META-ANALYSIS OF THE INFLUENCE OF PATIENT CHARACTERISTICS ON THE EFFECTIVENESS AND HARMS OF RECOMBINANT HUMAN BONE MORPHOGENETIC PROTEIN-2 IN LUMBAR SPINAL FUSION

by

Amber L. Laurie

A THESIS

Presented to the Department of Public Health & Preventive Medicine

and the Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements for the degree of

Master of Science

June 2014

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Amber L. Laurie

has been approved

Mentor/Advisor

Member

Member

TABLE OF CONTENTS

Contents

LIST OF TABLES iii
ACKNOWLEDGMENTS v
ABSTRACT 1
INTRODUCTION 3
Background 3
Spinal Fusion Techniques 4
Purpose 6
Patient Characteristics and Outcomes in Lumbar Spinal Fusion 6
METHODS 8
Outcomes 9
Patient Characteristics 10
THEORY 11
Random Effects 12
Random Effects in Logistic Models 13
Random Effects in Continuous Models 14
Estimation Methods 15
Longitudinal Modeling 17
Denominator Degrees of Freedom 19
Comparison of Calculated DDFM 21
Model Building 23
RESULTS 27
Preliminary Analyses 27
Outcomes 27
Patient Characteristics 29
Final Model Results 35
Primary Effectiveness Outcomes: Fusion and Success 35
Secondary Effectiveness Outcomes: Self-Reported Outcomes 36
Harms Outcomes: Adverse Events 37
Results of Sensitivity Analysis 39

Estimate of Trial-Specific Random Effects 40
DISCUSSION 41
Strengths 42
Limitations and Future Directions 42
Lack of Blinding 42
Differences between Trials 43
Limitation of BMI 45
Statistical Limitations 45
Missing Data 46
Considerations Regarding Subgroup Analyses 48
SUMMARY AND CONCLUSIONS 49
REFERENCES 50
TABLES 56
FIGURES 81
ABBREVIATIONS 88

LIST OF TABLES

Table 1. Study details 56

Table 2. Definition of outcomes 57

Table 3. Table of AIC values for variance-covariance matrices in full main effects models for longitudinal repeated measures 58

Table 4. Table of denominator degrees of freedom and p-values for main effects in full main effects model for longitudinal repeated measures 59

Table 5. Distribution of patient characteristics by type of surgery 60

Table 6. Frequency and percent of missing data by study period 61

Table 7. Frequency and percent of patients with desired outcomes by follow-up period 62

Table 8. Table of change in fusion and success outcomes since last follow-up 63

Table 8. Cumulative frequency and percent of patients with adverse events by follow-up period and type of adverse event 64

Table 10. Summary statistics for continuous, self-reported outcomes over time 65

Table 11. Overall sample participant characteristics 66

Table 12. Percent of patients with fusion or success within each patient characteristic for patient characters that are significantly related ($p<0.05$) to adverse events 67

Table 13. Percent of patients with an adverse event within each patient characteristic group for patient characters that are significantly related ($p<0.05$) to adverse events. 68

Table 14. Average change in self-reported outcomes between baseline and the final follow-up at 24 months within each patient characteristic group for patient characters that are significantly related ($p<0.05$) to change in self-reported outcomes. 69

Table15. Percent of patients with fusion and success within each patient characteristic group at the 24 months follow-up. 70

Table 16. Table of odds ratios for fusion for comparing rhBMP-2 versus ICBG by patient characteristic 71

Table 17. Table of odds ratios for fusion for comparing levels within patient characteristics 72

Table 18. Table of odds ratios for success for comparing rhBMP-2 versus ICBG by patient characteristic 73

Table 19. Table of odds ratios for success for comparing levels within patient characteristics 74

Table 20. Table of estimated difference in self-reported outcomes between rhBMP-2 versus ICBG by period 75

Table 21a and b. Table of estimated difference in self-reported outcomes between different levels of patient characteristics 76-77

Table 22. Table of odds ratios for adverse events for comparing levels within patient characteristics 78

Table 23. Table of odds ratios for related and related, severe adverse events for comparing rhBMP-2 versus ICBG by previous surgery 79

Table 24. Table of odds ratios for related and related, severe adverse events for comparing levels within patient characteristics 80

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Rochelle Fu for involving me in this project. It was an especially enjoyable experience to work with her. I am deeply grateful for her guidance and encouragement.

I would like to thank the other members of my committee: Dr. Yiyi Chen for acting as my committee chair and for her expert statistical knowledge and Dr. Roger Chou for his adept editing and clinical input.

Finally, I would like to thank my husband and stable base, Jonathan Chin. I am amazed by his unconditional support, and I am fortunate to have his friendship.

ABSTRACT

Purpose

To determine how patient characteristics impact estimates of effectiveness and harms of rhBMP-2 versus iliac crest bone graft (ICBG) in lumbar spinal fusion.

Methods

Using individual patient data from 10 industry-sponsored randomized, controlled trials of rhBMP-2 including 1,255 patients, linear and generalized linear mixed models were used to assess effects of patient characteristics on estimates of effectiveness and harms of rhBMP-2.

Results

At 6 months, patients younger than 60 years had higher odds of fusion with rhBMP-2 versus ICBG (OR: 3.17, 95% CI: 1.20 to 8.37), but there was no treatment difference in persons over 60 years (OR: 1.12, 95% CI: 0.37 to 3.37). At 24 months, smokers had significantly higher odds of fusion with rhBMP-2 versus ICBG (OR: 4.90, 95% CI: 2.42 to 9.91), but there was no treatment difference in non-smokers (OR: 1.45, 95% CI: 0.85 to 2.47). In normal and overweight patients, rhBMP-2 was associated with higher odds of fusion versus ICBG at 24 months (Normal weight OR: 3.14, 95% CI: 1.46 to 6.73; Overweight OR: 2.60, 95% CI 1.29 to 5.22). For obese patients, effects of rhBMP-2 were smaller and not statistically significant (OR: 1.92, 95% CI: 0.82 to 4.50), and for severely obese patients, there was a non-significant trend for decreased odds of fusion in rhBMP-2 versus ICBG (OR: 0.36, 95% CI: 0.11 to 1.11). Similar results were found for the outcome of success, with normal weight patients having significantly higher odds

of success with rhBMP-2 versus ICBG at 24 months (OR: 2.17, 95% CI: 1.36 to 3.44), but no treatment difference for all other weight groups.

Patients without a previous back surgery had reduced odds for related or related, severe adverse events at 24 months with rhBMP-2 as opposed to ICBG (related OR: 0.22, 95% CI: 0.09 to 0.51; related, severe OR: 0.09 95% CI: 0.03 to 0.32). For patients with a previous back surgery, there was not a significant treatment difference in harms (related OR: 2.10, 95% CI: 0.86 to 5.09; related, severe OR: 1.41, 95% CI: 0.53 to 3.77).

Limitations

Analyses are limited by the data in that there was up to 12% missing data on outcomes at 24 months and patients and study staff was not blinded to the treatment.

Conclusions

The results indicate that baseline smoking status, age, BMI and presence of a previous surgery had an impact on estimates of effectiveness and harms for rhBMP-2 versus ICBG in lumbar spinal fusion. There is preliminary support for rhBMP-2 increasing fusion for smokers and individuals under the age of 60, and improving fusion and success for patients of normal weight, but not for patients that are obese or severely obese, where rhBMP-2 may be less effective than ICBG. Also, rhBMP-2 resulted in decreased device-related adverse events in individuals with no previous back surgeries but not within individuals with a previous back surgery. Future studies of rhBMP-2 should include planned subgroup analysis in patients over 60, smokers, patients that are obese and severely obese, and individuals with previous back surgeries.

INTRODUCTION

Background

The development of new medications and medical devices depends largely on industry-sponsored clinical trials. There are a number of ways in which industry-sponsored studies can be affected by biases that obscure the true benefit of the product being developed. In recent years, this topic has been actively discussed, and a recent meta-analysis found that industry-sponsored trials have 3.5 times the odds of reporting results that promote the industry sponsor versus non-industry sponsored trials [1]. One way to examine for biases in research is to allow for the external and independent analysis of individual patient data (IPD) to investigate the findings of industry sponsored projects [2].

The use of individual patient data has long been considered the gold standard in meta-analysis [3]. The traditional method of meta-analysis using only summary data from publications can provided biased results compared to IPD analysis. Furthermore, the use of IPD allows for additional analyses that are not able to be performed with summary data including analysis by patient characteristics, use of standardized outcomes across studies, and the use of sophisticated modeling procedures to control for and estimate trial-related heterogeneity.

The Yale University Open Data Access (YODA) project seeks to facilitate the analysis of IPD. The YODA project acts as a repository of raw data from industry-sponsored clinical trials. Independent investigators can apply to have access to this data, on which they can perform their own reviews. This model allows for the independent and external

validation (or refutation) of findings, and in addition, allows for the incorporation of analysis of data that has not been published.

The first realization of the YODA project was for a product, recombinant human bone morphogenetic protein-2 (rhBMP-2), manufactured by Wyeth-Genetics Institute (A Division of Pfizer Inc., St. Louis, MO), and used with INFUSE® Bone Graft/LT-CAGE® Device produced by Medtronic (Minneapolis, Minnesota). RhBMP-2 is a bone graft substitute for one's own bone (autogenous) material and is typically used in spinal fusion. Two meta-analyses on the use rhBMP-2 in spinal fusion have recently been completed as a result of funding from the YODA project [4,5]. Both of these reviews used IPD from Medtronic-sponsored, randomized, controlled trials as well as published literature to examine efficacy and safety of rhBMP-2 in comparison to the standard of care, iliac crest bone graft (ICBG). Both reviews found similar results in that rhBMP-2 has minimal to no clinical advantage over bone graft and that use of rhBMP-2 may be associated with an increase in risk of some complications. The meta-analyses performed by Fu et al. and Simmonds et al., established the overall effect and potential harms of rhBMP-2, and with the availability of IPD, the current analysis extends the prior meta-analyses to examine the effect of patient characteristics on efficacy and safety of rhBMP-2 as compared to ICBG.

Spinal Fusion Techniques

The current analysis was restricted to RCTs on lumbar spinal fusion which can vary by the surgical approach (anterior or posterior) and by whether or not the disc between vertebrae is removed. For this analysis, three specific surgical types of lumbar spinal fusion were included, anterior lumbar interbody fusion (ALIF), posterolateral fusion

lumbar fusion (PLF), and posterior lumbar interbody fusion (PLIF). In an anterior procedure, the surgical approach to the spine is made in the abdomen as opposed to the posterior and posterolateral approaches in which the surgical approach is from the back. For an interbody type of spinal fusion, a vertebral disc(s) is removed and the bone graft material is placed between the vertebrae in the remaining space. Both ALIF and PLIF use this fusion method. In spinal fusion with PLF, the bone graft is placed in the inter-transverse area and fixed in place with screws or wires.

Anterior lumbar interbody fusion (ALIF) is the only surgical approach for which rhBMP-2 is FDA approved [6]. ALIF has fewer risks than the posterior methods in that the back muscles and nerves are not contacted during surgery and that larger implant cages can be inserted. Placing the bone graft from the anterior approach results in a compressed state, and this has been related to increased fusion as opposed to PLIF. However, because of the anterior approach, this method may lead to an increase in retrograde ejaculation, a condition in which semen enters the bladder instead of exiting via the urethra [7,8].

In the posterior approaches the back nerves and muscles are directly affected and are involved in the healing process. In PLIF, the posterior approach leads to less space for the bone implant insertion and only smaller sizes of implants may be placed. As a result PLIF has been linked to neural compression and more complications [5] and is not commonly used. RhBMP-2 is not FDA approved in for PLF or PLIF procedures, but rhBMP-2 is commonly used off-label in PLF [5].

With the use of rhBMP-2, patients undergo one less surgical procedure as there is no need for autogenous ICBG. It has been reported that patients with rhBMP-2 as

compared to ICBG have less blood loss [9]; therefore this could be an indication for the use of rhBMP-2 in obese patients as it could potentially reduce the risk for surgery complications and blood loss by reducing the total time of surgery and procedure sites needed for the procedure.

Purpose

The current study aims to build on the recent systematic reviews and meta-analyses by using IPD obtained from the YODA project to focus on the effect of patient characteristics and determine if there are patient characteristics that change the effectiveness or risk of harm when comparing rhBMP-2 to ICBG. Previous research suggests that some patient characteristics may modify the effectiveness of rhBMP-2. These patient characteristics include age at baseline, body mass index (BMI), number of previous surgeries, fusion type, number of lumbar spinal levels treated, diabetes, pre-operative work status, smoking, and sex. The primary objective of this analysis is to determine how patient characteristics impact estimates of effectiveness-as measured by both objective and self-report measures of rhBMP-2 versus iliac crest bone graft (ICBG). A secondary aim of this analysis is to determine effects of patient characteristics on estimates of overall harms of rhBMP-2 versus iliac crest bone graft (ICBG) in lumbar spinal fusion.

Patient Characteristics and Outcomes in Lumbar Spinal Fusion

The effectiveness of spinal fusion has been linked to a number of patient characteristics via subgroup analyses in RCTs and meta-analyses. Two particular patient characteristics – age and smoking status – have been assessed to look at differential treatment effect between rhBMP-2 and ICBG.

Elderly patients are at an increased risk of complications from lumbar spinal fusion [10]. Previous work on the effectiveness rhBMP-2 in spinal fusion for the elderly found that for patients over 60 years old, the use of rhBMP-2 as compared to ICBG decreased costs and improved outcomes, but this study did not examine how age and treatment might have affected the outcomes simultaneously [11].

Both pre-operative and post-operative smoking status have been found to be associated with poorer outcomes in lumbar spinal fusion [12,13]. A study comparing rhBMP-2 and ICBG in smokers and non-smokers found that patients who smoked pre-operatively had significantly lower fusion rates than non-smokers, but within patients who smoked, patients with rhBMP-2 had non-significantly higher rates of fusion than those with ICBG [13].

Research on the relationship between rhBMP-2 and the other patient characteristics (BMI, number of previous surgeries, fusion type, number of lumbar spinal levels treated, diabetes, pre-operative work status, and sex), is limited, but there is research on the effect of these characteristics on outcomes in lumbar spinal fusion. Obesity is associated with increased risk of surgery complications in lumbar fusion, less improvement via some self-report measures, and more blood loss as a result of surgery [14-16]; though obese patients still experience significant improvement from baseline.

Patients with diabetes are a higher risk for many health complications. Two retrospective studies found an increase in complications for patients with diabetes undergoing fusion surgery, as well as lower fusion rates [17,18]. These increased complications mainly involve infections at the operation site, and may be related to impaired wound healing in diabetics.

For patients with repeated back surgeries, the risks of complications are increased and the outcomes are significantly worse than for individuals without a prior back surgery. One study found coexisting morbidity associated with poor outcomes in patients with repeated back surgeries, but not in patients with only without a previous back surgery [19].

Pre-operative work status has also been identified as a predictor of outcomes of lumbar spinal fusion. Specifically, it has been found that patients who were working pre-operatively had significantly better outcomes after lumbar spinal fusion [20].

Race and age have also been evaluated as predictors of outcomes of spinal fusion. In an analysis of gender and outcomes in spinal fusion, across all ages, women were less satisfied with the outcome of surgery, albeit this finding was non-significant (p=0.06) [21]. Within elderly patients of spinal fusion, women were significantly less satisfied with surgery, but also had a higher prevalence of comorbidities such as heart disease, diabetes, and depression than did men [22].

METHODS

The data included in this analysis was obtained from the YODA project for the explicit purpose of this analysis and previously summarized in the report by Fu et al. (2013). From the data provided by the YODA project, patients with IPD from ten randomized controlled trials of rhBMP-2 for lumbar spinal fusion were included. We restricted the analysis to studies of fusion with rhBMP-2 versus ICBG. Only outcomes collected through the first 24 months were used, as most studies did not report follow-up after 24

months. Table 1 lists the studies included in the meta-analysis with the number of patients included in each treatment arm.

Outcomes

For the primary objective related to effectiveness, the outcomes used for this analysis were overall success and spinal fusion. Overall success and fusion outcomes were defined dichotomously, as previously described in Fu et al (2013) (Table 2). For these variables, if some data was missing, patients were classified as failures, but if all data was missing, patients were excluded from the analysis. Fusion and overall success were assessed at 6, 12, and 24 months (the process of fusion typically takes up to 6 months).

Two sensitivity analyses were performed for fusion and success outcomes. In the first sensitivity analysis, individuals who had missing outcomes at 12 or 24 months were assumed to have positive outcomes if they had a positive outcome at the most recent non-missing outcome. In the second sensitivity analysis, individuals who had a positive outcome at 6 months were assumed to have a positive outcome at 12 and 24 months, and individuals with a positive outcome at 12 months were considered to have a positive outcome at 24 months.

The secondary outcomes for the primary objective are patient reported functional status measured using the Oswestry Disability Index (ODI) [23] and Physical Component Summary (PCS) and Mental Component Summary (MCS) of the Short Form (36) Health Survey Harms [24]. Thresholds for minimum clinically important differences were defined as 12.8 for the ODI and 4.9 for the SF-36 PCS [25]. We used the same threshold for a minimum clinically important difference for SF-36 MCS as for the SF-36 PCS. These

outcomes were assessed at all available time points, which included baseline, 6 weeks, 3 months, 6 months, 12 months and 24 months. A random sample of 50 participants was chosen and a spaghetti plot of ODI, SF-36 PCS and MCS scores over time was also plotted. Missing data were tabulated by outcome and period.

The outcomes of for the secondary objective related to harm including overall AEs, SAEs (SAEs), device-related AEs, and AEs that are both severe and device-related. SAEs are defined as AEs that put the patient at immediate risk of death, require the removal of the implant or device, or those that limit the patient's ability to perform routine activities. Related AEs are those that are deemed by study medical staff to be caused by the device. AEs were analyzed as present or absent at 4 weeks and 2 years with patients having had one or more AEs considered AE present. Missing data on AEs was presumed to indicate the lack of an AE. Distribution of AEs at 4 weeks and 2 years was plotted and tabulated to determine the average occurrence of the different types of AEs in the study over time.

Patient Characteristics

We selected patient characteristics for analysis based on past research of clinical relevance, presence of between-patient variability on the characteristic, the amount of missing data, and the possibility missing data introducing bias. For continuous patient characteristics (age and BMI), we defined categories for analysis. Age was dichotomized as <60 or ≥ 60 years of age based on previous research in spinal fusion [11]. BMI was categorized based on the U.S. Preventive Services Task Force with a modification to combine Class II and Class III obesity into a single category of severely obese due to small sample size of Class II obesity (Underweight < 18.0, Normal weight

18.0 to 24.9, overweight 25.0 to 29.9, obese 30.0 to 34.9, and severely obese BMI ≥35.0) [26]. We used chi-square tables to determine if the distribution of patient characteristics was equivalent across treatment groups. Associations among categorized patient characteristics were assessed using chi-square tables. Outcomes across all time points were assessed by patient characteristics using chi-square tables for the dichotomous outcomes and t-tests for continuous outcomes. We also performed chi-square analyses to evaluate the association between missingness status and patient characteristics. Also, to determine if the distribution of covariates was affected by dropout, the distribution of patient characteristics with non-missing outcomes was determined for the final follow-up compared to the distributions of patient characteristics at baseline. To assess for interaction effects at the most basic level, within each level of all patient characteristic, a chi-square analysis was done between outcome and treatment.

Since this study focused on the surgical procedures of ALIF, PLF, and PLIF, patients with laparoscopic surgery or without instrumentation were excluded.

THEORY

One important aspect of this project was the use of individual patient data, which allowed for analysis at the level of patient characteristics and permitted use of more sophisticated modeling procedures. A number of statistical procedures and concepts were used for this project and are discussed below. These include random effects, quadrature methods for estimation of the likelihoods, longitudinal data modeling, and estimation of denominator degrees of freedom.

Random Effects

To control for heterogeneity of the treatment effect between the studies, a random effects regression model was used for all outcomes. Random effects models, which are often referred to as variance component models, are a type of hierarchical linear modeling which assumes that all data belongs within the hierarchy and differences in the outcomes are related to the membership in that hierarchy [27,28]. The heterogeneity in outcomes that arises from this membership is unobserved and random effects models estimate this unobserved heterogeneity. Random effects models allow for the estimation of the variance of both the main effects and the unobserved (random) effects and the variance of the overall outcome is the sum of the variance of these two components- hence the nickname variance component models. As a result, models with random effects are generally more efficient resulting in decreased variance and increased power.

Differences between treatment effects were allowed to vary across studies, and these differences were blocked based on the type of fusion being performed in each study to provide further heterogeneity of the random effects parameters, or in other words, the effect for type of fusion was nested within the study effect. Since only one study included the use of PLIF, both posterior treatment modalities were combined into a single group for the purpose of estimation of random effects. Full independence was assumed across all studies and patients within each study. This model resulted in allowing the regression coefficients to vary from study to study.

Random Effects in Logistic Models

The specification of random effects differed based on the inclusion of interaction terms and whether or not the outcome was modeled at a single point or longitudinally. For dichotomous outcomes, all outcomes were modeled at a single time point. The one-step model for binary outcomes as described in Turner et al. [29] was used and is described below:

$$\text{Logit}(\pi_{ij}) = \alpha_i + (\theta + \delta_i)x_{ij} + \boldsymbol{\beta X}$$

$$\delta_i \sim N(0, \tau^2)$$

where π_{ij} is the predicted response probability for the jth subject in the ith study, α_i is the trial effect for the ith study, θ is the estimated overall treatment effect, δ_i is the deviation from the overall treatment effect for the ith study, x_{ij} is the treatment (0 or 1) for the jth subject in the ith study, β is a vector of coefficients of fixed effects (patient covariates), X is a vector of fixed effects (patient covariates). δ_i is distributed normally with mean 0 and varianceτ^2, which is the estimated between-study treatment variation.

With the additional of an interaction term between, say, a binary patient covariate and treatment, the model becomes

$$Logit(\pi_{ij}) = \alpha_i + (\theta_0 + \delta_i)x_{ij} + \gamma\, covariate_{ij} + \theta_1\, x_{ij}\, covariate_{ij} + \boldsymbol{\beta X}$$

$$\delta_i \sim N(0, \tau^2)$$

where $covariate_{ij}$ is a binary patient characteristic for the jth subject in the ith study involved in an interaction with treatment, and θ_0 is the estimated overall treatment effect

when smoking=0 (for example, non-smoker) and ($\theta_0 + \theta_1$) is the estimated overall treatment effect when, for example, smoking=1 (smoker).

Random Effects in Continuous Models

Within the continuous outcomes (ODI, SF-36 PCS and MCS) all data was modeled longitudinally, and the generalized model can be written as [30]

$$Y_{ijk} = \alpha_i + (\theta_k + \delta_{ik})x_{ijk} + \boldsymbol{\beta X} + \varepsilon_{ijk}$$

$$\delta_{ik} \sim N(0, \tau^2)$$

where Y_{ijk} predicted response for the jth subject in the ith study at the kth time point, α_i is the trial effect for the ith study, δ_{ik} is the deviation from the overall treatment effect for the ith study at the kth time point, θ_k is the estimated overall treatment effect at the *k*th time point, x_{ijk} is the treatment (0 or 1) for the jth subject in the ith study, $\boldsymbol{\beta}$ is a vector of coefficients of fixed effects (other patient characteristics), X is a vector of fixed effects (other patient characteristics), and δ_{ik} is modeled normally with mean zero and variance τ^2 , which is the estimated between-study treatment variation. ε_{ijk} is the residual which is modeled via the unstructured variance-covariance matrix. The above model implies an interaction between treatment and period by estimating a separate treatment effect at each time point. An important assumption is that the random effects are distributed $N(0, \tau^2)$ where τ^2 is the variance of the random effects. No significant interactions between treatment and patient characteristics were found for the continuous, longitudinal outcomes.

Estimation Methods

For complicated models- like those that include random effects- quadrature rules may be used to approximate a definite integral of a marginal log likelihood function. For the logistic models in this analysis, a Gauss-Hermite quadrature method was used to approximate the marginal log likelihood, as other methods did not converge. A Gaussian quadrature rule is a specific type of quadrature rule that can find an exact value for functions with $2n-1$ degrees from -1 to 1 by summing across the value of the function multiplied by a weight, w_i, and n is the number of quadrature points.

$$\int_{-1}^{1} f(x)dx \cong \sum_{i=1}^{n} w_i f(x_i).$$

The Gauss-Hermite approximation extends this so that it may be applied to functions of the exponential family form so that

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=1}^{n} w_i f(x)_i$$

where n is the number of quadrature points, w_i are the weights and each x_i is the root of the Hermite polynomial $H_n(x_i)$ with i from 1 to n, with n being the number of quadrature points. In the Gauss-Hermite approximation, the weights are calculated as

$$wi = \frac{2^{n-1} n! \left(\sqrt{p_i}\right)}{n^2 [H_{n-1}(x_i)]^2}.$$

For a Gauss-Hermitte approximation to be appropriate there must be conditional independence and the absence of residual random effects. Also, the model must have subject-level effects, either that of random effects and/or repeated measures.

This method was further specified as being an adaptive Gauss-Hermite quadrature approximation [31,32], which means that a rule is used to determine when the number of quadrature points is large enough to ensure that the error of the approximation is minimized past a threshold. The process of adaptive quadrature estimation is as follows. The starting values for the estimates of the fixed effects are estimated with the first quadrature approximation and are fit with a generalized linear model. Given the fixed effects produced by the generalized linear model, the covariance parameters are estimated. Then the pseudo-likelihood is updated to improve these estimates for the fixed effects and obtain estimates for the random effects. At this point, if the number of quadrature points is not specified, the procedure will determine how many quadrature points are required to minimize the difference of the log-likelihood approximation. If the relative difference does not fall below a required threshold, then the model will fail to converge. Otherwise, the process will continue with repetitions of estimations for fixed effects, covariance parameters, and random effects until convergence is achieved.

The number of random effects greatly increases the number of conditional log likelihoods that are estimated, as the number of conditional log likelihoods estimated are n^r with r being the number of random effects. For this analysis, quadrature points were always specified, so in effect, an adaptive process was not utilized. Models were estimated with a varying number of quadrature points and the point at which the model become stable was determined to be N=5 quadrature points. Stability was determined by a difference of less than 1/10 of a percent difference in all coefficients between models. For the purpose of this analysis, quadrature estimation was specified to use N=15 quadrature points to be over-conservative in ensuring stable estimates.

Longitudinal Modeling

For the models of continuous, self-reported outcomes, longitudinal mixed models were used as the self-reported outcomes were shown to have a linear change over time. Using notation from SAS PROC MIXED [33] the longitudinal mixed model can be written as follows,

$$y = X\beta + Z\gamma + \varepsilon$$

Where y is a vector of the outcome, γ is a vector of random effects, X is a matrix of main effects, Z is a design matrix of random effects, with γ and ε distributed normally. The variance matrix of y, is modeled jointly by an additive combination of the covariance structures of both repeated, residual and/or random effects. There are a number of choices for how to model the covariance structures of repeated measures, and the optimal choice will well characterize relationships in the data over time, minimize bias due to missing outcomes, and maintain parsimony. With the selected random effects and covariance structure repeated effects, a restricted/residual maximum likelihood method is used to obtain estimates of the function.

There are a number of ways to model the covariance structure, and SAS's PROC MIXED offers over 20 methods for creation of a covariance structure. For this project, 8 common variance-covariance matrices were tested for the full main effects model for ODI, SF-36 PCS, and SF-36 MCS and compared via AIC values (Table 3). Structures assessed were the homo- and heterogeneous autoregressive, compound symmetry, and Toeplitz matrices, as well as the variance components and unstructured matrices. The number of parameters estimated for each structure ranged from 21 with the unstructured

matrix to the variance components in which only one parameter was estimated. Models were compared by the use of AIC which is a measure of goodness of fit that is corrected for model complexity with the best model being a well-fitting and parsimonious model. AIC is calculated by

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model and L is the maximum likelihood of the model. When comparing models, the best model will be the model with the lowest AIC value. Another measure, AICc includes an even stronger correction for a model with many parameters [34]. AICc is calculated by

$$AICc = \frac{2k(k+1)}{n-k-1} + AIC$$

where n is the sample size. When the sample size is large relative to the number of parameters in the model, AIC and AICc will be very similar. Since the sample size was large, in all cases AIC and AICc were within 1 percentage; therefore, the more familiar AIC was used. For all three outcomes, the unstructured variance-covariance matrix was superior to any other method with all other methods having an AIC that was at least 151 points higher. Although the unstructured variance-covariance matrix was the most complex, it provided the best fit among the options and was used throughout for the final models.

Denominator Degrees of Freedom

Within the bounds of the random effects and longitudinal models, there were a number of choices to be made around the model specifications. One such specification was around the denominator degrees of freedom (DDFM).

In a simple setting, take a one-way ANOVA, the DDFM are the degrees of freedom for the within subjects component or the difference between the N number of subjects and k number of groups being compared. These degrees of freedom are used as the basis for the distribution on which p-values are calculated. As models become more complex with various groups being compared and various nesting procedures occurring with random effects, the calculation of DDFM become less clear. There are a number of methods for calculation of DDFM in these complex models, and the best method will be the one that minimizes Type I error (error of rejecting the null hypothesis when the null hypothesis is true) and increases power, i.e., the probability of rejecting the null hypothesis when the null hypothesis is not true. In this analysis, six calculation methods for DDFM were considered and compared.

The containment method is the default when random effects are included in γ or the $(r\ x\ 1)$ vector of random effects (i.e. random effects that are not considered residual). It is not recommended for unbalanced data regarding patient characteristics. For random effects models, this method chooses the smallest contribution of random effects. Where A is a vector of fixed effects, and B is a vector of random sources of variation, the containment method chooses the lower of (a-1) versus (b-1). When there are no random effects, the containment method reduces to the residual method for DDFM.

The residual method for calculation of DDFM uses the sum of the frequencies used minus the rank of the X matrix of $(n\ x\ \ p)$ of main effects. This method of calculation is superior for studies where there is unbalanced data within the patient characteristics [32]. The residual method is not changed by the presence of random effects.

The between-within method for calculation of DDFM uses the residual degrees of freedom which are divided into between and within components. If there are no random effects or repeated effects by subject, then this method will result in DDFM equivalent to that of the residual procedure.

The Satterthwaite method for estimation of variance was proposed by Satterthwaite in 1946 [35]. In the context of this analysis, the degrees of freedom are found by the creation of a linear combination of the mean squares. The Kenward-Roger estimation method uses the Satterthwaite procedure, except in that it uses an inflated variance matrix which makes the estimation more conservative than Satterthwaite. The Kenward-Roger and Satterthwaite methods cannot be used along with the quadrature method of estimation, so they were not able to be tested within the context of the logistic models.

In certain statistical procedures, it is possible to use an infinite number of degrees of freedom. When this is done p-values for the estimates for the fixed effects are calculated via the standard normal distribution instead of the t-distribution, and p-values for the overall tests for fixed effects are calculated via the chi-square distribution instead of the F-distribution, with the degrees of freedom for the chi-square distribution equal to the numerator degrees of freedom. This is the most liberal of the methods and will result in the lowest p-values. Logistic models were performed in PROC GLIMMIX, and models were estimated for the purpose of comparisons with an infinite number of degrees of

freedom, however, continuous models were performed with PROC MIXED which does not support infinite number of degrees of freedom, so these models were not estimated with this particular parameter setting.

Comparison of Calculated DDFM

In order to compare the various methods of DDFM, the continuous, longitudinal outcome of ODI was used as all types of DDFM methods were available except that of an infinite number of DDFM. Therefore a repeated mixed effect model for ODI was performed using residual, containment, between-within, Satterthwaite and Kenward-Roger methods for calculation of DDFM. Table 4 shows a comparing of the denominator degrees of freedom used for the overall test of fixed effects for each variable in the model.

For the residual method, DDFM is the same for all variables and is the largest DDFM at 7176. The residual DDFM was estimated by the calculating the difference between the number of observations and the covariance parameters estimated (7199-23=7176).

For this model, the containment method calculates the DDFM by taking the number of observations, subtracting the number of covariance parameters, and then splitting the DDFM between the random effect variables and the fixed effects. This results in a DDFM of 7176 being split between 7167 for all fixed effects and 9 for the effect of treatment, which resulted from 10- 1 studies on which the random effects are based.

The between-within method had the same overall DDFM as the residual method, but it was split into between and within components. The variables with the covariate of period (period and period x treatment interaction) had 5935 DDFM (within) and the

remaining variables had 1241 DDFM (between) (5935+1241=7176; within DDFM + between DDFM= total residual DDFM).

A comparison of the Satterthwaite and Kenward-Roger methods for estimation of variance components found that the DDFM were very similar between the two with a maximum difference of less than 2%. In fact, the DDFM estimates were the same for all variables except those in which there was more than 2 levels (period, tx *x* period interaction, type of fusion, and BMI). The Kenward-Roger and Satterthwaite procedures are not able to be used when adaptive quadrature, which was necessary for convergence with all logistic models.

In a comparison of DDFM methods, all estimates were equivalent for estimate standard error except in the Kenward-Roger estimation which as mentioned above uses an inflated variance covariance matrix, resulting in the standard error being larger. For this analysis of ODI, the difference between the other estimation procedures and Kenward-Roger model resulted in a median difference of 0.2% in standard error with a maximum difference of 8% for the estimate for treatment and the treatment (SE 0.7277 in other models compared to 0.7871 in the Kenward-Roger model).

In comparing the p-values for the overall test for significance there were no changes in results of significance of estimates for this particular analysis, except in that treatment was significant to $p<0.05$ level in all estimation models except containment, which uses the smallest DDFM for treatment at DDFM=9. The maximum difference between p-values for overall effect was found to be 0.027 for the variable treatment between the Containment (p=0.053) and the residual method (p=0.026). It is worth noting that the

estimated coefficients were unaffected by varying these procedures since the estimation of DDFM does not affect this part of the estimation of coefficients.

The DDFM method chosen for this analysis was the residual method. It was desired that the DDFM method was the same for all models, which excluded the use of Satterthwaite, Kenward-Roger, and infinite DDFM. The containment method was not chosen as there was considerable unbalance within patient characteristics across the studies. The residual method was chosen as it is considered appropriate for unbalanced data.

Model Building

As stated above the dichotomous outcomes were success, fusion, and presence of AEs. Generalized linear mixed models were built for success and fusion separately for follow-up at 6, 12, and 24 months. This outcome is not stable once achieved so it was not deemed appropriate to use a survival model. AEs were modeled separately for two time points, 4 weeks and 2 years. The AE outcomes were determined by whether or not the participant had experienced one or more AEs up to the time point being modeled. Models were built separately for each time point for overall AEs, severe AEs, related AEs, and severe and related AEs. Longitudinal continuous models were built for the ODI, SF-36 PCS, and MCS outcomes. For this model, ANCOVA was considered but not used due to desire to keep model interpretation as clear as possible and due to baseline scores being equivalent between treatment groups.

Using the models with the specifications described in the theory section, a manual stepwise backward selection procedure was used. Within all models, treatment type and type of surgery (ALIF, PLIF, and PLF) were considered as a main effect to control for the

treatment and type of surgery, regardless of the significance of these variables. Within the longitudinal model for self-reported outcomes, an interaction between period and treatment was always included to control for the effect of time on treatment, regardless of the significance of the interaction.

Since the main purpose of this analysis was to examine for the presence of interaction effects between treatment and patient characteristics, model building initiated with a test for interactions. For each patient characteristic, tests for interaction effects between each patient characteristic was performed within the context of a full model with all patient covariates as well as a simple model with only the patient characteristic, treatment and the interaction between the two. For longitudinal models, as well as an interaction between treatment and covariate, a three way interaction between period, treatment, and covariate was used to determine if a treatment and covariate interaction differed across time. When the three way interaction between period, treatment, and covariate was found to be non-significant, then a two-way interaction between treatment and covariate was considered. Interactions were evaluated as quantitative or qualitative [36]. A quantitative interaction occurs when the effect of treatment is in the same direction for each subgroup, but the magnitude of effect differs. A qualitative interaction occurs when the effect is in the opposite direction between subgroups.

Any covariates with an interaction effect with treatment of $p<0.20$ proceeded to next model. The resulting model with all covariates and all interaction effects $p<0.20$ is subsequently referred to as the original model. Once the original model was constructed, variables were removed sequentially until $p<0.05$ for all interaction terms and covariates. Covariates were always retained in the model if there was an interaction

term with that covariate still present in the model. This reduced model is referred to as the preliminary model.

Confounding was tested on the preliminary model by individually adding all covariates not included preliminary model back into the model and assessing the change in all other covariates. All variables whose presence in the model resulted in a >10% change in a coefficient of a covariate were considered confounders and were added back to the preliminary model. For logistic models, the transformed coefficient of an odds ratio was evaluated for confounding.

All possible interaction terms that were not included in the preliminary model were added back to the model individually to test for significance in the presence of other terms. All main effects that were not in the preliminary model were also added to the preliminary model one-by-one and tested for significance. If a previously removed covariate obtained significance after being added to the preliminary model, it was included in the final model.

The final model included all covariates and interaction terms that were significant ($p<0.05$) or were a confounder to any of the variables that were significant. Again, the final model always included type (ALIF, PLIF, vs PLF), treatment and all longitudinal models included an interaction between period and treatment.

When multiple interactions between treatment and patient characteristics were included in a single final model, an issue with interpretation developed. For each estimate comparing rhBMP-2 with ICBG within a level of a patient characteristic was required to be interpreted in the context of the referent level of all other patient characteristics that interacted with treatment. For example, if both BMI and age had a significant interaction

with treatment, then the effect of treatment within individuals over 60 only applied to individuals that were of normal weight as well. For the sake of simplicity, it was desired to correct this so that interaction terms with treatment were able to be interpreted with each patient characteristic, independently of the levels of the other patient characteristics. To accomplish this, when calculating the interaction effect between treatment and a patient characteristic, the values of the other patient characteristics that interacted with treatment were held constant at the average value for that patient characteristic with the use of linear combinations

For sake of comparing models across time points, models were combined so that there were similar models across all time points for an outcome. The result is that the final models for fusion and success were consistent across all time points for that outcome. For the outcomes of fusion and overall success, preliminary models were built separately for each time point (6, 12 and 24 months), and the final models were constructed from all variables that were included in the preliminary models for each time point for fusion and success separately.

For the continuous, self-reported outcomes, preliminary models were built separately for ODI, SF-36 PCS, and SF-36 MCS, and the final models were constructed from all variables that were included in the primary models for any of the three outcomes. The result is a final, consistent model for all self-reported outcomes across ODI, SF-36 PCS and SF-36 MCS.

For AEs, preliminary models were built separately for overall, severe, related and severe/related AEs separately at 4 weeks and 2 years. Final models for overall and severe AEs were combined so that all variables that were significant in the individual

models were included in the overall model so that there was one consistent model across overall and severe AEs.

Final models for related and related and severe AEs were combined so that the final combined model was consistent for related and related, severe AEs.

Model building was repeated separately for ALIF and PLF models using the same procedure as above, except that surgery type was not included as a covariate, and random study effects were not grouped by surgery type. The final models built for the separate surgery types were compared and assessed against with the final model from the overall modeling building.

All analyses were performed with SAS 9.3 (Cary, NC). PROC GLIMMIX was used for logistic mixed models, and PROC MIXED was used and for the repeated measures mixed models for continuous outcomes.

RESULTS

Preliminary Analyses

Outcomes

This analysis included 622 patients with rhBMP-2 and 592 patients with ICBG across ten randomized, controlled trials (Table 1). Most fusions were of the PLF type (n=722), followed by the ALIF type (n=466), with the fewest patients with PLIF (n=67). All patient characteristics except sex were significantly different between ALIF and PLF fusion types (Table 5). Missing data for the outcomes are presented in Table 6. The percent of

missing data increased over time from the first follow-up to the last. Maximum missing data was between 10 and 12% at the 24 month follow-up.

The proportion of patients with fusion increased over time from 75% at 6 months to 88% at 24 months (Table 7; Figure 1). The proportion of patients with overall success increased over time from 46% at 6 months to 53% at 24 months (Table 7; Figure 1). A within-patient examination of fusion and success was performed to examine the stability of fusion or success once achieved. There were a number of patients who achieved fusion or success at one time point who did not criteria for these outcomes at subsequent time points. For example, at the 12 month follow-up, 6% of patients who had achieved fusion at 6 months were determined to not have achieved fusion as of 12 months. Table 8 shows fusion and success rates for the 12 month and 24 month follow-up periods. The majority of individuals that achieved fusion and success did so by 6 months and most of those individuals maintained fusion or success as time progressed.

Table 9 and Figure 2 show the percentage and frequency of AEs by 4 weeks and 2 years. The proportion of participants with any AE at 4 weeks was 31% and at 2 years was 55%. The proportion of patients with SAEs was 6% at 4 weeks and 12% at 2 years. The proportion of patients that experienced device-related AEs 1% was at 4 weeks and 5% at 2 years. The proportion of participants with a severe and device-related AEs was <1% at 4 weeks and 4% at 2 years.

Mean ODI scores improved (decreased) over time, with a mean improvement of 27 points (Table 10). Mean SF-36 PCS and SF-MCS scores improved (increased) over time, with a mean improvement of 12 points. For SF-36 MCS scores, the greatest improvement (6 points) was observed between baseline and 3 months, with little change

after 3 months (change of <1 point). Figures 3, 4, and 5 show SF-36 PCS, SF-36 MCS, and ODI scores for 50 random patients over time. Figure 6 shows mean SF-36 PCS and MCS scores over time, and Figure 7 shows mean ODI scores over time.

Patient Characteristics

Patient characteristics suspected of having an impact on benefits and harms of spinal fusion surgery include age, gender, BMI, smoking, presence of diabetes, presence of a previous back surgery, and preoperative work status. Table 11 presents the overall distribution of patient characteristics. For this analysis, age was categorized as a dichotomous variable with a cutoff of 60 years to replicate previous work [11]. 20% of all patients in the analysis were 60 years or old, compared with 7% 70 years or older.

Body mass index (BMI) was available as pre- and post-surgery measures. Pre-surgery BMI was used in this analysis, given prior evidence on its association with outcomes of fusion surgery. An analysis on pre-operative and post-surgical BMI showed no significant change. From the operative measurement to 6 weeks post-operatively, the median BMI measurement decreased from 27.4 to 26.6 kg/$m^{2,}$, and increased back to a median of 27.2 kg/m^2 at 24 months. This was determined to be a clinically non-significant change over time. In addition, we could not calculate postoperative BMI in a substantial proportion of patients due to missing data. The percent of missing BMI data ranged between 43 and 48% for each follow up period with the most missing data of 48% missing at 24 months.

For pre-operative BMI, only three participants were missing data. For one subject, the first post-surgery weight measurement was used to impute the missing pre-surgery BMI measurement. The other 2 participants were categorized as having missing BMI

measures. Both patients with missing (and not imputed) BMI had AEs at 2 years and both were considered failures for fusion and overall success at the 24 month follow-up. One of the patients with missing BMI had a related and severe AE and was a smoker. The other individual with missing BMI had diabetes. In order to protect the integrity of the model, these two participants were included in the analysis by creating a missing category for BMI. There were 8 (< 1% of the entire sample) individuals with underweight BMI, so we included them with the normal weight patients.

Within the dataset, there was smoking status and number of cigarettes per day at the pre-operative measurement as well as for post-operative measures. The researchers aimed to categorize pre-operative smoking status as light (<20 cigarettes per day) or heavy (20+ cigarettes per day). However, 38% of patients that reported smoking pre-operatively were missing number of cigarettes. Post-operative smoking status and number of cigarettes per day was analyzed preliminarily. Smoking status was missing for between 1 to 11% of patients at each follow-up period and of individuals reporting smoking at each follow-up, and 45 to 52% were missing number of cigarettes per day. Furthermore, among smokers at baseline, those with missing post-operative smoking data were significantly more likely to have fusion and success at 24 months than those without missing post-operative smoking data. For the sake of parsimony and to minimize missing data and bias, pre-operative smoking status (smoker versus nonsmoker) was used.

Pre-operative work status was used as a measure of severity of disease before the surgery. Data available were dichotomous working or non-working status, and for those who reported working, full time versus part time and full duty versus light duty was also

reported. Full-time status and full duty status were not used due to significant missing data. Forty percent of participants who reported working pre-operatively were missing responses for both time and duty questions. Post-operative work status was not used in the analysis as it is an outcome potentially related to success of spinal fusion.

The dataset included the number of levels being treated in with the spinal fusion, but only one study in this dataset (Study 13) allowed multiple-level fusion. In this study, 28 or 2% of all the patients had spinal fusion at more than one level. Therefore, the number of levels being treated was not included in the analysis as a patient characteristic.

Surgical procedures also varied by dose, concentration and carrier of rhBMP-2. All interbody fusion procedures (ALIF and PLIF) utilized an absorbable collagen sponge (marketed as INFUSE ® Bone Graft) as the carrier with a concentration of 1.5 mg/mL and doses ranged between 3.9 to 11.7 mg. PLF procedures included the use of concentrations ranging between 1.5 and 2.1 mg/mL, doses between 12 and 63 mg in one of three types of carriers: absorbable collagen sponge, biphasic calcium phosphate and compression-resistant matrix (marketed as AMPLIFY^{TM} rhBMP-2 Matrix). It was the researchers' wish to include dose of rhBMP-2 in the analysis, but a complete separation between doses within fusion procedures existed, limiting the ability to separate the effects of dose from the effect of surgery type.

Balance of randomization in terms of patient characteristics was tested by performing chi-square analysis by treatment group. The distribution of patient characteristics was similar across treatment groups, with the exception of diabetes. 60% of diabetics were randomized to ICBG and 40% to rhBMP-2. However, the overall number of patients with diabetes was small (71 patients or 6% of the total sample were patients with diabetes).

Missing status for outcomes at the final follow-up period was analyzed for a relationship with patient characteristics. Smokers were more likely to have missing outcomes ($p<0.05$ for success, fusion, and self-reported outcomes) than non-smokers. For each outcome, 4 to 5% more smokers than non-smokers had missing data No other patient characteristics was associated with missing outcomes at 24 months. The distribution of patient characteristics with non-missing outcomes at the final follow-up was also very similar to that of the distribution of patient characteristics at the baseline.

Patient characteristics were assessed for potential interactions with other patient characteristics. Patients with diabetes were more likely to be over 60 (44% of diabetics versus 19 % of non-diabetics, $p<0.01$), severely obese and obese (24% and 33% of diabetics were severely obese and obese versus 10% and 21%, $p<0.01$), and were less likely to be working (18%, versus 40% of non-diabetics, $p<0.01$). Patients that were over 60 were more likely to be female (61%, versus 54% of people under 60, $p=0.04$), non-smokers (89%, versus 65% of people under 60, $p<0.01$), not working (82%, versus 56% of people under 60, $p<0.01$), and were less likely to have had a previous back surgery (23%, versus 33% of people under 60, $p<0.01$). Participants with a previous back surgery were less likely to working (33% of patients with a previous back surgery were working as opposed to 41% of patients without a previous back surgery, $p<0.01$), and more likely to be male (35% of patients with a previous back surgery were male as opposed to 28% were female, $p<0.01$). Males were more likely to be smokers (35% of males were smokers as opposed to 27% of females, $p<0.01$), and more likely to be working (43% of males were working as opposed to 35% of females, $p<0.01$). Individuals that were of a higher BMI were less likely to have had a previous back surgery (26% of patients that had a BMI 30 or more had a previous back surgery as

opposed to 34% of patients with a BMI less than 30 had a previous back surgery, p=0.03), less likely to be smokers (27% of patients that had a BMI over 30 or were smokers as opposed to 32% of patients with a BMI less than 30 smokers, p<0.01), and more likely to be female (58% of patients that had a BMI of 30 more were female as opposed to 53% of patients with a BMI less than 30 were female, p<0.01).

Prior to an analysis in the regression setting, the relationship between each patient characteristic and outcomes was assessed independently in preliminary analyses. Significant findings represent a relationship that is not corrected for other covariates. Significant relationships between patient characteristics and fusion and success are presented in Table 12. Across all time points for fusion and 2 of 3 time points for success, smokers were less likely to have positive outcomes than non-smokers. Having a severely obese BMI was found to be related to a lower proportion of patients with fusion. Diabetics had a lower percentage of fusion at 6 months than did non-diabetics. The group with a previous back surgery had a higher percentage of patients with fusion at 6 months and 24 months than patients without any previous back surgeries. There were a higher proportion of individuals who were working at baseline with success at 12 and 24 months than individuals not working at baseline. . Age and sex were not independently related to fusion or success at any time point.

Significant relationships between AEs and patient characteristics are shown in Table 13. Individuals that were over 60 were more likely to have a serious AE at 4 weeks than individuals under 60. Individuals that were severely obese were less likely to have any AE at 2 years than were any other BMI group; however, the group with the most SAEs at 4 weeks was those that were obese, but not severely obese. Individuals that were

diabetic were more likely to have any AE at 2 years than non-diabetics. Serious and related AEs were rare, but individuals with a previous surgery were more likely to have a serious and related AE at 4 weeks than those without a previous back surgery. Females were more likely to have a SAE at 4 weeks than males. Smokers were more likely to have any AE or a related AE at 2 years, and were more likely to have a serious and related AE at 4 weeks and 2 years. Finally, individuals that were not working at baseline were more likely to have a SAE than those who were working at baseline.

For continuous, self-reported patient outcomes, the change over time was examined by level of patient characteristic via t-test or ANOVA (significant differences shown in Table 14). Patients over 60 year reported greater improvement in SF-36 PCS score. Smokers reported less improvement in ODI and SF-36 PCS outcomes than non-smokers. Those not working at baseline reported less improvement than those working at baseline.

A preliminary analysis of interaction between patient characteristics and treatment for outcome was assessed for all outcomes. This was assessed by performing chi-square analyses for dichotomous outcomes and t-test for continuous outcomes comparing outcomes by patient characteristic separately for each treatment group (Table 15). The results were examined for any discrepancies between the treatment groups. For example, at the 24 month follow-up period, there was no relationship between smoking status and fusion for rhBMP-2, but there was a significant relationship ($p<0.01$) between smoking status and fusion for ICBG. The remaining preliminary analyses of interactions between patient characteristics and treatment for all other outcomes was analyzed but is not shown.

Final Model Results

Primary Effectiveness Outcomes: Fusion and Success

Across the three follow-up periods for fusion, there were three significant interaction effects with treatment (age, smoking status, and BMI, see Table 16), as well as two main effects of previous surgery and type of surgery. However, the significance of the interaction effects was not consistent across the time periods. At 6 months, there was a significant interaction between age and treatment with participants under the age of 60 having significantly higher odds of fusion with rhBMP-2 versus ICBG (OR: 3.17, 95% CI: 1.20 to 8.37), but there was no difference in persons over 60 years of age (OR: 1.12, 95% CI: 0.37 to 3.37). At 24 month, smokers at baseline had significantly higher odds of fusion when on rhBMP-2 versus ICBG (OR: 4.90, 95% CI: 2.42 to 9.91), but there was no difference in non-smokers (OR: 1.45, 95% CI: 0.85 to 2.47). At the 24 month follow-up, there was a significant interaction between treatment and BMI. In normal weight and overweight patients, rhBMP-2 was associated with higher odds of fusion versus ICBG (Normal weight OR: 3.14, 95% CI: 1.46 to 6.73; Overweight OR: 2.60, 95% CI 1.29 to 5.22). For obese patients, effects of rhBMP-2 were smaller and not statistically significant (OR: 1.92, 95% CI: 0.82 to 4.50), and for severely obese patients, there was a non-significant trend for decreased odds of fusion (OR: 0.36, 95% CI: 0.11 to 1.11).

There were a number of significant findings between patient characteristics and outcomes in which there was not a significant interaction with treatment. Across all time points at which fusion was analyzed, the PLF and PLIF surgical techniques were associated with decreased odds of fusion compared with ALIF (Table 17). At the 24 month follow-up period only, patients with a previous back surgery had increased odds

of fusion compared with patients without a previous back surgery (OR: 1.81, 95% CI 1.16 to 2.84) (Table 17).

For the outcome of overall success, there was an interaction between BMI and treatment, as well as significant main effects of smoking and baseline work status. At the 24 month follow-up period, normal weight patients experienced significantly higher odds of success with rhBMP-2 versus ICBG (OR: 2.17, 95% CI: 1.36 to 3.44) (Table 18). For all other BMI groups, there was no benefit associated with the use of rhBMP-2. Baseline smoking and work status were significantly related to success without a significant interaction with treatment. Across all time points for success, smokers had lower odds of success than non-smokers (OR at 24 months: 0.56, 95% CI 0.43 to 0.73) (Table 19). At the12 and 24 month follow-up, patients who were working at baseline had higher odds of success than those who were not (OR at 24 months: 1.41, 95% CI: 1.10 to 1.81) (Table 19).

Secondary Effectiveness Outcomes: Self-Reported Outcomes

Across all self-reported outcomes, there were no significant interaction effects involving patient characteristics and treatment. The final models for self-reported outcomes included an interaction between study period and treatment, type of surgery, age, BMI, previous surgery, sex, smoking and baseline work status (Tables 20 and 21a,b). The only patient characteristic that was not found to be significant in the self-reported outcomes was diabetes. However, the estimates of difference of outcomes by patient characteristics and treatment did not exceed the minimum clinically important difference of 12.8 for the ODI and 4.9 for the SF-36 measures.

For ODI and SF-36 PCS, there was a significant interaction between treatment and period with patients on rhBMP-2 experiencing more improvement in the ODI and SF-36 PCS than patients with ICBG at each time point. The type of surgery was also significantly associated with outcomes for the ODI and SF-36 PCS, with PLF having better ODI outcomes as compared to ALIF. However, for the SF-36 PCS, patients on PLF and PLIF experienced less improvement than those on ALIF, and PLIF patients had less improvement. Patients over 60 experienced better outcomes than patients under 60 on the ODI and SF-36 MCS

As compared to normal weight patients, obese patients experienced poorer outcomes on all three outcomes, while severely obese patients experienced poorer outcomes on both SF-36 measures. For the ODI and SF-36 PCS, male and patients with previous back surgery experienced worse outcomes. Smokers have significantly worse outcomes for the SF-36 MCS only. For all self-reported outcomes, those working at baseline experienced more improvement.

Harms Outcomes: Adverse Events

Effects of patient characteristics on the odds of AEs and severe AEs are presented in Table 22. There were no significant interactions between treatment and patient characteristics for overall AEs and severe AEs. For overall AEs at 4 weeks and 2 years, PLIF was associated with significantly higher odds of AEs than did patients with ALIF. Patients with PLF had significantly lower odds of AEs than did patients with ALIF. Patients that were 60 or more also had higher odds of an AE than patients less than 60 and patients that were working at baseline had lower odds of an AE than patients that were not working as of baseline.

Due to the small number of related and related, severe AEs at 4 weeks, models with these as outcomes did not converge, and results are not reported. Therefore for the outcomes of related AEs and related and severe AEs, only 2 year outcomes were modeled. Outcomes for related and related, severe AEs are presented in Tables 23 and 24. There was a significant interaction between the presence of a previous back surgery and treatment. For related and related, severe AEs at 2 years, patients without a previous back surgery had significantly reduced odds for an AE when on rhBMP-2 as opposed to ICBG, but for patients with a previous back surgery, there was no benefit to the use of rhBMP-2.

Type of surgery and smoking status were significantly related to related and related, severe AEs, without the presence of an interaction with treatment. For related and related, severe AEs at 2 years, patients with PLF and PLIF had lower odds of AEs as compared to patients with ALIF. Patients that were smokers had significantly higher odds of related and related, severe AEs than non-smokers.

Baseline work status was not significant in the model but was a negative confounder in that the relationship between previous surgery and the interaction between previous surgery and treatment became stronger with the inclusion of work as a covariate for related only (not related and severe AEs). There was an 18% increase for the coefficient for previous surgery and a 12% increase in the coefficient for the interaction between treatment and previous surgery. When work was included, there was a non-significant trend for baseline work status to be related to the occurrence of AEs with those working at baseline having lower odds of AEs than those that were not working at baseline (OR: 0.52, 95% CI 0.26 to 1.02).

Across all outcomes and all time points, there were no interactions between surgery type and treatment. Model building separately for ALIF and PLF resulted in different estimates and different variable being significant, but for coefficients that were significant ($p<0.05$), the difference was in magnitude, not in direction. In addition, there were no significant treatment by surgery type interactions in any of the final models. Therefore, we chose to only model the data all together and not separate by surgery type.

Results of Sensitivity Analysis

For fusion at 24 months, the significance of all variables was the same in the original and first sensitivity analysis. The model did not converge under the second sensitivity analysis, so a comparison was not able to be made at that time point. For the second sensitivity analysis, the treatment*BMI interaction was significant 6 and 12 months ($p=0.03$ for both), the treatment*age interaction was significant at 12 months ($p=0.03$), and the main effect of previous surgery also became significant at 6 months ($p=0.03$). Finally, the main effect of surgical type was non-significant under the first sensitivity analysis, but was significant in the main analysis and second sensitivity analysis.

The main effect of smoking at 12 months and the main effect of work at 24 months were non-significant in the second sensitivity analysis, but were significant in the main analysis and the first sensitivity analysis. Furthermore, the interaction between treatment and BMI was significant in the main analysis and first sensitivity analysis, but was not significant in the second sensitivity analysis.

Estimate of Trial-Specific Random Effects

The trial-specific effect α was calculated for each trial and these effects were compared informally for each outcome. For fusion outcomes, PLIF and PLF studies had marginally better outcomes than ALIF (PLIF/PLIF OR: 1.19 versus ALIF OR: 0.89). Study 13 had the highest trial effect (OR 1.94) and Study 9 had the lowest (OR 0.31). Study 1 had the highest prediction error. At 24 months, PLIF/PLF random effects were estimated to be zero with an OR of 1.00 indicating that there was no between-trial heterogeneity for PLIF and PLF trials at that time point.

For the outcome of success, there was no difference between average trial effects for ALIF and PLIF/PLF (PLIF/PLIF OR: 1.01 versus ALIF OR: 0.98). For 12 and 24 months, all random effects were estimated to be 0 indicating a lack of trial effect. At 6 months, study 5 had the highest trial-specific effect (OR 1.09) and Study 9 had the lowest (OR 0.93). Study 4 had the highest prediction error.

For SF-36 MCS all random effects were estimated to be 0. For the ODI, all random effects for ALIF were 0. For the ODI and SF-36 PCS, there was not difference between PLIF/PLF and ALIF (PLIF/PLIF OR: 1.01 versus ALIF OR: 0.98). Study 13 had the best outcomes (OR 1.10) and Study 2 had the worst (OR 0.79). Study 1 had the highest prediction error.

For AEs and SAEs PLIF and PLF studies had an average trial effect that indicated an increased risk for AEs and SAEs over ALIF (PLIF/PLIF OR: 1.37 versus ALIF OR: 0.96). Study 13 had the highest trial effect (OR 1.93) and Study 14 had the lowest (OR 0.70). Study 6 had the highest prediction error. For SAEs at 4 weeks, ALIF random effects were estimated to be zero.

For related and related SAEs, there was not difference in average trial effect between PLIF/PLF and ALIF (PLIF/PLIF OR: 1.03 versus ALIF OR: 1.00).Study 13 had the highest average trial effect (OR 1.24) and Study 14 had the lowest (OR 0.87). Study 8 had the highest prediction error. For all related SAEs at 2 years random effects were estimated to be 0. For ALIF related AEs at 2 years random effects were estimated to be 0.

DISCUSSION

Based on an analysis of IPD, our results suggest differential effects of rhBMP-2 versus ICBG based on the presence of certain patient characteristics. RhBMP-2 was associated with increased likelihood of fusion versus ICBG for smokers over non-smokers, patients of normal weight versus patients obese or severely obese (in the latter group, rhBMP-2 may be less effective than ICBG). RhBMP-2 was also associated with decreased device-related AEs versus ICBG in individuals with no previous back surgeries compared with individuals with previous back surgeries.

The findings regarding age were less clear. Although rhBMP-2 was associated with increased likelihood of fusion versus ICBG in patients under 60, the likelihood of fusion was only marginally decreased for those over 60 through 12 months of follow-up, and at 24 months, rhBMP-2 significantly increased the likelihood of fusion over ICBG. It is possible that any benefit toward fusion of rhBMP-2 is delayed in the elderly. In addition, patients over 60 reported better outcomes based on the ODI and SF-36 MCS (albeit none of the differences met the clinically meaningful threshold), despite the lower likelihood of fusion. It is possible that there are unmeasured confounders that are contributing to this unexpected finding.

Future studies on rhBMP-2 should include subgroup analyses, particularly on smoking status, age, and weight to further examine these findings. If these relationships are replicated in future analyses, this could inform the use of rhBMP-2 over ICBG in lumbar spinal fusion for smokers and potentially discourage use of rhBMP-2 in obese and severely obese patients.

Strengths

This analysis has a number of strengths due to the use of IPD. In a traditional meta-analysis, analyses are performed from summary data. The inclusion of IPD in a meta-analysis allows for the standardization of outcomes across studies, the ability to model longitudinal trends, the use of random treatment effects for trial, check for bias in treatment randomization, and the ability to do subgroup analyses.

Additionally, this analysis meets a number of elements of a good subgroup analysis. Yusuf [36] outlines a number of recommendations to follow when conducting subgroup analyses in studies for which the subgroup analyses were not originally planned. The current project meets many of these recommendations in that subgroup membership was collected at baseline, had very little missing data, was not affected by treatment, included data from multiple studies and all patient characteristics analyzed had scientific evidence supporting a difference in outcomes.

Limitations and Future Directions

Lack of Blinding

Potentially the largest limitation of the evidence is the fact that for all trials, the patients and medical professionals conducting the study were not blinded to the treatment

received by each patient. This is a challenge to any analysis of this data. Since patients receiving ICBG experienced an extra surgery to obtain autogenous bone material, the participants were not blinded to the procedure. The only portion of the study that was blinded was the assessment of radiographic outcomes in which radiologists not aware of the patients' treatment status. Future studies of rhBMP-2 should blind all study personnel involved in post-operative management of patients and assessment of patient outcomes.

Differences between Trials

There were a number of limitations surrounding the patient characteristics. While there was minimal missing data on patient characteristics there were differences in inclusion and exclusion criteria for some studies, and the distribution of patient characteristics varied between the studies. The use of multiple variable regression and random treatment effects by trial was used to control for this, and differences in studies are outlined below.

There were a number of significant differences in study exclusion and inclusion criteria among the ten studies included in this meta-analysis. The most important difference was that only one study (Study 14) allowed patients with weight greater than 40% over ideal for age and height, which limited the number severely obese patients in this analysis. As the power of a subgroup analysis using IPD is limited by sample size, this creates a challenge for subgroups with small membership. Across the four studies that used the PLF surgical approach, those studies together included a wider distribution of patient characteristics than did the ALIF or PLIF studies (Table 5). An analysis of patient characteristics by ALIF versus PLF indicated that the distribution of patient

characteristics was significantly different for all patient characteristics except sex. Within both surgical approaches patient characteristics were similar with what would be expected for spinal fusion patients. Another difference in exclusion criteria was that two PLF studies (Studies 8 and 14) tested for and excluded patients with a high risk of osteoporosis, while the remaining studies did not test for risk of osteoporosis, which may biased to the selection of younger, healthier participants.

Three studies required females of child-bearing potential to use contraception for significantly longer than other studies (16 weeks versus 1 year for Studies 8, 13, 14). The effect of this is likely small; however 67% of women in this meta-analysis were within child bearing age overall (less than 55 years), and this may have biased the sample toward being older if younger women were less likely to join the study because of this requirement. This effect would be very small as the population for which spinal fusion is indicated tends to be older. The clinical significance of this is that for the three studies that required women to be on birth control for at least a year, there were a significant number of these women that were smokers (32% of women under 55 were smokers), this could have put them at a much greater risk of blood clots.

Finally, this meta-analysis included five pilot studies (Study 1, 4, 8, 9, 12) that utilized more stringent enrollment criteria (e.g., enrollment of only healthy individuals in Study 1), which may have limit the ability to evaluate the effect of treatment on a number of patient characteristics due to a reduced sample size.

The use of a random treatment effect by trial allowed for the comparison of trial-specific effects. In comparing average trial-related treatment effects in ALIF to PLIF/PLF, PLIF/PLF had a better average trial-specific effect for fusion and a worse average trial-

specific effect for AEs and SAEs. For a number of outcomes, the trial-related effects for PLIF/PLF were estimated to be zero which indicates that the trial-specific effect was not different from the overall treatment effect. Study 13 had the best trial-specific effect for fusion, ODI, SF-36 PCS, but it also had the worst trial-specific effect for AEs, SAEs, and related AEs. Study 14 had the best trial-specific effect for AEs and SAEs. Study 9 had the worst trial-specific effect for both fusion and success.

Limitation of BMI

A potential limitation with the interpretation of findings around BMI is that there were 8 individuals who were underweight that were included in the analysis with individuals that were normal weight. Due to the small sample size in the underweight group, it was not possible to analyze this group separately.

Statistical Limitations

When random effects were added to the logistic models, there were consistent issues with estimation. A high number of initial iterations were used as the models consistently failed to find valid starting points for values of the fixed effects. The default for initial iterations in PROC GLIMMIX is 4 and this model and for the purposes of this analysis, a very high value of 100 was used to allow for many iterations.

While the risk ratio was of more interest in this study and is easier to interpret, a log link did not converge for the models of fusion, success and AEs. Therefore, a logit link was used for all dichotomous outcomes, and therefore, odds ratios are reported.

In a number of models, the random effects variance parameters for ALIF, PLIF, and/or PLF were estimated to be very close to zero (but still non-zero), which resulted in the

variance-covariance matrix not being positive definite. It was decided to combine the posterior procedures since the number of individuals with PLIF was small, as well as was in only one study. One way to remove this issue would be to eliminate the random effects grouping of surgery type. However, this covariate was determined to be of significant clinical importance and controlling for it was required. Since this can be caused by collinearity, this was examined by running a linear regression model with all main effects included in this analysis. The maximum VIF was found to be 1.23, far below the cutoff of 10.0. The final determination was that the random effect of surgery type remained in the model even though it resulted in the variance covariance matrix having estimates very close to zero.

Missing Data

Another limitation was missing outcome data (Table 6); however, this limitation was modest with maximum 11.5% missing at the 24 month follow-up (less than the commonly used 20% threshold). Also, smoking status was related to missing data with smokers having more missing outcome data than non-smokers. This could have biased the analysis, as it is likely that the smokers had poorer outcomes and therefore missing data was more likely to be undesirable outcomes. Methods of imputation of outcomes or sensitivity analyses could address these concerns.

There is research on more detailed data regarding smoking and smoking cessation before spinal fusion [12, 37]. Specifically, the amount of time before surgery an individual stops smoking as well as the amount they smoke each day has been related to outcomes. However, in this dataset there was significant missing data around number of cigarettes per day and time before surgery smoking was stopped. This data could be

included in the analysis by including missing as a category, so that missing data is not removed from the dataset or these values could be imputed using one of a number of imputation methods.

While any bias from the 4 patients with missing patient characteristics (3 missing baseline BMI and 1 missing baseline work status) would be decidedly minimal, it would be possible to perform a sensitivity analysis to assess for bias. This is performed by repeating any significant findings by conducting the model by placing the participants with missing subgroup membership into each subgroup and re-performing the analysis. If the finding remains for all replications of the analysis, then it is relatively certain that the subgroup relationship exists. However, if the finding disappears for at least one of these re-modeling, then the significant finding may be related to bias of the missing patient data. Another way to address the missing patient characteristics would be to use one of numerous imputation methods to obtain estimates for the missing patient characteristics.

Another type of sensitivity analysis could be performed on missing outcome data which could help determine if any biases exist around missing outcome. For patients who are missing success or fusion outcomes, the missing outcome would be imputed as a failure, and all models would be rerun with the updated outcomes. Any significant findings from the primary analysis could be compared to the findings of the sensitivity analysis, and if large difference occurs, this might be evidence for bias in missing outcomes.

Considerations Regarding Subgroup Analyses

To assess whether the effectiveness of rhBMP-2 versus ICBG differ by patient characteristics, linear models for self-reported outcomes and generalized linear models for fusion and success were used with interaction terms between treatment and patient characteristics. These analyses amount to a number of subgroup analyses, and these analyses were limited to pre-specified patient characteristics to reduce the change of an inflated Type I error [36,38].

There exist special considerations when subgroup analyses are being performed in studies in which they were not originally planned. One consideration occurs when a main effect of treatment is significant, but upon subgroup analysis (e.g., by gender) it is only significant within one level (e.g., effects of treatment are only significant in males). When this occurs, Yusuf (1991) cautions that this finding does not mean that the treatment effect is only significant in men, but that it is potentially stronger in men. He cautions that the lack of significance in a subgroup analysis may be due to a lack of power and not a result of a qualitative interaction. The overall main effect should be considered more reliable than post-hoc subgroup findings.

Since these subgroup analyses are being conducted post-hoc, any findings that contradict overall findings should be considered with caution as the overall findings should be considered primary and more accurate than subgroup findings. It is possible that a differential subgroup effect exists, which occurs when the a subgroup effect differs from the overall treatment effect, but this can only be confirmed with further studies that are adequately and intentionally powered for such a subgroup analysis. As Rothwell

says (2005), “The best test of the validity of subgroup analyses is not significance but replication.”

SUMMARY AND CONCLUSIONS

The findings from this meta-analysis of independent patient data from 10 industry-sponsored, randomized controlled trials suggest that a different treatment effect exists for some patient characteristics. For the primary outcome of effectiveness, rhBMP-2 increases fusion for smokers, and increases fusion and success for normal weight patients, but not obese or severely obese patients in which rhBMP-2 may be less effective than ICBG. For the secondary outcome of harms, rhBMP-2 was related to a decrease in device-related AEs in individuals with no previous back surgeries but not within individuals with a previous back surgery. Future research on rhBMP-2 should include planned subgroup analysis that specifically consider patients that are elderly, smokers and obese or severely obese. These findings, if replicated, could provide patient-specific recommendations regarding the use of rhBMP-2 in lumbar spinal fusion.

REFERENCES

[1] Ross JS, Gross CP, Krumholz HM. Promoting transparency in pharmaceutical industry-sponsored research. Am J Public Health 2012;102:72–80.

[2] Ross JS, Lehman R, Gross CP. The importance of clinical trial data sharing: toward more open science. Circ Cardiovasc Qual Outcomes 2012;5:238–40.

[3] Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: Is there a difference? Lancet 1993;341:418–22.

[4] Simmonds MC, Brown JVE, Heirs MK, Higgins JPT, Mannion RJ, Rodgers MA, et al. Safety and effectiveness of recombinant human bone morphogenetic protein-2 for spinal fusion: a meta-analysis of individual-participant data. Ann Intern Med 2013;158:877–89.

[5] Fu R, Selph S, McDonagh M, Peterson K, Tiwari A, Chou R, et al. Effectiveness and harms of recombinant human bone morphogenetic protein-2 in spine fusion: a systematic review and meta-analysis. Ann Intern Med 2013;158:890–902.

[6] Health C for D and R. Recently-Approved Devices - InFUSE™ Bone Graft/LT-CAGE™ Lumbar Tapered Fusion Device - P000058 n.d.

[7] Carragee EJ, Mitsunaga KA, Hurwitz EL, Scuderi GJ. Retrograde ejaculation after anterior lumbar interbody fusion using rhBMP-2: a cohort controlled study. Spine J 2011;11:511–6.

[8] Tiusanen H, Seitsalo S, Osterman K, Soini J. Retrograde ejaculation after anterior interbody lumbar fusion. Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc 1995;4:339–42.

[9] Burkus JK, Gornet MF, Dickman CA, Zdeblick TA. Anterior lumbar interbody fusion using rhBMP-2 with tapered interbody cages. J Spinal Disord Tech 2002;15:337–49.

[10] Lee MJ, Hacquebord J, Varshney A, Cizik AM, Bransford RJ, Bellabarba C, et al. Risk Factors for Medical Complication after Lumbar Spine Surgery: a multivariate analysis of 767 patients. Spine 2011;36:1801–6.

[11] Glassman SD, Carreon LY, Djurasovic M, Campbell MJ, Puno RM, Johnson JR, et al. RhBMP-2 versus iliac crest bone graft for lumbar spine fusion: a randomized, controlled trial in patients over sixty years of age. Spine 2008;33:2843–9.

[12] Andersen T, Christensen FB, Laursen M, Høy K, Hansen ES, Bünger C. Smoking as a predictor of negative outcome in lumbar spinal fusion. Spine 2001;26:2623–8.

[13] Glassman SD, Dimar JR 3rd, Burkus K, Hardacker JW, Pryor PW, Boden SD, et al. The efficacy of rhBMP-2 for posterolateral lumbar fusion in smokers. Spine 2007;32:1693–8.

[14] Djurasovic M, Bratcher KR, Glassman SD, Dimar JR, Carreon LY. The effect of obesity on clinical outcomes after lumbar fusion. Spine 2008;33:1789–92.

[15] Vaidya R, Carp J, Bartol S, Ouellette N, Lee S, Sethi A. Lumbar spine fusion in obese and morbidly obese patients. Spine 2009;34:495–500.

[16] Rosen DS, Ferguson SD, Ogden AT, Huo D, Fessler RG. Obesity and self-reported outcome after minimally invasive lumbar spinal fusion surgery. Neurosurgery 2008;63:956–960; discussion 960.

[17] Browne JA, Cook C, Pietrobon R, Bethel MA, Richardson WJ. Diabetes and early postoperative outcomes following lumbar fusion. Spine 2007;32:2214–9.

[18] Glassman SD, Alegre G, Carreon L, Dimar JR, Johnson JR. Perioperative complications of lumbar instrumentation and fusion in patients with diabetes mellitus. Spine J Off J North Am Spine Soc 2003;3:496–501.

[19] Herno A, Airaksinen O, Saari T, Sihvonen T. Surgical results of lumbar spinal stenosis. A comparison of patients with or without previous back surgery. Spine 1995;20:964–9.

[20] Anderson PA, Schwaegler PE, Cizek D, Leverson G. Work status as a predictor of surgical outcome of discogenic low back pain. Spine 2006;31:2510–5.

[21] Thornes E, Ikonomou N, Grotle M. Prognosis of Surgical Treatment for Degenerative Lumbar Spinal Stenosis: A Prospective Cohort Study of Clinical Outcomes and Health-Related Quality of Life Across Gender and Age Groups. Open Orthop J 2011;5:372–8.

[22] Shabat S, Folman Y, Arinzon Z, Adunsky A, Catz A, Gepstein R. Gender differences as an influence on patients' satisfaction rates in spinal surgery of elderly patients. Eur Spine J 2005;14:1027–32.

[23] Fairbank JC, Couper J, Davies JB. The Oswestry Low Back Pain Questionnaire. Physiotherapy 1980; 66: 271-273.

[24] Ware J, Keller S, Kosinski M. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston: The Health Institute; 1994.

[25] Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. Spine J Off J North Am Spine Soc 2008;8:968–74.

[26] U.S. Preventive Services Task Force. Screening for obesity in adults: Recommendations and rationale. Ann Intern Med 2003;139:930-932.

[27] DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: An update. Contemp Clin Trials 2007;28:105–14.

[28] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.

[29] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000;19:3417-32.

[30] Higgins JP, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Stat Med 2001;20:2219-41.

[31]Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random effects model: an example. J R Stat Soc Series C 2001;50:325-35.

[32] SAS Institute Inc. The GLIMMIX Procedure. SASSTAT® 131 User's Guide, Cary, NC: SAS Institute Inc.; 2013.

[33] SAS Institute Inc. The MIXED Procedure. SASSTAT® 922 Users Guide, Cary, NC: SAS Institute Inc.; 2008.

[34] Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika;76:297-307.

[35] SATTERTHWAITE FE. An approximate distribution of estimates of variance components. Biometrics 1946;2:110–4.

[36] Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA J Am Med Assoc 1991;266:93–8.

[37] Glassman SD, Anagnost SC, Parker A, Burke D, Johnson JR, Dimar JR. The effect of cigarette smoking and smoking cessation on spinal fusion, Spine 2000;25:2608-15.

[38] Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.

[39] Boden SD, Zdeblick TA, Sandhu HS, Heim SE. The use of rhBMP-2 in interbody fusion cages. Definitive evidence of osteoinduction in humans: a preliminary report. Spine 2000;25:376–81.

[40] Burkus JK, Transfeldt EE, Kitchel SH, Watkins RG, Balderston RA. Clinical and radiographic outcomes of anterior lumbar interbody fusion using recombinant human bone morphogenetic protein-2. Spine 2002;27:2396–408.

[40] Haid RW Jr, Branch CL Jr, Alexander JT, Burkus JK. Posterior lumbar interbody fusion using recombinant human bone morphogenetic protein type 2 with cylindrical interbody cages. Spine J Off J North Am Spine Soc 2004;4:527–538; discussion 538–539.

[41] Dawson E, Bae HW, Burkus JK, Stambough JL, Glassman SD. Recombinant human bone morphogenetic protein-2 on an absorbable collagen sponge with an osteoconductive bulking agent in posterolateral arthrodesis with instrumentation. A prospective randomized trial. J Bone Joint Surg Am 2009;91:1604–13.

[42] Boden SD, Kang J, Sandhu H, Heller JG. Use of recombinant human bone morphogenetic protein-2 to achieve posterolateral lumbar spine fusion in humans: a prospective, randomized clinical pilot trial: 2002 Volvo Award in clinical studies. Spine 2002;27:2662–73.

[43] Dimar JR 2nd, Glassman SD, Burkus JK, Pryor PW, Hardacker JW, Carreon LY. Clinical and radiographic analysis of an optimized rhBMP-2 formulation as an autograft replacement in posterolateral lumbar spine arthrodesis. J Bone Joint Surg Am 2009;91:1377–86

TABLES

Table 1. Study details

Study Number[†]	Fusion Type	Study Publication	Number and Percent of Participants					
			ICBG		rhBMP-2		Total	
			n	%	n	%	n	%
1	ALIF	Boden et al., 2000[39]	3	30.0%	7	70.0%	10	100%
2	ALIF	Burkus et al., 2002[9]	136	48.6%	144	51.4%	280	100%
4	ALIF	Burkus et al., 2002[40]	22	47.8%	24	52.2%	46	100%
5	ALIF	Burkus et al., 2005[39]	30	35.3%	55	64.7%	85	100%
6	PLIF	Haid et al., 2004[40]	33	49.3%	34	50.7%	67	100%
8	PLF	Dawson et al., 2009[41]	21	45.7%	25	54.3%	46	100%
9	ALIF	Unpublished	20	44.4%	25	55.6%	45	100%
12	PLF	Boden et al., 2002[42]	5	31.3%	11	68.8%	16	100%
13	PLF	Unpublished	99	50.3%	98	49.7%	197	100%
14	PLF	Dimar et al., 2009[43]	224	48.4%	239	51.6%	463	100%
Total			593	47.3%	662	52.7%	1255	100%

[†] Study number as used in Fu et al. 2014.

Table 2. Definition of outcomes (Summarized from Fu et al. [5])

Outcome Variable	Surgical Approach (Study Numbers)	Definition and Criteria
Overall Success	All Approaches (All Studies)	All of the following criteria must be satisfied: ● Fusion (see below) ● ≥15-point improvement in ODI score for low back pain at each visit post-operatively as compared to preoperative score. ● Maintenance or improvement in neurological status as measured by having the same or better score in 4 tests of neurological status (motor function, sensory function, deep tendon reflexes, and sciatic tension signs) as compared with preoperative score. ● No serious adverse events classified as device- or device, surgical related ● No additional procedure classified as "failure"
Fusion	ALIF (1)	● Bone growing continuously through the cage and connecting with vertebral bodies above and below 1 cage as determined by CT scan or radiographs if CT scans were not available.
Fusion	ALIF/PLIF (2,4-6,9)	All of the following criteria must be satisfied: ● Evidence of continuous trabecular bone growth connecting the vertebral bodies and/or through 1 or both implants ● Absence of radiolucency covering >50% of implant ● Translation of ≤3 mm and angulation of <5 degrees All were determined by CT scan or radiographs if CT scans were not available
Fusion	PLF (8,12-14)	All of the following criteria must be satisfied: ● Evidence of continuous trabecular bone growth connecting the transverse processes ● Absence of radiolucent lines through the fusion mass ● Translation of ≤3 mm and angulation of <5 degrees All were determined by CT scan or radiographs if CT scans were not available

Table 3. Table of AIC values for variance-covariance matrices in full main effects models for longitudinal repeated measures

	Number of Covariance Parameters and AIC Values for Variance – Covariance Matrices							
	Unstructured	**Auto-regressive**	**Heterogeneous Autoregressive**	**Compound Symmetry**	**Heterogeneous Compound Symmetry**	**Toeplitz**	**Heterogeneous Toeplitz**	**Variance Components**
	Required $\theta^{\dagger}$	**Required $\theta^{\dagger}$**	**Required $\theta^{\dagger}$**	**Required $\theta^{\dagger}$**	**Required $\theta^{\dagger}$**	**Required $\theta^{\dagger}$**	**Required $\theta^{\dagger}$**	**Required $\theta^{\dagger}$**
	21	2	7	2	7	6	11	1
Outcome	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**
ODI	56,608.8	58,326.6	58,121.6	58,043.3	57,875.8	57,679.8	57,556.2	60,802.9
SF-36 PCS	47,534.9	49,152.6	48,751.8	48,981.5	48,475.1	48,591.2	48,213.6	51,119.8
SF-36 MCS	50,982.3	51,796.6	51,783.6	51,278.9	51,248.3	51,158.0	51,133.6	53,912.6

† Required number of parameters to be estimated for each variance-covariance matrix

Table 4. Table of denominator degrees of freedom and p-values for main effects in full main effects model for longitudinal repeated measures

	Denominator Degrees of Freedom (DDFM) and p-values for DDFM Estimation Methods									
	Residual		Containment		Between-Within		Kenward-Roger		Satterthwaite	
Fixed Effects	**DDFM**	**p-value†**	**DDFM**	**p-value†**	**DDFM**	**p-value†**	**DDFM**	**p-value†**	**DDFM**	**p-value†**
Treatment	7176	0.03	9	0.05	1241	0.03	58.2	0.04	58.2	0.03
Study Period	7176	<0.01	7167	<0.01	5935	<0.01	1208	<0.01	1214	<0.01
Surgery Type	7176	<0.01	7167	<0.01	1241	<0.01	67.7	0.01	66.4	<0.01
Age	7176	<0.01	7167	<0.01	1241	<0.01	1241	<0.01	1241	<0.01
BMI	7176	0.15	7167	0.15	1241	0.16	1286	0.16	1273	0.16
Previous Back Surgery	7176	<0.01	7167	<0.01	1241	<0.01	1237	<0.01	1237	<0.01
Sex	7176	<0.01	7167	<0.01	1241	<0.01	1239	<0.01	1239	<0.01
Smoking Status	7176	0.07	7167	0.07	1241	0.07	1237	0.07	1237	0.07
Baseline Work Status	7176	<0.01	7167	<0.01	1241	<0.01	1236	<0.01	1236	<0.01
Treatment by Period Interaction	7176	0.03	7167	0.03	5935	0.03	1201	0.03	1207	0.03

† p-values are from a test of significance of fixed effects based on the F-distribution

Table 5. Distribution of patient characteristics by type of surgery

		Number and Percent of Participants by Surgery Type							
	p-value[†]	ALIF		PLIF		PLF		Total	
Age	<0.01	**n**	%	**n**	%	**n**	%	**n**	%
Under 60		446	95.7%	61	91.0%	493	68.3%	1000	79.7%
60 and Over		20	4.3%	6	9.0%	229	31.7%	255	20.3%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%
BMI	<0.01	**n**	%	**n**	%	**n**	%	**n**	%
Underweight or Normal		159	34.1%	20	29.9%	172	23.8%	351	28.0%
Overweight		185	39.7%	28	41.8%	290	40.2%	503	40.1%
Obese		93	20.0%	15	22.4%	159	22.0%	267	21.3%
Severely Obese		27	5.8%	4	6.0%	101	14.0%	132	10.5%
Missing		2	0.4%	0	0.0%	0	0.0%	2	0.2%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%
Diabetes	<0.01	**n**	%	**n**	%	**n**	%	**n**	%
Non Diabetic		454	97.4%	65	97.0%	665	92.1%	1184	94.3%
Diabetic		12	2.6%	2	3.0%	57	7.9%	71	5.7%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%
Previous Back Surgeries	<0.01	**n**	%	**n**	%	**n**	%	**n**	%
None		289	62.0%	42	62.7%	533	73.8%	864	68.8%
1 or more		177	38.0%	25	37.3%	189	26.2%	391	31.2%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%
Sex	0.10	**n**	%	**n**	%	**n**	%	**n**	%
Female		243	52.1%	35	52.2%	412	57.1%	690	55.0%
Male		223	47.9%	32	47.8%	310	42.9%	565	45.0%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%
Smoking Status	<0.01	**n**	%	**n**	%	**n**	%	**n**	%
Non-smoking		308	66.1%	34	50.7%	533	73.8%	875	69.7%
Smoking		158	33.9%	33	49.3%	189	26.2%	380	30.3%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%
Baseline Work Status	<0.01	**n**	%	**n**	%	**n**	%	**n**	%
Not working		244	52.4%	43	64.2%	481	66.6%	768	61.2%
Working		221	47.4%	24	35.8%	241	33.4%	486	38.7%
Missing		1	0.2%		0.0%	0	0.0%	0	0.0%
Total		466	100.0%	67	100.0%	722	100.0%	1255	100.0%

[†] p-value is for a chi-square test of relationship between type of surgery and patient characteristics which tested if the distribution of a patient characteristic was the same across surgery type.

Table 6. Frequency and percent of missing outcome variables by study period

	Frequency and Percent of Missing Data by Study Period											
	Operative		6 Weeks		3 month		6 Month		12 Month		24 Month	
	n	%	n	%	n	%	n	%	n	%	n	%
Fusion	--	--	--	--	--	--	51	4.1%	69	5.5%	140	11.2%
Success	--	--	--	--	--	--	47	3.8%	70	5.6%	126	10.0%
ODI	1	0.1%	33	2.6%	22	1.8%	48	3.8%	76	6.1%	138	11.0%
SF-36 PCS	7	0.6%	126	10.1%	112	9.0%	57	4.5%	79	6.3%	144	11.5%
SF-36 MCS	7	0.6%	126	10.1%	112	9.0%	57	4.5%	79	6.3%	144	11.5%

Table 7. Frequency and percent of patients with desired outcomes by follow-up period

	Frequency and Percent of Participants by Study Period					
	6 Month		12 Month		24 Month	
	n	%	n	%	n	%
With Fusion	905	75.2%	970	81.8%	977	87.6%
With Overall Success	556	46.0%	614	51.8%	598	53.0%

Table 8. Table of change in fusion and success outcomes since last follow-up

	Frequency and Percent of Participants with Change between Follow-up Period					
	Change between baseline and 6 months		Change between 6 and 12 Months		Change between 12 and 24 Month	
Fusion	**n**	**%**	**n**	**%**	**n**	**%**
Lost Fusion	0	0.0%	73	6.3%	68	6.2%
No Change	299	24.8%	937	80.4%	902	82.2%
Achieved Fusion	905	75.2%	155	13.3%	127	11.6%
Overall Success	**n**	**%**	**n**	**%**	**n**	**%**
Lost Success	0	0.0%	121	10.3%	135	12.1%
No Change	652	54.0%	860	73.4%	838	75.3%
Achieved Success	556	46.0%	191	16.3%	140	12.6%

Table 9. Cumulative frequency and percent of patients with adverse events by follow-up period and type of adverse event

	Frequency and Percent of Participants with Adverse Events by Follow-up Period			
	4 weeks		2 years	
Type of Adverse Event	**n**	**%**	**n**	**%**
Overall Adverse Event	385	30.7%	695	55.4%
Severe Adverse Event	79	6.3%	153	12.2%
Device-Related Adverse Event	18	1.4%	64	5.1%
Severe and Device-Related Adverse Event	3	0.2%	46	3.7%

Table 10. Summary statistics for continuous, self-reported outcomes over time

	Mean and Standard Deviation of Self-Reported Outcomes over Time											
	Operative		6 Weeks		3 month		6 Month		12 Month		24 Month	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ODI	52.3	12.6	40.4	17.5	31.5	17.4	27.3	18.1	25.9	19	25.3	20.2
SF-36 PCS	27.8	6.4	31.6	7.6	35.9	9.4	38.8	10.8	39.9	11.5	40.4	11.9
SF-36 MCS	43.7	12.4	47.6	11.6	49.6	11.9	49.6	11.7	49.3	11.9	49.6	11.6

Table 11. Overall sample participant characteristics

	Frequency and Percent of Participants	
Age of Participant	**n**	**%**
Under 60	1000	79.7%
60 and Over	255	20.3%
Baseline Body Mass Index	**n**	**%**
Underweight (< 18.5)	8	0.6%
Normal (18.5 to 24.9)	343	27.3%
Overweight (25.0 to 29.9)	503	40.1%
Obese (30.0 to 34.9)	266	21.2%
Severely Obese (≥ 35)	132	10.5%
Missing	3	0.2%
Diabetes	**n**	**%**
Non Diabetic	1184	94.3%
Diabetic	71	5.7%
Previous back surgeries	**n**	**%**
None	864	68.8%
1 or more	391	31.2%
Sex of Participant	**n**	**%**
Female	690	50. %
Male	565	45.0%
Baseline Smoking Status	**n**	**%**
Non-smoking	875	69.7%
Smoking	380	30.3%
Baseline Work Status	**n**	**%**
Not working	768	61.2%
Working	486	38.7%
Missing	1	0.1%

Table 12. Percent of patients with fusion or success within each patient characteristic for patient characters that are significantly related (p<0.05) to adverse events

		Percent of Patients with Desirable Outcome within Each Patient Characteristic					
		Fusion			Success		
		6 months	12 months	24 months	6 months	12 months	24 months
Age	**Under 60**	NR	NR	NR	NR	NR	NR
	Over 60						
BMI	**Underweight**	NR	NR	86%	NR	NR	NR
	Overweight			90%			
	Obese			88%			
	Severely Obese			84%			
Diabetes	**Non-Diabetic**	76%	NR	NR	NR	NR	NR
	Diabetic	64%					
Previous Back Surgery	**None**	74%	NR	86%	NR	NR	NR
	1 or More	79%		91%			
Sex	**Female**	NR	NR	NR	NR	NR	NR
	Male						
Smoking	**Non-Smoker**	48%	83%	90%	NR	54%	57%
	Smoker	40%	78%	82%		47%	43%
Work Status	**Not-Working**	NR	NR	NR	NR	48%	49%
	Working					58%	59%

NR indicates there was not a significant relationship between the patient characteristic and occurrence of fusion or success. A relationship between patient characteristic and occurrence of fusion of success was determined to be significant if the p-value for a chi-square test for association or a Fisher's exact test where appropriate was $p < 0.05$.

Table 13. Percent of patients with an adverse event within each patient characteristic group for patient characters that are significantly related (p<0.05) to adverse events.

		Percent of Participants with Adverse Events within Each Patient Characteristic							
		Adverse Events		Severe Adverse Events		Related Adverse Events		Severe and Related Adverse Events	
		4 weeks	2 years	4 weeks	2 years	4 weeks	2 years	4 weeks	2 years
Age	**Under 60**	NR	NR	5%	NR	NR	NR	NR	NR
	Over 60			11%					
BMI	**Underweight**	NR	58%	7%	NR	NR	NR	NR	NR
	Overweight		56%	5%					
	Obese		55%	10%					
	Severely Obese		43%	5%					
Diabetes	**Non-Diabetic**	NR	56%	NR	NR	NR	NR	NR	NR
	Diabetic		44%						
Previous Back Surgery	**None**	NR	NR	NR	NR	NR	NR	0%	NR
	One or More							1%	
Sex	**Female**	NR	NR	8%	NR	NR	NR	NR	NR
	Male			5%					
Smoking	**Non-Smoker**	NR	53%	NR	NR	NR	4%	0%	2%
	Smoker		60%				8%	1%	7%
Work Status	**Not-Working**	NR	NR	8%	NR	NR	NR	NR	NR
	Working			4%					

NR indicates there was not a significant relationship between the patient characteristic and occurrence of adverse events. A relationship between patient characteristic and occurrence of adverse events was determined to be significant if the p-value for a chi-square test for association or a Fisher's exact test where appropriate was <0.05.

Table 14. Average change in self-reported outcomes between baseline and the final follow-up at 24 months within each patient characteristic group for patient characters that are significantly related (p<0.05) to change in self-reported outcomes.

		Average Change in Outcome between Baseline and 24 months within Each Patient Characteristic		
		ODI (decrease in score is improvement)	**PCS (increase in score is improvement)**	**MCS (increase in score is improvement)**
Age	**Under 60**	NR	12	NR
	Over 60		14	
BMI	**Underweight**	NR	NR	NR
	Overweight			
	Obese			
	Severely Obese			
Diabetes	**Non-Diabetic**	NR	NR	NR
	Diabetic			
Previous Back Surgery	**None**	NR	NR	NR
	1 or More			
Sex	**Female**	NR	NR	NR
	Male			
Smoking	**Non-Smoker**	-28	13	NR
	Smoker	-24	11	
Work Status	**Not-Working**	-25	11	NR
	Working	-29	15	

NR indicates there was not a significant relationship between the patient characteristic and change in self-reported outcomes. A relationship between patient characteristic and occurrence of self-reported outcomes was determined to be significant if the p-value for a t-test or an ANOVA where appropriate was <0.05.

Table 15. Percent of patients with fusion and success within each patient characteristic group at the 24 months follow-up.

		Percent of Patients with Fusion and Success by Treatment							
		rhBMP-2		ICBG		rhBMP-2		ICBG	
		% with fusion	p-value†	% with fusion	p-value†	% with success	p-value†	% with success	p-value†
Age	**Under 60**	90%	NS	83%	NS	56%	NS	48%	NS
	60 and Over	94%		85%		55%		56%	
BMI	**Underweight or Normal**	92%	<0.01	78%	0.02	64%	<0.01	46%	NS
	Overweight	94%		85%		56%		53%	
	Obese	91%		85%		52%		52%	
	Severely Obese	79%		91%		40%		49%	
	Missing BMI	NA		0%		NA		0%	
Diabetes	**Non Diabetic**	91%	NS	83%	NS	56%	NS	50%	NS
	Diabetic	87%		89%		50%		50%	
Previous Back Surgery	**No**	89%	<0.01	82%	NS	56%	NS	50%	NS
	Yes	96%		86%		54%		49%	
Sex	**Female**	91%	NS	83%	NS	57%	NS	49%	NS
	Male	91%		85%		54%		51%	
Smoking	**Non-smoking**	91%	NS	88%	<0.01	58%	0.04	56%	<0.01
	Smoking	91%		71%		49%		36%	
Work Status	**Not working**	91%	NS	81%	NS	51%	<0.01	47%	NS
	Working	92%		87%		63%		54%	

† Reported p-values are for chi-square test of association between outcomes and patient characteristics which were performed separately for each treatment group to assess for an interaction effect. NS indicates a non-significant (>0.05) p-value.

Table 16. Table of odds ratios for fusion for comparing rhBMP-2 versus ICBG by patient characteristic

		Odds Ratio and 95 % Confidence Intervals (CI) for Fusion for rhBMP-2 versus ICBG											
		6 Months				12 Months				24 Months			
Age		**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Under 60	**ICBG**	1.00	.	.	<0.01	1.00	.	.	0.75	1.00	.	.	0.24
	rhBMP-2	**3.17**	**1.20**	**8.37**		**2.07**	**1.15**	**3.72**		**1.81**	**1.10**	**3.00**	
60+	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	1.12	0.37	3.37		1.81	0.75	4.36		**3.47**	**1.31**	**9.22**	
Smoking		**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Non-Smoker	**ICBG**	1.00	.	.	0.13	1.00	.	.	0.71	1.00	.	.	0.01
	rhBMP-2	2.53	0.95	6.72		**2.10**	**1.13**	**3.91**		1.45	0.85	2.47	
Smoker	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	2.67	0.94	7.56		1.80	0.88	3.71		**4.90**	**2.42**	**9.91**	
BMI		**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Normal	**ICBG**	1.00	.	.	0.87	1.00	.	.	0.65	1.00	.	.	<0.01
	rhBMP-2	2.52	0.88	7.24		1.82	0.86	3.83		**3.14**	**1.46**	**6.73**	
Overweight	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	**3.14**	**1.13**	**8.72**		**2.45**	**1.23**	**4.87**		**2.60**	**1.29**	**5.22**	
Obese	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	2.77	0.92	8.34		1.92	0.83	4.43		1.92	0.82	4.50	
Severely Obese	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	1.06	0.32	3.56		1.39	0.49	3.94		0.36	0.11	1.11	

† Reported p-values are based on a test for fixed effects of an interaction between the patient characteristic and treatment within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

Table 17. Table of odds ratios for fusion for comparing levels within patient characteristics

	Odds Ratio and 95 % Confidence Intervals (CI) for Fusion within Levels of Patient Characteristics											
	6 Months				12 Months				24 Months			
Type	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
ALIF	1.00	.	.	<0.01	1.00	.	.	0.05	1.00	.	.	0.01
PLF	**0.45**	**0.29**	**0.67**		**0.58**	**0.38**	**0.90**		0.64	0.38	1.08	
PLIF	1.14	0.47	2.79		0.72	0.31	1.67		**0.31**	**0.14**	**0.66**	
Previous Back Surgery	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
None	1.00	.	.	0.08	1.00	.	.	0.54	1.00	.	.	<0.01
1 or More	1.32	0.97	1.80		0.90	0.65	1.25		**1.81**	**1.16**	**2.84**	

† Reported p-values are based on a test for fixed effects of the patient characteristic within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

Table 18. Table of odds ratios for success for comparing rhBMP-2 versus ICBG by patient characteristic

		Odds Ratios and 95 % Confidence Intervals (CI) for Success for rhBMP-2 versus ICBG									
		6 Months				12 Months				24 Months			
BMI		**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Normal	**ICBG**	1.00	.	.	0.26	1.00	.	.	0.33	1.00	.	.	0.03
	rhBMP-2	**1.64**	**1.02**	**2.63**		**1.57**	**1.00**	**2.44**		**2.17**	**1.36**	**3.44**	
Overweight	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	**1.95**	**1.29**	**2.95**		1.24	0.86	1.80		1.15	0.79	1.68	
Obese	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
	rhBMP-2	1.35	0.80	2.29		1.01	0.62	1.66		1.03	0.62	1.72	
Severely	**ICBG**	1.00	.	.		1.00	.	.		1.00	.	.	
Obese	**rhBMP-2**	0.90	0.43	1.88		0.77	0.38	1.55		0.67	0.32	1.39	

† Reported p-values are based on a test for fixed effects of an interaction between the patient characteristic and treatment within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

Table 19. Table of odds ratios for success for comparing levels within patient characteristics

	Odds Ratio and 95 % Confidence Intervals (CI) for Success within Levels of Patient Characteristics											
	6 Months				12 Months				24 Months			
Type	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
ALIF	1.00	.	.	0.28	1.00	.	.	0.14	1.00	.	.	0.19
PLF	0.78	0.58	1.06		0.84	0.65	1.07		0.96	0.74	1.25	
PLIF	0.78	0.43	1.40		0.62	0.36	1.07		0.60	0.35	1.05	
Smoking	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Non-Smoker	1.00	.	.	<0.01	1.00	.	.	0.02	1.00	.	.	<0.01
Smoker	**0.69**	**0.54**	**0.90**		**0.74**	**0.58**	**0.96**		**0.56**	**0.43**	**0.73**	
Baseline Work Status	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Not Working	1.00	.	.	0.16	1.00	.	.	<0.01	1.00	.	.	<0.01
Working	1.19	0.94	1.51		**1.49**	**1.17**	**1.89**		**1.41**	**1.10**	**1.81**	

† Reported p-values are based on a test for fixed effects of the patient characteristic within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

Table 20. Table of estimated difference in self-reported outcomes between rhBMP-2 versus ICBG by period

Period		Estimated Difference and 95 % Confidence Intervals (CI) for Self-Reported Outcomes Comparing rhBMP-2 with ICBG										
		Oswestry Disability Index (ODI) (decrease is the desired outcome)				**Short Form-36: Physical Component Survey (increase is the desired outcome outcome)**				**Short Form-36: Mental Component Survey (increase is the desired outcome outcome)**			
		Δ	**95% CI**		**p-value†**	Δ	**95% CI**		**p-value†**	Δ	**95% CI**		**p-value†**
	ICBG	0	.	.	0.02	0	.	.	<0.01	0	.	.	0.20
Operative	rhBMP-2	-0.85	-2.28	0.57		-0.09	-0.82	0.64		1.23	-0.11	2.57	
	ICBG	0	.	.		0	.	.		0	.	.	
6 Weeks	rhBMP-2	0.10	-1.87	2.07		0.39	-0.51	1.29		0.84	-0.46	2.13	
	ICBG	0	.	.		0	.	.		0	.	.	
3 Months	rhBMP-2	**-2.23**	**-4.16**	**-0.30**		**1.40**	**0.32**	**2.48**		0.89	-0.43	2.20	
	ICBG	0	.	.		0	.	.		0	.	.	
6 Months	rhBMP-2	**-2.74**	**-4.73**	**-0.74**		**2.10**	**0.89**	**3.30**		0.47	-0.79	1.74	
	ICBG	0	.	.		0	.	.		0	.	.	
12 Months	rhBMP-2	**-2.46**	**-4.57**	**-0.34**		**2.31**	**1.04**	**3.58**		0.00	-1.30	1.30	
	ICBG	0	.	.		0	.	.		0	.	.	
24 Months	rhBMP-2	**-2.49**	**-4.79**	**-0.20**		**1.49**	**0.14**	**2.84**		**1.43**	**0.14**	**2.73**	

† Reported p-values are based on a test for fixed effects of the patient characteristic within a generalized linear mixed model. Δ represents the difference in change over time between rhBMP-2 and ICBG. Bolded estimates indicate that the estimate is significantly different from 0.

Table 21a. Table of estimated difference in self-reported outcomes between different levels of patient characteristics

		Estimated Difference and 95 % Confidence Intervals (CI) for Self-Reported Outcomes Comparing Levels of Patient Characteristics											
		Oswestry Disability Index (ODI) (decrease is the desired outcome)				**Short Form-36: Physical Component Survey (increase is the desired outcome outcome)**				**Short Form-36: Mental Component Survey (increase is the desired outcome outcome)**			
Type		Δ	**95% CI**		**p-value**†	Δ	**95% CI**		**p-value**†	Δ	**95% CI**		**p-value**†
	ALIF	0	.	.	<0.01	0	.	.	<0.01	0	.	.	0.73
	PLF	**-2.41**	**-3.81**	**-1.04**		**-0.74**	**-1.48**	**-0.01**		-0.45	-1.56	0.64	
	PLIF	-0.72	-3.51	1.99		**-2.11**	**-3.52**	**-0.75**		-0.35	-2.63	1.86	
Age		Δ	**95% CI**		**p-value**†	Δ	**95% CI**		**p-value**†	Δ	**95% CI**		**p-value**†
	<60	0	.	.	<0.01	0	.	.	0.06	0	.	.	<0.01
	60 or more	**-4.93**	**-6.52**	**-3.39**		-0.76	-1.56	0.02		**4.80**	**3.45**	**6.11**	
BMI		Δ	**95% CI**		**p-value**†	Δ	**95% CI**		**p-value**†	Δ	**95% CI**		**p-value**†
	Normal	0	.	.	0.15	**0**	**.**	**.**	<0.01	**0**	**.**	**.**	0.04
	Overweight	0.68	-0.76	2.09		**-0.05**	**-0.79**	**0.66**		**-1.81**	**-3.04**	**-0.61**	
	Obese	1.77	0.10	3.40		**-1.38**	**-2.23**	**-0.55**		**-1.45**	**-2.87**	**-0.07**	
	Severely Obese	1.84	-0.29	3.91		**-1.80**	**-2.88**	**-0.75**		**-1.87**	**-3.68**	**-0.11**	

† Reported p-values are based on a test for fixed effects of the patient characteristic within a generalized linear mixed model.

Δ represents the difference in change over time between each level of patient characteristic and the referent level of that patient characteristic. Bolded estimates indicate that the estimate is significantly different from 0.

Table 21b. Table of estimated difference in self-reported outcomes between different levels of patient characteristics

	Estimated Difference and 95 % Confidence Intervals (CI) for Self-Reported Outcomes Comparing Levels of Patient Characteristics											
	Oswestry Disability Index (ODI) (decrease is the desired outcome)				**Short Form-36: Physical Component Survey (increase is the desired outcome outcome)**				**Short Form-36: Mental Component Survey (increase is the desired outcome outcome)**			
Previous Back Surgery	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]
None	**0**	**.**	**.**	<0.01	**0**	**.**	**.**	0.01	0	.	.	0.30
1 or More	**2.19**	**0.91**	**3.42**		**-0.84**	**-1.49**	**-0.22**		-0.57	-1.65	0.48	
Sex	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]
Female	**0**	**.**	**.**	<0.01	**0**	**.**	**.**	<0.01	0	.	.	0.16
Male	**-3.05**	**-4.24**	**-1.90**		**1.92**	**1.31**	**2.50**		0.73	-0.28	1.71	
Smoking	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]
Non-Smoker	0	.	.	0.07	0	.	.	0.10	**0**	**.**	**.**	<0.01
Smoker	1.22	-0.09	2.49		-0.56	-1.22	0.09		**-2.25**	**-3.36**	**-1.17**	
Baseline Work Status	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]	Δ	**95% CI**		**p-value**[†]
Non-Working	**0**	**.**	**.**	<0.01	**0**	**.**	**.**	<0.01	**0**	**.**	**.**	<0.01
Working	**-7.72**	**-8.95**	**-6.52**		**2.45**	**1.82**	**3.06**		**4.63**	**3.58**	**5.65**	

[†] Reported p-values are based on a test for fixed effects of the patient characteristic within a generalized linear mixed model. Δ represents the difference in change over time between each level of patient characteristic and the referent level of that patient characteristic. Bolded estimates indicate that the estimate is significantly different from 0.

Table 22. Table of odds ratios for adverse events for comparing levels within patient characteristics

	Odds Ratio and 95 % Confidence Intervals (CI) for Overall and Severe Adverse Events within Levels of Patient Characteristics														
	Overall 4 Weeks				**Overall 2 Years**				**Severe 4 Weeks**				**Severe 2 Years**			
Treatment	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
ICBG	1.00	.	.	0.33	1.00	.	.	0.46	1.00	.	.	0.51	1.00	.	.	0.45
rhBMP-2	0.80	0.52	1.23		0.75	0.36	1.56		0.81	0.43	1.49		0.82	0.49	1.35	
Type	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
ALIF	**1.00**	**.**	**.**	<0.01	**1.00**	**.**	**.**	<0.01	1.00	.	.	0.27	1.00	.	.	0.27
PLF	**0.47**	**0.33**	**0.67**		**0.29**	**0.20**	**0.42**		1.48	0.79	2.72		0.79	0.50	1.25	
PLIF	**3.22**	**1.66**	**6.15**		**13.71**	**2.16**	**82.49**		2.56	0.69	9.14		1.59	0.65	3.80	
Age	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
<60	**1.00**	**.**	**.**	<0.01	1.00	.	.	0.25	**1.00**	**.**	**.**	0.04	1.00	.	.	0.10
60 or more	**1.67**	**1.20**	**2.31**		1.16	0.85	1.59		**1.76**	**1.03**	**2.97**		1.44	0.93	2.20	
Baseline Work Status	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Not Working	**1.00**	**.**	**.**	0.05	1.00	.	.	0.10	**1.00**	**.**	**.**	0.05	1.00	.	.	0.17
Working	**0.76**	**0.58**	**0.99**		0.81	0.62	1.03		**0.58**	**0.33**	**0.98**		0.77	0.53	1.10	

† Reported p-values are based on a test for fixed effects of the patient characteristic within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

Table 23. Table of odds ratios for related and related, severe adverse events for comparing rhBMP-2 versus ICBG by previous surgery

		Odds Ratios and 95 % Confidence Intervals (CI) for Related and Related, Severe Adverse Events for rhBMP-2 versus ICBG							
		Related AEs at 2 Years				**Related, Severe AEs at 2 Years**			
Previous Back Surgery		**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
None	**ICBG**	**1.00**	**.**	**.**	**<0.01**	**1.00**	**.**	**.**	**<0.01**
	rhBMP-2	**0.22**	**0.09**	**0.51**		**0.09**	**0.03**	**0.32**	
1 or More	**ICBG**	**1.00**	**.**	**.**		**1.00**	**.**	**.**	
	rhBMP-2	2.10	0.86	5.09		1.41	0.53	3.77	

† Reported p-values are based on a test for fixed effects of an interaction between the patient characteristic and treatment within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

Table 24. Table of odds ratios for related and related, severe adverse events for comparing levels within patient characteristics

	Odds Ratio and 95 % Confidence Intervals (CI) for Related and Related, Severe Adverse Events within Levels of Patient Characteristics							
	Related AEs at 2 Years				Related, Severe AEs at 2 Years			
Type	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
ALIF	**1.00**	**.**	**.**	0.01	**1.00**	**.**	**.**	0.03
PLF	**0.41**	**0.23**	**0.73**		**0.42**	**0.22**	**0.79**	
PLIF	**0.84**	**0.29**	**2.36**		**0.87**	**0.28**	**2.59**	
Smoking	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Non-Smoker	**1.00**	**.**	**.**	<0.01	**1.00**	**.**	**.**	<0.01
Smoker	**2.04**	**1.20**	**3.40**		**2.94**	**1.59**	**5.33**	
Baseline Work Status*	**OR**	**95% CI**		**p-value†**	**OR**	**95% CI**		**p-value†**
Not Working	1.00	.	.	0.20	1.00	.	.	0.06
Working	0.69	0.39	1.20		0.52	0.26	1.02	

† Reported p-values are based on a test for fixed effects of the patient characteristic within a logistic mixed model. Bolded odds ratios indicate that the odds ratio is significant and the 95% confidence interval does not include 1.00.

FIGURES

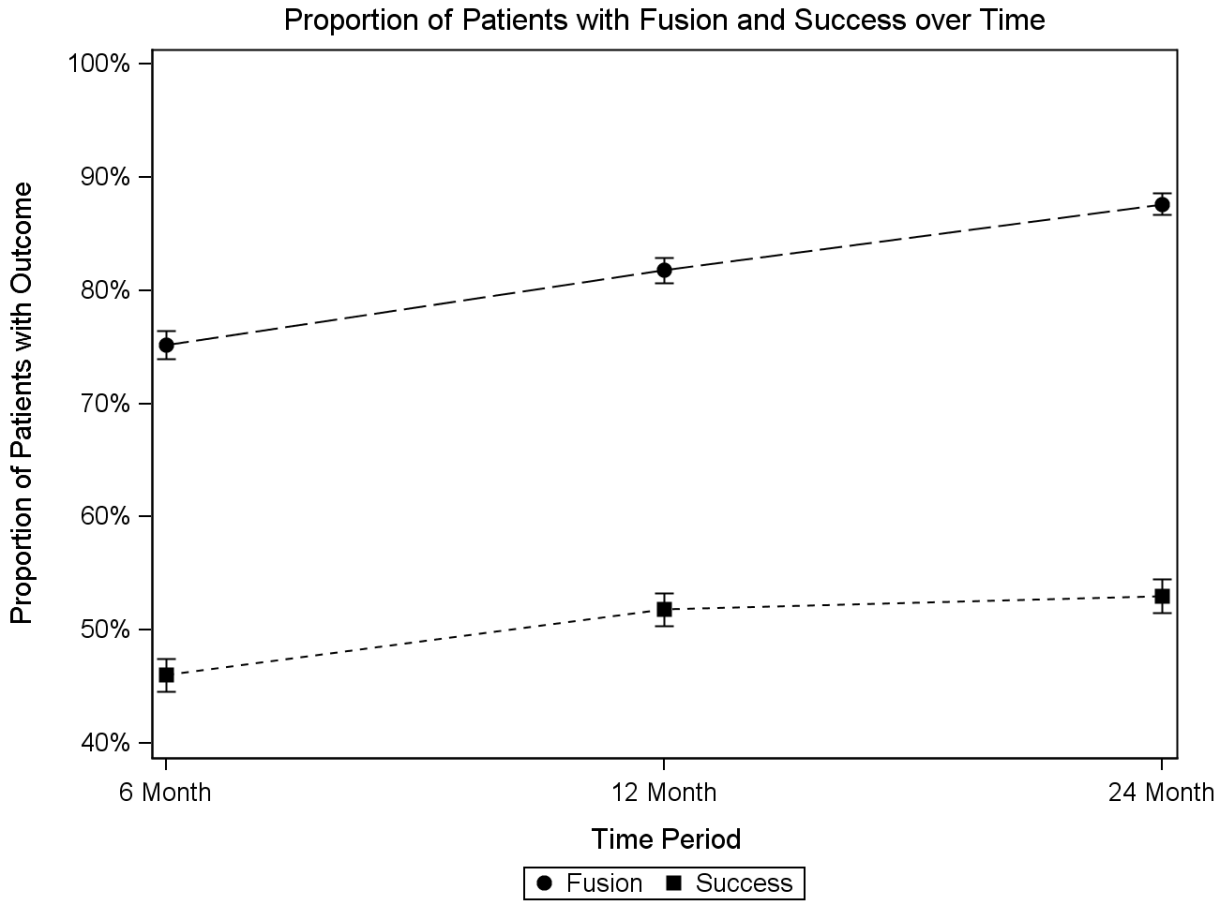

Figure 1.

Proportion of patients with fusion and success with standard error bars after lumbar spinal fusion after 6, 12 and 24 months.

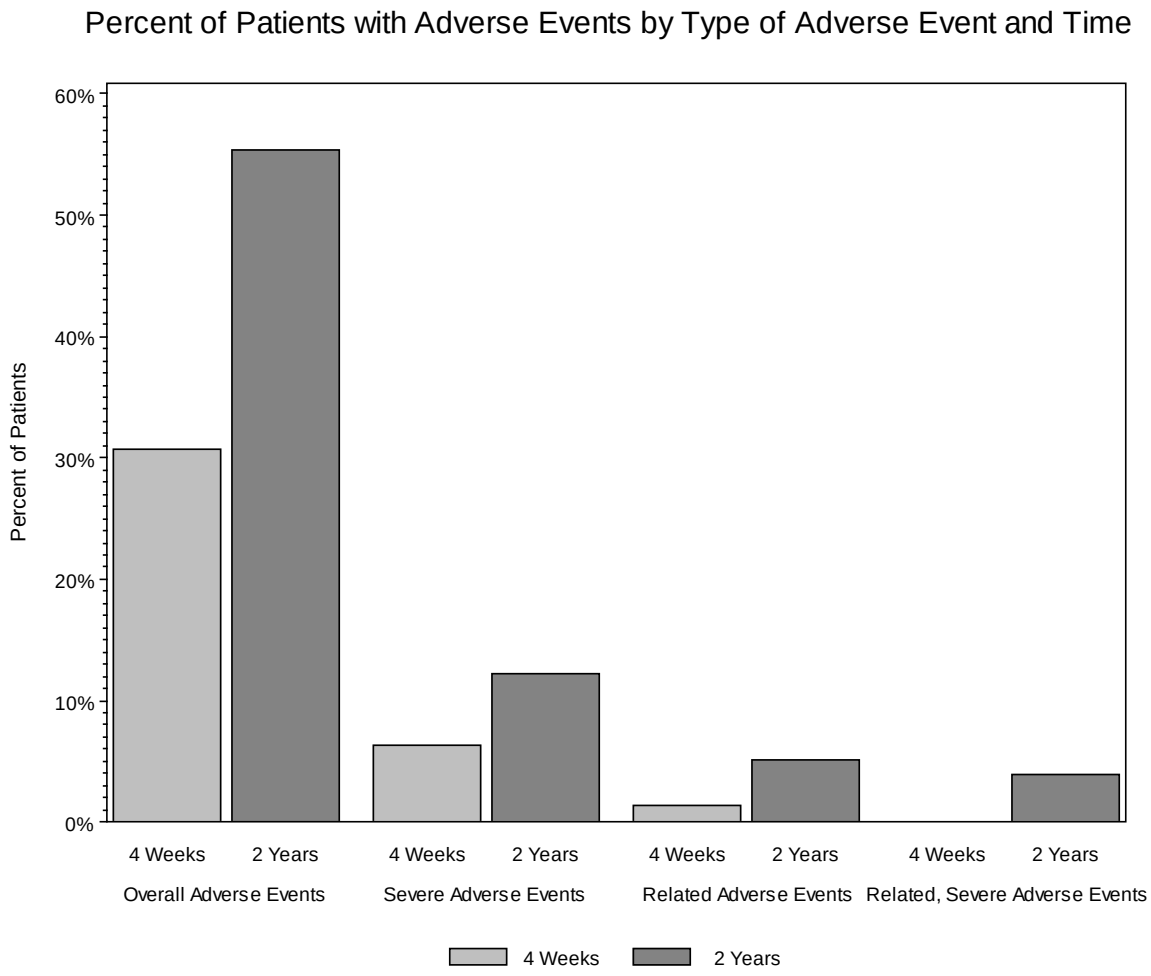

Figure 2.

Proportion of patients with adverse events after 4 weeks and 2 years. Adverse events are categorized as overall (any adverse event), severe, related, and related and severe.

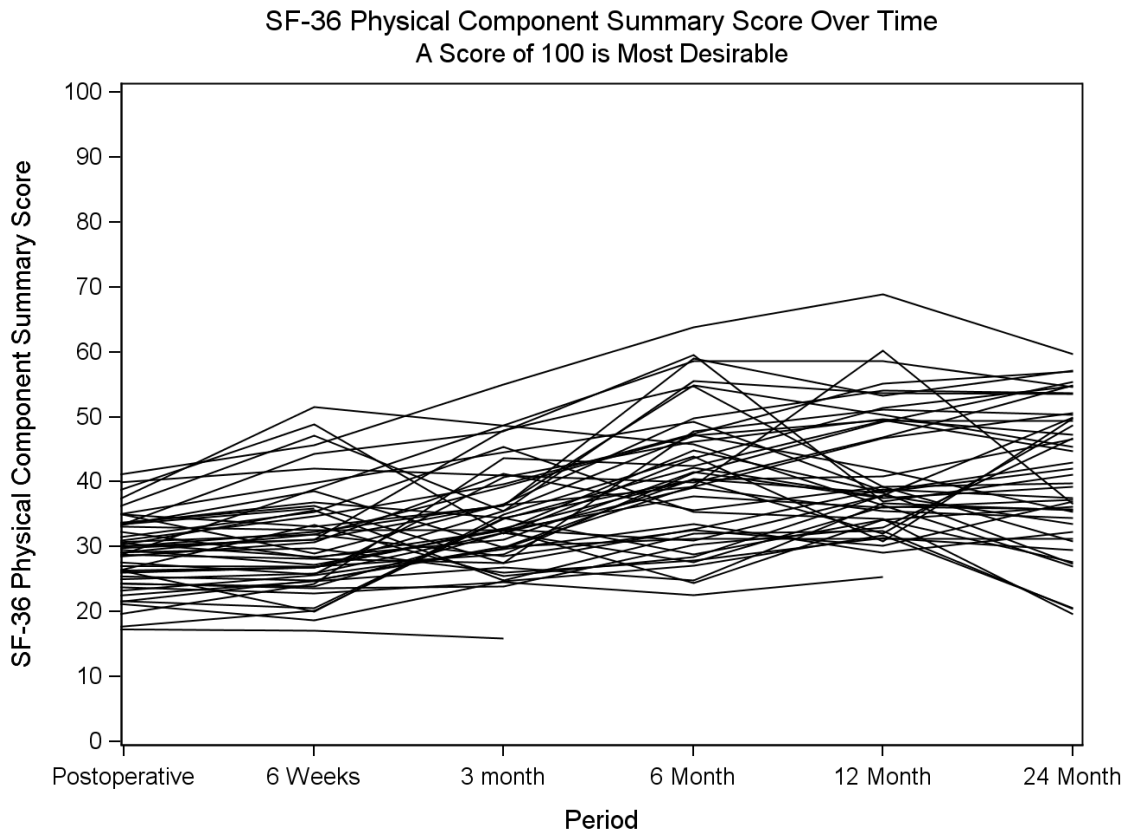

Figure 3.

SF-36 PCS score over time after lumbar spinal fusion for 50 random patients. An increase in SF-36 PCS score is considered improvement.

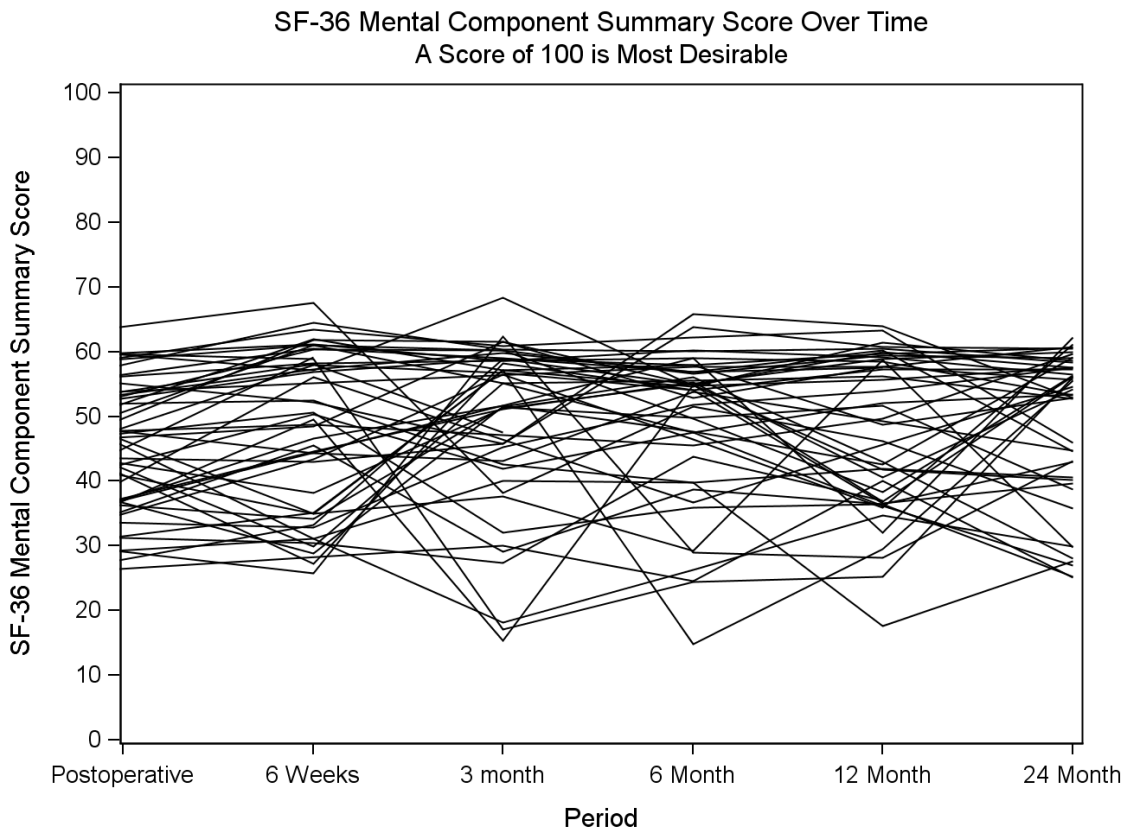

Figure 4.

SF-36 MCS score over time after lumbar spinal fusion for 50 random patients. An increase in SF-36 MCS score is considered improvement.

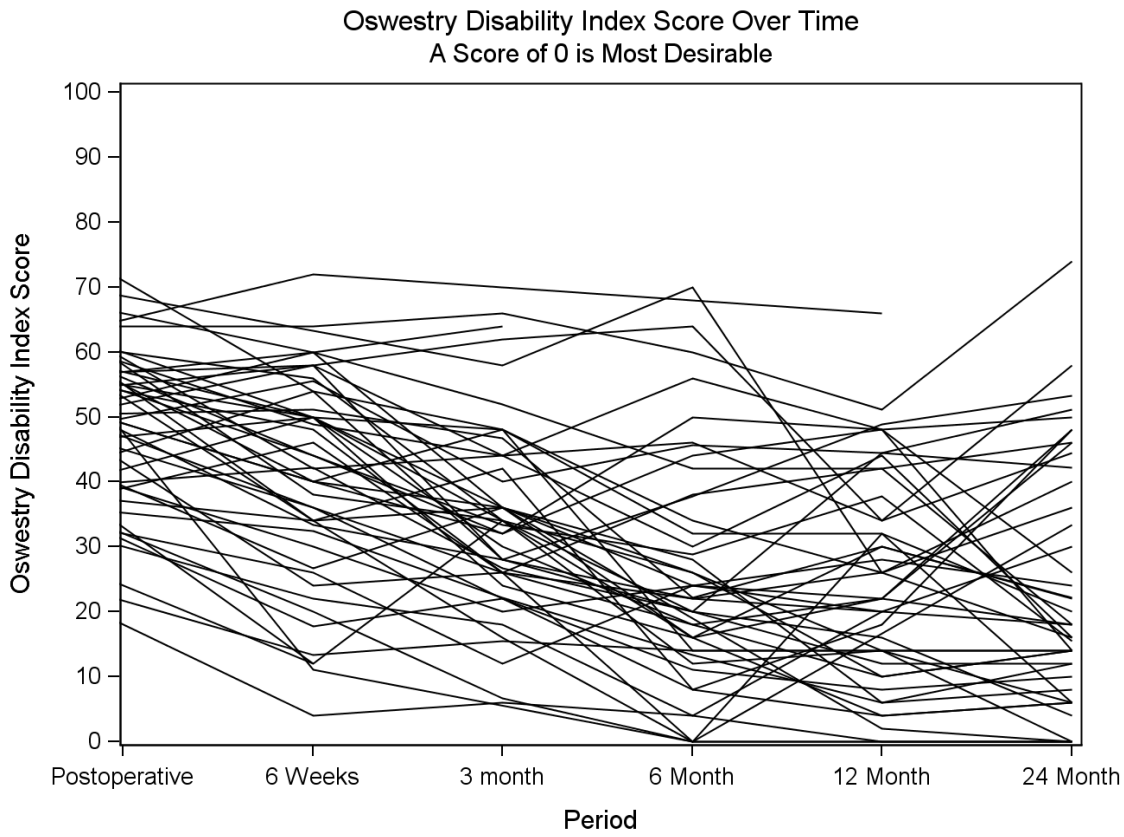

Figure 5.

ODI score over time after lumbar spinal fusion for 50 random patients. A decrease in ODI score is considered improvement.

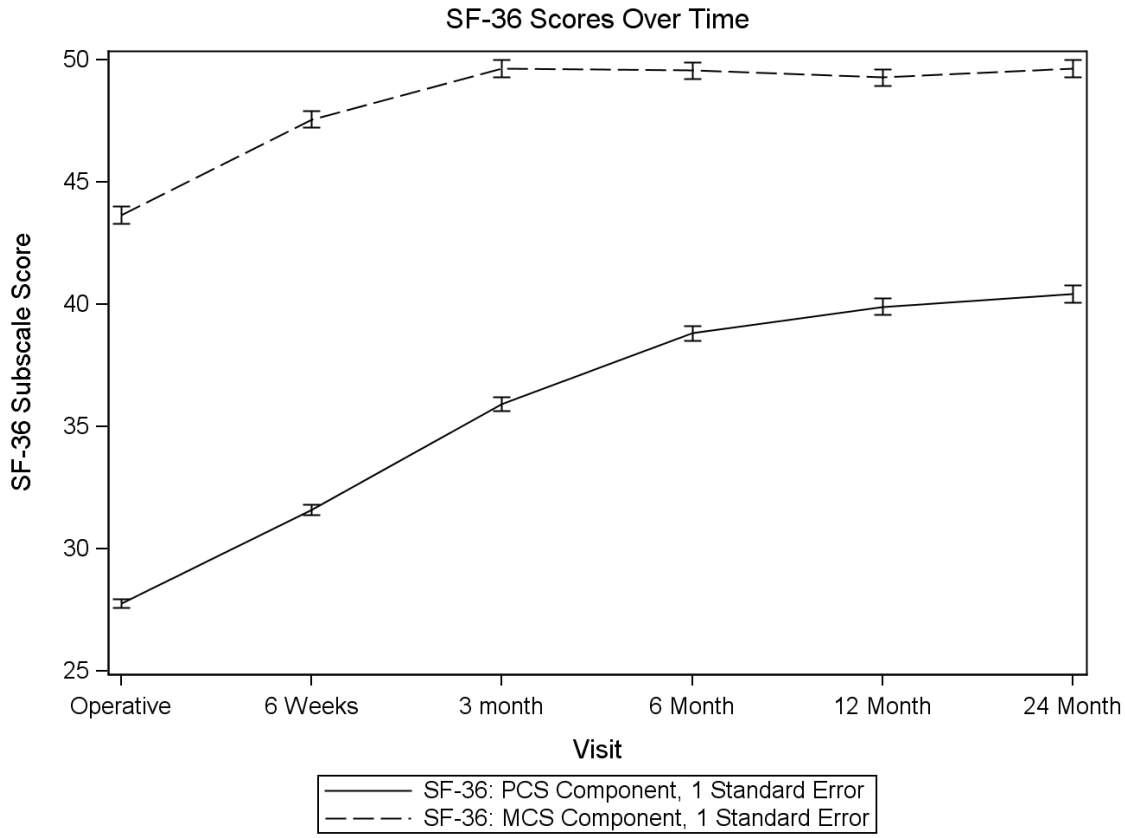

Figure 6

Mean SF-36 physical component summary and mental component summary scores over time after lumbar spinal fusion with standard error bars. An increase in SF-36 subscale score is considered improvement.

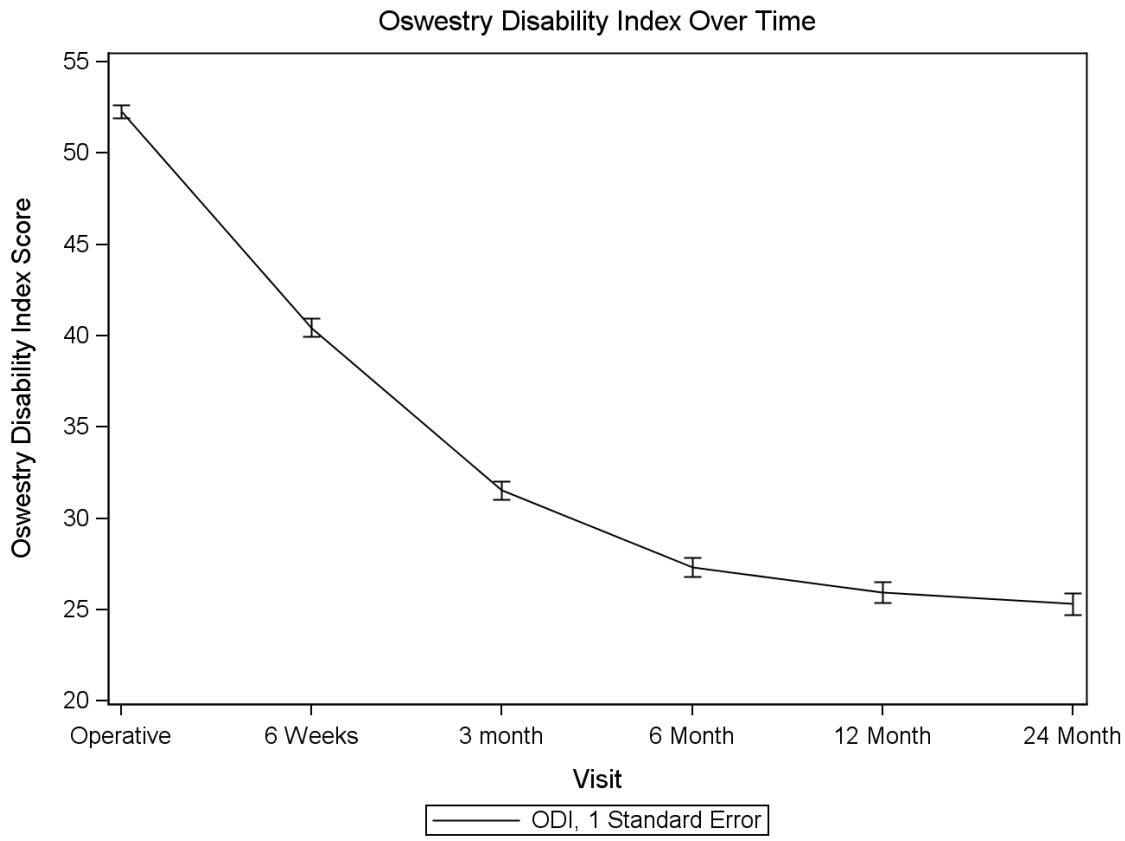

Figure 7

Mean Oswestry Disability Index scores with standard error bars over time after lumbar spinal fusion. A decrease in Oswestry Disability Index score is considered improvement.

ABBREVIATIONS

AE	adverse event
ALIF	anterior lumbar interbody fusion
ANOVA	analysis of variance
BMI	body mass index (kg/m^2)
DDFM	denominator degrees of freedom
ICBG	iliac crest bone graft
IPD	individual patient data
ODI	Oswestry Disability Index
OR	odds ratio
PLF	posterolateral lumbar fusion
PLIF	posterior lumbar interbody fusion
rhBMP-2	recombinant bone morphogenetic protein-2
SAE	serious adverse event
SF-36 PCS	Short Form 36: Physical Component Summary
SF-36 MCS	Short Form 36: Mental Component Summary
YODA	Yale University Open Data Access