JB&JS OPEN ACCESS

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[1-3]. 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[4]. 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[5-8]. As a result, reducing the question burden of this

Disclosure: The authors indicated that no external funding was received for any aspect of this work. On the **Disclosure of Potential Conflicts of Interest** forms, *which are provided with the online version of the article*, one or more of the authors checked "yes" to indicate that the author had a relevant financial relationship in the biomedical arena outside the submitted work (including employment by, and other relationships with, OBERD, the developer of the software system used for outcome data collection in this study) (http://links.lww.com/JBJSOA/A147).

JBJS Open Access • 2020:e0052. **openaccess.jbjs.org**

TABLE I Questions and Response Options for Veterans RAND 12 Item Health Survey (VR-12)*

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?	
a. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf.	Yes, limited a lot Yes, limited a little No, not limited at all
b. Climbing **several** flights of stairs.	Yes, limited a lot Yes, limited a little No, not limited at all
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**?	
a. **Accomplished less** than you would like.	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
b. Were limited in the **kind** of work or other activities.	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
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)?	
a. **Accomplished less** than you would like.	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
b. Didn't do work or other activities as **carefully** as usual.	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
5. During the past 4 weeks, how much did **pain** interfere with your normal work (including both work outside the home and housework)?	Not at all 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.	

continued

TABLE I (*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.)

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 on 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[9,10].

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[11], 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

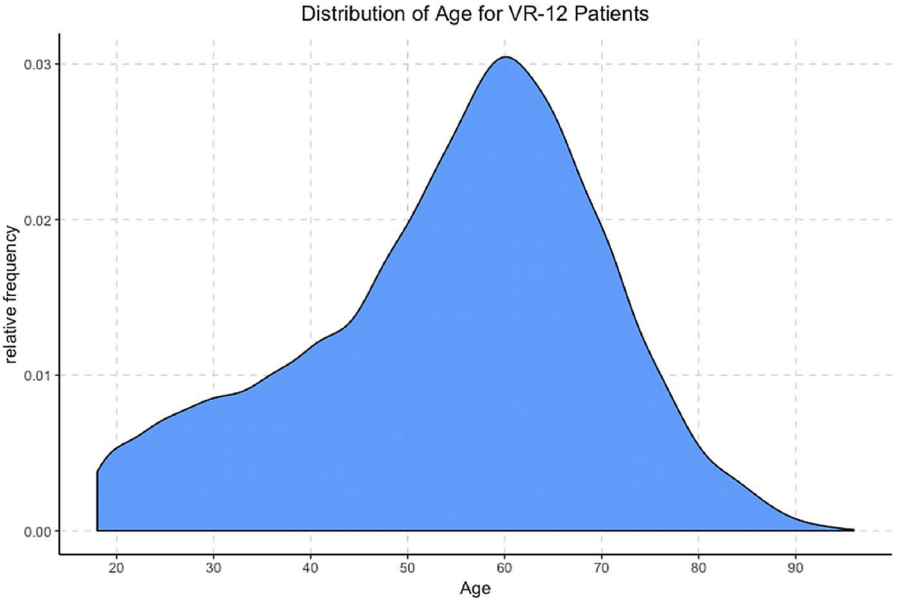

Fig. 1

Distribution of ages of patients included in the CAT model analysis of the VR-12.

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[12]. 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

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		R	ICC*
		Full VR-12	CAT VR-12		
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	7,593	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		R	ICC*
		Full VR-12	CAT VR-12		
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	7,593	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.

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[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

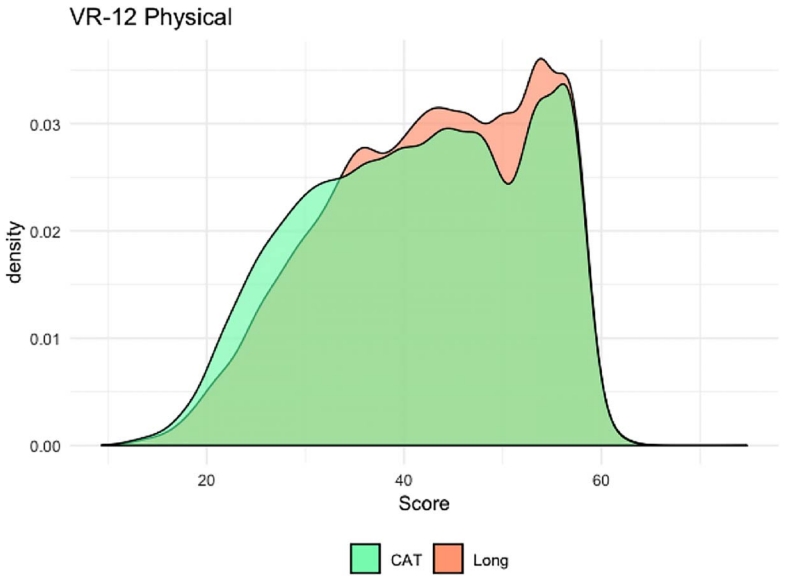

Fig. 2

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.

JBJS Open Access • 2020:e0052. openaccess.jbjs.org

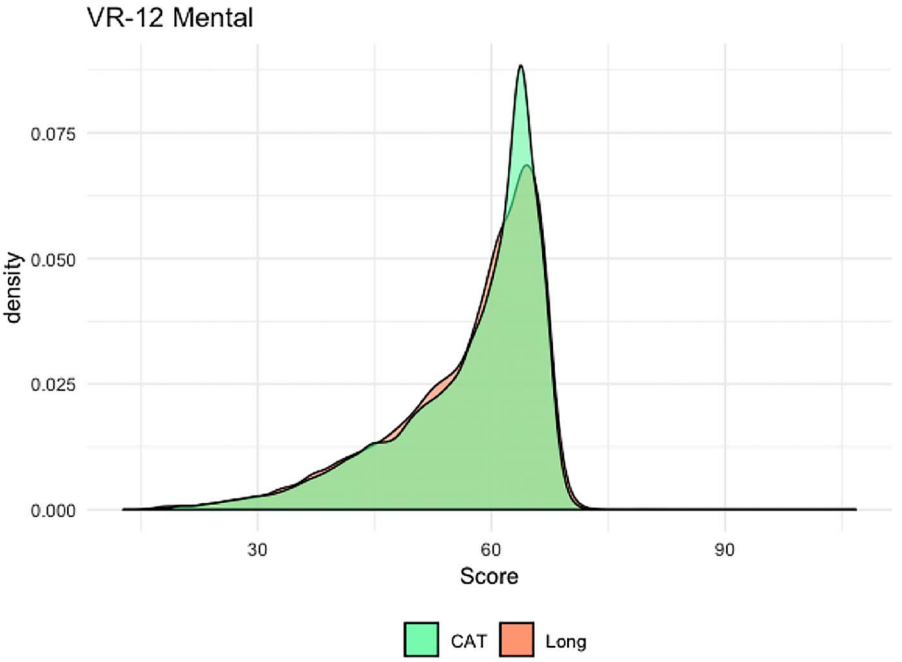

Fig. 3

Distribution of MCS scores on the full (long) VR-12 (orange) overlaid with the distribution of the MCS 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

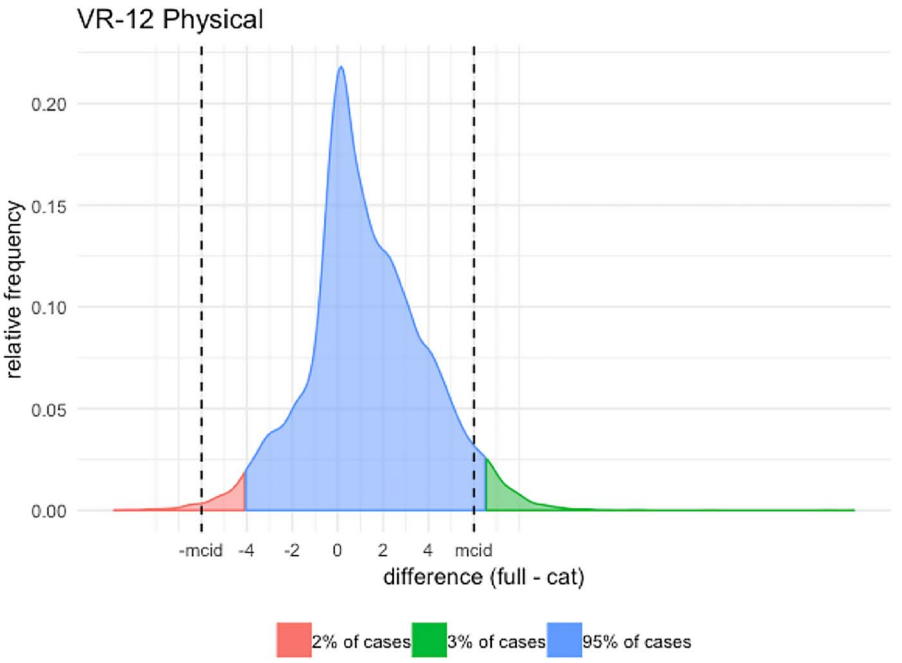

Fig. 4

Distribution of the differences between the full VR-12 and CAT PCS scores. Less than 5% of the absolute values of score differences were greater than the MCID. The differences are clustered around zero with very slight bias for lower CAT scores relative to full scores.

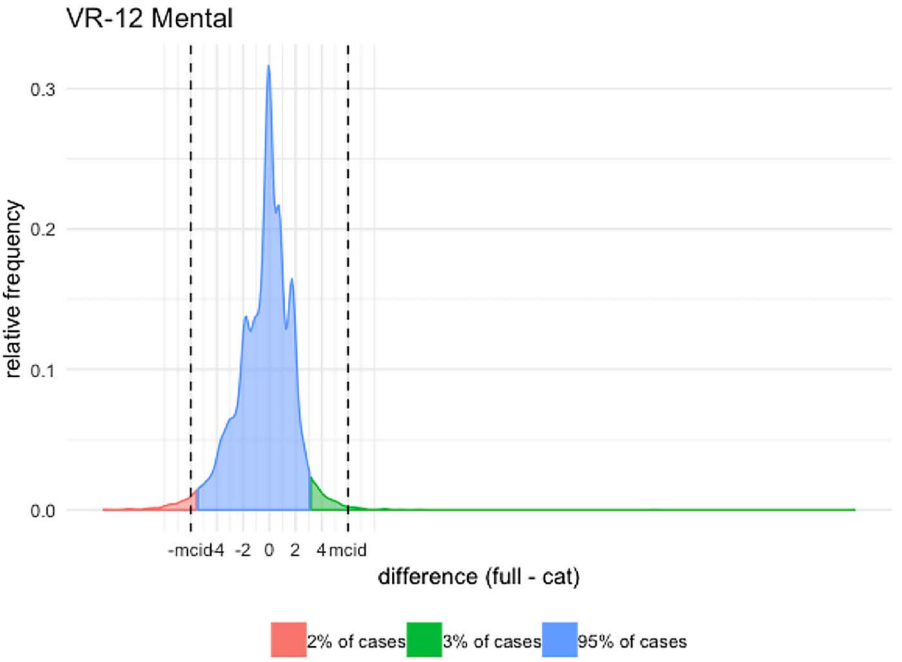

Fig. 5

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.

increasingly popular measure for characterizing outcomes of hip and knee arthroplasty[15-17] as well as various arthroscopic procedures including rotator cuff repair and SLAP (superior labral tear from anterior to posterior) repair[18-20]. 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[21,22]. In this role, it is used to

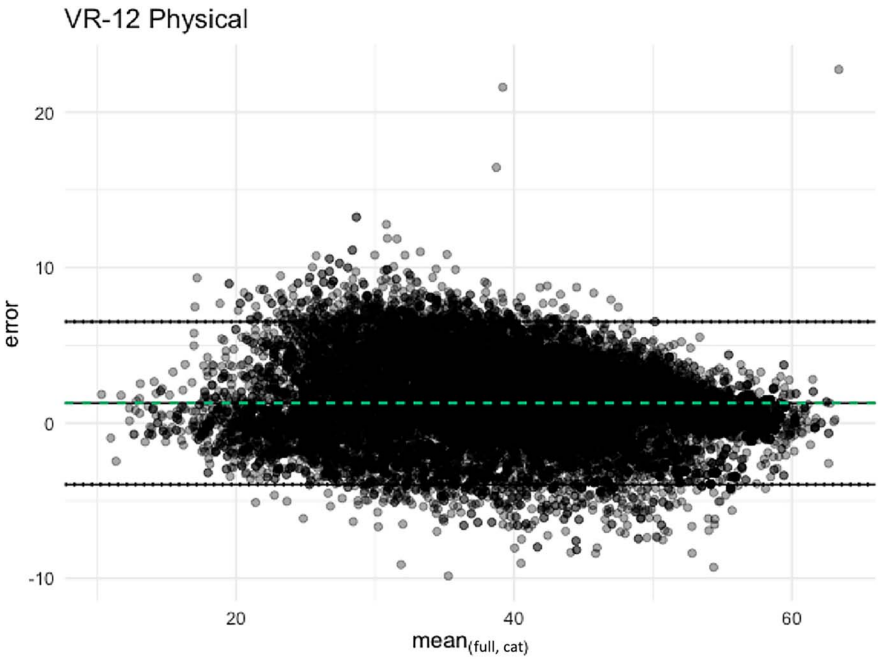

Fig. 6

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.

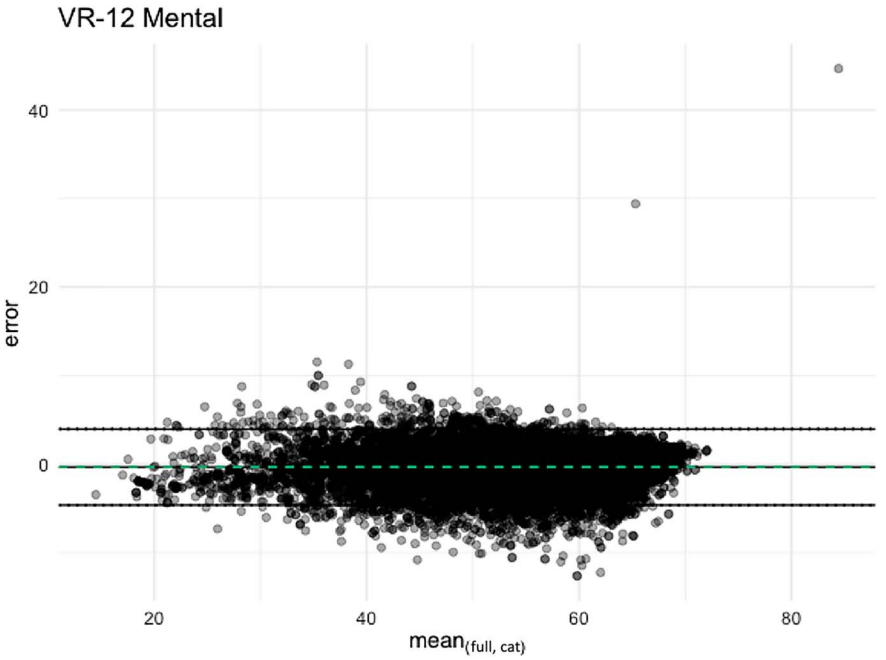

Fig. 7

Bland-Altman plot of the difference between the CAT and full VR-12 MCS 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 independent of the overall score (i.e., no bias of greater differences at higher or lower scores).

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[13,14,23]. 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[24]. 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")[24,25]. The OBERD CAT, however, relies on machine learning methods rather than item response theory. It utilizes the questions of the historical forms rather than separate item banks to collect outcomes, making them interchangeable with the existing PROMs such as the VR-12. In its practical application from a patient perspective, the administration of the CAT strongly resembles that of the full questionnaire. Both questionnaires are completed using a tablet device in a private clinic room, and questions are delivered to the patient's screen 1 at a time. The patient is unaware whether the questions are a fixed set or generated by a system of algorithms such as the CAT model, and thus the patient experience remains unchanged. This consistency is important as it has been demonstrated that the method of administration of general health surveys can affect the outcome score[26].

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[27]. Importantly, however, it has not been shown that a change in item order would affect responses within the VR-12. In fact, multiple methods have been validated for converting item-based responses of the VR-12 to relevant counterparts through scoring algorithms, which have supported the basis of item independence for this survey[21,25,27].

A few more limitations must be considered to place these results in proper context. By nature of being a retrospective comparative analysis, the model could be used only on previously stored responses and not in a live setting (while the patients were actually responding to the survey). Only comparative prospective studies will be able to confirm the assumptions of certain measure requirements, including item independence, and truly validate the CAT model's fidelity to the full survey. Additionally, although our cohort included patients with a range of demographic and diagnostic characteristics to maximize generalizability of the CAT model, our population had a higher mean PCS (42.8) and MCS (57.0) compared with a recent report of the contemporary U.S. population (40.1 and 50.2, respectively)[7]. Therefore, our patient population likely represents a slightly healthier cohort than the general population. Additionally, although this study did not compare outcomes on the basis of demographic or diagnostic factors, there was a consistent level of agreement for both the MCS and the PCS at multiple clinic sites, each of which treated patients with a distinct demographic and diagnostic make-up (Table II). Determining whether these factors impact the accuracy of the CAT model likely requires a prospective comparative approach.

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. ■

Liam T. Kane, BS[1]
Surena Namdari, MD, MSc[1]
Otho R. Plummer, PhD[2]
Pedro Beredjiklian, MD[1]
Alexander Vaccaro, MD, PhD, MBA[1]
Joseph A. Abboud, MD[1]

[1]Rothman Orthopaedic Institute, Philadelphia, Pennsylvania

[2]Universal Research Solutions, Columbia, Missouri

Email address for J.A. Abboud: Joseph.abboud@rothmanortho.com

ORCID iD for L.T. Kane: 0000-0003-1524-8378
ORCID iD for S. Namdari: 0000-0002-8226-0310
ORCID iD for O.R. Plummer: 0000-0002-6037-4069
ORCID iD for P. Beredjiklian: 0000-0001-7625-6270
ORCID iD for A. Vaccaro: 0000-0002-8073-0796
ORCID iD for J.A. Abboud: 0000-0002-3845-7220

References

1. Brogan AP, DeMuro C, Barrett AM, D'Alessio D, Bal 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 before knee arthroscopy. Orthop J Sports Med. 2019 May 15;7(5):2325967119844265.

6. Resnik L, Ekerholm S, 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, Rothendler 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, Kliethermes SA, Norman RS, Blyholder L, Sanfilippo J, McGuine 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):675-82. 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. Hsueh IP, Chen JH, Wang CH, 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 JA, 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 measures. Qual Life Res. 2017 Aug;26(8):2093-102. Epub 2017 Mar 29.

14. Zhou L, Natarajan M, Miller BS, Gagnier JJ. Establishing minimal important differences for the VR-12 and SANE scores in patients following treatment of rotator cuff tears. Orthop J Sports Med. 2018 Jul 26;6(7):2325967118782159.

15. Mata-Fink A, Philipson DJ, Keeney BJ, Ramkumar DB, Moschetti WE, Tomek IM. Patient-reported outcomes after revision of metal-on-metal total bearings in total hip arthroplasty. J Arthroplasty. 2017 Apr;32(4):1241-4. Epub 2016 Oct 11.

16. Prohaska MG, Keeney BJ, Beg HA, Swarup I, Moschetti WE, Kantor SR, Tomek IM. Preoperative body mass index and physical function are associated with length of stay and facility discharge after total knee arthroplasty. Knee. 2017 Jun;24(3):634-40. Epub 2017 Mar 20.

17. Bienstock DM, Snyder DJ, Kroshus TR, Ahn A, Koenig KM, Molloy IB, Jevsevar DS, Poeran J, Moucha CS. Relationship between baseline patient-reported outcomes and demographic, psychosocial, and clinical characteristics: a retrospective study. J Am Acad Orthop Surg Glob Res Rev. 2019 May 9;3(5):e039.

18. 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.

19. Thigpen CA, Floyd SB, Chapman C, Tokish JM, Kissenberth MJ, Hawkins RJ, Brooks JM. Comparison of surgeon performance of rotator cuff repair: risk adjustment toward a more accurate performance measure. J Bone Joint Surg Am. 2018 Dec 19;100(24):2110-7.

20. McIntyre LF, Bishai SK, Brown PB 3rd, Bushnell BD, Trenhaile SW. Patient-reported outcomes after use of a bioabsorbable collagen implant to treat partial and full-thickness rotator cuff tears. Arthroscopy. 2019 Aug;35(8):2262-71. Epub 2019 Jul 23.

21. Gornet MF, Copay AG, Sorensen KM, Schranck FW. 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.

22. 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.

23. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ. Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion: clinical article. J Neurosurg Spine. 2013 Feb;18(2):154-60. Epub 2012 Nov 23.

24. Lapin BR, Kinzy TG, Thompson NR, Krishnaney A, Katzan IL. Accuracy of linking VR-12 and PROMIS Global Health scores in clinical practice. Value Health. 2018 Oct;21(10):1226-33. Epub 2018 Apr 26.

25. Schalet BD, Rothrock NE, Hays RD, Kazis LE, Cook KF, Rutsohn JP, Cella D. Linking physical and mental health summary scores from the Veterans RAND 12-Item Health Survey (VR-12) to the PROMIS(®) Global Health Scale. J Gen Intern Med. 2015 Oct;30(10):1524-30. Epub 2015 Jul 16.

26. McHorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care. 1994 Jan;32(1):40-66.

27. Selim A, Rogers W, Qian S, Rothendler JA, Kent EE, Kazis LE. A new algorithm to build bridges between two patient-reported health outcome instruments: the MOS SF-36® and the VR-12 Health Survey. Qual Life Res. 2018 Aug;27(8):2195-206. Epub 2018 Apr 19.