Patient-reported outcomes 3 months after spine surgery: is it an accurate predictor of 12-month outcome in real-world registry platforms?

Scott L. Parker, MD,1 Anthony L. Asher, MD,2 Saniya S. Godil, MD,1 Clinton J. Devin, MD,3 and Matthew J. McGirt, MD2

Departments of 1Neurological Surgery and 2Orthopaedic Surgery, Vanderbilt University Medical Center, Nashville, Tennessee; and 3Carolina Neurosurgery & Spine Associates, Charlotte, North Carolina

OBJECT The health care landscape is rapidly shifting to incentivize quality of care rather than quantity of care. Quality and outcomes registry platforms lie at the center of all emerging evidence-driven reform models and will be used to inform decision makers in health care delivery. Obtaining real-world registry outcomes data from patients 12 months after spine surgery remains a challenge. The authors set out to determine whether 3-month patient-reported outcomes accurately predict 12-month outcomes and, hence, whether 3-month measurement systems suffice to identify effective versus noneffective spine care.

METHODS All patients undergoing lumbar spine surgery for degenerative disease at a single medical institution over a 2-year period were enrolled in a prospective longitudinal registry. Patient-reported outcome instruments (numeric rating scale [NRS], Oswestry Disability Index [ODI], 12-Item Short Form Health Survey [SF-12], EQ-5D, and the Zung Self-Rating Depression Scale) were recorded prospectively at baseline and at 3 months and 12 months after surgery. Linear regression was performed to determine the independent association of 3- and 12-month outcome. Receiver operating characteristic (ROC) curve analysis was performed to determine whether improvement in general health state (EQ-5D) and disability (ODI) at 3 months accurately predicted improvement and achievement of minimum clinical important difference (MCID) at 12 months.

RESULTS A total of 593 patients undergoing elective lumbar surgery were included in the study. There was a significant correlation between 3-month and 12-month EQ-5D (r = 0.71; p < 0.0001) and ODI (r = 0.70; p < 0.0001); however, the authors observed a sizable discrepancy in achievement of a clinically significant improvement (MCID) threshold at 3 versus 12 months on an individual patient level. For postoperative disability (ODI), 11.5% of patients who achieved an MCID threshold at 3 months dropped below this threshold at 12 months; 10.5% of patients who did not meet the MCID threshold at 3 months continued to improve and ultimately surpassed the MCID threshold at 12 months. For ODI, achieving MCID at 3 months accurately predicted 12-month MCID with only 62.6% specificity and 86.8% sensitivity. For postoperative health utility (EQ-5D), 8.5% of patients lost an MCID threshold improvement from 3 months to 12 months, while 4.0% gained the MCID threshold between 3 and 12 months postoperatively. For EQ-5D (quality-adjusted life years), achieving MCID at 3 months accurately predicted 12-month MCID with only 87.7% specificity and 87.2% sensitivity.

CONCLUSIONS In a prospective registry, patient-reported measures of treatment effectiveness obtained at 3 months correlated with 12-month measures overall in aggregate, but did not reliably predict 12-month outcome at the patient level. Many patients who do not benefit from surgery by 3 months do so by 12 months, and, conversely, many patients reporting meaningful improvement by 3 months report loss of benefit at 12 months. Prospective longitudinal spine outcomes registries need to span at least 12 months to identify effective versus noneffective patient care.

http://thejns.org/doi/abs/10.3171/2015.9.FOCUS15356

KEY WORDS registry science; outcomes; spine

ABBREVIATIONS ASA = American Society of Anesthesiologists; MCID = minimum clinical important difference; ODI = Oswestry Disability Index; PRO = patient-reported outcome; ROC = receiver operating characteristic; SF-12 = 12-Item Short Form Health Survey; VSR = Vanderbilt Spine Registry.


INCLUDE WHEN CITING DOI: 10.3171/2015.9.FOCUS15356.
THE current growth rate of health care costs in the United States is simply not sustainable, and therefore the ability to define real-world effectiveness of health care delivery is of vital importance. To bend the cost curve and ultimately achieve sustainability in health care, medical providers and surgical treatments of the highest quality and effectiveness must be preferentially used and purchased. The Institute of Medicine defines health care quality as “The degree to which health care services for individuals and populations increase the likelihood of desired health outcomes.” Safety and effectiveness (quality), patient-centered outcomes, as well as cost of care are each central components of the health care value equation (benefit/cost).

The Institute of Medicine, Agency for Healthcare Research and Quality, and the Patient Protection and Affordable Care Act call for prospective, well-designed registries for the measurement of real-world effectiveness to facilitate evidence-based health care reform. Prospective registries are able to provide the necessary statistical power and real-world health delivery setting required for true value measurement in individuals and populations.

Value-based purchasing, in which providers are held accountable for both the quality and cost of health care services, will soon dominate most incentive models in an increasingly resource-restricted medical economy. As such, prospective registries that measure both safety and effectiveness of care have been deemed the preferred method to identify health care waste and low value care, which is instrumental in the attempt to define value. However, such large-scale registry solutions create economic burdens of their own, and evolving registry science is needed to make it a practical and economically sustainable way to empower value-based health care reform. In particular, longitudinal patient follow-up is the most costly and challenging aspect of patient-reported outcome registries. Understanding the utility of long-term follow-up is of vital importance in the evolution of registry science and its ability to power evidence-based economic reform. If longitudinal measurement of patient benefit beyond 3 months is not mandatory to discriminate between effective and ineffective care for patients undergoing lumbar spine surgery, then significant economic savings may be realized as registries begin to power value-based measurement and reform. Accordingly, in the current study we set out to determine whether 3-month patient-reported outcomes (PROs) accurately predict 12-month outcomes in patients undergoing elective lumbar spine surgery; and hence, whether 3-month measurement systems suffice to identify effective versus noneffective spine care.

Methods

All patients undergoing elective spinal surgery at our institution are enrolled into the Vanderbilt Spine Registry (VSR). The VSR is composed of data from a multidisciplinary spine care center composed of providers from the Departments of Neurosurgery, Orthopaedic Surgery, Physical Medicine & Rehabilitation, and Pain Management. Patient demographics, clinical presentation, indications for surgery, radiological studies, and operative variables are assessed prospectively for each case. Baseline, 3-month, and yearly postoperative pain, disability, and quality of life are assessed via phone interview by an independent investigator not involved with clinical care. Validated questionnaires are used to collect data on the following outcome measures: 1) pain—numeric rating scale for back, neck, arm, and leg pain;2,15 2) disease-specific physical disability—Oswestry Disability Index (ODI)6,7,28,29 3) mood/emotional disability—Zung Self-Rating Depression Scale;26,11 4) generic health-related physical and mental quality of life—Short Form-12 (SF-12) Physical Component Summary, and Mental Component Summary;9,20 and 5) preference-based health status—EQ-SD,3,13,14 All data are collected and managed using REDCap (Research Electronic Data Capture) web-based data capture tools hosted at Vanderbilt University.12

Patient Selection

For the purposes of this study, all patients undergoing elective lumbar surgery for degenerative spine disease over a 2-year period were assessed. A total of 593 patients undergoing spinal surgery for lumbar disc herniation, stenosis, spondylolisthesis, or failed–back surgery syndrome were included in this analysis.

Statistical Analysis

All analyses were performed using STATA version 12.0 (StataCorp). Summary data were presented as means ± SD for parametric data and as medians with interquartile range for nonparametric data.

Linear regression was performed to adjust for patient, disease, and radiographic variance (race, smoking, employment status, revision surgery, number of levels treated, American Society of Anesthesiologists [ASA] Physical Status Classification System class, hypertension, arthritis, diabetes, estimated blood loss, length of surgery, length of hospital stay, 90-day morbidity, and preoperative ODI and EQ-5D score) to determine the independent association of 3- and 12-month outcomes. The first model was run using EQ-5D as a dependent variable with race, smoking, employment status, revision surgery, number of levels, ASA class, hypertension, arthritis, diabetes, estimated blood loss, length of surgery, length of hospital stay, 90-day morbidity, and preoperative ODI and EQ-5D score as independent variables. A second linear regression model was run using ODI as a dependent variable and race, smoking, employment status, revision surgery, number of levels treated, ASA class, hypertension, arthritis, diabetes, estimated blood loss, length of surgery, length of hospital stay, 90-day morbidity, and preoperative ODI and EQ-5D score as independent variables.

Receiver operating characteristic (ROC) curve analysis was performed to determine whether improvement in general health state (EQ-5D) and disability (ODI) at 3 months accurately predicted improvement and achievement of minimum clinical important difference (MCID) at 12 months. MCID thresholds were derived from previously published values.20 The concept of MCID has been introduced as a measure for the critical threshold needed to achieve treatment effectiveness. Using this measure,
treatment effects reaching the MCID threshold value imply clinical significance and justification for implementation into clinical practice. In other words, MCID can be thought of as the smallest change in outcome measure that is important to patients. In determining MCID, anchor-based approaches are the most accepted and often used calculation method. These approaches allow for the comparison of the change in PRO scores following surgery to another external anchor measurement, such as patient-perceived improvement following surgery or patient satisfaction with surgery. The literature contains multiple anchor-based calculation methods, and there are several studies that have established MCID thresholds for different spine surgeries.19–23

The ROC curve is a quantitative method of measuring the accuracy of a test and compares the diagnostic test with a perfect test (100% sensitivity and specificity equals an AUC of 1.0).11,32 In our analysis, the ROC curve (sensitivity and specificity) was used to determine whether 3-month patient-reported outcomes were able to accurately predict 12-month outcomes in patients undergoing elective lumbar spine surgery. ROC curves were generated using adjusted EQ-5D and ODI scores. In ROC analysis for this study, the change score of each PRO at 3 months (EQ-5D and ODI) represented the diagnostic test and “achievement of MCID at 12-months” was taken as the “gold standard” or external criterion.

Results
Patient Population
A total of 593 patients undergoing elective lumbar surgery were included in the study. The mean age ± SD of the cohort was 57.3 ± 14.0 years with 271 (45.7%) males. One hundred fifty-eight (26.6%) patients had leg dominant symptoms, 225 (39.3%) patients were employed preoperatively, 330 (55.6%) underwent lumbar fusion surgery, and 200 (36.8%) underwent revision surgery. One hundred twenty (21.7%) patients were smokers, 123 (22.7%) had diabetes, 304 (55.9%) had hypertension, 353 (64.9%) had arthritis, and 91 (16.7%) had coronary artery disease. The mean body mass index of the cohort was 30.2 ± 7.2.

Correlation Between 3-Month Versus 12-Month Outcomes
There was a significant correlation between 3-month and 12-month EQ-5D (r = 0.71; p < 0.0001) and ODI (r = 0.70; p < 0.0001). After adjusting for baseline patient, disease, and radiographic variance via linear regression, 3-month EQ-5D and ODI scores remained independently associated with 12-month scores (p < 0.001; adjusted R² = 0.45 and 0.59, respectively).

3-Month Versus 12-Month ODI Scores
With regard to improvement in disability (ODI), only 430 (78.0%) patients remained concordant in achieving (or not achieving) the MCID threshold at 3 months and 12 months postoperatively. Sixty-three (11.5%) patients who achieved the MCID threshold at 3 months subsequently worsened and reported not achieving MCID at 12 months. Additionally, 58 (10.5%) patients who did not achieve MCID at 3 months continued to improve and achieved MCID by 12 months (Fig. 1). A scattergram comparing 3- and 12-month disability scores (ODI) demonstrated a substantial extent of variation in ODI score at these 2 follow-up times (movement away from the reference line denoting no change from 3 months to 12 months postoperatively) (Fig. 2). Therefore, for ODI, achieving MCID at 3 months accurately predicted 12-month MCID with only 62.6% specificity and 86.8% sensitivity for lumbar surgery (Fig. 3).

3-Month Versus 12-Month EQ-5D (quality-adjusted life year) Scores
With regard to improvement in health utility (EQ-5D), 492 (87.5%) patients remained concordant in achieving (or not achieving) the MCID threshold at 3 months and 12 months postoperatively. However, 48 (8.5%) patients who achieved the MCID threshold at 3 months subsequently worsened and lost MCID at 12 months. Additionally, 22 (4.0%) patients who did not achieve MCID at 3 months continued to improve and achieved MCID by 12 months (Fig. 1). A scattergram comparing 3- and 12-month health utility scores (EQ-5D) demonstrated a substantial extent of variation in EQ-5D score at these 2 follow-up times (movement away from the reference line denoting no change from 3 months to 12 months postoperatively [Fig. 2]). Therefore, for EQ-5D (quality-adjusted life years), achieving MCID at 3 months accurately predicted 12-month MCID with only 87.7% specificity and 87.2% sensitivity for patients undergoing elective lumbar spinal procedures (Fig. 3).

Subset Analyses
One hundred fifty-eight (26.6%) patients had leg dominant symptoms, 225 (39.3%) patients were employed preoperatively, 330 (55.6%) underwent fusion, 200 (36.8%) underwent revision surgery, and 111 (18.7%) of the patients were depressed preoperatively. Subset analysis was performed individually on each of these subcohorts. Three-month outcome remained a poorly accurate (< 85% specific/sensitive) predictor for 12-month outcome for each of these patient subsets (Fig. 4).

Discussion
In the current study, we set out to understand whether PRO metrics assessed at 3 months postoperatively were accurately predictive of outcomes at 12 months postoperatively. Although there was a significant correlation between absolute outcome scores at 3 and 12 months postoperatively in aggregate via linear regression analysis, we observed a wide variation in the extent of change between 3 and 12 months after surgery. Nearly as many patients demonstrated worsening PROs as reported improving PROs between 3 and 12 months. Furthermore, the extent of 3- to 12-month improvement or worsening also varied widely among patients. Even at its peak reliability in ROC analysis, the 3-month PRO score remained with greater than 10% inaccuracy. Whether one defines surgical treatment response as 12-month MCID improvement or any degree of PRO improvement at 12 months, 3-month scores did not reliably identify treatment responders from nonre-
Patients achieving MCID at both 3 months and 12 months
Patients achieving MCID at 3 months but reporting loss of benefit at 12 months
Patients who fail to achieve MCID at 3 months, but continue to improve and achieve MCID at 12 months
Patients who fail to achieve MCID at any time point

FIG. 1. Scatter plots showing achievement of a clinically significant improvement (MCID) at 3 and 12 months postoperatively in patients undergoing elective lumbar spine surgery. We observed a sizable discrepancy in achievement of MCID threshold at 3 versus 12 months on an individual patient level. For both disability (ODI) and health utility (EQ-5D), many patients who failed to reach the MCID threshold by 3 months continued to improve and surpassed the MCID threshold at 12 months (green dots). There were also a sizable number of patients that achieved the MCID threshold at 3 months but subsequently dropped below this threshold improvement at 12 months (blue dots).

FIG. 2. Scatter plots showing variation in 3-month and 12-month outcomes in patients undergoing elective lumbar spine surgery. The dotted line represents no difference between 3-month and 12-month outcomes. There was substantial movement away from the line in both directions. Nearly the same proportions of patients continue to improve between time points as worsen between time points.
sponders. In the assessment of postoperative improvement in disability (ODI), 11.5% of patients who achieved an MCID threshold at 3 months subsequently dropped below this threshold improvement at 12 months. Furthermore, 10.5% of patients who did not meet an MCID threshold at 3 months continued to improve and ultimately surpassed the MCID threshold at 12 months. This degree of patient crossover in clinically significant improvement for disability (22.0% crossover) and health utility (12.5% crossover) suggests that a patient’s status at 3 months cannot reliably predict their outcome status at 12 months following elective lumbar spinal procedures; and therefore, 3-month outcomes do not appear to be a valid proxy for determining the long-term effectiveness of lumbar spine surgery.

Historically, in the literature, 2-year follow-up has been considered an appropriate long-term follow-up to assess outcomes of spine surgery. However, taking the feasibility of collecting data prospectively, there have been many studies in the literature to support that 12-month data are adequate to report long-term outcomes in this patient population. Therefore, we compared 3-month (short-term outcome) and 12-month outcomes (long-term outcome) in this study. But, there is lack of evidence to establish whether 12 months or 24 months is the gold standard for reporting long-term outcomes in patient populations undergoing spine surgery.

Prospective registries have emerged as a feasible means to rapidly and efficiently capture vast amounts of clinical data pertaining to real-world care across large patient populations with the statistical power necessary to assess true value measurement at both the individual and population level. With the proper design and implementation rigor, registries provide the potential to create reliable and expeditious access to outcomes data such as patient satisfaction, pain, disability, quality of life, and cost-effectiveness. National and international spine registries have been created to assess both outcome and economic impacts of health care delivery. While not an alternative to a randomized controlled trial, prospectively organized registries are more feasible and closer to daily clinical situations. Furthermore, prospective registries measure quality and effectiveness on real-world patient populations in real-world health care delivery settings that are not limited, constrained, and ultimately biased by the inherent artificial research setting and structure criteria of a clinical trial. It is for these reasons that registries may represent the “next disruptive technology” in clinical research.

Critical to the determination and improvement of the effectiveness and efficiency of health care delivery is the acquisition of outcome data that are both valid and reliable. Such data can be used by providers to guide care and clinical decisions in real time and also examine the outcomes of care thereafter to identify opportunities for improvement. Within prospective registry data collection systems, PRO instruments are increasingly being used to evaluate effectiveness of care. Secondary to the inherent costs of such data collection platforms, more robust and cumbersome outcome metrics must be balanced with what is feasible and scalable in larger registry efforts. Therefore, the use of outcome instruments that are most simple and feasible, while not compromising accuracy or validity, is preferred.

Another principal aspect of large-scale registry efforts that consumes a sizable amount of time and resources is longitudinal long-term follow-up. For the VSC registry, the cost per patient has been calculated to be approximately $160 for 1-year follow-up. Hospital systems and surgeon groups with limited resources may initially face challenges with participating in high-quality longitudinal patient-reported outcome registries. For national or regional registries tracking tens of thousands of patients, an additional or nonessential patient contact point could cost several million dollars of potential savings. It is these very factors that provided the impetus for the current study. Reduced follow-up times could dramatically lower the operation costs of large-scale registries; however, there must be consistent evidence that shorter-term follow-up can reliably and accurately predict long-term outcomes at an individual patient level. Regardless, registries remain supe-
prior to many controlled trial designs from a cost efficiency perspective. To provide perspective, the cost per enrolled patient in randomized controlled trials can exceed $6000.\textsuperscript{5}

The results of the current study suggest that outcome assessments at 3 months do not provide a reliable indicator for a patient’s outcome at 12 months following elective lumbar spine surgery. This is evident by the sizeable degree of patient crossover of the MCID threshold between 3 and 12 months postoperatively. We observed a 22.0% crossover incidence in MCID threshold for disability (ODI) and 12.5% crossover incidence in health utility (EQ-5D) in this patient population. Subset analysis revealed a similar degree of discordance, and 3-month outcome remained a poorly accurate (< 85% specific/sensitive) predictor for 12-month outcome for each of the patient subcohorts analyzed. This degree of discordance in clinically significant improvement demonstrates that 3-month outcomes do not provide a valid assessment of a patient’s long-term outcome status or the long-term effectiveness of lumbar spine surgery. As such, longitudinal follow-up to at least 1-year postoperatively should be performed to accurately evaluate long-term outcomes in patients undergoing elective spine surgery for lumbar degenerative diseases.

The findings of this study should be interpreted in the light of inherent study limitations. One of the main limitations is that it is a single-center study, and therefore results from one institutional registry cannot be extrapolated across the country. Additionally, the study findings can be a factor of small sample size as we cannot perform multiple subset analyses to determine predictive factors leading to these results. Therefore, nationwide prospective registries like N2QOD can provide similar analysis in the future to establish a lack of correlation between 3- and 12-month outcomes.

Conclusions

In a prospective, longitudinal registry, patient-centered measures of treatment effectiveness obtained at 3 months did not reliably predict outcomes 12 months after elective lumbar spine surgery. Many patients who did not benefit from surgery by 3 months did so by 12 months; conversely, many patients reporting meaningful improvement by 3 months go on to lose this benefit at 12 months. Even with the associated additional expenditures required to provide accurate and reliable data that can be used to drive evi-
dence-based health care reform, prospective longitudinal spine outcomes registries should span at least 12 months to identify effective versus noneffective patient care.

References

Disclosures
The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions
Conception and design: Devin, Parker, Asher, McGirt. Acquisition of data: Devin, Parker, Godil. Analysis and interpretation of data: Devin, Parker, Asher, McGirt. Drafting the article: Parker. Critically revising the article: Devin, Asher, McGirt. Reviewed submitted version of manuscript: McGirt. Statistical analysis: Godil.

Correspondence
Clinton J. Devin, Department of Orthopaedic Surgery, Vanderbilt University Medical Center, 1215 21st Ave. S, Nashville, TN 37232. email: clinton.j.devin@vanderbilt.edu.