HHS Public Access

Author manuscript

J Arthroplasty. Author manuscript; available in PMC 2018 April 01.

Published in final edited form as:

J Arthroplasty. 2017 April ; 32(4): 1153–1158.e1. doi:10.1016/j.arth.2016.11.007.

External Validation of a Prognostic Model for Predicting Nonresponse Following Knee Arthroplasty

Daniel L. Riddle, PT, PhD, FAPTA,
Departments of Physical Therapy, Orthopaedic Surgery and Rheumatology, West Hospital, Room B-100, Virginia Commonwealth University, Richmond, Virginia 23298-0224

Gregory J. Golladay, MD,
Department of Orthopaedic Surgery, Virginia Commonwealth University

William A. Jiranek, MD, and
Department of Orthopaedic Surgery, Virginia Commonwealth University

Robert A. Perera, PhD
Department of Biostatistics, Virginia Commonwealth University

Abstract

Background—Instruments designed to predict extent of pain and function following knee arthroplasty (KA) recovery have strong potential to guide patients and clinicians in shared decision-making. Our purpose was to test the external validity of a recently developed prognostic instrument designed to estimate the probability of nonresponse following knee arthroplasty.

Methods—We used data from the Osteoarthritis Initiative (OAI), a nine-year multi-site National Institutes of Health study designed to examine the natural history of knee osteoarthritis in 4,796 subjects. A total of 427 subjects underwent KA over the study period. Dowsey and colleagues examined the prognostic role of obesity, general mental health, pain and function, and Kellgren and Lawrence knee osteoarthritis (OA) grades. Calibration of the prognostic model was determined using a calibration curve. The c-statistic was used to indicate discrimination of the model.

Results—In the primary analysis 63 (19.3%) of 326 subjects in OAI were classified as non-responders. The calibration curve generated from OAI data indicated poor calibration relative to the recently developed instrument. Discrimination as measured by the c-statistic was 0.76.

Conclusion—The external validity of the prognostic instrument was partially supported. While discrimination of the model was very similar to the recently developed instrument, calibration was poor indicating poor agreement between actual versus predicted probabilities of nonresponse. WOMAC and Kellgren and Lawrence grades show strong potential for use in future prognostic

corresponding author: Phone: 804-828-0234, Fax: 804-828-8111, dlriddle@vcu.edu.

model development. Measurements of general mental health and obesity were not prognostic for nonresponse.

Keywords

knee; arthroplasty; nomogram; prognosis; pain; function

INTRODUCTION

A hallmark of contemporary prognostic research is the development of simple-to-use instruments that determine the extent to which multiple variables predict an important outcome [1]. The nomogram developed by Dowsey and colleagues for predicting risk of nonresponse following knee arthroplasty (KA) is an example [2]. Development of these instruments, however, is inadequate to assure generalizability to different clinical sites. There are, for example, differences in healthcare systems, methods of obtaining measurements, and patient characteristics that could explain why a prognostic model may not generalize beyond the sample in which it was developed [3]. Validation on a sample independent from the development study is required before the instrument can be recommended for widespread use [3,4].

Estimating the risk of poor outcome following KA is a highly important preoperative prognostic assessment. Patients who are better informed about their prognosis can engage in more robust shared decision-making. Additionally, if patients have modifiable factors which place them at high risk for poor outcome prior to surgery, these factors can be optimized prior to surgery to reduce poor outcome risk. Given the recent focus by Medicare [5] and other stakeholders [6] on outcome monitoring for patients undergoing KA, this type of research is a high priority.

Dowsey and colleagues examined the prognostic value of several pre-operative variables in predicting one-year pain and physical function outcomes in a sample of 562 patients with KA [2]. Poor outcome was defined using the term “nonresponse” as specified by Outcome Measures in Rheumatology - Osteoarthritis Research Society International (OMERACT-OARSI) responder criteria [7]. These criteria use WOMAC pain and function scores in various combinations to classify patients as either responding or not responding to treatment. A total of 11 pre-operative predictor variables were measured and four were found in the multivariable model to estimate risk of nonresponse. Key pre-operative predictor variables were body mass index (BMI) classified as $<40 kg/m^2$ or ≥ 40 kg/m2, Kellgren and Lawrence (K&L) grade, classified as grade 4 or ≤ grade 3, the Total WOMAC score, and the SF-12 Mental Component Summary score classified into quartiles. Results were converted to an easy-to-use nomogram designed to speed translation in clinical settings.

Our objective was to conduct an external validation study of the work by Dowsey and colleagues [2]. We examined the extent to which the multivariable model and nomogram generalized to an independent dataset.

MATERIAL AND METHODS

Data source

The Osteoarthritis Initiative (OAI) is a nine-year prospective community-based National Institutes of Health and privately funded multicenter natural history longitudinal study of persons with radiographic knee osteoarthritis (OA) or who are at high risk for developing knee OA. Knee OA risk was determined by one or more of the following: overweight or obesity, prior knee injury or surgery, knee symptoms or family history of knee replacement surgery. The OAI study was approved by the Institutional Review Boards and all subjects signed informed consent at one of four sites: (1) University of Maryland, Baltimore, Maryland, (2) Ohio State University, Columbus, Ohio, (3) University of Pittsburgh, Pittsburgh, Pennsylvania, and (4) Memorial Hospital of Rhode Island, Pawtucket, Rhode Island.

A total of 17,457 men and women aged 45 to 79 years were screened and 4,796 were enrolled. Among the more common exclusion criteria were magnetic resonance imaging height and weight limitations (n=2,328), recruitment thresholds for age and gender (n=2,954) and dropouts prior to the enrollment visit (n=4,381). The OAI website provides detailed information (http://www.oai.ucsf.edu/).

Subjects

Over the nine-year study period, a total of 427 subjects had at least one primary KA. In the OAI, subjects indicate during yearly visits when a KA was done in the prior year. OAI investigators adjudicate all KAs using medical record data and in our study 412 KAs were confirmed and 15 were pending adjudication.

Much like the study of Dowsey et al, some subjects in our sample had more than one arthroplasty. In the OAI, 107 of 427 subjects had right and left knee arthroplasty procedures over the nine-year period and in 56 of the 107, two KAs were done in the same year. When a subject had bilateral KA in different years or in the same year but different days we used data associated with the earliest procedure and when bilateral KA was done on the same day, we randomly selected either the right or left knee for study. We used only on knee per subject so that each subjects' data was independent.

The diagnosis was OA in the majority of cases (381 of 427), 2 had rheumatoid arthritis, 1 had osteonecrosis, 16 had post-traumatic OA and in 27, the pre-operative diagnostic data were missing. There were 383 total KAs and 34 partial KAs in our sample while for 10 knees the procedural data were missing. For the 34 knees with partial KAs, there were 3 with a lateral KA, 29 with medial KA and 2 with patellofemoral KA.

Measurement of outcome

Each year from baseline to the nine-year follow-up, subjects completed the WOMAC Pain (scores ranging from 0 to 20 with higher scores equating to worse pain during activity), Physical Function (scores ranging from 0 to 68 with higher scores equating to more difficulty with activity) and Stiffness (scores ranging from 0 to 8 with higher scores equating

to worse stiffness) scales. Scores on these scales are summed for the WOMAC Total Score. The WOMAC scale has been used extensively in the arthroplasty literature and has been validated in multiple studies [8,9]. In accordance with the method used by Dowsey and colleagues, [2] the three WOMAC subscales to calculate responder scores using the OMERACT-OARSI responder criteria [7]. In brief, these criteria classify subjects as responders if they demonstrate a 50% reduction and a 20 point absolute change (using scores transformed to a 0 to 100 scale) in either WOMAC Pain or Function, or a 20% improvement and a 10 point absolute improvement in any 2 of the 3 subscales of pain, function or total scores. Subjects are classified as non-responders if they do not meet any of these criteria.

In the OAI, the time between yearly study visits and the surgery date varied for each subject. Because OAI has time varying data relative to KA surgery, we used two postoperative windows of time to examine responder criteria. For the pre-operative data we used the visit prior to surgery. For the postoperative follow-up we used two windows of time. For the primary analysis, we used the postoperative follow-up period (described as Model #1) between 365 days and 730 days post-surgery. For a secondary analysis (described as Model #2) the postoperative visit occurred between 180 days and 540 days post-surgery. We chose these time periods because pain and function outcome from 6 months to 2 years following KA has been shown to be reasonably stable in several outcome studies [10–12]. Using two windows of time maximized the number of subjects for each analysis and allowed for a sensitivity analysis.

Predictors of outcome

We used the four pre-operative predictor variables identified by Dowsey and colleagues in their multivariable model for predicting nonresponse. These variables were BMI, dichotomized as either $<40 kg/m^2$ or $\geq 40 kg/m^2$, Kellgren and Lawrence (K&L) grade [13] classified as grade 4 or ≤ grade 3, the Total WOMAC score, and the SF-12 Mental Component Summary (MCS) score, a validated general health status measure of generalized psychological distress [14], classified into quartiles as normal (≥50 points), mild psychological distress (40 to 49.9 points), moderate psychological distress (30 to 39.9 points) and severe psychological distress (<30 points). Because K&L grades and the MCS were not collected in OAI during follow-up years 5 and 7, we used preoperative scores within 2 years of surgery for subjects who had TKA in OAI years 7 and 9, respectively.

Knee radiographs were obtained on all subjects using a highly standardized, reliable and valid approach [15]. Two central site readers and a third adjudicator, all either a rheumatologist or a musculoskeletal radiologist with extensive training and very high reliability [16] read all radiographs.

Data analysis

Validation of a prognostic model requires use of variables and coefficients from the original prognostic model including, in this case, the multiple logistic regression model estimates by Dowsey and colleagues [2]. The four variables predicting nonresponse were BMI dichotomized to ≥40 or <40, K&L knee OA grade, dichotomized to <4 or 4, MCS scores divided into quartiles, and the WOMAC Total score. Because only 2 subjects had MCS

scores in the bottom (worst) quartile, we included these two subjects in the third quartile of scores (<40).

To develop a prognostic model, both calibration and discrimination of the model are tested [17]. “Calibration” describes how well the observed and predicted probabilities of the outcomes of interest (i.e., non-response following KA) align. Calibration is typically tested by plotting the predicted probabilities versus the observed outcome. A line is created to represent the agreement between the predicted and observed probabilities where a line at 45 degrees indicates perfect agreement. “Discrimination” in our study describes the ability of the model to assign higher probabilities for true non-responders as compared to true responders. Discrimination is commonly tested using the concordance statistic (c-statistic), also known as the area under the receiver operating characteristic curve [17].

Dowsey and colleagues assessed calibration by plotting the observed proportions of nonresponse and the predicted probabilities of nonresponse from their logistic regression model. The plot showed only slight deviation from perfect prediction across the range of probabilities, with slight overestimation at the end of the distribution and slight underestimation occurring at the beginning and middle of the distribution. To assess discrimination of the model (i.e., how well does the model differentiate among responders versus non-responders) Dowsey and colleagues reported a c-statistic of 0.74. This estimate indicates that the model will assign a higher probability to a randomly selected non-responder 74% of the time [18].

We used the coefficients from the model by Dowsey and colleagues (see Table 2) to assess calibration and discrimination in our external validation dataset [3]. Calibration was tested by plotting observed and predicted probabilities using OAI data and the coefficients from Dowsey and colleagues. The c-statistic was calculated for discrimination. Calibration and discrimination were tested both for the primary analysis (i.e., Model #1) and the sensitivity analysis (i.e., Model #2).

To aid interpretation, we conducted a multiple logistic regression using the four key variables from Dowsey et al. (i.e., BMI, K&L grade, MCS and WOMAC Total scores) to generate coefficients directly from OAI data. This approach was used to determine the extent to which the two datasets may have differed in their predictive qualities for each of the key predictor variables. All analyses were completed using the R statistical software [19].

Results

A total of 326 subjects had follow-up outcome data for the primary analysis (i.e., Model #1) and of these, 63 (19.3%) were classified as non-responders. The postoperative visit was, on average, 536 (sd=95) days post-surgery. Missing data on predictor variables further lowered the sample size to 272 with a total of 52 participants classified as non-responders (19.1%). For the sensitivity analysis (i.e., Model #2), a total of 349 subjects were included and 85 (24.4%) were classified as non-responders. The postoperative visit for the sensitivity analysis was a mean of 355 (sd=108) days following surgery. Similar to the primary

analysis, missing data on predictors reduced the total sample size to 284 with 69 (24.3%) being classified as non-responders.

The preoperative characteristics of the OAI samples for the key preoperative variables are reported in Table 1. Our validation calibration plot indicated poor calibration of the OAI data when applying the prognostic coefficients reported by Dowsey and colleagues. There was substantial underestimation at the lower and upper middle parts of the calibration curve (i.e. when the curve deviates above the diagonal line) (See Figure 1). The calibration curve for the sensitivity analysis (Model #2) was essentially identical to that found for Model #1 (see appendix). To test discrimination, the c-statistic for the primary analysis was 0.76 (95%CI = 0.69, 0.83) and 0.76 (95%CI = 0.70, 0.83) for Model #2. Dowsey and colleagues reported a c-statistic of 0.74. The coefficients from the multiple logistic regression estimated from OAI data and from the work of Dowsey and colleagues are reported in Table 2.

Discussion

The prognostic nomogram developed by Dowsey and colleagues [2] is, to our knowledge, the first prognostic instrument designed to aid both clinicians and patients in estimating risk of nonresponse to KA. As Dowsey and colleagues acknowledge, however, external validation on an independent dataset is necessary before broad implementation of the nomogram can be endorsed.

Our study tested the external validity of the Dowsey et al. [2] prognostic nomogram and our findings were mixed. On the one hand, discrimination, as measured by the c-statistic was very similar to that reported by Dowsey and colleagues. A c-statistic discrimination index of 0.76 indicates that 76% of the time, non-responders would be expected to have higher probabilities of non-response as compared to responders. Discrimination as a test of validation was therefore supported. In addition, our nonresponse rate of 19.3% for the primary analysis is similar to the nonresponse rate of 15% reported by Dowsey and colleagues [7]. On the other hand, our examination of model calibration, an indicator of the stability of the prediction of nonresponse across a range of probabilities, indicated instability in estimating actual probability of nonresponse. In particular, the nomogram substantially underestimated the observed proportions on the lower and middle to upper parts of the curve (see Figure 1). Dowsey and colleagues found that calibration was much closer to ideal with only slight deviations from perfect prediction.

It is the calibration test that describes the utility of the nomogram for estimating probability of non-response as determined with the nomogram. In our validation testing of the nomogram, the calibration line fitting our data deviated substantially from the diagonal line indicating poor validation. Based on this analysis, the nomogram developed by Dowsey and colleagues appears to require additional development and study prior to broad clinical application.

Our study has several strengths. First, data were collected from four sites in different regions of the country. This may support generalizability of our findings. In addition, though our sample is not as large as that reported by Dowsey and colleagues, the sample sizes are

nonetheless relatively large and loss to follow-up is relatively low. We had an 88% follow-up for 2 to 3 year post surgery visits when accounting for subjects who had surgery in years 8 or 9 and who were not eligible for follow-up visits at the time of our study. When considering all subjects independent of surgery date, we had 76% follow-up at the 2- to 3-year visit window and 82% follow-up at the 6- to 18-month visit window.

We included both unicompartmental and total knee arthroplasty cases which may have influenced our findings. We compared the preoperative WOMAC Pain and Function scores of these two subgroups of KAs and found that scores were similar. For example, patients with unicompartmental KA had preoperative WOMAC Pain mean scores of 8.2 points compared to 7.5 points for persons with total KA. Differences among unicompartmental and total KAs preoperatively were not significantly different (p values of 0.329 for WOMAC Pain (t = 0.977) and 0.969 for WOMAC Function (t=0.038)).

We conducted a second post hoc analysis to determine if the main findings were maintained if we excluded persons with unicompartmental KA. The area under the curve estimates were 0.774 (0.698, 0.850) for the Model #1 analysis and 0.777 (0.712, 0.840) for the Model #2 analysis. These estimates are almost identical to the full analyses as were the calibration plots (see Appendix). These data, provide substantial evidence to suggest that inclusion of unicompartmental KA in our sample did not influence our findings.

Our sample, much like sample studied by Dowsey and colleagues included persons with unilateral and bilateral KA. Because bilateral KA may be associated with poor outcome, we determined the proportion of persons with bilateral KA and the proportion of persons with unilateral KA in the non-responder group. For the primary analysis (described as Model #1), 16.3% of the persons with bilateral arthroplasty were classified as non-responders while 20.4% of persons with unilateral knee arthroplasty were non-responders. We consider this difference to be minor and supportive of the argument that the findings are not generally influenced by subjects with bilateral versus unilateral knee arthroplasty.

One limitation of our study is that the OAI study sample is relatively healthy at baseline, and only a minority of the subjects have undergone arthroplasty. However, the study specifically assessed nonresponse in persons who had undergone KA. In addition, a small number of subjects (n = 34) in OAI underwent unicompartmental KA which may influence generalizability. While, improvements following unicompartmental knee replacement are similar to that reported for total knee arthroplasty [20], and indications are overlapping [21,22], inclusion of subjects with unicompartmental KA may have influenced the results. This concern is lessened by our post hoc analyses showing no significant difference in preoperative scores among unicompartmental versus total KA recipients and the analyses showing essentially no difference in findings when the unicompartmental cases were excluded. Since one of our goals was to design this study to mimic routine clinical practice, we included persons who underwent either total or unicompartmental KAs. Similarly, we included persons who had multiple KA surgeries over the study period, much like the study of Dowsey and colleagues [2].

Our study also is limited because our data was time varying unlike the data from Dowsey and colleagues. As a result, we conducted a sensitivity analysis in addition to the primary analysis to account for the fact that subjects in OAI were followed up at different times relative to the date of KA. One approach used 2- to 3-year postoperative data while the other used 6-month to 18-month postoperative data. Our findings were very similar for both time periods suggesting that these time windows did not contribute error to our analyses. The time from the pre-operative OAI visit to surgery averaged 169 days (sd=95) and this variation also may have contributed error to our estimates (see Table 2). In two post hoc analyses we examined: 1) number of days between the preoperative visit and surgery, used as a covariate in the logistic regression model, and 2) a subgroup of only those with a preoperative visit within 6 months of surgery (n = 231). In post hoc analysis 1, number of days was not a significant predictor of responder status and, in post hoc analysis 2, findings were not appreciably changed relative to the findings for the entire sample (See Appendix).

To determine why the calibration test for the Dowsey et al. prognostic model did not generalize to OAI data, we conducted two post hoc analyses. First, we generated a calibration plot using multiple logistic regression coefficients estimated from OAI data. Because these post hoc analyses were exploratory, and the calibration was not performed on an external sample, we used a random sampling boostrapping procedure to estimate the probability distributions, much like the methods by Dowsey and colleagues [2]. This plot (see Figure 2) showed a better fit as compared to the validation test but probabilities of nonresponse deviated from ideal at the ends and middle of the curve, suggesting a non-linear association. When examining the calibration plot generated only from the baseline WOMAC total score (see Figure 3) there were more substantial deviations at the ends of the curve relative to the plot shown in Figure 2. These findings suggest that the WOMAC total score was the likely source of non-linearity in our calibration test of the Dowsey et al. model. In addition, our sample had lower (better) WOMAC scores prior to surgery as compared to the Dowsey et al. data, which also may have contributed to the poor calibration in our data.

Compared to the sample studied by Dowsey and colleagues [2], our sample had: 1) a lower percentage of subjects with BMI ≥ 40kg/m2, 2) a lower percentage with moderate or severe MCS scores, 3) a similar percentage with K&L scores of 3 or less and 4) lower (better) WOMAC Total scores. Neither the BMI or MCS variables in OAI contributed substantively to the model unlike that of Dowsey et al. For example, our sample's lower percentages of morbidly obese subjects and fewer psychologically distressed subjects relative to the sample in the Dowsey et al. study likely contributed to the poor calibration in the OAI dataset. We had, for example, only 17 (4%) subjects in our study with a BMI of 40 or greater compared to 101 (16.4%) in the Dowsey et al. study. Our small sample of morbidly obese subjects was likely influenced by weight restrictions imposed in OAI because of MRI requirements. However, our morbidly obese subjects were not at higher risk for nonresponse relatively to the non-morbidly obese sample (see Table 2).

As demonstrated in the multiple logistic regression analysis, the coefficients in Table 2 suggest that WOMAC Total scores were more powerful predictors of nonresponse in the OAI data as compared to the Dowsey et al. data. The contribution of K&L grades to the prediction of nonresponse appeared to be very similar in the two datasets while BMI and

MCS made non-significant contributions to prognostic prediction. Overall, the substantive differences in BMI, MCS, and WOMAC Total scores between the two datasets likely explain the weak calibration findings in OAI relative to the findings by Dowsey et al. and reflect fundamental differences in sample characteristics. Importantly, in spite of these differences, WOMAC Total and K&L grades appear to be important prognostic predictors across datasets.

Additional prognostic variables also warrant further study and may add to the predictive power of WOMAC scores and index knee K&L grades. For example, instruments designed to quantify pain catastrophizing, anxiety and depressive symptom severity may provide more granular and therefore more stable and more predictive measures of psychological health [23]. Contralateral knee OA and pain status [24] as well as index knee quadriceps strength[25], fibromyalgia symptoms[26] and overall bodily pain also have been associated with poor outcome following KA and may add to the power of a prognostic model. Finally, greater comorbidity[27] and African American race[28] also are associated with poor outcome.

Conclusions

The prognostic nomogram developed by Dowsey and colleagues was partially supported in our external validation study. The prognostic role of WOMAC Total scores and K&L grades were substantiated in our study but the importance of BMI and SF-12 MCS scores were not supported. Differences in sample characteristics between our study and that of Dowsey and colleagues [2] likely contributed to the differences. Additional development and validation of prognostic instruments to assist both clinicians and patients is warranted prior to broad application. Our study and the study of Dowsey and colleagues support the use of self-reported preoperative pain and function measures as well as K&L grades in the further development of KA prognostic outcome instruments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

Reference List

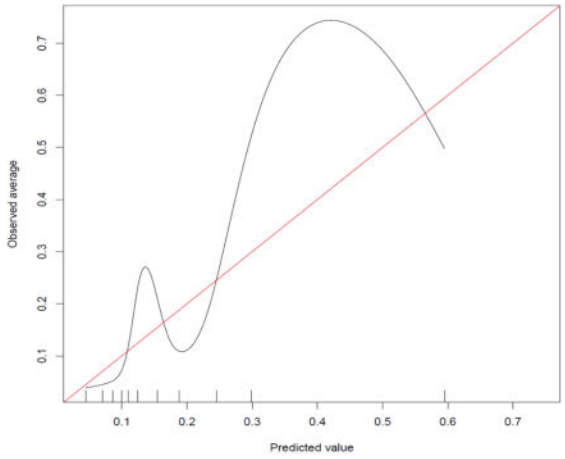
1. Tevis SE, Weber SM, Kent KC, Kennedy GD. Nomogram to Predict Postoperative Readmission in Patients Who Undergo General Surgery. JAMA Surg. 2015; 150(6):505–510. [PubMed: 25902340]
2. Dowsey MM, Spelman T, Choong PF. Development of a Prognostic Nomogram for Predicting the Probability of Nonresponse to Total Knee Arthroplasty 1 Year After Surgery. J Arthroplasty. 2016

3. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009; 338:b605. [PubMed: 19477892]
4. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015; 350:g7594. [PubMed: 25569120]
5. Center for Medicare and Medicaid Services. Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty: Hospital-Level Performance Measure(s) Phase 3 Measure Methodology Report. May 1. 2015 https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html
6. Schilling PL, Bozic KJ. Development and Validation of Perioperative Risk-Adjustment Models for Hip Fracture Repair, Total Hip Arthroplasty, and Total Knee Arthroplasty. J Bone Joint Surg Am. 2016; 98(1):e2. [PubMed: 26738909]
7. Pham T, van der HD, Altman RD, et al. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. Osteoarthritis Cartilage. 2004; 12(5):389–399. [PubMed: 15094138]
8. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes. 2003; 1(1): 17. [PubMed: 12801417]
9. Collins NJ, Roos EM. Patient-reported outcomes for total hip and knee arthroplasty: commonly used instruments and attributes of a “good” measure. Clin Geriatr Med. 2012; 28(3):367–394. [PubMed: 22840304]
10. Fortin PR, Penrod JR, Clarke AE, et al. Timing of total joint replacement affects clinical outcomes among patients with osteoarthritis of the hip or knee. Arthritis Rheum. 2002; 46(12):3327–3330. [PubMed: 12483739]
11. Lingard EA, Katz JN, Wright EA, Sledge CB. Predicting the outcome of total knee arthroplasty. J Bone Joint Surg Am. 2004; 86-A(10):2179–2186. [PubMed: 15466726]
12. Nilsdotter AK, Toksvig-Larsen S, Roos EM. A 5 year prospective study of patient-relevant outcomes after total knee replacement. Osteoarthritis Cartilage. 2009; 17(5):601–606. [PubMed: 19091604]
13. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957; 16(4):494–502. [PubMed: 13498604]
14. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996; 34(3):220–233. [PubMed: 8628042]
15. Kothari M, Guermazi A, von IG, et al. Fixed-flexion radiography of the knee provides reproducible joint space width measurements in osteoarthritis. Eur Radiol. 2004; 14(9):1568–1573. [PubMed: 15150666]
16. Felson, DT. Central Reading of Knee X-rays for K-L Grade and Individual Radiographic Features of Knee OA. 2011.
17. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009; 338:b604. [PubMed: 19336487]
18. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic biomarkers (Part I): discrimination. Nephrol Dial Transplant. 2010; 25(5):1399–1401. [PubMed: 20139066]
19. R Development Core Team. R: A language and environment for tatistical computing. R Foundation for Statistical Computing; Jan 1. 2008 http://www.R-project.org
20. Baker PN, Petheram T, Jameson SS, et al. Comparison of patient-reported outcome measures following total and unicondylar knee replacement. J Bone Joint Surg Br. 2012; 94(7):919–927. [PubMed: 22733946]
21. Carr AJ, Robertsson O, Graves S, et al. Knee replacement. Lancet. 2012; 379(9823):1331–1340. [PubMed: 22398175]
22. Beard D, Price A, Cook J, et al. Total or Partial Knee Arthroplasty Trial - TOPKAT: study protocol for a randomised controlled trial. Trials. 2013; 14:292. [PubMed: 24028414]

23. Lungu E, Vendittoli PA, Desmeules F. Preoperative Determinants of Patient-reported Pain and Physical Function Levels Following Total Knee Arthroplasty: A Systematic Review. Open Orthop J. 2016; 10:213–231. [PubMed: 27398109]
24. Maxwell J, Niu J, Singh JA, Nevitt MC, Law LF, Felson D. The influence of the contralateral knee prior to knee arthroplasty on post-arthroplasty function: the multicenter osteoarthritis study. J Bone Joint Surg Am. 2013; 95(11):989–993. [PubMed: 23780536]
25. Zeni JA Jr, Snyder-Mackler L. Preoperative predictors of persistent impairments during stair ascent and descent after total knee arthroplasty. J Bone Joint Surg Am. 2010; 92(5):1130–1136. [PubMed: 20439658]
26. Brummett CM, Urquhart AG, Hassett AL, et al. Characteristics of fibromyalgia independently predict poorer long-term analgesic outcomes following total knee and hip arthroplasty. Arthritis Rheumatol. 2015; 67(5):1386–1394. [PubMed: 25772388]
27. Hawker GA, Badley EM, Borkhoff CM, et al. Which patients are most likely to benefit from total joint arthroplasty? Arthritis Rheum. 2013; 65(5):1243–1252. [PubMed: 23459843]
28. Goodman SM, Parks ML, McHugh K, et al. Disparities in Outcomes for African Americans and Whites Undergoing Total Knee Arthroplasty: A Systematic Literature Review. J Rheumatol. 2016; 43(4):765–770. [PubMed: 26834217]

Appendix 1

Method #2 Calibration Curve



Post hoc analysis 1

Multiple Logistic regression coefficients for primary outcome including number of days measurement was obtained prior to surgery

	Coefficient	S.E.	Wald Z	p value
Intercept	0.36	0.598	0.61	0.544
BMI ≥40	−0.05	0.931	−0.05	0.960
K&L Grade ≤3	0.96	0.365	2.62	0.009
WOMAC Total	−0.82	0.150	−5.49	<0.001
MCS (Mild)	−1.06	0.616	−1.72	0.086
MCS (Mod/Severe)	1.27	0.879	1.44	0.149

	Coefficient	S.E.	Wald Z	p value
Pre-surgical Day	−0.00	0.002	−0.54	0.591

OAI, Osteoarthritis Initiative; SE, standard error; BMI, body mass index; K&L, Kellgren and Lawrence knee osteoarthritis grade; WOMAC, Western Ontario and McMaster Universities Arthritis Index, MCS, mental component score of the Short Form Health Survey; Pre-surgical Day, Number of days measurement obtain prior to surgery.

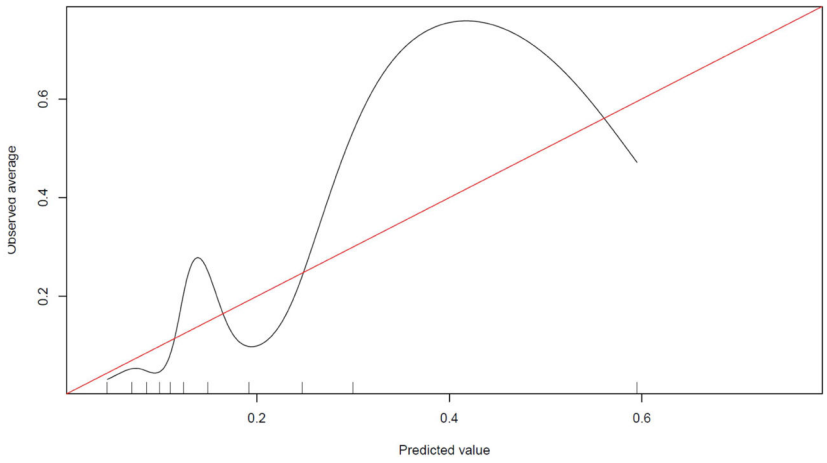
Post hoc analysis 2

Multiple Logistic regression coefficients for primary outcome including only participants with baseline assessment within six months of surgery (n=231).

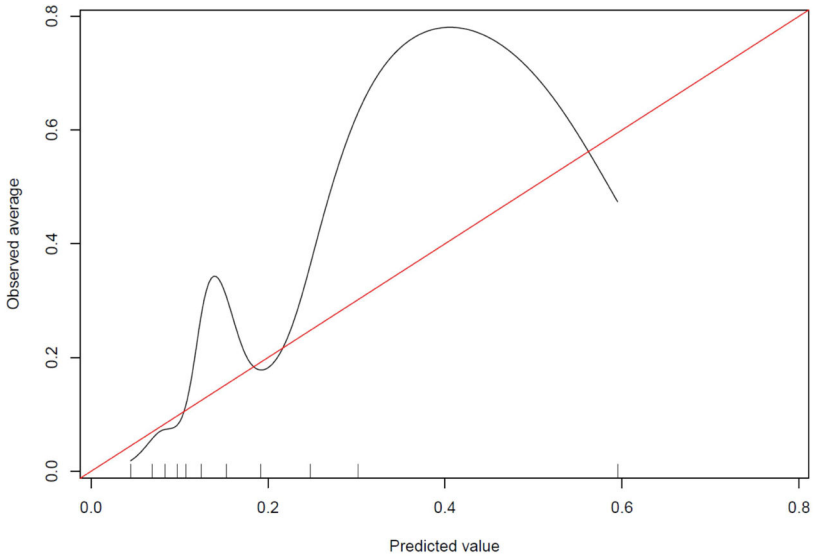
	Coefficient	S.E.	Wald Z	p value
Intercept	0.57	0.450	1.27	0.203
BMI ≥40	0.05	0.910	0.05	0.959
K&L Grade ≤3	0.98	0.361	2.72	0.007
WOMAC Total	−0.83	0.148	−5.62	<0.001
MCS (Mild)	−1.05	0.617	−1.71	0.089
MCS (Mod/Severe)	1.24	0.875	1.41	0.158

OAI, Osteoarthritis Initiative; SE, standard error; BMI, body mass index; K&L, Kellgren and Lawrence knee osteoarthritis grade; WOMAC, Western Ontario and McMaster Universities Arthritis Index, MCS, mental component score of the Short Form Health Survey.

Post hoc analysis 3



Calibration curve analysis with unicompartmental KAs removed from the dataset and using Model #1.



Calibration curve analysis with unicompartmental KAs removed from the dataset and using Model #2.

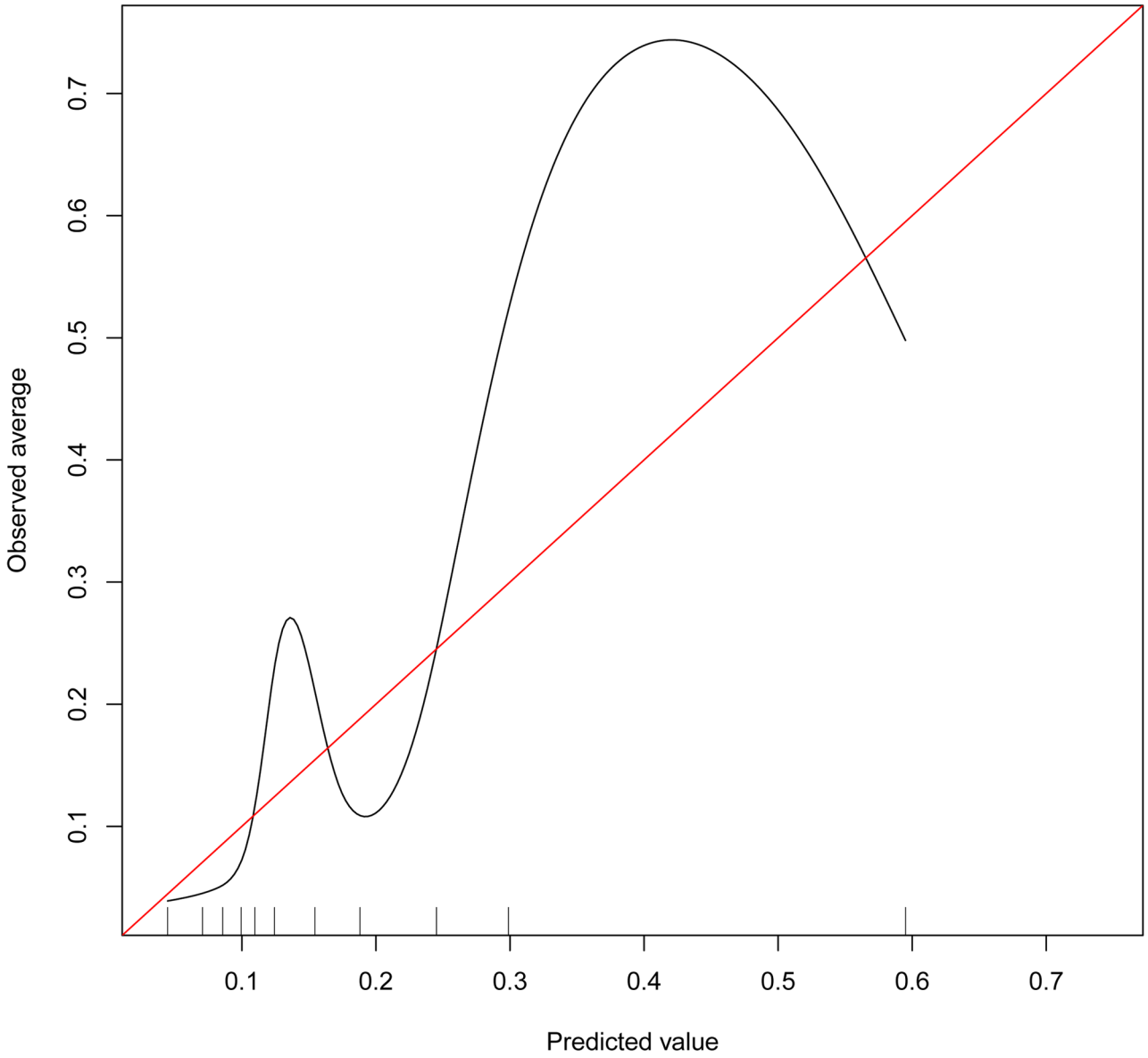


Figure 1.

The external validation curve is illustrated. The observed probability of nonresponse is presented on the y-axis and the regression-predicted probability of non-response is presented on the x-axis. The diagonal line indicates perfect agreement between the observed and predicted probability of nonresponse.

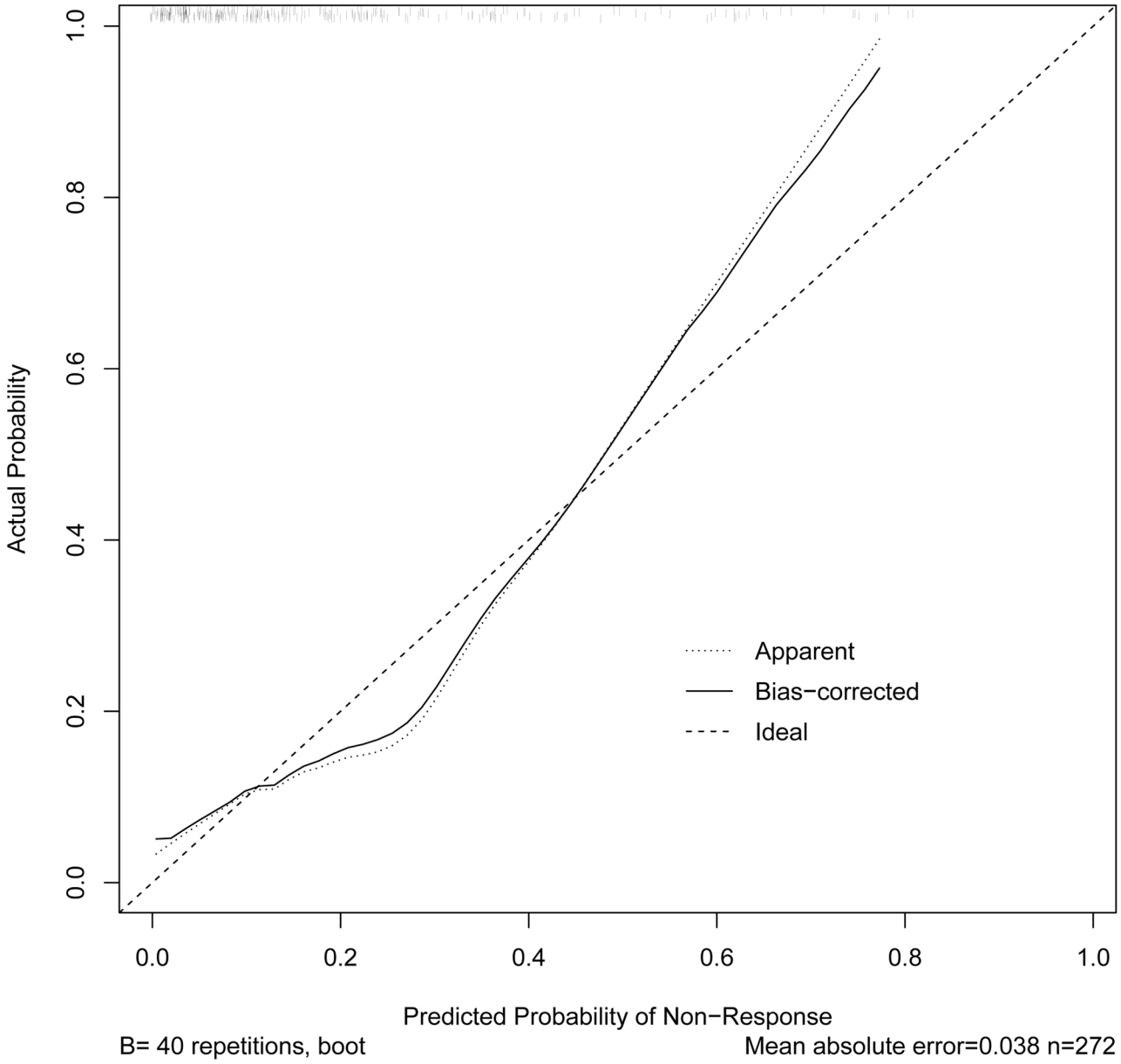


Figure 2.

The OAI data were used to generate the calibration curve. The observed probability of nonresponse is presented on the y-axis and the regression-predicted probability of non-response is presented on the x-axis. The diagonal line indicates perfect agreement between the observed and predicted probability of nonresponse. Deviation from perfect agreement occurs at beginning, middle and end of the curve.

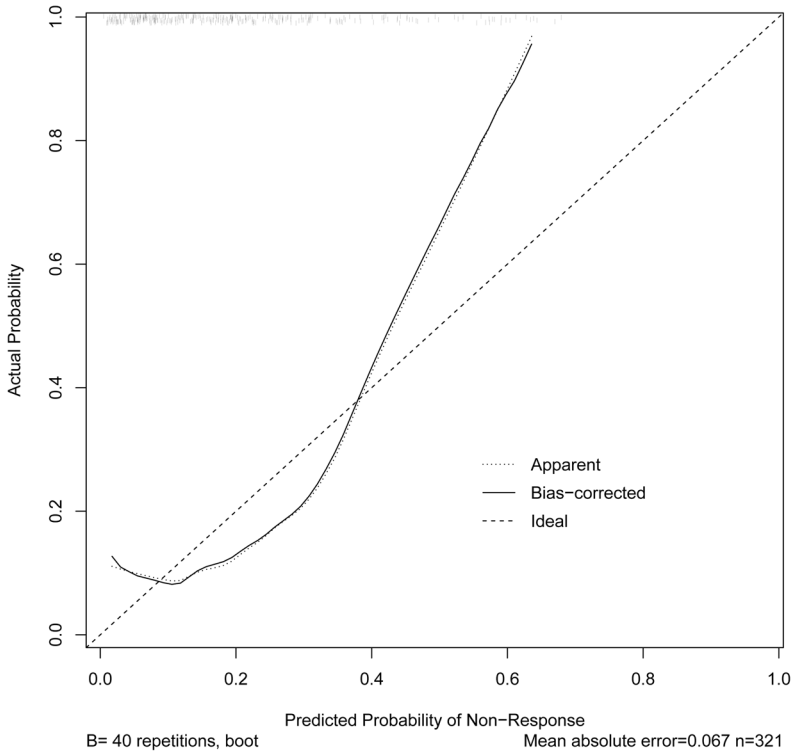


Figure 3.

The preoperative WOMAC Total scores from the OAI data were used to generate the calibration curve. The observed probability of nonresponse is presented on the y-axis and the regression-predicted probability of non-response is presented on the x-axis. The diagonal line indicates perfect agreement between the observed and predicted probability of nonresponse. A more substantial deviation from perfect agreement as compared to Figure 2.

Table 1

Pre-operative Characteristics of the Sample Stratified by Responder Status *

Variable	Overall Sample (n = 427) [missing]	Responder Method #1		Responder Method #2	
		Responders (n = 263)	Non-responders (n = 63)	Responders (n = 264)	Non-responders (n = 85)
Female, n (%)	258 (60.4%) [0]	159 (60.5%)	39 (61.9%)	143 (58.9%)	36 (62.1%)
Age, mean years (sd)	68.24 (8.4) [0]	68.3 (8.3)	67.0 (9.0)	68.4 (8.4)	67.3 (9.2)
Obesity BMI category, n (%)	[10]				
<30 kg/m^2	206 (49.4%)	133 (51.0%)	29 (47.5%)	126 (52.3%)	25 (44.6%)
≥30 to <35 kg/m^2	144 (34.5%)	85 (32.6%)	25 (41.0%)	74 (30.7%)	24 (42.9%)
≥35 to <40 kg/m^2	50 (12.0%)	33 (12.6%)	5 (8.2%)	31 (12.9%)	5 (8.9%)
≥40 kg/m^2	17 (4.0%)	10 (3.8%)	2 (3.3%)	10 (4.1%)	2 (3.6%)
Charlson Comorbidity, n (%)	[26]				
≤1	347 (86.5%)	218 (85.5%	57 (91.9%)	222 (86.4%)	79 (94.0%)
≥2	54 (13.5%)	37 (14.5%)	5 (8.1%)	35 (13.6)	5 (6.0%)
WOMAC Total, mean (SD)	35.4 (16.4) [29]	38.2 (14.2)	22.9 (18.5)	37.9 (14.3)	21.3 (17.2)
SF-12 MCS, mean (SD)	55.9 (8.6) [1]	55.6 (8.6)	56.2 (7.7)	55.4 (8.7)	56.6 (7.7)
K&L Grade, n(%)	[86]				
≤3	133 (39.0%)	77 (34.5%)	32 (61.5%)	71 (34.3%)	28 (58.3%)
4	208 (61.0%)	146 (65.5%)	20 (38.5%)	136 (65.7%)	20 (41.7%)
Days from pre-op visit to surgery (SD)	176 (103) [18]	162 (96)	194 (88)	159 (101)	191 (94)

* Responder status is presented for Responder Method #1 which used outcomes measured between 365 days to 730 days following surgery and Responder Method #2 which used outcomes measured between 180 days and 540 days following surgery.

Table 2

Multiple Logistic regression coefficients for four key variables for predicting nonresponse from Dowsey and colleagues and from the OAI data

Variable	Data from Dowsey et al			Data from OAI		
	Coefficient (se)	Odds Ratio	P value	Coefficient (se)	Odds Ratio	P value
Intercept	−1.53 (0.50)		0.002	0.57 (0.45)		0.20
BMI ≥40	1.25 (0.29)	3.48	<0.001	0.04 (0.91)	1.04	0.95
K&L Grade ≤3	0.95 (0.25)	2.59	<0.001	0.98 (0.36)	2.66	0.007
WOMAC Total	−0.21 (0.09)	0.81	0.02	−0.83 (0.15)	0.44	<0.001
MCS (Mild)	−0.08 (0.35)	0.93	0.83	−1.05 (0.62)	0.35	0.09
MCS (Moderate)	0.55 (0.31)	1.74	0.075	1.24 (0.87)	3.46	0.15
MCS (Severe)	1.19 (0.42)	3.30	0.005	-	-	-

OAI, Osteoarthritis Initiative; SE, standard error; BMI, body mass index; K&L, Kellgren and Lawrence knee osteoarthritis grade; WOMAC, Western Ontario and McMaster Universities Arthritis Index, MCS, mental component score of the Short Form Health Survey.