Use of Computerized Adaptive Testing to Develop More Concise Patient-Reported Outcome Measures

Liam T. Kane, BS, Surena Namdari, MD, MSc, Otho R. Plummer, PhD, Pedro Beredjiklian, MD, Alexander Vaccaro, MD, PhD, MBA, and Joseph A. Abboud, MD

Investigation performed at the Rothman Orthopaedic Institute, Philadelphia, Pennsylvania

Background: Patient-reported outcome measures (PROMs) are essential tools that are used to assess health status and treatment outcomes in orthopaedic care. Use of PROMs can burden patients with lengthy and cumbersome questionnaires. Predictive models using machine learning known as computerized adaptive testing (CAT) offer a potential solution. The purpose of this study was to evaluate the ability of CAT to improve efficiency of the Veterans RAND 12 Item Health Survey (VR-12) by decreasing the question burden while maintaining the accuracy of the outcome score.

Methods: A previously developed CAT model was applied to the responses of 19,523 patients who had completed a full VR-12 survey while presenting to 1 of 5 subspecialty orthopaedic clinics. This resulted in the calculation of both a full-survey and CAT-model physical component summary score (PCS) and mental component summary score (MCS). Several analyses compared the accuracy of the CAT model scores with that of the full scores by comparing the means and standard deviations, calculating a Pearson correlation coefficient and intraclass correlation coefficient, plotting the frequency distributions of the 2 score sets and the score differences, and performing a Bland-Altman assessment of scoring patterns.

Results: The CAT model required 4 fewer questions to be answered by each subject (33% decrease in question burden). The mean PCS was 1.3 points lower in the CAT model than with the full VR-12 (41.5 ± 11.0 versus 42.8 ± 10.4), and the mean MCS was 0.3 point higher (57.3 ± 9.4 versus 57.0 ± 9.6). The Pearson correlation coefficients were 0.97 for PCS and 0.98 for MCS, and the intraclass correlation coefficients were 0.96 and 0.97, respectively. The frequency distribution of the CAT and full scores showed significant overlap for both the PCS and the MCS. The difference between the CAT and full scores was less than the minimum clinically important difference (MCID) in >95% of cases for the PCS and MCS.

Conclusions: The application of CAT to the VR-12 survey demonstrated an ability to lessen the response burden for patients with a negligible effect on score integrity.

In modern orthopaedic medicine, patients are burdened by the administration of several questionnaires that are designed as data collection tools to obtain patient-reported outcome measures (PROMs). PROMs have several functions, including to aid research by assigning an overall function score, to evaluate the value of care through objective outcomes, and to direct treatment and reimbursement rates. While the benefits of PROMs to the medical and scientific community are clear, there is demand from both the patient and the physician standpoint to develop more efficient PROMs that improve compliance while maintaining the integrity of the outcome score.

The Veterans RAND 12 Item Health Survey (VR-12) is an example of a widely used PROM that calculates physical and mental health outcome scores for patients receiving orthopaedic care (Table I). The VR-12 was developed from the Veterans RAND 36 Item Health Survey (VR-36) with use of extensive research to identify the 12 most important questions with the greatest influence on scoring variability. Currently, it is one of the most popularly used general outcome measures, with applications across several orthopaedic subspecialties, and has been used to characterize subjects in numerous population-based studies. As a result, reducing the question burden of this
| Question                                                                 | Response Options                                      |
|------------------------------------------------------------------------|-------------------------------------------------------|
| **1. In general, would say your health is:**                           | Excellent  
Very good  
Good  
Fair  
Poor |
| **2. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?** | Yes, limited a lot  
Yes, limited a little  
No, not limited at all  
Yes, limited a lot  
Yes, limited a little  
No, not limited at all |
| a. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf. |                                                      |
| b. Climbing several flights of stairs.                                  |                                                      |
| **3. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?** | No, none of the time  
Yes, a little of the time  
Yes, some of the time  
Yes, most of the time  
Yes, all of the time  
No, none of the time  
Yes, a little of the time  
Yes, some of the time  
Yes, most of the time  
Yes, all of the time |
| a. **Accomplished less** than you would like.                           |                                                      |
| b. Were limited in the **kind** of work or other activities.           |                                                      |
| **4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems** (such as feeling depressed or anxious)? | No, none of the time  
Yes, a little of the time  
Yes, some of the time  
Yes, most of the time  
Yes, all of the time  
No, none of the time  
Yes, a little of the time  
Yes, some of the time  
Yes, most of the time  
Yes, all of the time |
| a. **Accomplished less** than you would like.                           |                                                      |
| b. Didn’t do work or other activities as **carefully** as usual.        |                                                      |
| **5. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?** | A little bit  
Moderately  
Quite a bit  
Extremely |

These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling.
particular PROM could have a widespread impact on streamlining patient care.

Advances in data science have shown that the score of an outcome measure can be accurately predicted from fewer questions if the correct questions are asked. Predictive models, developed through a process known as computerized adaptive testing (CAT), offer a potential solution. The goal of CAT is to identify the correct subset of questions selected from the full questionnaire to ask each patient on the basis of his/her previous responses. CAT is trained through so-called machine learning programs, also described as artificial intelligence, that analyze how response patterns affect overall outcome scores. The CAT model then uses its own recognition of these patterns to self-improve its efficiency and minimize question burden in an accurate manner. Technology for this purpose has been successfully developed and applied in other fields, demonstrating potential to effectively improve the patient experience.20

Prior to real-time use, CAT models specific to each PROM must be validated by comparing the accuracy of scores generated using fewer questions with that of scores generated using the full questionnaires. A CAT version of the VR-12 was recently developed within the OBERD software system (Universal Research Solutions, www.oberd.com), a general tool used for outcome data collection. A CAT model developed for a specialty-specific PROM using the same software system was recently validated, but its application for other PROMs remains undetermined. The purpose of this study was to evaluate the success of this CAT model in improving VR-12 efficiency by (1) decreasing the question burden of responders and (2) maintaining accuracy of outcome scores.

Materials and Methods

This was a retrospective, single-institution analysis evaluating 19,523 patients presenting to 5 different orthopaedic clinics across 4 different subspecialties (shoulder and elbow, hand, sports medicine, and spine). Subjects with a variety of ages and diagnoses were included in the analysis (Fig. 1, Table II). Data were collected with OBERD, which we have used for several years to collect outcome data from patients. Through

---

### Table 1 (continued)

| Question                                                                                                                                                                                                 | Response Options                  |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| **6. How much of the time during the past 4 weeks:**                                                                                                                                                     |                                   |
| a. Have you felt calm and peaceful?                                                                                                                                                                      | All of the time                    |
|                                                                                            | Most of the time                   |
|                                                                                            | A good bit of the time             |
|                                                                                            | Some of the time                   |
|                                                                                            | A little of the time               |
|                                                                                            | None of the time                   |
| b. Did you have a lot of energy?                                                                                                                                                                         | All of the time                    |
|                                                                                            | Most of the time                   |
|                                                                                            | A good bit of the time             |
|                                                                                            | Some of the time                   |
|                                                                                            | A little of the time               |
|                                                                                            | None of the time                   |
| c. Have you felt downhearted and blue?                                                                                                                                                                   | All of the time                    |
|                                                                                            | Most of the time                   |
|                                                                                            | A good bit of the time             |
|                                                                                            | Some of the time                   |
|                                                                                            | A little of the time               |
|                                                                                            | None of the time                   |
| 7. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)? | All of the time                    |
|                                                                                            | Most of the time                   |
|                                                                                            | Some of the time                   |
|                                                                                            | A little of the time               |
|                                                                                            | None of the time                   |

*The Veterans RAND 12 Item Health Survey was developed from the Veterans RAND 36 Item Health Survey, which was developed and modified from the original RAND version of the 36-item Health Survey version 1.0 (also known as the “MOS SF-36”); “VR-12: How to create VR-12 scales and PCS/MCS summaries” © 2014 by Trustees of Boston University. All rights reserved. (All questions should be directed to Professor Lewis Kazis, Boston University School of Public Health. E-mail: lek@bu.edu.)
this process, the database has provided extensive resources for training and evaluating predictive models. During the initial visit to our orthopaedic surgery clinics, each patient completed a full VR-12 survey using OBERD by means of a tablet device (iPad; Apple), resulting in the baseline calculation of both a physical component summary score (PCS) and a mental component summary score (MCS). This CAT model developed for the VR-12 was trained using a random sample of 27,800 de-identified administrations of the VR-12 collected routinely by users of the software. The algorithms constructed by the CAT system through this training were then retrospectively applied to each set of patient responses stored on the instrument (i.e., not live while the patients were answering the survey). The 19,523 responses were not part of the original training set that helped develop the algorithms. Beginning with the first VR-12 item, the CAT takes each set of responses through a range of VR-12 items guided by previous answers. This resulted in the calculation of a CAT-specific score of both the PCS and the MCS for each individual subject.

The decrease in question burden was measured by assessing the percentage difference in the number of questions between the CAT model and the full survey. Several statistical methods were used to compare the accuracy of the CAT scores with the full scores, derived by analyses recommended by Bland and Altman. These methods included (1) comparing the means and standard deviations (SDs) of both sets of scores, (2) calculating a Pearson correlation coefficient to measure the strength of the linear correlation between scores, (3) calculating an intraclass correlation coefficient to determine the extent to which score differences were explained by inherent variability of the VR-12, (4) plotting the frequency distributions of scores for the CAT and full model against one another, (5) plotting the distribution of the score differences (full score minus CAT score) for analysis, and finally (6) generating a Bland-Altman plot to assess the patterns in score differences. Analyses were performed with the R software suite (version 3.4.2; R Foundation for Statistical Computing), with the Python programming language (version 3.4.5; Python Software Foundation), or using Microsoft Excel spreadsheets.

The accuracy of CAT was viewed in the context of the minimal clinically important difference (MCID) for the VR-12, which is the minimum deviation in score that must occur for a

| Table II Diagnostic Information of Patients Whose Stored Responses to the Full VR-12 Were Applied to the CAT Model |
|---|---|---|
| Site  | No. of Patients | Common Diagnoses |
| 1     | 3,163 | Degenerative disc disease, intervertebral disc disorder, spinal stenosis |
| 2     | 2,860 | Rotator cuff sprain, osteoarthritis of shoulder, rupture of rotator cuff |
| 3     | 7,593 | Carpal tunnel syndrome, trigger finger, tenosynovitis |
| 4     | 2,478 | Tear of meniscus, transient synovitis of knee, disruption of knee ligament |
| 5     | 3,429 | Rotator cuff sprain, rupture of rotator cuff, osteoarthritis of shoulder |
noticeable change in health to be present. Previous literature has supported use of a 6-point change in the MCS and PCS as the MCID\textsuperscript{13,14}.

### Results

While the full VR-12 form comprises 12 questions, the CAT model required 8 questions to be answered in its application for all subjects, representing a 33% decrease in question burden. The model found the most useful initial question to be Question 6c (Table I), so the algorithm started with this question for each application. Questions 2b, 3a, 4b, and 6a were eliminated by CAT.

The mean CAT score was 1.3 points lower than the mean full score for the PCS (41.5 ± 11.0 versus 42.8 ± 10.4) and 0.3 point higher for the MCS (57.3 ± 9.4 versus 57.0 ± 9.6), with very similar SDs. The Pearson correlation coefficients were 0.97 for the PCS and 0.98 for the MCS, representing strong linear relationships between scores, and the intraclass correlation coefficients were 0.96 and 0.97, respectively, indicating strong agreement between scores as well. For each individual practice cohort, these values varied no more than 0.02. The mean scores for the individual sites are provided in Tables III and IV. The distribution of the CAT and full scores showed significant overlap for both the PCS (Fig. 2) and the MCS (Fig. 3). For the PCS, the difference between the CAT score and the full score was less than the MCID in >95% of cases (Fig. 4), and demonstrated a slight skew of the CAT to underestimate the score. For the MCS, the difference between the CAT score and the full score was also less than the MCID in >95% of cases (Fig. 5), with the differences evenly clustered around zero.

The Bland-Altman plot demonstrated that the differences between the CAT and the full VR-12 scores were largely independent of the overall score for both the PCS (Fig. 6) and the MCS (Fig. 7), although a slight decrease in the score difference was demonstrated at the overall score extremes (highest and lowest scores). This pattern is more identifiable in the PCS

| Table III | Summary of Statistical Data Comparing Accuracy of PCS of Full VR-12 with CAT Model at Different Sites |
|-----------|----------------------------------------------------------------------------------------------------|
| Site      | No. of Patients | Mean ± SD | Full VR-12 | CAT VR-12 | R  | ICC* |
|-----------|----------------|-----------|------------|----------|----|------|
| 1         | 3,163          | 37.1 ± 10.7 | 36.0 ± 10.8 | 0.97     | 0.97 |
| 2         | 2,860          | 43.0 ± 9.7  | 41.0 ± 10.5 | 0.97     | 0.95 |
| 3         | 7,093          | 45.8 ± 9.7  | 44.7 ± 10.5 | 0.97     | 0.96 |
| 4         | 2,478          | 41.8 ± 10.1 | 41.6 ± 10.1 | 0.97     | 0.97 |
| 5         | 3,429          | 41.7 ± 10.0 | 39.7 ± 10.7 | 0.97     | 0.95 |
| Overall   | 19,523         | 42.8 ± 10.4 | 41.5 ± 11.0 | 0.97     | 0.96 |

* ICC = intraclass correlation coefficient.

| Table IV | Summary of Statistical Data Comparing Accuracy of MCS of Full VR-12 with CAT Model at Different Sites |
|----------|-----------------------------------------------------------------------------------------------------|
| Site     | No. of Patients | Mean ± SD | Full VR-12 | CAT VR-12 | R   | ICC* |
| 1        | 3,163           | 52.1 ± 11.2 | 52.5 ± 11.0 | 0.98     | 0.98 |
| 2        | 2,860           | 57.4 ± 8.7  | 58.0 ± 8.5  | 0.97     | 0.97 |
| 3        | 7,093           | 58.0 ± 8.6  | 59.1 ± 8.3  | 0.98     | 0.97 |
| 4        | 2,478           | 57.5 ± 9.1  | 57.8 ± 8.8  | 0.97     | 0.97 |
| 5        | 3,429           | 56.2 ± 9.6  | 56.8 ± 9.5  | 0.97     | 0.96 |
| Overall  | 19,523          | 57.0 ± 9.6  | 57.3 ± 9.4  | 0.98     | 0.97 |

* ICC = intraclass correlation coefficient.

---

![Fig. 2](image-url)  
Distribution of PCS scores on the full (long) VR-12 (orange) overlaid with the distribution of the PCS scores on the CAT model (blue). Green shows where the full and CAT scores are the same.
Bland-Altman plot than in the MCS plot. However, a greater score error was unbiased toward either single extreme. In other words, greater score errors were not seen with higher scores compared with lower scores, or vice versa, for either the PCS or the MCS.

**Discussion**

PROMs are an essential part of orthopaedic care. The VR-12 is particularly useful as a well-validated, non-proprietary, and relatively short outcome measure compared with other PROMs. Recent literature has shown that it has become an
increasingly popular measure for characterizing outcomes of hip and knee arthroplasty as well as various arthroscopic procedures including rotator cuff repair and SLAP (superior labral tear from anterior to posterior) repair. It has additional practical value due to its integration in the Medicare Health Outcomes Survey and its use by the Centers for Medicare & Medicaid Services as a quality-of-life measure and source of performance assessment. In this role, it is used to

Fig. 5
Distribution of the differences between the full VR-12 and CAT MCS scores. Less than 5% of the absolute values of score differences were greater than the MCID. Most of the differences are clustered around zero.

Fig. 6
Bland-Altman plot of the difference between the CAT and full VR-12 PCS scores versus the mean of the 2 scores for each case. Most of the score differences are less than the MCID of 6 points, indicating that the CAT would not affect clinical interpretation of the outcomes. The differences in scores are shown to be slightly decreased at the overall score extremes, but bias is not seen toward larger versus smaller scores or vice versa.
assess the quality of programs including the Veterans Affairs, Medicare Advantage, and other health-care plans, and thereby contributes to the direction of reimbursements and other financial incentives for various plans and providers. Given this utility and widespread application, countless patients are asked to complete the VR-12 survey in its full form every day as a fixture of participation in the health-care system. For this reason, the CAT model was developed in an effort to help reduce the question burden placed on patients, thereby improving the patient experience by shifting focus away from data collection instruments and toward patient-driven goals and patients’ relationship with their physician.

In order to determine the value of the CAT model for this purpose, we first evaluated its ability to decrease question burden for potential responders. Our analysis showed that the model required a fixed number of questions for each response set with a uniform decrease in question burden (33%) across the board compared with the full VR-12. It remains unclear, however, to what degree the actual gross reduction in question burden would improve the patient experience. On the basis of public reporting of survey completion time, we can estimate that removal of 4 questions saves an average of about 140 seconds in response time. While these time savings may seem trivial when the survey is viewed as a singular event, this perception fails to take into account the demands placed on a patient during a typical health-care visit. In this setting, patients are frequently asked to provide various categories of data, from personal biographical information to medical history to problem-specific questionnaire responses. Every engagement to streamline these individual burdens may benefit the process as a whole, and combined efforts could certainly create more substantial improvements in health-care efficiency.

The second part of our evaluation of CAT concerned the accuracy of its score outputs relative to the scores generated from the full VR-12 survey. The accuracy of these scores was viewed in the context of the MCID to ensure that the analysis was anchored in the reality of subjective patient experience. The MCID for the VR-12 has been shown to vary somewhat based on the method of calculation (i.e., distribution-based versus anchor-based approach) and the patient cohort, but we estimated the MCID to be 6 points for both the PCS and the MCS on the basis of a review of the available literature. In this context, our results demonstrate that the scores strongly resemble each other not only in terms of summary statistics, including mean, SD, and Pearson correlation coefficient, but more importantly in terms of individual score breakdowns. The essential recreation of the score distributions by the CAT model was an important finding especially given that the score frequencies were not distributed normally. Additionally, for both the PCS and the MCS, the difference between the CAT score and the full VR-12 was less than the MCID in >95% of cases. This indicates that the CAT model outcome scores are faithful to the full version not just at the population level but also at the individual patient level.

Lastly, in terms of test-retest reliability, the intraclass coefficients demonstrated a stronger agreement between the CAT and full VR-12 scores than between scores of the same measure administered to the same individual twice. This suggests that there is likely more variability within the full VR-12 itself than...
between the full VR-12 and the CAT model. Taken together, these findings demonstrate support for the implementation of CAT in a live setting (while patients are responding to the survey) to elicit VR-12 outcome scores.

The CAT used by OBERD is distinct from alternative CAT systems, including those developed in PROMIS (Patient-Reported Outcomes Measurement Information System), which use methodology based on item response theory that requires a separate set of questions (i.e., “item bank”)

Because CAT was able to eliminate 4 questions from the VR-12, it would seem that the remaining 8 questions could generate a satisfactory “VR-8 short form,” but rigorous assessment of such a PROM awaits further study. An important element to consider for CAT is that, unlike standard PROMs, the order of questions may change from patient to patient on the basis of the earlier responses that they provided. As a result, theoretically the CAT model is valid only for PROMs in which the individual questions are independent from one another; in other words, when outcome scores would be the same regardless of the question order, which is not always the case. For example, it has been demonstrated that there may be a difference in how respondents answer a certain “energy” item depending on whether it is integrated in a 12-item (VR-12) or 36-item (Short Form [SF]-36) questionnaire

In conclusion, the CAT system was designed to incorporate machine learning algorithms into PROM collection in order to improve PROM efficiency and improve patient experience. In this study, the application of the CAT model to the VR-12 survey demonstrated an ability to lessen the response burden for patients and had very little impact on score integrity. Additional studies that validate CAT-generated scores for specialty-specific questionnaires will help determine the potential scope of the model’s application.

References

1. Brogan AP, DeMuro C, Barrett AM, D’ Alessio D, Bai V, Hogue SL. Payer perspectives on patient-reported outcomes in health care decision making: oncology examples. J Manag Care Spec Pharm. 2017 Feb;23(2):125-34.

2. Jenkinson C, Morley D. Patient reported outcomes. Eur J Cardiovasc Nurs. 2016 Apr;15(2):112-3. Epub 2015 Dec 17.

3. Wolfe F, Michaud K. Proposed metrics for the determination of rheumatoid arthritis outcome and treatment success and failure. J Rheumatol. 2009 Jan;36(1):27-33.

4. Kazis LE, Selim A, Rogers W, Ren XS, Lee A, Miller DR. Dissemination of methods and results from the Veterans Health Study: final comments and implications for
future monitoring strategies within and outside the veterans healthcare system. J Ambul Care Manage. 2006 Oct-Dec;29(4):310-9.

5. Bessette MC, Westermann RW, Davis A, Farrow L, Hagen MS, Miniaci A, Nickodem R, Parker R, Rosneck J, Saluan P, Spindler KP, Stearns K, Jones MH; Cleveland Clinic Sports Knee Group. Predictors of pain and function level before knee arthroscopy. Orthop J Sports Med. 2019 May 15;7(5):2325967119842465.

6. Resnik L, Ekerholm L, Borgia M, Clark MA. A national study of veterans with major upper limb amputation: survey methods, participants, and summary findings. PLoS One. 2019 Mar 14;14(3):e0213578.

7. Selim AJ, Rogers W, Fleishman JA, Qian SX, Fincke BG, Rotthandler JA, Kazis LE. Updated U.S. population standard for the Veterans RAND 12-Item Health Survey (VR-12). Qual Life Res. 2009 Feb;18(1):43-52. Epub 2008 Dec 3.

8. Snedden TR, Scerpella J, Klieathers NA, Norman RS, Byholder L, Sanfilippo J, McGuire TA, Heiderscheit B. Sport and physical activity level impacts health-related quality of life among collegiate students. Am J Health Promot. 2019 Jun;33(5):875-82. Epub 2018 Dec 26.

9. Chien TW, Wu HM, Wang WC, Castillo RV, Chou W. Reduction in patient burdens with graphical computerized adaptive testing on the ADL scale: tool development and simulation. Health Qual Life Outcomes. 2009 May 5;7:39.

10. Hsieh IP, Chen JH, Hou WH, Hsieh CL. Development of a computerized adaptive test for assessing activities of daily living in outpatients with stroke. Phys Ther. 2013 May;93(5):681-93. Epub 2013 Jan 17.

11. Plummer OR, Abboud BJ, Bell JE, Murthi AM, Romeo AA, Singh P, Zmistowski BM. A concise shoulder outcome measure: application of computerized adaptive testing to the American Shoulder and Elbow Surgeons Shoulder Assessment. J Shoulder Elbow Surg. 2019 Jul;28(7):1273-80. Epub 2019 Mar 2.

12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307-10.

13. Kronzer VL, Jerry MR, Ben Abdallah A, Wildes TS, McKinnon SL, Sharma A, Avidan MS. Changes in quality of life after elective surgery: an observational study comparing two methods. Lancet. 1986 Feb 8;1(8476):307-10.

14. Bessette MC, Westermann RW, Davis A, Farrow L, Hagen MS, Miniaci A, Nickodem R, Parker R, Rosneck J, Saluan P, Spindler KP, Stearns K, Jones MH; Cleveland Clinic Sports Knee Group. Predictors of pain and function level before knee arthroscopy. Orthop J Sports Med. 2019 May 15;7(5):2325967119842465.

15. Bents EJ, Brady PC, Adams CR, Tokish JM, Higgins LD, Denard PJ. Patient-reported outcomes of knotted and knotless glenohumeral labral repairs are equivalent. Am J Orthop (Belle Mead NJ). 2017 Nov;Dec;46(6):279-83.

16. Bents EJ, Brady PC, Adams CR, Tokish JM, Higgins LD, Denard PJ. Patient-reported outcomes of knotted and knotless glenohumeral labral repairs are equivalent. Am J Orthop (Belle Mead NJ). 2017 Nov;Dec;46(6):279-83.

17. Biomechanics of rotator cuff tears. Orthop J Sports Med. 2018 Jul;26(7):1292-7. Epub 2018 Feb 28.

18. Kazis LE, Selim AJ, Rogers W, Qian SX, Brazier J. Monitoring outcomes for the Medicare Advantage program: methods and application of the VR-12 for evaluation of plans. J Ambul Care Manage. 2012 Oct-Dec;35(4):263-76.

19. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ. Assessment of health-related quality of life in spine treatment: conversion from SF-36 to VR-12. Spine J. 2018 Jul;18(7):1292-7. Epub 2018 Feb 28.

20. Avidan MS. Changes in quality of life after elective surgery: an observational study comparing two methods. Lancet. 1986 Feb 8;1(8476):307-10.