24 Months</td>
<td>Fusion, 24 Months</td>
<td>Cumulative Number of Adverse Events up to 24 Months</td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------</td>
<td>-----------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPD Results</td>
<td>Published Results*</td>
<td>IPD Results</td>
<td>Published Results*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rhBMP-2 (%)</td>
<td>ICBG (%)</td>
<td>RR (95% CI)</td>
<td>rhBMP-2 (%)</td>
</tr>
<tr>
<td>rhBMP-2/BCP Canada Pivotal (13)</td>
<td>99 98</td>
<td>48/97 (49%)</td>
<td>40/95 (42%)</td>
<td>1.18 (0.86, 1.60)</td>
<td>89/96 (93%)</td>
</tr>
<tr>
<td>Unpublished RCT/Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/MASTER GRAFT® Pilot (8)</td>
<td>25 21</td>
<td>15/24 (63%)</td>
<td>10/20 (50%)</td>
<td>1.25 (0.73, 2.14)</td>
<td>17/21 (81%)</td>
</tr>
<tr>
<td>Dawson, 2009 (67) RCT/Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMPLIFY™ (rhBMP-2/CRM) Pivotal (14)</td>
<td>239 224 22 118/211 (56%)</td>
<td>105/186 (56%)</td>
<td>0.99 (0.83, 1.18)</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Dimar, 2009 (68) RCT/Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rhBMP-2/CRM 2-level Pilot (15)</td>
<td>29 12/26 (46%)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>18/26 (69%)</td>
</tr>
<tr>
<td>Unpublished IS/Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Anterior cervical discectomy and fusion – off-label use**

| INFUSE® CORNER STONE® ACDF Pilot (7) | 18 15 | 10/12 (83%) | 10/12 (83%) | 1.00 (0.70, 1.43) | NR | NR | 11/12 (92%) | 12/12 (100%) | 0.92 (0.77, 1.09) | 45 13 2 1 |
| Baskin, 2003 (5) RCT/Fair          |                |              |                      |              |              |                      |              |          |

ACDF = anterior cervical discectomy and fusion; ACS = absorbable collagen sponge ALIF = anterior lumbar interbody fusion; BCP = biphasic calcium phosphate; C = comparator group (ICBG group); Cl = confidence interval; CRM = compression resistant matrix; I = investigational group (rhBMP-2 group); ICBG = iliac crest bone graft; IDE = investigational device exemption; IPD = Individual patient data; IS = intervention series; NA = not applicable; NR = not reported; NS = not significant; NSR = not separately reported; PEEK = polyetheretherketone; PLF = posterior lumbar fusion; PLIF = posterior lumbar interbody fusion; RCT = randomized controlled trial; rhBMP-2 = recombinant human bone morphogenetic protein-2; US = United States.

* For unpublished studies, cells are blank.
† More information about the type and number of specific adverse effects can be found in Appendix F of the full report (17). These numbers don’t include non-union and non-union pending.
‡ The type and number of specific adverse effects reported by each journal publication can be found in Table 16 of the full report (17).
¶ Study 5 data not published independently. Burkus, 2005 (65) contains pooled data from Studies 4 and 5.
† The comparison group in this study received artificial disk, not ICBG. Discrepancy in numbers between published trial and IPD was partially due to an updated Medtronic data set provided to the authors.
** n not reported; results reported only as percentages.
†† The Mexico pilot was an intervention series with two cohorts.
‡‡ I1 = rhBMP-2 without internal fixation; I2 = rhBMP-2 + TSRH (Texas Scottish Rite Hospital) pedicle screw instrumentation; C = autograft + TSRH. This study only followed patients for 12 months, so there were no data at 24 months.
§§ The cumulative number of adverse events up to 12 months.
¶¶ The table in this publication reports a slightly higher percentage (97.3%).
Figure 1. Comparison of Cancer Risk Between rhBMP-2 Versus Control

<table>
<thead>
<tr>
<th>Study (Number)</th>
<th>Control Rate</th>
<th>Risk Ratio (95% CI)</th>
<th>Events, rhBMP-2</th>
<th>Events, Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>24 month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infuse LT Cage Pivotal (2)</td>
<td>0.74%</td>
<td>1.90 (0.17, 20.74)</td>
<td>2/143</td>
<td>1/136</td>
</tr>
<tr>
<td>Infuse Bone Dowel Pilot (4)</td>
<td>0%</td>
<td>2.76 (0.12, 64.41)</td>
<td>1/24</td>
<td>0/22</td>
</tr>
<tr>
<td>Infuse Bone Dowel Pivotal (5)</td>
<td>0%</td>
<td>1.66 (0.07, 39.55)</td>
<td>1/55</td>
<td>0/30</td>
</tr>
<tr>
<td>Maverick Disc Pivotal (10)</td>
<td>0.74%</td>
<td>2.35 (0.48, 11.55)</td>
<td>3/172</td>
<td>3/405</td>
</tr>
<tr>
<td>Amplify Pivotal (14)</td>
<td>0.89%</td>
<td>4.69 (1.04, 21.15)</td>
<td>10/239</td>
<td>2/224</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td></td>
<td><strong>3.45 (1.34, 8.85)</strong>*</td>
<td>17/633</td>
<td>6/817</td>
</tr>
<tr>
<td><strong>48 month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infuse Bone Dowel Pilot (4)</td>
<td>0</td>
<td>2.76 (0.12, 64.41)</td>
<td>1/24</td>
<td>0/22</td>
</tr>
<tr>
<td>Maverick Disc Pivotal (10)</td>
<td>1.24%</td>
<td>1.41 (0.34, 5.85)</td>
<td>3/172</td>
<td>5/405</td>
</tr>
<tr>
<td>BCP Canada (12)</td>
<td>2.04%</td>
<td>0.34 (0.01, 8.15)</td>
<td>0/48</td>
<td>1/49</td>
</tr>
<tr>
<td>Amplify Pivotal (14)</td>
<td>2.23%</td>
<td>2.25 (0.81, 6.28)</td>
<td>12/239</td>
<td>5/224</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td></td>
<td><strong>1.82 (0.84, 3.95)</strong>*</td>
<td>16/483</td>
<td>11/700</td>
</tr>
</tbody>
</table>

*The combined risk ratio (RR) was obtained using a generalized fixed effects model with binomial distribution and log link without correction for zero events. The RR from each study was estimated, when there is zero event, by adding a continuity correction of 0.5, for illustrative purposes.

Rongwei Fu, Ph.D
3181 SW Sam Jackson Park Rd.
#CB669
Portland, OR 97239
USA

REF: M12-2731

Dear Dr. Fu:

Thank you for submitting to *Annals of Internal Medicine* your revised manuscript, "Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-Analysis." Your revisions have improved the manuscript and we remain interested in considering it for publication. However, before we can consider it further, you must prepare a revision that satisfactorily addresses the remaining concerns outlined in this letter.

Understanding the project team's desire to publish this paper quickly, we will need to have your completed revision returned within 2 weeks. As we are not requesting any adjustments to your analyses, this should be readily accomplished.

1. Abstract: We unfortunately cannot accommodate an abstract of 395 words. We do appreciate the more complex nature of this study and the need to relax our normal limit of 275 words. You will, however, need to shorten your abstract to no more than 350 words.

2. Similarly, the text of the manuscript is longer than we usually are able to accommodate (limit for systematic reviews 3000 words). We recognize, however, that your paper needs to describe a more complex combination of data sources and present the results of two large questions. Careful line editing should enable you to reduce the total length substantially. Doing so almost always improves readability and thus impact for readers, and can be done without losing important content. You should try to reduce the manuscript length to less than 4500 words.

Here is a single example of the type of line editing to which I refer (reducing from 41 to 30 words at current line 120):

“Trials sponsored by the sole manufacturer of rhBMP-2 devices, Medronic, and published before 2009 reported beneficial effects in spinal fusion with no or few adverse and no “unanticipated device-related” events (5-8).”
3. Line 219: please clarify if 48 months was the longest period of follow-up you found, or if you chose to limit the assessed follow-up in your report to 48 months. We assume the former as it would not seem appropriate to ignore other available information, but felt this requires confirmation.

4. Lines 335 and 338: “…was no difference…” Given the wide confidence intervals in the point estimates, similar to those presented in the prior section addressing ALIF, you should similarly qualify your statement with language similar to that used at line 311 (“…though estimates were frequently imprecise, precluding strong conclusions.”). The same issue applies at line 390.

5. Line 381 – 382: Are you referring here to the results of your sensitivity?

6. Line 384: “Estimates were similar…” – to what? To each other? To the estimates presented above?

7. We had trouble finding the references in the table that corresponded to the results you describe in your narrative discussion related to Table 3. For example, we could not find reference 58/Kleeman in the table. Perhaps we have misunderstood. Please review the results sections and all tables to ensure they are accurate / consistent.

8. Your presentation of Aim #2 frequently seems to editorialize. While I realize it is difficult to avoid “interpreting” the reporting that you found, as to some extent you are reporting your interpretation of the published papers relative to what you found in your data collection and analysis, try as best possible to restrict your narrative results section to merely presenting what you found. Avoid interpretation and editorializing, as such may leave the reader with the impression that s/he is reading a “biased report of whether there was bias.” Such interpretation and editorializing should be restricted to the discussion of your paper. For example, refrain from language such as “stack the deck” or “dramatically”. Although I realize it may be “splitting hairs”, a single example of language that might merely report what you found as opposed to interpreting / editorializing would be at line 466 to state, “A complete accounting of reasons for study termination could not be found.” rather than, “…and they gave an incomplete account of the reasons…” Finally, pointing out a few examples of where reporting seemed accurate may also help assure the reader of a balanced presentation.

9. Line 423: “…the investigators republished this integrated analysis…” - please clarify if you are indicating that duplicative publishing (self-plagiarism) occurred (i.e, that the authors misrepresented previously published analyses as new.

10. Line 438: I suspect many readers will want to know why the study was terminated early.

11. Line 532: To what “journal practices” do you refer?

12. Appendix Figure 1, bottom box (“Included articles by approach and study design”): As you note the sponsor for some but not all studies, it is unclear the related status of the 6 trials of ALIF, and of the cohort studies. Please revise for clarity.
13. Appendix table 2: Please make clear for readers why there are no baseline data entered for the INFUSE/LT-CAGE Pilot study.

14. Please provide in an appendix the search strategy used.

**Statistical Comments**

We need you to clarify a couple of points for us in fitting the mixed-effect models.

1. You say you fit models with random intercepts and heterogeneous variances, but the IPD data is longitudinal in nature. That implies that you also fitted a random slope (at least in some models). Is this the case? Or, did you simply fit different models to the data reported at 6 weeks, 3 months, 6 months and 12 months? We are not asking you to reanalyze the data, but simply report what you did more clearly.

2. Please provide a copy of your SAS code with your revision (We usually request copies of code for more complicated analyses, but I forgot to ask for it in our e-mail exchange prior to receiving your revised manuscript.)

3. Always tell the reader the specific SAS procedure used to fit the various models: Proc GLM, Proc Genmod, Proc Mixed, Proc NLMixed, etc.

4. The Management of missing data section is a bit confusing without reference to the details provided in Appendix Table 1. Each outcome was assessed by examining of multiple criteria. Success was defined as meeting all the designated criteria. So, it makes sense to treat cases where data is provided for some, but not all requisite criteria as failures. You need to clarify this for the reader. Perhaps a sentence like, “Overall success and fusion success are composite assessments determined from multiple clinical criteria, all of which must be satisfied to classify a case as a success,” inserted at the beginning of this section would be helpful. You could then delete the opening phrase on line 195, and start a new sentence with “Patients meeting some criteria, but missing data for other criteria, were classified as failures in our primary analyses.”

5. The “pseudo-trial” approach tends to overcorrect, shrinking the test statistics toward the null. While there are other techniques for incorporating zero-event studies into study- or patient-level meta-analyses. The fixed-effect MH method you asked about in our e-mail exchange can include zero-event studies, without employing a continuity correction. Or you could consider a more complicated non-linear mixed-effect model, such as presented in Stijnen, et al. Stat Med 2010, 29: 3046-3067 where they fit a Poisson-Normal model that incorporates zero-event trials in a meta-analysis using Proc NLMixed (see Appendix Example 6 for SAS code). However, it is doubtful that implementation of any of these methods will alter the findings.

a. So, at a minimum, please include reference to a peer review methods paper that supports the use of a pseudo-trial to allow inclusion of zero-event studies.
b. If none can be found, then simply delete this sentence and the associated sensitivity analysis. Note how may zero-event studies were found, and the impact that their exclusion may have for your findings.

Please send your revised manuscript and cover letter to us within two weeks of receiving this letter. Number the lines of your revised manuscript consecutively throughout. In the cover letter, group and number your responses to correspond to the comments in this letter. Please restate each comment and follow it with your response indicating what you did, why you did it, and the line number of the revised manuscript where the changes may be found. Submit your revised manuscript and cover letter at https://www.acponline.org/authors/ by clicking the "Revise Paper" link listed below your "Tasks."

Please keep all editorial correspondence confidential, and refrain from sharing either the correspondence itself or the essence of its content with individuals who are not your collaborators. Doing so helps ensure we can offer you advice that is in the best interests of your paper, without concern for how it might be considered or used by others.

We look forward to receiving your revised manuscript.

Sincerely,

Darren Taichman, MD, PhD

Executive Deputy Editor
12. Authors: Response Letter 2 (10 April 2013)

April 10, 2013

Darren Taichman, MD, PhD
Executive Deputy Editor, Annals of Internal Medicine

REF: M12-2731

Dear Dr Taichman,

We are submitting a revised version of our manuscript, "Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-analysis." Our responses to the comments from the Editor and the Statistical Editor are detailed below.

We would be happy to discuss any additional changes you may suggest. Thank you for the opportunity to resubmit our manuscript.

Sincerely,

Rongwei Fu, Ph.D.
3181 SW Sam Jackson Park Rd.
MC: CB669
Portland, OR 97239
USA

Editor’s comments:

1. Abstract: We unfortunately cannot accommodate an abstract of 395 words. We do appreciate the more complex nature of this study and the need to relax our normal limit of 275 words. You will, however, need to shorten your abstract to no more than 350 words.

Response: We have shortened the abstract to no more than 350 words.

2. Similarly, the text of the manuscript is longer than we usually are able to accommodate (limit for systematic reviews 3000 words). We recognize, however, that your paper needs to describe a more complex combination of data sources and present the results of two large questions. Careful line editing should enable you to reduce the total length substantially. Doing so almost always improves readability and thus impact for readers, and can be done without losing important content. You should try to reduce the manuscript length to less than 4500 words.
Here is a single example of the type of line editing to which I refer (reducing from 41 to 30 words at current line 120):
“Trials sponsored by the sole manufacturer of rhBMP-2 devices, Medronic, and published before 2009 reported beneficial effects in spinal fusion with no or few adverse and no “unanticipated device-related” events (5-8).”

Response: We have line edited the manuscript and reduced the word length to less than 4500 words.

3. Line 219: please clarify if 48 months was the longest period of follow-up you found, or if you chose to limit the assessed follow-up in your report to 48 months. We assume the former as it would not seem appropriate to ignore other available information, but felt this requires confirmation.

Response: We want to clarify that 48 months was not the longest period of follow-up we found; there were four studies (across different surgical approaches) with follow-up longer than 48 months in one arm (Study 2 and study 3) or both arms (Study 10 and study 14) (please see Appendix Table 2). In the manuscript, we now refer the readers to Appendix Table 2 for information about studies with longer follow-up.

We did not ignore information longer than 48 months and reported results at longer follow-up time points in our report. For example, for cancer, two cancers occurred in the ICBG group and one cancer occurred in the rhBMP-2 group after 48 months. We don’t have enough data to do a separate analysis for after 48 months, but we did an analysis combining 48 months data with these three cancers as a sensitivity analysis, which produced similar results to that from 48 months.

However, data after 48 months get very sparse with a large proportion of patients missing differentially in the rhBMP-2 and ICBG groups (so the results are more likely to be biased). Therefore we did not include the data after 48 months in the manuscript.

4. Lines 335 and 338: “…was no difference…” Given the wide confidence intervals in the point estimates, similar to those presented in the prior section addressing ALIF, you should similarly qualify your statement with language similar to that used at line 311 (“…though estimates were frequently imprecise, precluding strong conclusions.”). The same issue applies at line 390.

Response: For lines 335-338, we have revised the text based on the editor’s comments to say (Current Page 13 lines 314-317) “Similar to ALIF, there was no significant difference between rhBMP-2 and ICBG groups in adverse events (Table 2 and Appendix Figure 2), but estimates were frequently imprecise, precluding strong conclusions.”
For line 390, the revised text added (Current Page 15 lines 360-361) “but the event rates were low and estimates of RR were imprecise.”

5. Line 381 – 382: Are you referring here to the results of your sensitivity?

Response: We were referring to a sensitivity analysis including zero event trials; however, we have deleted the results of the sensitivity analysis based on the comment from the statistical editor.

6. Line 384: “Estimates were similar…” – to what? To each other? To the estimates presented above?

Response: Estimates were similar to previously presented results. We clarified accordingly in the text to say (Current Page 15 lines 353) “Excluding non-SEER cancers resulted in similar estimates to those including non-SEER cancers…”

7. We had trouble finding the references in the table that corresponded to the results you describe in your narrative discussion related to Table 3. For example, we could not find reference 58/Kleeman in the table. Perhaps we have misunderstood. Please review the results sections and all tables to ensure they are accurate / consistent.

Response: Unfortunately the table listed only the primary publication for each trial. We have added citations for secondary publications (integrated analyses, reviews, republications) to a footnote of the table.

8. Your presentation of Aim #2 frequently seems to editorialize. While I realize it is difficult to avoid “interpreting” the reporting that you found, as to some extent you are reporting your interpretation of the published papers relative to what you found in your data collection and analysis, try as best possible to restrict your narrative results section to merely presenting what you found. Avoid interpretation and editorializing, as such may leave the reader with the impression that s/he is reading a “biased report of whether there was bias.” Such interpretation and editorializing should be restricted to the discussion of your paper. For example, refrain from language such as “stack the deck” or “dramatically”. Although I realize it may be “splitting hairs”, a single example of language that might merely report what you found as opposed to interpreting / editorializing would be at line 466 to state, “A complete accounting of reasons for study termination could not be found.” rather than, “…and they gave an incomplete account of the reasons…” Finally, pointing out a few examples of where reporting seemed accurate may also help assure the reader of a balanced presentation.

Response: We have reviewed and revised the text with these comments in mind. Note that the names of some forms of reporting bias (e.g., “misrepresentation of facts”), while standard, can seem prejudicial.

The manuscript focuses on the years just before and after approval of rhBMP-2 (2000-2005). Some Medtronic trials published 2009 and later were well-reported.
We now mention these, and Table 3 indicates when published results and IPD results agree.

9. Line 423: “…the investigators republished this integrated analysis…” - please clarify if you are indicating that duplicative publishing (self-plagiarism) occurred (i.e., that the authors misrepresented previously published analyses as new.

Response: According to the URM, a re-publication should indicate a statement such as “This article is based on a study first reported in the [title of journal, with full reference].” The 2nd of the 3 articles includes the statement “these results have been published elsewhere” with an appropriate citation to the first article. However, the abstract of the 2nd article gives no indication that the results were published previously. The 3rd paper was a review by the same author that cites the first paper.

We were less interested in the possible ethical violation of duplicative publishing than in the way the 2nd and 3rd articles presented the study. The 2nd and 3rd have more unreserved claims about the superiority of rhBMP-2 and less accurate descriptions of the study design. We have revised the section accordingly.

10. Line 438: I suspect many readers will want to know why the study was terminated early.

Response: Medtronic provided a document that says Study #5 was terminated “for business reasons.” We have no additional information about the termination. We added that reason in the text. (Current Page 11 lines 280)

11. Line 532: To what “journal practices” do you refer?

Response: We revised the sentence to specify some of the practices we mean, citing relevant literature:

1. Policies on commercial supplements enabled lower-quality, biased articles to be published in major specialty journals, in this case *Spine*. Two seminal studies in this area are


Quoting Richard Smith (*BMJ 2003;326:1202*): “Another way in which journals become entangled is through publishing supplements. ..If a journal is willing to publish every paper presented at a symposium that was funded by a single company and that dealt with one drug, then it can charge a substantial fee. Often these papers
will be set pieces by, to be crude for a moment, “paid industry hacks” and will have been published many times. If, however, the journal wants to peer review every study and take only those that are original and pass review then the fee will be smaller. Studies have shown that papers published in supplements are of poorer quality than those published in the main journal.”

2. Lack of the following: requirements for trial registration, disclosure of primary outcome measures, or statements that authors had access to the data and control over the contents of the articles.

3. Standards for conflict of interest disclosure were lax.

Now the sentence reads (Current Page 20 lines 472–475) “Journal practices regarding sponsored supplements, trial registration, and conflict of interest disclosure may have contributed to publication of an incomplete and sometimes misleading evidence base (78-80).”

12. Appendix Figure 1, bottom box (“Included articles by approach and study design”): As you note the sponsor for some but not all studies, it is unclear the related status of the 6 trials of ALIF, and of the cohort studies. Please revise for clarity.

Response: We have revised the box for clarity and added sponsorship for all included studies.

13. Appendix table 2: Please make clear for readers why there are no baseline data entered for the INFUSE/LT-CAGE Pilot study.

Response: The data were omitted by mistake, and we have inserted them. We apologize for the oversight.

14. Please provide in an appendix the search strategy used.

Response: We have provided the search strategy in an appendix (Appendix 1).

Statistical Comments

We need you to clarify a couple of points for us in fitting the mixed-effect models.

1. You say you fit models with random intercepts and heterogeneous variances, but the IPD data is longitudinal in nature. That implies that you also fitted a random slope (at least in some models). Is this the case? Or, did you simply fit different models to the data reported at 6 weeks, 3 months, 6 months and 12 months? We are not asking you to reanalyze the data, but simply report what you did more clearly.
Response: We revised the Methods to be clearer that we fit different models to the data at each time point, and we added a sentence, (Current Page 8 line 214) “We fitted a separate model for each time point”.

2. Please provide a copy of your SAS code with your revision (We usually request copies of code for more complicated analyses, but I forgot to ask for it in our e-mail exchange prior to receiving your revised manuscript.)

Response: A copy of our SAS code for meta-analysis of IPD for continuous outcome, common binary outcome and rare binary outcome is provided. (Appendix 2)

3. Always tell the reader the specific SAS procedure used to fit the various models: Proc GLM, Proc Genmod, Proc Mixed, Proc NLMixed, etc.

Response: This information has been added to the text (Current Page 9 lines 225-227) as “IPD meta-analyses for continuous, common and rare binary outcomes were performed using PROC MIXED, PROC NLMIXED and PROC GENMOD respectively using SAS® software 9.2...”.

4. The Management of missing data section is a bit confusing without reference to the details provided in Appendix Table 1. Each outcome was assessed by examining of multiple criteria. Success was defined as meeting all the designated criteria. So, it makes sense to treat cases where data is provided for some, but not all requisite criteria as failures. You need to clarify this for the reader. Perhaps a sentence like, “Overall success and fusion success are composite assessments determined from multiple clinical criteria, all of which must be satisfied to classify a case as a success,” inserted at the beginning of this section would be helpful. You could then delete the opening phrase on line 195, and start a new sentence with “Patients meeting some criteria, but missing data for other criteria, were classified as failures in our primary analyses.”

Response: We have revised the text as suggested, by adding the suggested sentence to start the section and a reference to the Appendix Table 1. Now the section starts with (Current Page 7 starting from line 184) “Overall success and fusion are determined using multiple criteria; all had to be satisfied to classify a case as a success (Appendix Table 1). In the primary analysis, patients meeting some criteria but missing data for others were...”.

5. The “pseudo-trial” approach tends to overcorrect, shrinking the test statistics toward the null. While there are other techniques for incorporating zero-event studies into study- or patient-level meta-analyses. The fixed-effect MH method you asked about in our e-mail exchange can include zero-event studies, without employing a continuity correction. Or you could consider a more complicated non-linear mixed-effect model, such as presented in Stijnen, et al. Stat Med 2010, 29: 3046-3067 where they fit a Poisson-Normal model that incorporates zero-event trials in a meta-analysis using Proc NLMixed (see Appendix
Example 6 for SAS code). However, it is doubtful that implementation of any of these methods will alter the findings.

a. So, at a minimum, please include reference to a peer review methods paper that supports the use of a pseudo-trial to allow inclusion of zero-event studies.

b. If none can be found, then simply delete this sentence and the associated sensitivity analysis. Note how many zero-event studies were found, and the impact that their exclusion may have for your findings.

Response: We have revised the text based on the editor’s comment b. Specifically, we deleted the sentence about the sensitivity analysis. We agree that the “pseudo-trial” approach tends to overcorrect, shrinking the test statistics towards the null. However, since our sensitivity analysis using the “pseudo-trial” approach still shows a significant difference at 24 months after potential overcorrection, we think it helps the argument that excluding trials with zero cancer events in both groups will not affect the results, and hence kept it in our report. In the discussion section, we refer to the sensitivity analysis as follows (Current page 19 lines 462-464) “Seven Medtronic-sponsored trials (total n = 429) with zero cancer events in both groups were not included in the meta-analysis but were not expected to affect the results (17).”
Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-analysis

Rongwei Fu, PhD
Shelley Selph, MD
Marian McDonagh, PharmD
Kimberly Peterson, MS
Arpita Tiwari, MHS
Roger Chou, MD
Mark Helfand, MD, MS

Author affiliations:
Rongwei Fu: Department of Public Health & Preventive Medicine, and Department of Emergency Medicine, Oregon Health & Science University; Portland, OR
Shelley Selph, Marian McDonagh, Kimberly Peterson: Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University; Portland, OR
Arpita Tiwari: PhD Student, Oregon State University, Corvallis, OR
Roger Chou: Department of Medicine, and Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University; Portland, OR
Mark Helfand: Portland VA Medical Center; Portland, OR

Running title: Effectiveness and harms of rhBMP-2 in spine fusion

Key words: rhBMP-2; reporting bias; meta-analysis of individual patient data; spinal fusion; systematic review

Corresponding author: Rongwei Fu, PhD; 3181 SW Sam Jackson Park Road, Mail Code: CSB669, Portland, OR 97239; Email: fur@ohsu.edu; Phone 503-494-6069

Text word count: 4499 (excluding headings); Abstract word count: 349 (excluding subheadings); Figures: 1; Tables: 3; Appendices: 2. Appendix Tables: 3. Appendix Figures: 2.
Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Yale University Open Access Data Project or Medtronic, Inc.

Contract support: Research Subcontract to Oregon Health & Science University under Sponsored Research Agreement between Yale University and Medtronic, Inc.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Rongwei Fu, PhD, Oregon Health & Science University, 3181 S.W. Sam Jackson Park Rd., Mail code CSB669, Portland, OR 97239.

Current Author Addresses:
Rongwei Fu: 3181 S.W. Sam Jackson Park Rd, Mail code CSB669, Portland, OR 97239
Shelley Selph, Marian McDonagh, Kimberly Peterson, Roger Chou, Mark Helfand: 3181 S.W. Sam Jackson Park Rd, Mail code BICC, Portland, OR 97239
Arpita Tiwari: OSU College of Public Health and Human Sciences, 401 Waldo Hall, Corvallis, OR 97331
Abstract

Background: Recombinant human bone morphogenetic protein-2 (rhBMP-2) is used as a bone graft substitute in spinal fusion, which unites (fuses) bones in the spine. The accuracy and completeness of journal publications of industry-sponsored trials on the effectiveness and harms of rhBMP-2 has been called into question.

Purpose: To independently assess the effectiveness and harms of rhBMP-2 in spinal fusion and reporting bias in industry-sponsored journal publications.

Data Sources: Individual patient data (IPD) of 17 industry-sponsored studies, related internal documents, and searches of MEDLINE (1996 to August 2012), other databases, and reference lists.

Study Selection: Randomized controlled trials (RCTs) and cohort studies of rhBMP-2 versus any control, and uncontrolled studies of harms.

Data Extraction: Effectiveness outcomes in IPD were recalculated using consistent definitions. Study characteristics and results were abstracted by one investigator and confirmed by another. Two investigators independently assessed quality using predefined criteria.

Data Synthesis: Thirteen RCTs and 30 cohort studies were included. RhBMP-2 and iliac crest bone graft had similar effects on overall success, fusion, and other effectiveness measures, and on the risk of experiencing any adverse event in lumbar spine fusion, though rates were high across interventions (78% to 88% at 24 months from surgery). For anterior lumbar interbody fusion, rhBMP-2 was associated with non-significantly increased risk of retrograde ejaculation and urogenital problems. RhBMP-2 was associated with increased risk of wound complications and dysphagia in anterior cervical spine fusion. At 24 months the cancer risk was increased with use of rhBMP-2 (RR 3.45; 95% CI 1.98 to 6.00), but event rates were low and cancer were heterogeneous. Early journal publications misrepresented the effectiveness and harms through selective reporting, duplicate publication, and underreporting.

Limitations: Outcome assessment was not blinded, with poor ascertainment of harms in trials. No trials were truly independent of industry sponsorship.

Conclusion: In spinal fusion, rhBMP-2 has no proven clinical advantage over bone graft and may be associated with important harms, making it difficult to identify clear indications for rhBMP-2. Earlier disclosure of all relevant data would have better informed clinicians and the public than the journal publications did.

Primary Funding Source: Research Subcontract to Oregon Health & Science University under Sponsored Research Agreement between Yale University and Medtronic, Inc.
The most common surgery for chronic low back pain with lumbar disc degenerative conditions is vertebral fusion (1) to restrict spinal motion and remove the presumed cause of pain. An interbody fusion, involving removal of a degenerated intervertebral disc and fusion of the adjacent vertebral bodies, can be performed via an anterior (anterior lumbar interbody fusion, ALIF), posterior (posterior lumbar interbody fusion, PLIF), or transforaminal (transforaminal lumbar interbody fusion, TLIF) approach. Posterolateral lumbar fusion (PLF) involves adjacent transverse processes.

Traditionally, spinal fusions use graft material from the patient’s iliac crest to promote fusion. In 2002, the U.S. Food and Drug Administration (FDA) approved recombinant human bone morphogenetic protein-2 (rhBMP-2), a genetically engineered protein with bone-growth stimulating properties, as a bone graft substitute in conjunction with a device implant (LT-CAGE™) for single-level ALIF (2). In December 2003, the FDA approved the use of rhBMP-2 with another implant (INTER FIX™) for similar indications (3). In clinical practice, rhBMP-2 has primarily been used “off-label” in PLF and TLIF (4).

Before 2009, trials sponsored by Medtronic, the sole manufacturer of rhBMP-2 devices, reported beneficial effects with no or few adverse events (5-8). Subsequent observational studies reported serious complications associated with rhBMP-2 in cervical spine fusion, (9-12) and FDA documents summarizing Medtronic-sponsored trials appeared to indicate substantially more adverse events than reported in journal publications (13). In 2008, the FDA issued a public health notification of life-threatening complications associated with off-label use of rhBMP-2 in cervical spine fusion—swelling of the neck and throat resulting in compression of the airway and other structures (14).
Selective reporting or underreporting of outcomes in journal publications may have misrepresented the balance of benefits and harms of rhBMP-2 (13, 15). Our study aimed to 1) estimate effectiveness and harms of rhBMP-2 in spinal fusion in a systematic review, using individual patient data (IPD) when available and 2) assess reporting biases in published articles of industry-sponsored studies.

**Methods**

We registered a short version of the review protocol at the PROSPERO registry of systematic reviews (16) on February 23, 2012, and deposited the full protocol with the Yale University Open Data Access (YODA) Project. Detailed methods and additional analyses are available elsewhere (17).

**Data Sources**

We used four sources of data: 1) Medtronic IPD, related protocols, and data dictionaries; 2) Medtronic internal reports; 3) documents from the FDA Web site, and 4) a broad-based literature search (Appendix 1 available at www.annals.org) to identify a) additional studies on rhBMP-2 and b) publications related to Medtronic-sponsored studies. For aim 1, we used data from sources 1), 2) and 4 a), and for aim 2, we compared the journal publications to other sources.
For data sources 1) and 2), the YODA Project provided de-identified patient-level data, protocols, data dictionaries, and Medtronic internal reports for all 17 Medtronic-funded studies of rhBMP-2 in spinal fusion completed or terminated by December, 2011. The internal reports included summaries of study data and adverse event case histories. We also received 1,229 MedWatch adverse event reports.

For data sources 3) and 4), we searched MEDLINE® (1996 to August 2012), Embase®, the Cochrane Library® (3rd Quarter 2012), Scopus, Clinicaltrials.gov, the FDA website, and manually searched reference lists.

For aim 1, two reviewers independently assessed each article for eligibility. For effectiveness and harms, we included controlled clinical trials and cohort studies of rhBMP-2 in spinal fusion. For harms, we also included uncontrolled intervention series. We excluded studies that combined results of rhBMP-2 with other bone morphogenetic proteins, unless we could determine rhBMP-2 was predominantly used. For aim 2, we identified publications in peer-reviewed journals that reported results from one or more Medtronic trials.

**Data Abstraction and Quality Rating**

One investigator abstracted patient and study characteristics and results, and a second reviewed abstracted data for accuracy. For Medtronic-funded studies, quality assessment was based on information from trial protocols and internal reports. Two investigators independently rated the quality of each study as good, fair, or poor, using criteria adapted from the Cochrane Back Review Group (18) and the US Preventive Services Task Force (19). Discrepancies were resolved through consensus.
Definitions and calculations of endpoints for individual patient data

We used the study protocols and ClinicalTrials.gov entries to determine pre-specified primary outcomes. In nine studies, the primary effectiveness measure was “overall success” (at 24 months); fusion was the primary endpoint in the remainder. Other effectiveness outcomes included pain, disability, neurologic status, function, and return to work. Studies differed slightly in defining effectiveness outcomes. To standardize effectiveness measures, we applied consistent definitions (Appendix Table 1, available at www.annals.org) and recoded and recalculated effectiveness outcomes using IPD.

We obtained adverse events directly from IPD except for urinary retention, wound infection, wound dehiscence, and possible lumbar radiculitis, which we identified by reviewing internal report case histories. We applied four alternative definitions (Appendix Table 1) for lumbar radiculitis, an outcome that was not pre-specified in the protocol.

Management of missing data

Overall success and fusion are determined using multiple criteria; all had to be satisfied to classify a case as a success (Appendix Table 1). In the primary analysis, patients meeting some criteria but missing data for others were classified as failures, and patients without data for any criteria were excluded. We also performed two sensitivity analyses: in one, patients with missing data for some or all criteria were excluded; in the other, such patients were included as failures. For other binary effectiveness outcomes, patients with missing data were excluded in the primary analysis but included as failures in the sensitivity analysis. For adverse events, all patients were included since we analyzed cumulative adverse events from the time of surgery.
Data Synthesis and Analysis

Aim #1: Effectiveness and harms of rhBMP-2

We stratified analyses by spinal area (lumbar, cervical) and surgical approach (e.g., ALIF, PLF) for all outcomes except cancer and death, for which we combined all surgical approaches because these rare outcomes were not necessarily affected by surgical technique. Only the ALIF and PLF trials provided sufficient data for meta-analyses, which were based on IPD from Medtronic-sponsored trials. We identified one additional trial without corresponding IPD (20) and qualitatively compared its results with IPD results.

For effectiveness endpoints, we calculated outcomes at 6 weeks and at 3, 6, 12, and 24 months after surgery, based on the time points typically evaluated in the trials. For harms, we aggregated data into two periods: 1) operative and up to four weeks post-operative, and 2) up to 24 months post-operative. Data beyond 24 months were sparse (see Appendix Table 2) and are reported elsewhere (17) except that for cancer and death, we also reported results through 48 months.

We used mixed effects models to combine IPD. For continuous outcomes, we used a linear mixed effects model to obtain a combined mean difference between rhBMP-2 and control groups after adjusting for baseline values and individual study effects (21). We assumed random treatment effects and heterogeneous residual variance across included studies. For common binary outcomes, we used a generalized linear mixed effects model assuming random treatment effects and binomial distribution with log link to obtain a combined risk ratio (RR). For rare binary outcomes, we used a generalized linear fixed effects model assuming binomial distribution with log link. We fitted a separate model for each time point. When the generalized
linear model with log link could not produce a combined estimate due to ill-fitting data, we provided combined estimates from a two-step approach (described elsewhere (17)).

We assessed statistical heterogeneity using the estimated between-study variance from the mixed effects model (21). We evaluated baseline age, sex, smoking status, diabetes status, previous back surgery, and employment status as potential sources of heterogeneity. We also performed sensitivity analyses by excluding poor quality studies and studies utilizing a lower rhBMP-2 concentration, and by excluding graft-site-related adverse events in analyses of harms. For cancer, we performed sensitivity analyses by excluding events not reportable to the National Cancer Institute Surveillance Epidemiology and End Results (SEER) Program (skin cancers with low propensity to metastasize). Results of sensitivity analyses were generally similar and, except for cancer and lumbar radiculitis, not reported separately. IPD meta-analyses for continuous, common and rare binary outcomes were performed using PROC MIXED, PROC NLMIXED and PROC GENMOD respectively using SAS® software 9.2 (SAS Institute Inc., Cary, NC, USA) (Appendix 2, sample SAS codes available at www.annals.org).

We rated the strength of evidence by outcome based on the aggregate risk of bias, consistency, directness, and precision of the evidence (22).

Aim #2: Assessment of reporting and related biases

We assessed publication and outcome reporting biases and quality of reporting (23) by comparing journal publications with corresponding study protocols, reports, and data dictionaries provided by Medtronic. We used a previously published protocol to classify publications as primary or secondary and to categorize potential sources of reporting bias (24, 25).
Role of the Funder

The YODA Project proposed the aims for the review and served as the intermediary for data and information requests to Medtronic. Medtronic provided comments on our draft report. Neither the YODA Project nor Medtronic influenced the conduct of our analyses or the content of this article.

Results

Aim #1: Benefits and Harms of rhBMP-2

Overview of included studies

We included 13 randomized controlled trials (RCTs), 12 sponsored by Medtronic (n=1,879) and one by Norton Healthcare (n=106) (20) (Appendix Figure 1, available at www.annals.org). All RCTs compared rhBMP-2 with iliac crest bone graft (ICBG) except for Study 10, which compared artificial disc replacement to fusion with rhBMP-2 (see Appendix Table 2 for study identification numbers). The trials applied similar eligibility criteria and enrolled similar populations within each surgical approach. We excluded one very small (n=3) Medtronic trial. Eight studies enrolled fewer than 100 subjects (sample sizes ranged 14 to 85). At 24 months, follow-up rates were >90% in both groups in 9 of the 12 trials.

While there were some baseline differences between patients receiving ICBG and rhBMP-2, we did not detect a pattern favoring rhBMP-2. The main risks for bias were lack of blinding of surgeons, patients, and outcome assessors (except for radiologic endpoints). The quality of ascertainment varied across outcomes. Effectiveness outcomes (e.g., pain, function, fusion) were generally ascertained with well-designed questionnaires or scales. For harms, the studies used broad classifications for many adverse events, and events were generally not
actively elicited using specific symptom questionnaires or objective tests. For example, for retrograde ejaculation, it was unclear how the outcome was defined or whether investigators asked about specific symptoms. No trial defined radiculitis, and adverse events consistent with possible radiculitis were variously classified as back and leg pain, neurological events, or spinal events. Cancer was not a pre-specified endpoint and only captured by voluntary reporting. Local effects, such as inflammation, heterotopic bone formation, or osteolysis, were seldom reported.

We also identified 30 cohort studies comparing rhBMP-2 with bone graft, 47 intervention series, (17) and 34 case series or reports. Four were prospective Medtronic studies (Appendix Table 2, Studies 3, 11, 15, and 16). Most others were small and provided little information on patient characteristics. Most of the cohort studies reported baseline differences between groups or did not report baseline characteristics, had unclear blinding of outcome assessors, and failed to adjust for potential confounding.

**Anterior lumbar interbody fusion (ALIF)**

Five Medtronic-sponsored trials with IPD (Studies 1, 2, 4, 5, and 9, overall fair-quality) evaluated rhBMP-2 vs. ICBG in ALIF. Studies 1, 2, and 9 used rhBMP-2 with the approved LT-Cage or the INTER FIX devices. Studies 4 and 5 used rhBMP-2 with an off-label bone dowel. According to Medtronic, Study 5 was terminated early for business reasons with less than half of the projected sample (n = 180) enrolled (27).

The five RCTs (n=465) provided moderate strength evidence of no consistent differences between rhBMP-2 and ICBG in overall success, fusion rates, or other effectiveness measures from the immediate postoperative period through 24 months (Table 1). An exception was a 3-point higher SF-36 physical component summary score in the rhBMP-2 group at 3, 6, 12, and 24
months. At 24 months, fusion rates ranged from 60% to 100% and the average overall success rate was 61% for the rhBMP-2 group and 53% for the ICBG group.

Adverse events were common. Through 4 weeks after surgery, 38% of rhBMP-2 and 45% of ICBG patients experienced at least one adverse event; by 24 months, 80% in each group had (Table 2 and Appendix Figure 2). Meta-analysis showed no significant differences between groups for any specific adverse event including lumbar radiculitis, though estimates were frequently imprecise, precluding strong conclusions (Table 2). For retrograde ejaculation, subsidence (defined as sinking or settling of the device into bone), and urogenital problems, risk estimates favored ICBG but the differences were not statistically significant and confidence intervals were wide.

Two small single-center cohort studies found higher rates of subsidence and similar or worse fusion rates in patients who received rhBMP-2. (28, 29) Two single-center retrospective cohort studies evaluated retrograde ejaculation in ALIF. The first found a higher rate of retrograde ejaculation in rhBMP-2 patients compared with ICBG (5/69 vs. 1/174). (30) In the other, rates were similar for patients who received rhBMP-2 (4/54) and those who had an artificial disk implant without rhBMP-2 (4/41) (31).

**Posterolateral fusion (PLF) in the lumbar spine**

Four Medtronic-sponsored randomized trials with IPD (Studies 8, 12, 13, and 14, overall fair quality) and one other trial without IPD (20) evaluated rhBMP-2 for PLF. Studies 12, 13, and 14 used a higher dose and concentration of rhBMP-2 than used in ALIF trials. The non-Medtronic trial did not report dosage.
Meta-analysis based on IPD (n=722) provided moderate strength evidence of no consistent difference between rhBMP-2 and ICBG in effectiveness outcomes through 24 months (Table 1). The fusion rate at 24 months ranged from 70% to 90% in the ICBG group and 86% to 93% in the rhBMP-2 group; the rate of overall success ranged from 40% to 60% in both groups. The additional trial (20) also found no difference in fusion rates at 24 months (rhBMP-2 86% vs. ICBG 71%; RR 1.12, 95% CI 0.98 to 1.29).

Similar to ALIF, there was no significant difference between rhBMP-2 and ICBG groups in adverse events (Table 2 and Appendix Figure 2), but estimates were frequently imprecise, precluding strong conclusions. The only exception is that the rhBMP-2 group had increased risk of back and leg pain through 4 weeks, though heterogeneous events (e.g., radiculopathy, Baker’s cyst, arthritic knee pain, or ankle pain) were included and may be unrelated to fusion surgery.

Results from cohort studies (32-39) and intervention series (40-46) appeared consistent with the randomized trials, though few studies (32, 37, 38, 42) reported specific adverse events.

**Other lumbar spine fusion techniques**

We were not able to reach conclusions on effectiveness or harms of rhBMP-2 for other lumbar fusions. Except for one small Medtronic-sponsored fair-quality trial of PLIF (n=67) (Study 6), only low-quality observational studies were available (17).

**Cervical spine fusion**

In a small (n=33) Medtronic trial (Study 7), there was no difference between rhBMP-2 and ICBG in effectiveness endpoints. Three cohort studies also found no clear differences in effectiveness (11, 12, 29).
In Study 7, rhBMP-2 was associated with greater risk of adverse events than ICBG at 24 months (45 adverse events in 18 patients vs. 13 adverse events in 15 patients; RR 2.88, 95% CI 1.30 to 6.41). A large, fair quality cohort study (n=27,067) found rhBMP-2 associated with increased risk of complications (OR 1.43, 95% CI 1.12 to 1.70), dysphagia/dysphonia (OR 1.63; 95% CI 1.30 to 2.05), and wound complications (OR 1.67, 95% CI 1.10 to 2.53) (9). Smaller cohort studies (total n=346) were consistent with these results (10-12). Intervention series reported 5% to 60% of patients developed dysphagia, with differences in dysphagia definitions (47-51).

In posterior cervical spine fusion, there were no controlled trials of rhBMP-2, and one cohort study showing no difference in rates of major complications (9).

Cancer and death

Five Medtronic-sponsored trials with IPD (Studies 2, 4, 5, 10, 14) reported at least one cancer through 24 months and were included in our meta-analysis (see Appendix Table 3 for detailed information about cancer events). Compared with ICBG, rhBMP-2 was associated with increased risk of cancer (RR 3.45, 95% 1.98 to 6.00 and absolute difference 1.9 percentage points, 95% CI 0.5 to 3.2), with a number needed to harm of 53 (95% CI 31 to 200) (Figure 1). Data were insufficient to determine the effect of rhBMP-2 dose on estimates of cancer risk. Although 10 of 17 cancers with rhBMP-2 occurred in the largest high-dose trial (n=239, Study 14), another high-dose study (Study 13) reported no cancers with rhBMP-2 (n=98). At 48 months, the increased risk was no longer statistically significant (4 studies; RR 1.82; 95% CI 0.84 to 3.95).
Excluding non-SEER cancers resulted in similar estimates to those including non-SEER cancers (RR 2.92 through 24 months; 95% CI 1.75 to 4.87 and RR 1.92 through 48 months; 95% CI 0.86 to 4.32). One cohort study of 125 patients (24 rhBMP-2, 101 ICBG) reported a statistically non-significant increased risk of cancer (RR 2.10; 95% CI 0.69 to 6.41). (52) Overall, the strength of evidence is low due to sparse data.

There was no difference between rhBMP-2 and ICBG in risk of death through 24 months (9 trials, RR 0.67; 95% CI 0.28 to 1.63—Studies 2, 4, 6-10, 13-14) or 48 months (4 trials, RR 0.65; 95% CI 0.33 to 1.30—Studies 4, 10, 13-14), but the event rates were low and estimates of RR were imprecise.

**Aim #2: Quality of Reporting in Published Trials**

**On-label use in anterior lumbar interbody fusion (ALIF)**

In 2002, the FDA approved rhBMP-2 with the LT-Cage in ALIF based on three premarketing studies (Studies 1, 2, and 3) (2). The primary publications of the pivotal trials did not report the primary endpoint, overall success at 24 months (rates ranging from 50% to 60% for both groups) (Table 3, Studies 2 and 3) (6, 53, 54).

By 2004, at least 12 articles and reviews reporting results from these studies had been published in major orthopedic journals (6, 53-62). In contrast with reports to the FDA, many of these articles presented the results of the pivotal trials as demonstrating better fusion rates than ICBG. For example, the primary publication for Study 2 reiterated high fusion rates (94.5% vs. 88.7%) in the abstract, results, and conclusion sections while failing to mention that the difference was not statistically significant (6). Another publication reported results for one site in Study 3 (22 of the 137 subjects), stating a 100% rate of fusion and “improvement in back pain,
leg pain, and function”, which did not represent the overall results for the study (Table 3, Study 3) (57). Seven other Medtronic-supported articles that referred to Study 3 cited this article instead of the overall results (5, 8, 54, 56, 60, 63, 64).

In 2003, Burkus and colleagues published a post hoc “integrated analysis” that promoted the idea that rhBMP-2 would have superior outcomes compared with ICBG with sufficient sample size (54). The authors combined the rhBMP-2 groups from Study 2 and Study 3 and compared them with a control group that combined the ICBG arm of Study 2 (n=136) with an older, unrelated, unpublished series of patients (n=266) who underwent laparoscopic surgery with the LT-CAGE (54). According to an internal Medtronic report, surgeons in the unrelated study were likely less skilled with the new laparoscopic cage technique, as evidenced by longer operative times, higher blood loss, and longer hospital stays (65). The authors did not mention this concern and concluded that rhBMP-2 “had statistically superior outcomes” for these outcomes and for fusion rates. In 2004, in another journal, stated “…the outcomes represent typical results from a wide variety of surgeons with different degrees of experience…” (60).

Articles by Medtronic-associated investigators underreported adverse events in both rhBMP-2 and ICBG groups (Table 3, rows 1-3). As noted previously (13), these articles reported “no adverse events due to rhBMP-2” (56) or “no unanticipated device-related” events (5-8). In the control group, the articles emphasized “donor site hip pain,” which was assessed only in the control group patients and only on the side of the iliac crest operation. The primary publication for Study 2 represented the hip pain scores in the rhBMP-2 group as zeroes even though hip pain was not measured in that group (Figure 1 of Burkus 2002) (6). Adverse events were well-reported in a 2011 publication in which on-label rhBMP-2 was the control group (Table 3,
Study 10). It reported that 7% of rhBMP-2 patients had a serious adverse event that was “possibly device-related” (66).

**Off-label use in anterior lumbar interbody fusion (ALIF)**

Two Medtronic studies of rhBMP-2 used bone dowels, an off-label lumbar application (Table 3, Studies 4, 5). In 2002, Burkus and colleagues reported that 24 out of 24 patients (100%) receiving rhBMP-2 achieved fusion at 24 months compared with 13 out of 19 in the control group (68%) (Table 3, Study 4) (7). The larger, pivotal bone dowel trial (Study 5) was terminated early. Study 5 was published only in an article that combined the pilot and pivotal trials, representing them as “a two-part, prospective, randomized, multicenter study” with “two sequential phases.” It reported that “fusion rates were significantly better in the study group (p<0.001)” without mentioning early termination (56), as did two additional articles by the same author (63, 67). In our analysis, fusion rates for Study 5 were 91% for rhBMP-2 vs. 95% for ICBG (Table 3, Study 5).

**Off-label use in posterior lumbar interbody fusion (PLIF)**

In December, 1999 (prior to FDA approval of rhBMP-2 for use in ALIF), Medtronic suspended enrollment in Study 6, a randomized trial of rhBMP-2 in PLIF, because of ectopic bone formation in two patients, potentially leading to radiculopathy from nerve root impingement. Medtronic followed the 67 enrolled patients to 24 months. In March 2002 Medtronic requested FDA permission to terminate the study. The same year, Medtronic sponsored a supplement in the journal *Spine* in which review articles were published along with conclusions from an “international panel of experts” that included outside experts, investigators...

Downloaded From: http://annals.org/pdfaccess.ashx?url=/data/journals/aim/927097/ on 09/13/2018
associated with Medtronic, and Medtronic employees. Two articles in the supplement discussed the concern about ectopic bone formation in Study 6. While noting that large randomized trials were needed to establish the safety of rhBMP-2 in off-label procedures, the supplement argued that ectopic bone formation, and complications it might cause, were due to poor technique (59, 61). No data from Study 6 were presented to support this argument. The international panel stated “when used properly, BMPs currently appear to be extremely safe for spine fusion” (68).

After Study 6 was terminated, an article published in 2004 (Table 3, Study 6) (8) reported data on ectopic bone formation (rhBMP-2 24/34 vs. ICBG 4/33, p<0.001) for the first time. Despite the small sample, the authors emphasized the lack of association between ectopic bone formation and leg pain and gave an incomplete account of the reasons for study termination (13, 69).

**Off-label use in posterolateral lumbar fusion (PLF)**

There were six studies of rhBMP-2 in PLF. Three (Studies 8, 12, 14) were published (70-72). Based on IPD from Study 14 (72), rhBMP-2 and ICBG did not differ in rates of overall success (56 % vs. 56%) and fusion (90% vs. 90%). In contrast, the journal publication and FDA summary reported that use of rhBMP-2 resulted in a higher fusion rate (96% vs. 89%, \( P=0.014 \)) (72, 73). This difference may be due to our classification of patients with partial data as failures, though it is not clear why this would differentially affect the rhBMP-2 group. Adverse events were underreported in two publications (Table 3, studies 8 and 12) (70, 71) but well-reported in the largest trial (Table 3, study 14) (72).
Discussion

In spinal fusion, rhBMP-2 and ICBG appear to be similarly effective when used in ALIF and PLF, though the current evidence does not allow definitive conclusions regarding the effectiveness in other surgical approaches. The SF-36 physical component summary scores were slightly better with rhBMP-2 than with ICBG in ALIF patients through 24 months, but the difference was only 2 to 3 points on a 100 point scale, failing to meet typical criteria for a clinically meaningful difference (74).

The use of rhBMP-2 in anterior cervical spine fusion was associated with statistically significant increases in overall adverse events, wound complications, and dysphagia or dysphonia. For lumbar fusion—both on-label and off-label—adverse events were common with both rhBMP-2 and ICBG. Although our analyses raises concerns regarding a possible increased risk of retrograde ejaculation, urinary retention, osteolysis/instability, subsidence, and heterotopic bone formation with rhBMP-2, the data on these harms were sparse and the quality of ascertainment was often poor. Our analysis underscores the need for more definitive evidence about harms before rhBMP-2 became widely used.

RhBMP-2 was associated with an increased risk of cancer, with or without non-SEER-reportable cancers. This finding should be interpreted with caution, as cancers were heterogeneous and, according to Medtronic, underreported (26). Seven Medtronic-sponsored trials (total n = 429) with zero cancer events in both groups were not included in the meta-analysis but were not expected to affect the results (17). Animal studies do not suggest that rhBMP-2 is carcinogenic (61), but BMPs are expressed by and promote the growth of some cancers (75-77). The development of cancers within 2 to 4 years also argues for a pro-oncogenic mechanism.
For both on-label and off-label indications, journal publications selected analyses and results that favored rhBMP-2 over ICBG. In their review, Carragee et al. demonstrated underreporting of adverse events in publications of five trials for which the FDA had made summary results public (13). Our study demonstrates that there was also underreporting of adverse events for off-label uses with results not previously available to the public. Journal practices regarding sponsored supplements, trial registration, and conflict of interest disclosure may have contributed to publication of an incomplete and sometimes misleading evidence base (78-80).

Meta-analysis of IPD offers several advantages over traditional, study-level meta-analysis (81). Compared with other reviews,(15, 82) we had a more complete and standardized evaluation of outcomes, with data from unpublished studies as well as data unreported or incompletely reported by published studies, thus reducing potential effects of publication and reporting bias. Additionally, we could recalculate and re-categorize outcome measures using consistent definitions, adjust for potential baseline imbalances, and perform sensitivity analyses to handle missing data.

Nevertheless, meta-analysis based on IPD requires substantially more time and resources than traditional study-level meta-analysis, and availability of IPD cannot compensate for flawed data collection or sparse data. Even with IPD on 1,879 patients from 12 trials, one additional trial, and many observational studies, the evidence base remains relatively small within each surgical approach. We found no published trials truly independent of the manufacturer. Additionally, there has been no prospective, well-designed, adequately-powered study specifically aimed to assess important harms using adequate ascertainment methods.
There was also insufficient information to adequately evaluate the effects of dose on risk of harms. Eleven Medtronic studies (Studies 1-11) used rhBMP-2 at a concentration of 1.5mg/mL, with total doses ranging from 0.6-16.8 mg. Higher and unapproved concentrations of rhBMP-2 (2.0-3.0 mg/mL) were used in five of the six PLF studies, with total doses ranging from 15.0-63.0 mg. Determining the effects of rhBMP-2 dosage was not possible due to differences in surgical approach, rhBMP-2 carrier, and fusion hardware.

Although we had unusual access to protocols and documents submitted by the manufacturer to the FDA, other information, such as operative notes and internal correspondence, might have helped assess the extent of design and reporting bias. Internal correspondence is essential to evaluate selective analysis reporting, ghostwriting, time lag bias, and misrepresentation of facts (25). Finally, we were not able to evaluate the integrity of adverse event adjudication.

In conclusion, we found substantial evidence of reporting bias and no evidence that rhBMP-2 is more effective than ICBG in spinal fusion, with some evidence of an association with important harms. More research is needed to provide more reliable estimates of risk of cancer and other adverse events and to identify patient populations in which use of rhBMP-2 may be beneficial, such as cases where use of bone graft alone is associated with a high risk of pseudoarthrosis. Based on the currently available evidence, it is difficult to identify clear indications for rhBMP-2 in spinal fusion.
References


38. **Taghavi CE, Lee K-B, Keorochana G, Tzeng S-T, Yoo JH, Wang JC.** Bone morphogenetic protein-2 and bone marrow aspirate with allograft as alternatives to autograft in instrumented revision posterolateral lumbar spinal fusion: a


47. Lanman TH, Hopkins TJ. Early findings in a pilot study of anterior cervical interbody fusion in which recombinant human bone morphogenetic protein-2 was used with poly(L-lactide-co-D,L-lactide) bioabsorbable implants. Neurosurg Focus. 2004;16(3):E6. [PMID:15198494]


65. **Medtronic.** II B: Report of pivotal clinical trial results (G960065) laparoscopic use of InFUSE bone graft/LT-CAGE lumbar tapered fusion device. Internal document:2072-423.


Acknowledgments: The authors acknowledge Robin Paynter, MLIS, for conducting literature searches; Howard Balshem, MS, Susan Carson, MPH, Elaine Graham, MLS, Allison Lowe, BA, Edwin Reid, MS, MAT, Katie Reitel, MPH, MSW, Sujata Thakurta, MA, Ngoc Wasson, MPH, and Leah Williams, BS, for their contributions to this article.

Figure Legends

Figure 1. Comparison of Cancer Risk Between rhBMP-2 Versus Control
Shows a forest plot of the comparison of cancer risk in 5 studies at 24 months and in 4 studies at 48 months.
Figure footnote:
*The combined risk ratio (RR) was obtained using a generalized linear fixed effects model with binomial distribution and log link without correction for zero events. The RR from each study was estimated, when there is zero event, by adding a continuity correction of 0.5, for illustrative purposes.

Appendix Figure 1. Literature Flow Chart.
Shows the literature retrieved and the numbers of abstract records and full-text articles excluded and included.
Figure footnote:
ALIF = anterior lumbar interbody fusion, PLF = posterolateral lumbar fusion, RCT = randomized controlled trial

Appendix Figure 2. Cumulative Proportion of Patients With At Least One Adverse Event: (a) ALIF and (b) PLF.
Shows the cumulative proportion of patients with at least one adverse event for ALIF and PLF at time points from operation.
Figure footnote:
AE = adverse event; ALIF = anterior lumbar interbody fusion; ICBG = iliac crest bone graft; PLF = posterolateral lumbar fusion; rhBMP-2 = recombinant human bone morphogenetic protein-2; SAE = serious adverse event.
*There was no significant difference between the rhBMP-2 versus ICBG groups at any time point for either outcome and surgery approach.
Figure 1.

<table>
<thead>
<tr>
<th>Study (Number)</th>
<th>Control Rate</th>
<th>Risk Ratio (95% CI)</th>
<th>Events, rhBMP-2</th>
<th>Events, Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>24 month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infuse LT Cage Pivotal (2)</td>
<td>0.74%</td>
<td>1.90 (0.17, 20.74)</td>
<td>2/143</td>
<td>1/136</td>
</tr>
<tr>
<td>Infuse Bone Dowel Pilot (4)</td>
<td>0%</td>
<td>2.76 (0.12, 64.41)</td>
<td>1/24</td>
<td>0/22</td>
</tr>
<tr>
<td>Infuse Bone Dowel Pivotal (5)</td>
<td>0%</td>
<td>1.66 (0.07, 39.55)</td>
<td>1/55</td>
<td>0/30</td>
</tr>
<tr>
<td>Maverick Disc Pivotal (10)</td>
<td>0.74%</td>
<td>2.35 (0.48, 11.55)</td>
<td>3/172</td>
<td>3/405</td>
</tr>
<tr>
<td>Amplify Pivotal (14)</td>
<td>0.89%</td>
<td>4.69 (1.04, 21.15)</td>
<td>10/239</td>
<td>2/224</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td></td>
<td><strong>3.45 (1.98, 6.00)</strong>*</td>
<td>17/633</td>
<td>6/817</td>
</tr>
<tr>
<td><strong>48 month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infuse Bone Dowel Pilot (4)</td>
<td>0%</td>
<td>2.76 (0.12, 64.41)</td>
<td>1/24</td>
<td>0/22</td>
</tr>
<tr>
<td>Maverick Disc Pivotal (10)</td>
<td>1.24%</td>
<td>1.41 (0.34, 5.85)</td>
<td>3/172</td>
<td>5/405</td>
</tr>
<tr>
<td>BCP Canada (12)</td>
<td>2.04%</td>
<td>0.34 (0.01, 8.15)</td>
<td>0/48</td>
<td>1/49</td>
</tr>
<tr>
<td>Amplify Pivotal (14)</td>
<td>2.23%</td>
<td>2.25 (0.81, 6.28)</td>
<td>12/239</td>
<td>5/224</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td></td>
<td><strong>1.82 (0.84, 3.95)</strong>*</td>
<td>16/483</td>
<td>11/700</td>
</tr>
</tbody>
</table>

Favor rhBMP-2   Favor ICBG
Appendix Figure 1.

14,697 records identified from database searches after removal of duplicates

45 additional records identified through other sources

14,742 records screened

13,969 records excluded at abstract level

773 full-text articles assessed for eligibility

636 full-text articles excluded
  • 44 non-English language
  • 75 ineligible outcome
  • 66 ineligible intervention
  • 16 ineligible population
  • 353 ineligible publication type
    o 338 general
    o 7 RCTs, abstract only
    o 2 pending trials*
  • 4 ineligible study design, including 1 Medtronic RCT with only three subjects
  • 81 ineligible area of the body

*1 active (NCT01013389, Actifuse ABX versus INFUSE in Posterolateral Instrumented Lumbar Fusion (PLF) with Interbody Fusion) and 1 completed, results not found (NCT00405600, Spine Fusion Instrumented with BMP-2 vs Uninstrumented with Infuse BMP-2 Alone)

Included Articles:

125 studies reported in 136 publications included in qualitative synthesis, plus Individual Patient Data provided by Medtronic

By study design:
  • 13 randomized controlled trials reported in 18 publications, including data from 2 unpublished trials
  • 30 cohort studies reported in 31 publications
  • 47 intervention series in 50 publications, including data from 3 unpublished trials
  • 34 case report/series in 35 publications
  • 1 combination studies reported in 2 publications

Included studies by approach and study design (sponsorship, approach if applicable):

<table>
<thead>
<tr>
<th>Trials</th>
<th>ALIF</th>
<th>PLF</th>
<th>Anterior Cervical</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 (Medtronic)</td>
<td>4 (Medtronic)</td>
<td>1 (Medtronic)</td>
<td>1 (Medtronic)</td>
<td>1 (Medtronic, PLIF)</td>
</tr>
<tr>
<td>4 (Medtronic)</td>
<td>1 (Other)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 (Other)</td>
<td>5 (Other)</td>
<td>7 (Other)</td>
<td>1 (Medtronic, PLIF)</td>
<td>28 (Other)</td>
</tr>
</tbody>
</table>

Downloaded From: http://annals.org/pdfaccess.ashx?url=/data/journals/aim/927097/ on 09/13/2018
Appendix Figure 2.

(a) ALIF

(b) PLF

Cumulative proportion of patients having at least one adverse event

Time period:
- Operative
- Postoperative
- 6 weeks
- 3 months
- 6 months
- 12 months
- 24 months

Legend:
- At least one AE, ICBG group
- At least one AE, rhBMP-2 group
- At least one SAE, ICBG group
- At least one SAE, rhBMP-2 group
Appendix 1. Search strategies

Searches were repeated in June 2012 to identify additional citations.

Database: Ovid MEDLINE(R) and Ovid OLDMEDLINE(R) <1946 to December Week 4, 2011>, Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations January 09, 2012

<table>
<thead>
<tr>
<th></th>
<th>Search Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bone Morphogenetic Protein 2/</td>
<td>3737</td>
</tr>
<tr>
<td>2</td>
<td>Bone Morphogenetic Proteins/</td>
<td>9745</td>
</tr>
<tr>
<td>3</td>
<td>bone morphogen$ protein-2.ti,ab.</td>
<td>2737</td>
</tr>
<tr>
<td>4</td>
<td>bone morphogen$ protein-ii.ti,ab.</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>bone morphogen$ protein2.ti,ab.</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>human recombinant BMP-2.ti,ab.</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>human recombinant BMP2.ti,ab.</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>recombinant human BMP-2.ti,ab.</td>
<td>146</td>
</tr>
<tr>
<td>9</td>
<td>recombinant human BMP2.ti,ab.</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>recombinant human bone morphogen$ protein-2.ti,ab.</td>
<td>793</td>
</tr>
<tr>
<td>11</td>
<td>recombinant human bone morphogenetic protein-2.rn.</td>
<td>749</td>
</tr>
<tr>
<td>12</td>
<td>BMP.ti,ab.</td>
<td>10098</td>
</tr>
<tr>
<td>13</td>
<td>BMPs.ti,ab.</td>
<td>2953</td>
</tr>
<tr>
<td>14</td>
<td>BMP-2.ti,ab.</td>
<td>3108</td>
</tr>
<tr>
<td>15</td>
<td>BMP-ii.ti,ab.</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>BMP2.ti,ab.</td>
<td>1373</td>
</tr>
<tr>
<td>17</td>
<td>rhBMP-2.ti,ab.</td>
<td>1106</td>
</tr>
<tr>
<td>18</td>
<td>rhBMP2.ti,ab.</td>
<td>62</td>
</tr>
<tr>
<td>19</td>
<td>rhBMP.ti,ab.</td>
<td>1230</td>
</tr>
<tr>
<td>20</td>
<td>rhBMPs.ti,ab.</td>
<td>34</td>
</tr>
<tr>
<td>21</td>
<td>rh-BMP.ti,ab.</td>
<td>62</td>
</tr>
<tr>
<td>22</td>
<td>hrBMP-2.ti,ab.</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>hr-BMP.ti,ab.</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>(infuse adj10 bone$).ti,ab.</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>---</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>'bone morphogenetic protein 2'/de</td>
<td>'bone morphogenetic protein'/de</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>'bone morphogenetic protein 2'/de</td>
<td>'bone morphogenetic protein'/de</td>
</tr>
</tbody>
</table>
'dibotermin alfa':tn
1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30


#31 and (1996:py or 1997:py or 1998:py or 1999:py or 2000:py or 2001:py or 2002:py or 2003:py or 2004:py or 2005:py or 2006:py or 2007:py or 2008:py or 2009:py or 2010:py or 2011:py or 2012:py) and 'human'/de and [embase]/lim not ('nonhuman'/de or 'animal model'/de or 'animal cell'/de or 'animal tissue'/de or 'human'/de)

Database: Ovid EBM Reviews - Cochrane Central Register of Controlled Trials <4th Quarter 2011>, Cochrane Database of Systematic Reviews <2005 to December 2011>, Database of Abstracts of Reviews of Effects <4th Quarter 2011>, Health Technology Assessment <4th Quarter 2011>

Search Strategy:
--------------------------------------------------------------------
1 bone morphogen$ protein-2.ti,ab.  37
2 recombinant human BMP-2.ti,ab.  1
3 recombinant human bone morphogen$ protein-2.ti,ab.  36
4 BMP.ti,ab.  67
5 BMPs.ti,ab.  9
6 BMP-2.ti,ab.  12
7 BMP2.ti,ab.  3
8 rhBMP-2.ti,ab.  49
9 rhBMP.ti,ab.  60
10 (infuse adj10 bone$).ti,ab.  9
11 Dibotermin alfa.ti,ab.  1
12 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11  127
13 limit 12 to yr="1996 -Current" [Limit not valid in DARE; records were retained]  123
--------------------------------------------------------------------

Database: Sciverse Scopus <01/11/2012>

Search Strategy:
--------------------------------------------------------------------
(TITLE-ABS-KEY({bone morphogen* protein*}) or bmp or bmp-2 or {BMP 2} or bmp-ii or {BMP ii} or bmpii or rhbmp or rhbmp-2 or rhbmp2 or {rhBMP 2} or rhbmp-ii or rhbmpii or {rhBMP ii} or rh-bmp or rh-bmp-2 or rh-bmp2 or {rh-BMP 2} or rh-bmp-ii or rh-bmpii or
Database: clinicaltrials.gov <01/11/2012>

Search Strategy:

"bone morphogenetic protein*" or BMP* or rh-BMP* or “Infuse Bone” or "Dibotermin alfa" or InductOS | Closed Studies | received from 01/01/1996 to 01/11/2012 (63)

Database: World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP), http://apps.who.int/trialsearch/AdvSearch.aspx <01/11/2012> (38)

Search Strategy:

Bone morphogenetic protein or bone morphogenic protein or BMP or rhBMP or rhBMP-2 or rhBMP2 or rhBMP-ii or rhBMPii or rh-BMP or rh-BMP-2 or rh-BMP2 or rh-BMP-ii or rh-BMPii or hrBMP or hr-BMP or hr-BMP-2 or hr-BMP2 or hr-BMP-ii or hr-BMPii or INFUSE Bone or Infuse Bone or InductOS or INDUCTOS or dibotermin alfa or Dibotermin Alfa AND date = 01/01/1996-11/01/2012

Database: Current Controlled Clinical Trials (ISRCTN Register), http://www.controlled-trials.com/isrctn/ <01/11/2012> (10)

Search Strategy:

Bone morphogenetic protein or bone morphogenetic proteins or bone morphogenic protein or bone morphogenetic proteins or BMP or BMPs or rhBMP or rhBMPs or rhBMP-2 or rhBMP2 or rhBMP-ii or rhBMPii or rh-BMP or rh-BMP-2 or rh-BMP2 or rh-BMP-ii or rh-BMPii or hrBMP or hr-BMPs or hrBMP-2 or hrBMP2 or hr-BMP-ii or hr-BMPii or hr-BMP or hr-BMPs or hrBMP-2 or hrBMP2 or hr-BMP-ii or hr-BMPii or Infuse Bone or InductOS or dibotermin alfa
Appendix 2: SAS Code

/*Sample SAS code for linear mixed effects model for combining continuous outcomes – Here Oswestry disability score (OSSCORE) is the outcome*/

proc mixed data = ALIF_combined;
class period study treat;
model OSSCORE = OSSCORE_b study treat/solution CL DDFM=RESIDUAL;
   /* OSSCORE_b is the baseline score of OSSCORE;
      study is the identification variable for each study;
      treat identifies rhBMP-2 vs. ICBG group. */
   random trtgrp /subject = study;   /* Specify random treatment effect */
repeated /group = study;   /* Specify heterogeneous residual variance for included studies */
where period = 7;   /* A separate model was fit for each follow-up time */
run;

/*Sample SAS code for generalized linear mixed effects model for combining common binary outcomes */
ods output ParameterEstimates = ParameterEstimates;
proc nlmixed data = ATLEASTONEAE;
   parms beta0= -0.3 beta1 = -0.2 sigma = 0.4;
   if treat = 0 then eta = beta0;
   if treat = 1 then eta = beta0 + beta1 + u;
   expeta = exp(eta); /* This corresponds to a log link */
   model Xevents ~ binomial(Nevents,expeta);
   random u ~ normal(0,sigma * sigma) subject=study;
where period = "Four Weeks";
   /* A separate model was fit for each follow-up time */
run;

data ParameterEstimates;
   set ParameterEstimates;
   RR = exp(Estimate);
   RR_low_normal = exp(Estimate - 1.96* StandardError);
   RR_upp_normal = exp(Estimate + 1.96* StandardError);
   /*Calculate 95% confidence interval based on normal approximation*/
run;

/*Sample SAS code for generalized linear fixed effects model for combining rare binary outcomes */
proc genmod data = cancer;
class study;
model Xevents/Nevents = study treat/dist = binomial link = log scale = deviance;
   /* A log link was used to produce a risk ratio;
      The option for scale = is used to correct over- or under-dispersion if necessary*/
estimate "RR for rhBMP-2 vs. ICBG" treat 1 ;
where period = "24mon";
    /* A separate model was fit for each follow-up time */
run;
### Table 1. Effectiveness Endpoints for ALIF and PLF With rhBMP-2 Versus ICBG

**Anterior lumbar interbody fusion***

<table>
<thead>
<tr>
<th>Endpoint (Scale)</th>
<th>6 weeks</th>
<th>3 months</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall success</strong></td>
<td>----</td>
<td>----</td>
<td>1.18 (0.93 to 1.50)</td>
<td>1.12 (0.95 to 1.33)</td>
<td>1.19 (0.99 to 1.42)</td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td>445 (4)</td>
<td>436 (4)</td>
<td>418 (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fusion</strong></td>
<td>----</td>
<td>----</td>
<td>1.10 (1.02 to 1.19)†</td>
<td>1.09 (0.95 to 1.24)†</td>
<td>1.05 (0.88 to 1.24)†</td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td>446 (5)</td>
<td>439 (5)</td>
<td>416 (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Neurological success</strong></td>
<td>1.02 (0.93 to 1.13)</td>
<td>1.06 (0.97 to 1.16)</td>
<td>1.01 (0.91 to 1.12)</td>
<td>1.04 (0.94 to 1.14)</td>
<td>1.08 (0.98 to 1.19)</td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td>434 (4)</td>
<td>442 (4)</td>
<td>433 (4)</td>
<td>420 (4)</td>
<td>400 (4)</td>
</tr>
<tr>
<td><strong>ODI success</strong></td>
<td>1.04 (0.83 to 1.29)</td>
<td>1.03 (0.87 to 1.23)</td>
<td>1.09 (0.95 to 1.25)</td>
<td>1.03 (0.92 to 1.15)</td>
<td>1.10 (0.97 to 1.24)</td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td>442 (4)</td>
<td>455 (5)</td>
<td>450 (5)</td>
<td>436 (5)</td>
<td>417 (5)</td>
</tr>
<tr>
<td><strong>Return to work‡</strong></td>
<td>1.21 (0.71 to 2.05)</td>
<td>0.97 (0.70 to 1.32)</td>
<td>1.02 (0.89 to 1.17)</td>
<td>1.01 (0.90 to 1.14)</td>
<td>1.06 (0.94 to 1.19)</td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td>211 (4)</td>
<td>210 (4)</td>
<td>207 (4)</td>
<td>201 (4)</td>
<td>196 (4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Weighted mean difference (95% CI)</strong></th>
<th><strong>Sample Size, n (Studies)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ODI (0-50)$§</strong></td>
<td>-2.36 (-6.91 to 2.19)</td>
</tr>
<tr>
<td><strong>Back pain (0-10)$§</strong></td>
<td>0.21 (-0.28 to 0.71)</td>
</tr>
<tr>
<td><strong>Leg pain (0-10)$§</strong></td>
<td>-0.57 (-1.12 to -0.02)</td>
</tr>
<tr>
<td><strong>SF-36® PCS (0-100)$¶</strong></td>
<td>0.55 (-1.02 to 2.11)</td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td>461 (5)</td>
</tr>
</tbody>
</table>

† p < 0.05; ‡ p < 0.01; § p < 0.001; ¶ p < 0.0001.
### Posterolateral lumbar fusion*

<table>
<thead>
<tr>
<th>Endpoint (Scale)</th>
<th>6 weeks</th>
<th>3 months</th>
<th>6 months</th>
<th>12 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk Ratio (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall success</td>
<td>----</td>
<td>----</td>
<td>1.34 (1.10 to 1.64)</td>
<td>1.07 (0.93 to 1.25)</td>
<td>1.05 (0.91 to 1.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>698 (4)</td>
<td>687 (4)</td>
<td>648 (4)</td>
</tr>
<tr>
<td>Fusion</td>
<td>----</td>
<td>----</td>
<td>1.37 (1.19 to 1.59)</td>
<td>1.29 (0.94 to 1.78)†</td>
<td>1.16 (0.96 to 1.41)†</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>694 (4)</td>
<td>686 (4)</td>
<td>637 (4)</td>
</tr>
<tr>
<td>Neurological success</td>
<td>1.03 (0.94 to 1.13)</td>
<td>1.0 (0.93 to 1.08)</td>
<td>1.02 (0.96 to 1.09) †</td>
<td>1.01 (0.95 to 1.07)</td>
<td>1.01 (0.92 to 1.10)</td>
</tr>
<tr>
<td></td>
<td>706 (4)</td>
<td>705 (4)</td>
<td>693 (4)</td>
<td>683 (4)</td>
<td>636 (4)</td>
</tr>
<tr>
<td>ODI success</td>
<td>1.00 (0.81 to 1.23)</td>
<td>1.03 (0.91 to 1.17)</td>
<td>1.07 (0.98 to 1.17)</td>
<td>1.01 (0.91 to 1.11)</td>
<td>1.01 (0.91 to 1.12)</td>
</tr>
<tr>
<td></td>
<td>707 (4)</td>
<td>704 (4)</td>
<td>693 (4)</td>
<td>683 (4)</td>
<td>640 (4)</td>
</tr>
<tr>
<td>Return to work‡</td>
<td>1.26 (0.71 to 2.21)</td>
<td>1.09 (0.85 to 1.40)</td>
<td>0.87 (0.67 to 1.14)</td>
<td>1.07 (0.96 to 1.19)</td>
<td>1.03 (0.94 to 1.14)</td>
</tr>
<tr>
<td></td>
<td>233 (3)</td>
<td>232 (3)</td>
<td>225 (3)</td>
<td>227 (3)</td>
<td>208 (3)</td>
</tr>
<tr>
<td><strong>Weighted mean difference (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sample Size, n (Studies)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ODI (0-50)§</strong></td>
<td>0.74 (-1.68 to 3.17)</td>
<td>-1.97 (-4.36, 0.42)</td>
<td>-2.40 (-4.85 to 0.04)</td>
<td>-2.09 (-5.28, 1.10)</td>
<td>-1.98 (-4.86 to 0.90)</td>
</tr>
<tr>
<td></td>
<td>718 (4)</td>
<td>714 (4)</td>
<td>703 (4)</td>
<td>694 (4)</td>
<td>650 (4)</td>
</tr>
<tr>
<td><strong>Back pain (0-10)§</strong></td>
<td>0.10 (-0.27 to 0.48)</td>
<td>-0.25 (-0.62 to 0.12)</td>
<td>-0.46 (-1.14 to 0.23)</td>
<td>-0.42 (-1.34 to 0.50)</td>
<td>-0.31 (-0.76 to 0.15)</td>
</tr>
<tr>
<td></td>
<td>716 (4)</td>
<td>713 (4)</td>
<td>702 (4)</td>
<td>693 (4)</td>
<td>649 (4)</td>
</tr>
<tr>
<td><strong>Leg pain (0-10)§</strong></td>
<td>0.23 (-0.21 to 0.66)</td>
<td><strong>-0.44 (-0.87 to -0.01)</strong></td>
<td>-0.27 (-0.71 to 0.17)</td>
<td>-0.29 (-0.75 to 0.16)</td>
<td>-0.34 (-0.82 to 0.13)</td>
</tr>
<tr>
<td></td>
<td>715 (4)</td>
<td>712 (4)</td>
<td>701 (4)</td>
<td>692 (4)</td>
<td>648 (4)</td>
</tr>
<tr>
<td><strong>SF-36® PCS (0-100)‖</strong></td>
<td>-0.10 (-1.15 to 0.96)</td>
<td>0.64 (-0.68 to 1.96)</td>
<td><strong>1.79 (0.27 to 3.31)</strong></td>
<td>1.83 (-0.19 to 3.85)</td>
<td>1.10 (-0.65 to 2.86)</td>
</tr>
<tr>
<td></td>
<td>709 (4)</td>
<td>708 (4)</td>
<td>696 (4)</td>
<td>689 (4)</td>
<td>644 (4)</td>
</tr>
<tr>
<td>SF-36® MCS (0-100)</td>
<td>0.52 (-0.94 to 1.98)</td>
<td>-0.05 (-1.59 to 1.50)</td>
<td>0.06 (-1.48 to 1.60)</td>
<td>-0.50 (-2.56 to 1.57)</td>
<td>0.54 (-3.16 to 4.25)</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td>709 (4)</td>
<td>708 (4)</td>
<td>696 (4)</td>
<td>689 (4)</td>
<td>644 (4)</td>
</tr>
</tbody>
</table>

ALIF = anterior lumbar interbody fusion, rhBMP-2 = recombinant human bone morphogenetic protein-2, ODI = Oswestry Disability Index, PCS = physical component summary, MCS = mental component summary, PLF = posterolateral lumbar fusion

Values in bold font are significant at 0.05 level.

* For ALIF, a total n = 465 was included in the analysis, excluding 4 patients who underwent open surgery in study 1; for PLF, a total n = 722 was included in the analysis, excluding 11 patients randomized to rhBMP-2 without instrumentation in study 12.

† These combined estimates were obtained using a two-stage approach.

‡ Includes only patients who worked before surgery. For ALIF, 221 patients worked before surgery; for PLF, the number was 241.

§ For ODI, back pain, and leg pain, high values represent worse outcomes and a negative difference favors rhBMP-2.

‖ For SF-36® PCS and MCS, low values represent worse outcomes and a positive difference favors rhBMP-2.
Table 2. Overall and Specific Adverse Events in Randomized Controlled Trials of ALIF and PLF with rhBMP-2 Versus ICBG

<table>
<thead>
<tr>
<th>Event type†</th>
<th>≤ 4 weeks</th>
<th>≤ 24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
</tr>
<tr>
<td>Anterior lumbar interbody fusion*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 1 Adverse event, any type</td>
<td>38% vs. 45%</td>
<td>0.84 (0.61 to 1.17)</td>
</tr>
<tr>
<td>≥ 1 Serious adverse event</td>
<td>9% vs. 8%</td>
<td>1.12 (0.72 to 1.74)</td>
</tr>
<tr>
<td>≥ 1 device-related adverse event</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Specific adverse events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomical/technical difficulty</td>
<td>0.9% vs. 3%</td>
<td>0.22 (0.04 to 1.05)</td>
</tr>
<tr>
<td>Back and/or leg pain</td>
<td>4% vs. 3%</td>
<td>1.05 (0.31 to 3.62)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>2% vs. 4%</td>
<td>0.56 (0.16 to 1.92)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>13% vs. 15%</td>
<td>0.86 (0.54 to 1.36)</td>
</tr>
<tr>
<td>Implant problems</td>
<td>2% vs. 1%</td>
<td>1.07 (0.10 to 11.75)</td>
</tr>
<tr>
<td>Infection (all types)</td>
<td>6% vs. 5%</td>
<td>1.10 (0.49 to 2.46)</td>
</tr>
<tr>
<td>Neurological</td>
<td>3% vs. 4%</td>
<td>0.81 (0.29 to 2.27)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (primary) ‡</td>
<td>3% vs. 3%</td>
<td>1.02 (0.35 to 2.99)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 2) ‡</td>
<td>2% vs. 3%</td>
<td>0.49 (0.11 to 2.07)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 3) ‡</td>
<td>3% vs. 3%</td>
<td>0.85 (0.23 to 3.04)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 4) ‡</td>
<td>0.8% vs. 2%</td>
<td>0.35 (0.07 to 1.78)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>2% vs. 3%</td>
<td>0.55 (0.21 to 1.41)</td>
</tr>
<tr>
<td>Event type†</td>
<td>≤ 4 weeks</td>
<td>≤ 24 months</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
</tr>
<tr>
<td>Retrograde ejaculation</td>
<td>4% vs. 1%</td>
<td>2.62 (0.28 to 24.56)</td>
</tr>
<tr>
<td>Spinal event</td>
<td>0% vs. 2%</td>
<td>0/167 vs. 3/158</td>
</tr>
<tr>
<td>Subsidence</td>
<td>2% vs. 1%</td>
<td>1.43 (0.24 to 8.41)</td>
</tr>
<tr>
<td>Urogenital</td>
<td>7% vs. 4%</td>
<td>1.96 (0.61 to 6.34)</td>
</tr>
<tr>
<td>Vertebral fracture</td>
<td>1% vs. 0%</td>
<td>2/168 vs. 0/156</td>
</tr>
<tr>
<td>Urinary retention‡</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>Wound infection‡</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>Wound dehiscence‡</td>
<td>---</td>
<td>----</td>
</tr>
<tr>
<td>Relevant additional surgeries</td>
<td>---</td>
<td>----</td>
</tr>
</tbody>
</table>

**Posterolateral lumbar fusion§**

**Overall adverse events**

| ≥ 1 Adverse event, any type | 51% vs. 49% | 0.99 (0.74 to 1.34) | 722 (4) | 88% vs. 87% | 1.02 (0.96 to 1.07) | 722 (4) |
| ≥ 1 Serious adverse event | 20% vs. 23% | 0.89 (0.67 to 1.18) | 722 (4) | 50% vs. 52% | 0.96 (0.83 to 1.11) | 722 (4) |
| ≥ 1 device-related adverse event | --- | ---- | ---- | 6% vs. 5% | 1.36 (0.57 to 3.23) | 722 (4) |

**Specific adverse events**

Anatomical/technical difficulty | 1% vs. 0% | 4/337 vs. 0/323 | 660 (2) | Same as four weeks |
<p>| Back and/or leg pain | 8% vs. 4% | 1.83 (1.15 to 2.93) | 706 (3) | 49% vs. 42% | 1.18 (0.91 to 1.52) | 722 (4) |
| Cardiovascular | 14% vs. 14% | 0.85 (0.40 to 1.81) | 706 (3) | 19% vs. 21% | 0.90 (0.57 to 1.40) | 722 (4) |
| Dural injury | 6% vs. 7% | 0.76 (0.55 to 1.04) | 722 (4) | 6% vs. 8% | 0.79 (0.50 to 1.32) | 722 (4) |</p>
<table>
<thead>
<tr>
<th>Event type†</th>
<th>≤ 4 weeks</th>
<th></th>
<th></th>
<th>≤ 24 months</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
<td>Sample Size, n (Studies)</td>
<td>Patients with rhBMP-2 vs. ICBG</td>
<td>Risk Ratio (95% CI)</td>
<td>Sample Size, n (Studies)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>7% vs. 10%</td>
<td>0.71 (0.36 to 1.44)</td>
<td>722 (4)</td>
<td>16% vs. 18%</td>
<td>0.88 (0.64 to 1.21)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Implant problems</td>
<td>2% vs. 0.6%</td>
<td>2.83 (0.87 to 9.26)</td>
<td>706 (3)</td>
<td>3% vs. 2%</td>
<td>1.58 (0.58 to 4.29)</td>
<td>706 (3)</td>
</tr>
<tr>
<td>Infection (all types)</td>
<td>9% vs. 10%</td>
<td>0.99 (0.57 to 1.73)</td>
<td>706 (3)</td>
<td>18% vs. 19%</td>
<td>0.96 (0.71 to 1.31)</td>
<td>706 (3)</td>
</tr>
<tr>
<td>Neurological</td>
<td>5% vs. 3%</td>
<td>1.53 (0.88 to 2.65)</td>
<td>722 (4)</td>
<td>26% vs. 23%</td>
<td>0.97 (0.62 to 1.51)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Primary) ‡</td>
<td>3% vs. 2%</td>
<td>1.30 (0.69 to 2.46)</td>
<td>722 (4)</td>
<td>24% vs. 26%</td>
<td>0.95 (0.73 to 1.22)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 2) ‡</td>
<td>3% vs. 2%</td>
<td>1.65 (0.62 to 4.40)</td>
<td>722 (4)</td>
<td>14% vs. 15%</td>
<td>0.90 (0.54 to 1.51)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 3) ‡</td>
<td>3% vs. 3%</td>
<td>1.32 (0.73 to 2.38)</td>
<td>722 (4)</td>
<td>24% vs. 26%</td>
<td>0.91 (0.71 to 1.18)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Possible lumbar radiculitis (Definition 4) ‡</td>
<td>2% vs. 1%</td>
<td>1.54 (0.45 to 5.20)</td>
<td>455 (4)</td>
<td>10% vs. 11%</td>
<td>0.89 (0.42 to 1.87)</td>
<td>455 (4)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>4% vs. 3%</td>
<td>1.38 (0.42 to 4.55)</td>
<td>706 (3)</td>
<td>7% vs. 5%</td>
<td>1.44 (0.87 to 2.39)</td>
<td>706 (3)</td>
</tr>
<tr>
<td>Spinal event</td>
<td>1% vs. 1%</td>
<td>1.05 (0.21 to 5.17)</td>
<td>676 (3)</td>
<td>9% vs. 10%</td>
<td>0.89 (0.61 to 1.29)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Urogenital</td>
<td>7% vs. 7%</td>
<td>1.03 (0.53 to 2.01)</td>
<td>722 (4)</td>
<td>13% vs. 12%</td>
<td>1.04 (0.60 to 1.82)</td>
<td>722 (4)</td>
</tr>
<tr>
<td>Vertebral fracture</td>
<td>2% vs. 0.9%</td>
<td>1.26 (0.23 to 6.94)</td>
<td>660 (2)</td>
<td>1% vs. 1%</td>
<td>0.94 (0.16 to 5.42)</td>
<td>660 (2)</td>
</tr>
<tr>
<td>Relevant additional surgeries</td>
<td>---</td>
<td>----</td>
<td>12% vs. 14%</td>
<td>0.72 (0.38 to 1.34)</td>
<td>722 (4)</td>
<td></td>
</tr>
</tbody>
</table>

ALIF = anterior lumbar interbody fusion; ICBG = iliac crest bone graft; rhBMP-2 = recombinant human bone morphogenetic protein-2; PLF = posterolateral lumbar fusion.

Values in bold font are significant at 0.05 level.

* For ALIF, a total n = 465 was included in the analysis, excluding 4 patients who underwent open surgery in study 1.

† Categories of adverse events are based on Medtronic datasets, except for those indicated otherwise.

‡ Based on individual adverse event case histories in the proprietary reports provided by Medtronic.

§ For PLF, a total n = 722 was included in the analysis, excluding 11 patients randomized to rhBMP-2 without instrumentation in study 12.
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name, Design, (Study #) (References*)</th>
<th>Sample Size, n</th>
<th>Overall Success, 24 Months</th>
<th>Fusion, 24 Months</th>
<th>Cumulative Number of Adverse Events up to 24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IPD Results</td>
<td>Published Results†</td>
<td>IPD Results‡</td>
</tr>
<tr>
<td></td>
<td>I  C</td>
<td>rhBMP-2 (%)</td>
<td>ICBG (%)</td>
<td>RR (95% CI)</td>
</tr>
<tr>
<td>Anterior lumbar interbody fusion – on-label use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pilot (1) Boden, 2000 (55) RCT/Poor</td>
<td>11 3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Pivotal (2) Burkus, 2002 (6) RCT/Fair</td>
<td>143 13 6</td>
<td>77/133 (58%)</td>
<td>68/123 (55%)</td>
<td>1.05 (0.84, 1.30)</td>
</tr>
<tr>
<td>INFUSE®/LT-CAGE® Lap Pivotal (Study 3) Burkus, 2003 (54) IS/Fair</td>
<td>134 70/115 (61%)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>INFUSE®/INTER FIX™ ALIF Pilot (Study 9) Unpublished RCT/Fair</td>
<td>25 20 11/23 (48%)</td>
<td>7/17 (41%)</td>
<td>1.16 (0.57, 2.36)</td>
<td>15/22 (68%)</td>
</tr>
<tr>
<td>MAVERICK™ Disc Pivotal (Study 10) Gornet, 2011 (66) RCT/Fair**</td>
<td>172 40 58/139 (42%)</td>
<td>233/371 (63%)</td>
<td>0.64 (0.53, 0.77)</td>
<td>57/103 (55.3%)</td>
</tr>
</tbody>
</table>

*p<0.001
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name, Design, (Study #) (References*)</th>
<th>Sample Size, n</th>
<th>Overall Success, 24 Months</th>
<th>Fusion, 24 Months</th>
<th>Cumulative Number of Adverse Events up to 24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IPD Results</td>
<td>Published Results†</td>
<td>IPD Results</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rhBMP-2 (%) (ICBG (%) RR (95% CI))</td>
<td>rhBMP-2 (%) (ICBG (%) RR (95% CI))</td>
<td>rhBMP-2 (%) (ICBG (%) RR (95% CI))</td>
</tr>
<tr>
<td>Anterior lumbar interbody fusion – off-label use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pilot RCT (Study 4) Burkus, 2002 (7)</td>
<td>24 22</td>
<td>17/24 (71%) 4/20 (20%) 3.54 (1.42, 8.83)</td>
<td>NR NR 24/24 (100%) 12/20 (60%) 1.65 (1.15, 2.35)</td>
<td>24/24 (100%) 13/19 (68.4%)</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pivotal (Study 5) Burkus, 2005 (63) ¶</td>
<td>55 30</td>
<td>33/50 (66%) 15/27 (56%) 1.19 (0.80, 1.76)</td>
<td>NR NR 43/47 (91%) 24/25 (96%) 0.95 (0.85, 1.07)</td>
<td>NSR NSR 95 76 0 0</td>
</tr>
<tr>
<td>Posterior lumbar interbody fusion – off-label use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/ INTERFIX™ PLIF RCT (Study 6) Haid, 2004 (8)</td>
<td>34 33</td>
<td>15/31 (48%) 10/31 (32%) 1.50 (0.80, 2.81)</td>
<td>NR NR 25/31 (81%) 21/30 (70%) 1.15 (0.86, 1.54)</td>
<td>112 120 29 35</td>
</tr>
</tbody>
</table>
|                                                        |                |                             |                   |                               |                   |                   |                   |§
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name, Design, (Study #) (References*)</th>
<th>Sample Size, n</th>
<th>Overall Success, 24 Months</th>
<th>Fusion, 24 Months</th>
<th>Cumulative Number of Adverse Events up to 24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IPD Results</td>
<td>Published Results†</td>
<td>IPD Results</td>
</tr>
<tr>
<td></td>
<td>I  C  rhBMP-2 (%)  ICBG (%)  RR (95% CI)</td>
<td>rhBMP-2 (%)  ICBG (%)  RR (95% CI)</td>
<td>rhBMP-2 (%)  ICBG (%)  RR (95% CI)</td>
<td>rhBMP-2 (%)  ICBG (%)  RR (95% CI)</td>
</tr>
<tr>
<td>INFUSE®/TELAMON PEEK PLIF Pilot IS (Study 11) Unpublished</td>
<td>30</td>
<td>13/25 (52%) NA NA NA</td>
<td>NA NA NA</td>
<td>24/25 (96%) NA NA NA</td>
</tr>
<tr>
<td>Posterior lumbar fusion – off-label use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rhBMP-2/BCP Mexico Pilot IS (Study 16 Unpublished)</td>
<td>I1: 7</td>
<td>I2: 8</td>
<td>NA NA NA</td>
<td>NA NA NA NA NA</td>
</tr>
<tr>
<td></td>
<td>I1: 5</td>
<td>I2: 4/11 (36%) 2/4 (50%)</td>
<td>I1 vs. C: 0.73 (0.21, 2.55)</td>
<td>I1: 10/10 (100%) 3/4 (75%)</td>
</tr>
<tr>
<td></td>
<td>I1: 5</td>
<td>I2: 4/10 (40%)</td>
<td>(100%)</td>
<td>(90%)</td>
</tr>
<tr>
<td></td>
<td>I1: 11</td>
<td>I2: 9/10 (90%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>48/97 (49%) 40/95 (42%)</td>
<td>1.18</td>
<td>89/96 (93%) 68/94 (72%)</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>(49%) 42%</td>
<td></td>
<td>(93%) 72%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Downloaded From: http://annals.org/pdfaccess.ashx?url=/data/journals/aim/927097/ on 09/13/2018
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name, Design, (Study #) (References*)</th>
<th>Sample Size, n</th>
<th>Overall Success, 24 Months</th>
<th>Fusion, 24 Months</th>
<th>Cumulative Number of Adverse Events up to 24 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>**Sample</td>
<td>rhBMP-2</td>
<td>ICBG</td>
<td>RR (95% CI)</td>
<td>rhBMP-2</td>
</tr>
<tr>
<td>INFUSE®/ MASTER GRAFT® Pilot RCT (Study 8) Dawson, 2009 (71)</td>
<td>25 21</td>
<td>15/24 (63%)</td>
<td>10/20 (50%)</td>
<td>1.25 (0.73, 2.14)</td>
</tr>
<tr>
<td>AMPLIFY™ (rhBMP-2/ CRM) Pivota RCTI (Study 14) Dimar, 2009 (72)</td>
<td>239 4</td>
<td>118/211 (56%)</td>
<td>105/186 (56%)</td>
<td>0.99 (0.83, 1.18)</td>
</tr>
<tr>
<td>rhBMP-2/ CRM 2-level Pilot IS (Study 15) Unpublished</td>
<td>29</td>
<td>12/26 (46%)</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Anterior cervical disectomy and fusion – off-label use**

| INFUSE®/ CORNER STONE® ACDF Pilot (Study 7) Baskin, 2003 (5) | 18 15 | 10/12 (83%) | 10/12 (83%) | 1.00 (0.70, 1.43) | NR | NR | 11/12 (92%) | 12/12 (100%) | 0.92 (0.77, 1.09) | 10/10 | 10/10 | 45 | 13 | 2 | 1 |

ACDF = anterior cervical disectomy and fusion; ACS = absorbable collagen sponge ALIF = anterior lumbar interbody fusion; BCP = biphasic calcium phosphate; C = comparator group (ICBG group); CI = confidence interval; CRM = compression resistant matrix; I = investigational group (rhBMP-2 group); ICBG = iliac crest bone graft; IDE = investigational device exemption; IPD = Individual patient data; IS = intervention series; NA = not applicable; NR = not reported; NS = not significant; NSR = not separately reported; PEEK = polyetheretherketone; PLF = posterior
lumbar fusion; PLIF = posterior lumbar interbody fusion; RCT = randomized controlled trial; rhBMP-2 = recombinant human bone morphogenetic protein-2; US = United States.

* The primary study publication is referenced in the table. Study results are also reported in the following publications—Study 1: Khan, 2002 (62), McKay, 2002 (59), Poynton, 2002 (61), Sandhu, 2003 (58); Study 2: McKay, 2002 (59), Burkus, 2003 (54), Sandhu, 2003 (58), Burkus, 2004 (56), Burkus, 2004 (60), Burkus, 2005 (64); Study 3: Kleeman, 2001 (57), Khan, 2002 (62), McKay, 2002 (59), Poynton, 2002 (61), Sandhu, 2003 (58), Burkus, 2004 (56), Burkus, 2004 (60), Burkus, 2005 (64); Study 4: Khan, 2002 (62), McKay, 2002 (59), Sandhu, 2003 (58), Burkus, 2004 (56), Burkus, 2005 (64), Medtronic (65); Study 4: Khan, 2002 (62), McKay, 2002 (59), Sandhu, 2003 (58), Burkus, 2004 (56), Burkus, 2005 (63), Burkus, 2005 (64), Burkus 2006, (67); Study 5: Burkus, 2004 (56), Burkus, 2005 (63), Burkus, 2005 (64), Burkus, 2005 (64), Study 5: Burkus, 2004 (56), Burkus, 2005 (63), Burkus, 2005 (64), Burkus, 2005 (64), Burkus, 2006 (67); Study 6: McKay, 2002 (59), Poynton, 2002 (61), Sandhu, 2003 (58), Burkus, 2005 (64); Study 7: McKay, 2002 (59), Study 8: Burkus, 2004 (56); Study 8: Burkus, 2004 (56), Burkus, 2005 (64); Study 12: Sandhu, 2003 (58); Study 16: McKay, 2002 (59), Burkus, 2005 (64).
† For unpublished studies, cells are blank.
‡ More information about the type and number of specific adverse effects can be found in Appendix F of the full report (17). These numbers do not include non-union and non-union pending.
§ The type and number of specific adverse effects reported by each journal publication can be found in Table 16 of the full report (17).
¶¶ Study 5 data not published independently. Burkus, 2005 (63) contains pooled data from Studies 4 and 5.
** The comparison group in this study received artificial disk, not ICBG. Discrepancy in numbers between published trial and IPD was partially due to an updated Medtronic data set provided to the authors.
†† n not reported; results reported only as percentages.
‡‡ The Mexico pilot was an intervention series with two cohorts.
§§ 1 I1 = rhBMP-2 without internal fixation; I2 = rhBMP-2 + TSRH (Texas Scottish Rite Hospital) pedicle screw instrumentation; C = autograft + TSRH. This study only followed patients for 12 months, so there were no data at 24 months.
The cumulative number of adverse events up to 12 months.
¶¶ The table in this publication reports a slightly higher percentage (97.3%).
**Appendix Table 1. Outcome Variable Definitions/Criteria from Medtronic Protocols Compared with Those in Published Studies and Individual Patient Data Analysis for Comparative Effectiveness and Harms**

<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Surgical Approach (Study Number)</th>
<th>Medtronic Protocol Definition/Criteria</th>
<th>Published Studies* Definition/Criteria</th>
<th>Individual Patient Data Analysis in This Review Definition/criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall success</td>
<td>ALIF/PLIF/PLF (2, 3, 4, 5, 6, 8, 9, 12, 13, 14; not defined in Study 1)</td>
<td>All of the following criteria need to be satisfied: Fusion Improvement in the ODI for low back pain (ODI success) Maintenance or improvement in neurologic status (neurologic success) No serious adverse event classified as implant- or implant/surgical-associated No additional surgical procedure classified as “failure”</td>
<td>Only reported for Study 8, which used the same definition as the Medtronic protocol.</td>
<td>Same definition as Medtronic protocol except for a minor difference in Study 8: In Study 8, definition of ODI success differed slightly. (See definition for ODI success below.) In primary analysis, patients had to satisfy all criteria; patients with data for some but not all criteria were categorized as failures. In sensitivity analyses, we categorized patients with data for some but not all criteria as missing and excluded them from the analysis.</td>
</tr>
<tr>
<td></td>
<td>Artificial disc (Maverick) (10)</td>
<td>Same as ALIF/PLF (see above), except: Fusion not a criterion Disc height is a criterion that postoperative disc height at each visit after 6 weeks was no more than 2mm shorter than postoperative disc height at 6 weeks</td>
<td>Same definition as Medtronic protocol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACDF (7)</td>
<td>Same as ALIF/PLF (see above), except: Success based on improvement in NDI</td>
<td>Definition not reported</td>
<td></td>
</tr>
<tr>
<td>Outcome Variable</td>
<td>Surgical Approach (Study Number)</td>
<td>Medtronic Protocol Definition/Criteria</td>
<td>Published Studies* Definition/Criteria</td>
<td>Individual Patient Data Analysis in This Review Definition/criteria</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------</td>
<td>-------------------------------------------------------------</td>
</tr>
<tr>
<td>Fusion</td>
<td>ALIF (1)</td>
<td>Bone growing continuously through the cage and connecting with vertebral bodies above and below through at least one cage,</td>
<td>Presence of continuous trabecular bone growth through both of the cages</td>
<td>Same definition as Medtronic protocols, except: If data from a CT scan were available, they were used first. If data from CT scan were not available but data from radiographs were available, radiographs data were used. In primary analysis, patients had to satisfy all criteria; patients with data for some but not all criteria were categorized as failures. In sensitivity analyses, we categorized patients with data for some but not all criteria as missing and excluded them from the analysis.</td>
</tr>
<tr>
<td>ALIF/PLIF</td>
<td>Artificial disc rhBMP-2 arm (2, 3, 4, 5, 6, 9,10)</td>
<td>All criteria must be met: 1) Evidence of continuous trabecular bone growth connecting the vertebral bodies and/or through either one or both implants; 2) Absence of radiolucency covering &gt;50% of implant 3) Translation of ≤ 3mm and angulation of &lt;5 degrees. Fusion assessed primarily with radiographs. CT scans used as a secondary method.</td>
<td>In addition to criteria used in the Medtronic protocols, patients who underwent a secondary surgery were considered as failed fusion in published studies (Study 2, combined analysis of Studies 2 and 3, Study 4, combined analysis of Studies 4 and 5, Study 6), and evidence of continuous trabecular bone growth was assessed generally using a CT scan. Study 6 used both CT and radiographs.</td>
<td>Same definition as Medtronic protocols: Study 8 used both CT and radiographs but did not say how they were used for assessing bridging trabecular bone. Study 12 and 14 used radiographs and CT scans as specified in the protocols.</td>
</tr>
<tr>
<td>PLF</td>
<td>(8, 12, 13, 14)</td>
<td>All criteria must be met: 1) Evidence of continuous trabecular bone growth connecting the transverse processes; 2) Absence of radiolucent lines through the fusion mass; 3) Translation of ≤ 3mm and angulation of &lt;5 degrees. Fusion assessed primarily with radiographs. CT scans used as a secondary method.</td>
<td>Same definition as Medtronic protocols: Study 8 used both CT and radiographs but did not say how they were used for assessing bridging trabecular bone. Study 12 and 14 used radiographs and CT scans as specified in the protocols.</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Surgical Approach (Study Number)</td>
<td>Medtronic Protocol Definition/Criteria</td>
<td>Published Studies* Definition/Criteria</td>
<td>Individual Patient Data Analysis in This Review Definition/criteria</td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>ODI success</td>
<td>ACDF (7)</td>
<td>All criteria must be met: 1) Evidence of bridging bone 2) Absence of radiolucency covering &gt;50% of superior or inferior surface of graft 3) Translation of ≤ 3mm and angulation of &lt;4 degrees. Radiographs and CT scans used to assess fusion.</td>
<td>Same definition as Medtronic protocol (The published study also used both CT and radiographs but did not say how they were used for assessing bridging trabecular bone.)</td>
<td>Same definition as Medtronic protocols, except: For Study 8, an increase of at least 15 points in ODI score was used (to be consistent with definitions used in all other studies).</td>
</tr>
<tr>
<td>ODI success</td>
<td>ALIF/PLIF/PLF, artificial disc (2, 3, 4, 5, 6, 8, 9, 10, 12, 13, 14)</td>
<td>At least a 15-point improvement in ODI score for back pain at each visit postoperatively compared with pre-operative index score (FDA’s recommendation, a 15-point improvement is clinically meaningful based on Copay, 2008 (74)). Study 8 used a 15% improvement instead of a 15-point increase.</td>
<td>Studies 1 and 12: at least 15% improvement Study 10: at least 15-point improvement Other studies reported results with improvement of at least 15% (Studies 2 and 8), or at least 20% (Study 8), or at least 15 points (Studies 4 and 6) without explicitly defining success, or only reported actual scores (combined analysis of Studies 2 and 3, and 4 and 5).</td>
<td></td>
</tr>
<tr>
<td>ODI success</td>
<td>ACDF (7)</td>
<td>At least a 15-point improvement in NDI score for neck pain at each visit postoperatively compared with pre-operative index score (FDA’s recommendation, a 15-point improvement is clinically meaningful based on Copay, 2008 ).</td>
<td>Same definition as Medtronic protocol</td>
<td>Same definition as Medtronic protocol</td>
</tr>
<tr>
<td>Outcome Variable</td>
<td>Surgical Approach (Study Number)</td>
<td>Medtronic Protocol Definition/Criteria</td>
<td>Published Studies* Definition/Criteria</td>
<td>Individual Patient Data Analysis in This Review Definition/criteria</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Neurologic success</td>
<td>ALIF/PLIF/PLF, artificial disc (2, 3, 4, 5, 6, 8, 9, 10, 12, 13, 14; not defined in Study 1)</td>
<td>Four neurologic tests evaluated motor function, sensory function, deep tendon reflexes, and sciatic tension signs (straight-leg raise). A score was developed for each test. Studies 2, 3, 5, 6, 9, 12, and 13: the scores of the four tests were totaled and an overall score was expressed as a percentage of the maximum possible score. A neurologic success was defined as a postoperative overall score no more than 10% worse than the pre-operative overall score. Studies 4, 8, 10, 14: neurologic success was defined as having the same or better score in all four tests compared to pre-operative score.</td>
<td>Not defined in Study 1; mean score reported. Studies 2, 4, 6, 10: used the same definition as the Medtronic protocols. Combined analysis of Studies 2 and 3, combined analysis of Studies 4 and Studies 5, 8, 14: neurologic success was not reported, neither were the mean scores. Study 12: scores not reported; outcome briefly mentioned.</td>
<td>Used definition from Medtronic protocols for studies 4, 8, 10, 14 for all studies.</td>
</tr>
<tr>
<td></td>
<td>ACDF (7)</td>
<td>Same as ALIF/PLF for Studies 2, 3, 5, 6, 9, 12, and 13 (see above), except: Sensory symptoms and the foraminal compression test were used in the place of sciatic tension signs.</td>
<td>Neurologic status of the patients was determined by evaluating two neurologic tests: motor and sensory function. Neurologic success was based on demonstrated maintenance or improvement in both tests.</td>
<td>Same as above, except: The four neurologic tests were motor function, sensory function, reflexes, and sensory symptoms; plus the foraminal compression test.</td>
</tr>
<tr>
<td>Surgical procedure “failure”</td>
<td>ALIF/PLIF/PLF, artificial disc (2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14)</td>
<td><strong>Surgical procedure classified as a failure when any of the following occurred:</strong> Supplemental fixation Device removal‡ Revision §</td>
<td>Combined analysis of Studies 4 and 5, Study 8 and 10: Same definition as Medtronic protocols. For Study 6: a second spinal surgery at the same level All other published studies: Not reported.</td>
<td>Same definition as Medtronic protocols</td>
</tr>
<tr>
<td>SF-36</td>
<td>(All studies)</td>
<td><strong>Standard definition†</strong></td>
<td><strong>Standard definition†</strong></td>
<td><strong>Standard definition†</strong></td>
</tr>
<tr>
<td>Outcome Variable</td>
<td>Surgical Approach (Study Number)</td>
<td>Medtronic Protocol Definition/Criteria</td>
<td>Published Studies* Definition/Criteria</td>
<td>Individual Patient Data Analysis in This Review Definition/criteria</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Back or leg pain</td>
<td>ALIF/PLIF/PLF (2, 3, 4, 5, 6, 8, 9, 12, 13, 14)</td>
<td>Sum of rating scores on intensity and duration of back or leg pain, both on a scale of 0 to 10. Back and leg pain not separately measured in Study 1</td>
<td>Same definition as Medtronic protocols</td>
<td>Only rating score on intensity of back or leg pain on a scale from 0 to 10 was used.</td>
</tr>
<tr>
<td></td>
<td>Artificial disc (10)</td>
<td>Multiplication of rating scores (0-10) on intensity and duration of back or leg pain</td>
<td>Same definition as Medtronic protocol</td>
<td>Only rating score on intensity of back or leg pain on a scale from 0 to 10 was used.</td>
</tr>
<tr>
<td>Neck or arm pain</td>
<td>ACDF (7)</td>
<td>Sum of rating scores on intensity and duration of neck or arm pain, both on a scale of 0 to 10</td>
<td>Same definition as Medtronic protocol</td>
<td>Only rating score on intensity of neck or arm pain on a scale from 0 to 10 was used.</td>
</tr>
<tr>
<td>Adverse event</td>
<td>(All studies)</td>
<td>No definition in protocols; “adverse event” listed in the data collection forms</td>
<td>Sparsely reported; as defined in Medtronic datasets</td>
<td>As defined in Medtronic datasets</td>
</tr>
<tr>
<td>Device-related adverse event</td>
<td>(All studies)</td>
<td>Reasonable possibility that the adverse event may have been caused by the implant(s) or by device and surgical procedure, as determined by Study investigators</td>
<td>Only reported in Study 10 and 14; as defined in Medtronic datasets</td>
<td>As defined in Medtronic datasets</td>
</tr>
<tr>
<td>Serious adverse event</td>
<td>(All studies)</td>
<td>For events defined in the WHO Recommendations for Grading of Acute and Subacute Toxic Effects, any adverse event with severity 3 or 4. For events not defined by the WHO Toxicity Scale, any adverse event if it limits the patient's ability to perform routine activities despite symptomatic therapy, if it results in the need to remove the implant, or if the patient is at immediate risk of death.</td>
<td>Only reported in Study 10, defined as WHO Grade 3 or 4 adverse event</td>
<td>Serious adverse events categorized as: An adverse event with a severity score of 3 or 4, based on Medtronic categorization of severity in Medtronic datasets.</td>
</tr>
<tr>
<td>Outcome Variable</td>
<td>Surgical Approach (Study Number)</td>
<td>Medtronic Protocol Definition/Criteria</td>
<td>Published Studies* Definition/Criteria</td>
<td>Individual Patient Data Analysis in This Review Definition/criteria</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Relevant additional surgeries</td>
<td>(All studies)</td>
<td>Not specifically defined what is “relevant” but classified additional surgeries as: supplemental fixation, device removal, revision and re-operation</td>
<td>Study 1: not reported. All other studies reported secondary/additional surgeries as classified in Medtronic protocols. Study 14 compared second surgeries including revision, non-elective removal, and revision only. Elective removal and reoperation was excluded.</td>
<td>Relevant additional surgery: Supplemental fixation Device removal Revision Re-operation Based on classification in Medtronic datasets Primary definition: back pain plus any leg or buttock pain or weakness (includes pain described as sciatica, radiculopathy or radicular pain, use of epidural steroids or decompression surgery). Sensitivity analysis looked at other definitions. Definition 2 similar to primary definition except back pain not required. Leg numbness and nerve root injections also included as indicating possible radiculitis. Definition 3 same as definition 2 except that any type of back and leg pain was included (e.g., osteoarthritis). Definition 4 defined possible lumbar radiculitis simply as back and/or leg pain with the use of epidural steroids or decompression surgery.</td>
</tr>
<tr>
<td>Possible lumbar radiculitis</td>
<td>ALIF/PLIF/PLF (1, 2, 4, 5, 6, 8, 9, 12, 13, 14)</td>
<td>Not defined as an outcome</td>
<td>Not defined as an outcome</td>
<td></td>
</tr>
</tbody>
</table>

ACDF = anterior cervical discectomy and fusion; ALIF = anterior lumbar interbody fusion; CT = computed tomography; FDA = U.S. Food & Drug Administration; NDI = Neck Disability Index (Vernon); ODI = Oswestry Disability Index/Oswestry Low Back Pain Disability Questionnaire; PLF = posterolateral lumbar fusion; PLIF = posterior lumbar interbody fusion; WHO = World Health Organization.

* Studies 9 and 13 were not published.
† Standard definition for Medical Outcomes Study 36-Item Short Form Health Survey (SF-36) was used for all studies in all sources (83)
‡ Studies 7, 8 and 14 used non-elective device removal.
§ Studies 6, 9, 12, and 13 used revision after two weeks of surgery. For revisions within two weeks of surgery, Medtronic determined whether or not these events were failures on a case-by-case basis, with input from FDA.
### Appendix Table 2. Included Medtronic Studies of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2)

<table>
<thead>
<tr>
<th>IDE Clinical Trial Name (Study #)</th>
<th>Study Design</th>
<th>Sample Size, n</th>
<th>rhBMP-2 Conc. Dose (mg)</th>
<th>Carriers</th>
<th>Baseline Characteristics</th>
<th>Duration of Followup, months</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study, Year (Reference)</td>
<td>Study Name, Design</td>
<td>I</td>
<td>C</td>
<td>Mean Age, years</td>
<td>Male, n (%)</td>
<td>Diabetes, n (%)</td>
<td>Smoking, n (%)</td>
</tr>
<tr>
<td>INFUSE®/ LT-CAGE® Pilot (Study 1)</td>
<td>RCT</td>
<td>11</td>
<td>3</td>
<td>1.5, 3.9-7.8</td>
<td>ACS</td>
<td>I: 42.5, 40.2</td>
<td>I: (6) 56% (55%)</td>
</tr>
<tr>
<td>Boden, 2000 (55)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 67%</td>
<td>C: 0</td>
</tr>
<tr>
<td>INFUSE®/ LT-CAGE® Pivotal (Study 2)</td>
<td>RCT</td>
<td>142</td>
<td>136</td>
<td>1.5, 4.2-8.4</td>
<td>ACS</td>
<td>I: 43.3, 42.3</td>
<td>I: 6 (4%)</td>
</tr>
<tr>
<td>Burkus, 2002 (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 68</td>
<td>C: 1</td>
</tr>
<tr>
<td>INFUSE®/ LT-CAGE® Lap Pivotal (Study 3)</td>
<td>IS</td>
<td>134</td>
<td></td>
<td>1.5, 4.2-8.4</td>
<td>ACS</td>
<td>I: 42.4</td>
<td>I: 3 (2%)</td>
</tr>
<tr>
<td>Burkus, 2003 (54) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 50</td>
<td>C: 49</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pilot (Study 4)</td>
<td>RCT</td>
<td>24</td>
<td>22</td>
<td>1.5, 8.1-11.7</td>
<td>ACS</td>
<td>I: 41.5</td>
<td>I: 8 (33%)</td>
</tr>
<tr>
<td>Burkus, 2002 (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 10</td>
<td>C: 1</td>
</tr>
<tr>
<td>INFUSE®/ Bone Dowel Pivotal (Study 5)</td>
<td>RCT</td>
<td>55</td>
<td>30</td>
<td>1.5, 8.1-11.7</td>
<td>ACS</td>
<td>I: 39.7</td>
<td>I: 24</td>
</tr>
<tr>
<td>Burkus, 2005 (63) †</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 42.1</td>
<td>C: 1</td>
</tr>
<tr>
<td>IDE Clinical Trial Name (Study #)</td>
<td>Study Design</td>
<td>Sample Size, n</td>
<td>rhBMP-2 Conc. (mg/cc)</td>
<td>Dose (mg)</td>
<td>Carrier</td>
<td>Baseline Characteristics</td>
<td>Quality</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>---------</td>
<td>--------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>INFUSE®/ INTER FIX™ ALIF Pilot (Study 9) Unpublished</td>
<td>RCT</td>
<td>25 20</td>
<td>1.5</td>
<td>8.4-16.8</td>
<td>ACS</td>
<td>I: 45.9</td>
<td>Fair</td>
</tr>
<tr>
<td>MAVERICK™ Disc Pivotal (Study 10)‡ Gornet, 2011 (66)</td>
<td>RCT</td>
<td>172 405</td>
<td>1.5</td>
<td>4.2-12.0</td>
<td>ACS</td>
<td>I: 40.2</td>
<td>Fair</td>
</tr>
<tr>
<td>Posterior lumbar fusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rhBMP-2/BCP Mexico Pilot § (Study 16) Unpublished</td>
<td>IS</td>
<td>I1: 7 8</td>
<td>2.23.0 15.0-40.0</td>
<td>BCP</td>
<td>I1: 53.9</td>
<td>12</td>
<td>Fair</td>
</tr>
<tr>
<td>rhBMP-2/BCP US Pivotal (Study 12) Boden, 2002 (70)</td>
<td>RCT</td>
<td>I1: 11 2 12: 8</td>
<td>2.1</td>
<td>42.0</td>
<td>C: 52.9</td>
<td>24</td>
<td>Poor</td>
</tr>
<tr>
<td>rhBMP-2/BCP Canada Pivotal (Study 13) Unpublished</td>
<td>RCT</td>
<td>99 98</td>
<td>2.1</td>
<td>42.0-63.0</td>
<td>BCP</td>
<td>I: 53.0</td>
<td>Fair</td>
</tr>
<tr>
<td>Study Name</td>
<td>Design</td>
<td>Sample Size, n</td>
<td>rhBMP-2 Conc. (mg/cc)</td>
<td>Dose (mg)</td>
<td>Carrier</td>
<td>Baseline Characteristics</td>
<td>Quality</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>--------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>---------</td>
<td>--------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>INFUSE®/ MASTER GRAFT® Pilot (Study 8) Dawson, 2009 (71)</td>
<td>RCT</td>
<td>25 21</td>
<td>1.5</td>
<td>12.0</td>
<td>ACS</td>
<td>I: 55.9, C: 56.9</td>
<td>Fair</td>
</tr>
<tr>
<td>AMPLIFY™ (rhBMP-2/ CRM) Pivotal (Study 14) Dimar, 2009 (72)</td>
<td>RCT</td>
<td>239 224</td>
<td>2.0</td>
<td>40.0</td>
<td>CRM</td>
<td>I: 53.2, C: 52.3</td>
<td>Fair</td>
</tr>
<tr>
<td>rhBMP-2/ CRM 2-level Pilot (Study 15) Unpublished</td>
<td>IS</td>
<td>29</td>
<td>2.0</td>
<td>40.0</td>
<td>CRM</td>
<td>I: 53.9, C: 52.3</td>
<td>Poor</td>
</tr>
<tr>
<td>Posterior lumbar interbody fusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFUSE®/ INTER FIX™ PLIF (Study 6) Haid, 2004 (8)</td>
<td>RCT</td>
<td>34 33</td>
<td>1.5</td>
<td>4.2-8.4</td>
<td>ACS</td>
<td>I: 46.3, C: 46.1</td>
<td>Fair</td>
</tr>
</tbody>
</table>

Circumferential posterior lumbar interbody fusion
<table>
<thead>
<tr>
<th>IDE Clinical Trial Name</th>
<th>Study Design</th>
<th>Sample Size, n</th>
<th>rhBMP-2 Conc. (mg/cc)</th>
<th>Dose (mg)</th>
<th>Carrier</th>
<th>Baseline Characteristics</th>
<th>Baseline Characteristics</th>
<th>Baseline Characteristics</th>
<th>Baseline Characteristics</th>
<th>Baseline Characteristics</th>
<th>Baseline Characteristics</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFUSE®/ TELAMON PEEK PLIF Pilot (Study 11) Unpublished</td>
<td>IS</td>
<td>30</td>
<td>1.5</td>
<td>8.4</td>
<td>ACS</td>
<td>I: 51.0</td>
<td>I: 12 (40%)</td>
<td>I: 2 (7%)</td>
<td>I: 8 (27%)</td>
<td>I: 14 (47%)</td>
<td>I: 9 (30%)</td>
<td>36</td>
</tr>
<tr>
<td>INFUSE®/ CORNER STONE® ACDF Pilot (Study 7) Baskin, 2003 (5)</td>
<td>RCT</td>
<td>18</td>
<td>15</td>
<td>1.5</td>
<td></td>
<td>I: 51.3</td>
<td>I: 8 (44%)</td>
<td>I: 0</td>
<td>I: 5 (28%)</td>
<td>I: 1 (6%)**</td>
<td>I: 12 (67%)</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C: 47.1</td>
<td>C: 7 (47%)</td>
<td>C: 0</td>
<td>C: 7 (47%)</td>
<td>C: 0**</td>
<td>C: 9 (60%)</td>
<td></td>
</tr>
</tbody>
</table>

ACDF = anterior cervical discectomy and fusion; ACS = absorbable collagen sponge, ALIF = anterior lumbar interbody fusion, BCP = biphasic calcium phosphate, C = comparator group (ICBG or artificial disk), CRM = compression resistant matrix, ICBG = iliac crest bone graft; I = investigational group (rhBMP-2 group), IDE = investigational device exemption, IS = intervention series, PEEK = polyetheretherketone, PLF = posterior lumbar fusion, PLIF = posterior lumbar interbody fusion, RCT = randomized controlled trial, rhBMP-2 = recombinant human bone morphogenic protein-2, US = United States

* Study 3 data not published independently. Burkus, 2003 (54) contains pooled data from Studies 3 and 2. Patients underwent laparoscopic ALIF in this study; patients in the other ALIF studies underwent open surgery except for 4 patients in the rhBMP-2 group of Study 1.
† Study 5 data not published independently. Burkus, 2005 (63) contains pooled data from Studies 4 and 5.
‡ Comparator is an artificial disk, not ICBG.
§ The Mexico pilot was an intervention series with two cohorts.
¶ I1 = rhBMP-2 without internal fixation, I2 = rhBMP-2 + TSRH (Texas Scottish Rite Hospital) pedicle screw instrumentation, C: autograft + TSRH
¶¶ 100 patients (including both intervention and control group) were followed for 24 months (rhbmp-2 vs. ICBG, using CD horizon spinal system) and 97 patients were followed for 48 months (rhbmp-2 vs. ICBG, using TSRH spinal system).
** Prior neck surgery
### Appendix Table 3. Cancer Occurrence at 24 and 48 Months in Randomized Trials*

#### Patients Receiving rhBMP-2

<table>
<thead>
<tr>
<th>Number of Cancers</th>
<th>Time Period from Surgery, months</th>
<th>Type of Surgery</th>
<th>rhBMP-2 Dose, mg</th>
<th>Type of Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basal cell carcinoma†</strong></td>
<td>2</td>
<td>1.5, 3</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Breast</strong></td>
<td>2</td>
<td>24</td>
<td>ALIF</td>
<td>4.2-8.4;8.1-11.7</td>
</tr>
<tr>
<td><strong>Carcinoid</strong></td>
<td>1</td>
<td>24</td>
<td>ALIF</td>
<td>4.2-12</td>
</tr>
<tr>
<td><strong>Colon</strong></td>
<td>0</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Larynx</strong></td>
<td>1</td>
<td>6</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Liver</strong></td>
<td>1</td>
<td>6</td>
<td>ALIF</td>
<td>4.2-12</td>
</tr>
<tr>
<td><strong>Lung</strong></td>
<td>1</td>
<td>6</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Melanoma</strong></td>
<td>1</td>
<td>24</td>
<td>ALIF</td>
<td>4.2-12</td>
</tr>
<tr>
<td><strong>Lymphoma</strong></td>
<td>1</td>
<td>24</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Ovarian</strong></td>
<td>1</td>
<td>12</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Pancreatic</strong></td>
<td>2</td>
<td>12</td>
<td>ALIF, PLF</td>
<td>4.2-8.4;40</td>
</tr>
<tr>
<td><strong>Prostate</strong></td>
<td>1</td>
<td>12</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Squamous cell carcinoma†</strong></td>
<td>2</td>
<td>12, 24</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Stomach</strong></td>
<td>1</td>
<td>24</td>
<td>PLF</td>
<td>40</td>
</tr>
<tr>
<td><strong>Thyroid</strong></td>
<td>1</td>
<td>12</td>
<td>ALIF</td>
<td>8.1-11.7</td>
</tr>
<tr>
<td><strong>Total cancers up to 24 months</strong></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(in 17 patients; total n = 633 patients)

#### Patients Receiving ICBG or Artificial Disc

<table>
<thead>
<tr>
<th>Number of Cancers</th>
<th>Time Period from Surgery, months</th>
<th>Type of Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breast</strong></td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><strong>Leukemia</strong></td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><strong>Melanoma</strong></td>
<td>2</td>
<td>36, 48</td>
</tr>
<tr>
<td><strong>Merkle cell carcinoma</strong></td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td><strong>Prostate</strong></td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><strong>Squamous cell carcinoma†</strong></td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td><strong>Thyroid</strong></td>
<td>1</td>
<td>36</td>
</tr>
</tbody>
</table>

### Cancers occurring between 24- and 48-month follow-ups‡:

<table>
<thead>
<tr>
<th>Number of Cancers</th>
<th>Time Period from Surgery, months</th>
<th>Type of Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breast</strong></td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td><strong>Leukemia</strong></td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><strong>Melanoma</strong></td>
<td>2</td>
<td>36, 48</td>
</tr>
<tr>
<td><strong>Merkle cell carcinoma</strong></td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td><strong>Prostate</strong></td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td><strong>Squamous cell carcinoma†</strong></td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td><strong>Thyroid</strong></td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td>----------------------</td>
<td>---</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Total cancers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>up to 48 months</strong></td>
<td>20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(in 16 patients; total n = 483 patients)</td>
<td>(in 11 patients; n = 700 patients)</td>
</tr>
</tbody>
</table>

ALIF = anterior lumbar interbody fusion; ICBG = iliac crest bone graft; rhBMP-2 = recombinant human bone morphogenetic protein-2; PLF = posterolateral lumbar fusion; AD = artificial disc.

* Does not include 1 pancreatic cancer and 1 renal cancer that were discovered during the study but determined to exist prior to the study.

† Non-SEER cancers (for which data is not reportable by the Surveillance, Epidemiology and End Results [SEER] Program, National Cancer Institute)

‡ Does not include cancers occurring after 48 months of followup or cancers occurring in rhBMP-2 arms of studies without a control arm (intervention series or when only the rhBMP-2 arm experienced continued followup); additional cancers reported in these rhBMP-2 patients were 1 each—colon cancer, breast cancer, squamous cell carcinoma, thyroid cancer, testicular cancer—and 2 basal cell carcinomas. Does not include 1 thyroid cancer and 1 leukemia in the control arms of studies after 48 months.

§ Total from studies following patients up to 48 months, excluding patients for whom only 24-month data were available.
April 15, 2013

Rongwei Fu, Ph.D
3181 SW Sam Jackson Park Rd.
#CB669
Portland, OR 97239
USA

Dear Dr. Fu:

Congratulations! We are pleased to accept for publication your manuscript, M12-2731, "Effectiveness and Harms of Recombinant Human Bone Morphogenetic Protein-2 (rhBMP-2) in Spine Fusion: A Systematic Review and Meta-Analysis."

As indicated in our letter requesting revisions, Annals routinely publishes a statement with original research articles (Article, Brief Communication) that indicates the availability to readers of each of the following three items: study protocol (original and amendments), computer code that authors used to generate the results reported in the published article, and the documented analytic dataset. If you have not already provided us with this information, please send an email to your manuscript representative that includes statements about the availability of these items. Description of availability should describe any terms on which availability is contingent. These terms may span from completely unrestrictive access (e.g., free availability of the entire raw data set via posting on an open access web site) to more restricted access (e.g., availability of certain portions of the analytic dataset to approved individuals through written agreements with the author or research sponsor). Below are two examples of how the information might appear.

Sample Reproducible Research Statements

Example 1
Protocol: not available
Statistical Code: Available to interested readers by contacting Dr. Smith at docsmith@med.edu
Data: not available

Example 2
Protocol: available at www.authorswebsite.org
Statistical Code: available at www.authorswebsite.org
Data: available for purchase from Dr. Smith docsmith@med.edu
Annals now gives authors the opportunity to submit brief videos about their articles. These videos can increase the attention an article receives from readers, reporters for national and international news venues, and researchers. Instructions for video development and submission can be found here. Please notify Rob Blackwell: rblackwell@acponline.org if you plan to submit a video.

A production editor will inform you when your paper has been scheduled for a specific issue and when you can expect proofs.

Please take care about the people that you talk with about your article and its planned publication in Annals. Doing so benefits you, the author, in several ways. First, it reduces the risk that another journal will learn about your article and publish a competing article before your article, which will make your work appear less novel and newsworthy. Second, some articles are of interest to the financial world. Premature disclosure of the content could appear as an attempt to aid and abet people who want to profit from insider knowledge. For these reasons, please do not talk about your article except to people that you know and trust. If you have any questions, please contact the editors.

Annals sends advance information about each issue to members of the news media and ask that they not release news stories before a specified embargo date of 5:00 p.m. Eastern time of the day before the publication date. If you are contacted prior to publication by members of the media who wish to discuss the content of your paper, we ask that you honor the embargo date. We also ask that you not distribute copies of your paper before publication. If the public relations office of your institution is planning its own publicity for your article, they are under the same constraints as the media regarding release of information and should contact the American College of Physicians' press office at 215-351-2653.

My colleagues and I very much appreciate the opportunity to publish your fine work.

Sincerely,

Darren Taichman, MD, PhD

Executive Deputy Editor
Hi Beth,

A total of 106 patients initially entered the study and 102 patients had results from the two-year follow-up. The results were reported on the 102 patients and the abstracts of the study quoted 102 patients, so we used 102 patients in our paper.

Thanks,
Rochelle

Beth Jenkinson
Supervisor, Editorial Production
Annals of Internal Medicine
Phone: 215.351.2645
Fax: 215.351.2644
E-mail: bjenkinson@acponline.org