Published in final edited form as:

Med Care. 2019 March ; 57(3): 180–186. doi:10.1097/MLR.0000000000001013.

Feasibility of Distinguishing Performance Among Provider Groups Using Patient-Reported Outcome Measures in Older Adults with Multiple Chronic Conditions

Adam J. Rose, MD, MSc^{1,2}, Elizabeth Bayliss, MD^{3,4}, Lesley Baseman⁵, Emily Butcher¹, Wenjing Huang, PhD⁶, and Maria Orlando Edelen, PhD¹

¹RAND Corporation, Boston, MA;

²Section of General Internal Medicine, Boston University School of Medicine, Boston, MA;

³Institute for Health Research, Kaiser Permanente Colorado, Denver, CO;

⁴Department of Family Medicine, University of Colorado School of Medicine, Aurora, CO;

⁵RAND Corporation, Arlington, VA;

⁶RAND Corporation, Santa Monica, CA.

Abstract

Objective: To examine minimum sample sizes and follow-up times required for patient reported outcome-based performance measures (PRO-based PMs) to achieve acceptable reliability as PMs.

Participants: We used two groups of patients age 65+ with at least 2 of 13 chronic conditions. The first was a sample of Medicare Advantage beneficiaries, who reported health-related quality of life (HRQoL) at baseline and two years. The second was a sample of primary care patients, who reported HRQoL at baseline and six months.

Measures: Medicare Advantage beneficiaries completed the VR-12, while the primary care sample completed the PROMIS-29. We constructed binary candidate PMs indicating stable or improved physical or mental HRQoL at follow-up, and continuous PMs measuring mean change over time.

Corresponding Author: Adam J. Rose, MD MSc, RAND Corporation, 20 Park Plaza, Suite 920, Boston, MA 02116, Phone: 617-338-2059 X 8672, Fax: 617-357-7470, arose@rand.org.

Complete Author Information

Adam J. Rose, MD MSc, RAND Corporation, 20 Park Plaza, Suite 920, Boston, MA 02116, Phone: 617-338-2059 X 8672, Fax: 617-357-7470, arose@rand.org

Elizabeth Bayliss, MD, Kaiser Permanente Colorado, 10065 E. Harvard Ave, Denver, CO 80207, Phone: 303-364-1328, Fax: 303-614-1305, Elizabeth.bayliss@kp.org

Lesley Baseman, RAND Corporation, 20 Park Plaza, Suite 920, Boston, MA 02116, Fax: 617-357-7470, lbaseman@rand.org

Emily Butcher, RAND Corporation, 20 Park Plaza, Suite 920, Boston, MA 02116, Phone: 617-338-2059, Fax: 617-357-7470, ebutcher93@gmail.com

Wenjing Huang, PhD, RAND Corporation, 1776 Main St., Santa Monica, CA 90401, Phone: 310-393-0411, Fax: 617-357-7470, whuang@rand.org

Maria Orlando Edelen, PhD, RAND Corporation, 20 Park Plaza, Suite 920, Boston, MA 02116, Phone: 617-338-2059, Fax: 617-357-7470, orlando@rand.org

Conflicts of Interest: The authors have no conflicts of interest to report.

Results: In the Medicare Advantage sample, with a sample size per entity profiled of 160, the most promising PM achieved a reliability of 0.32 as a PM. A sample size of 882 per entity would have been needed for this PM to achieve an acceptable reliability of 0.7. In the prospective sample, with a sample size of 27 per clinic, the most promising PM achieved a reliability of 0.16 as a PM. A sample size of 341 patients (at the clinic level) would have been needed for this PM to achieve a reliability of 0.7.

Conclusions: Achieving acceptable reliability for these PMs and conditions would have required minimum sample sizes of 341 at the clinic level or 880 at the health plan level. These estimates can guide the design of future PRO-based PMs.

Keywords

research methods; patient-reported outcomes; performance measures; quality of life; geriatrics; comorbidity

INTRODUCTION

As efforts to measure health care quality become more sophisticated, the goals of patient care extend beyond ensuring survival to helping patients optimize functional status and well-being. Though most performance measures (PMs) focus on processes of care or outcomes such as survival, it may be equally or more important to measure quality of life among older adults, especially those with multiple chronic conditions (MCCs).^{1,2} To be meaningful, PMs must be aligned with the goals of the patients whose care they purport to measure.³ It is clear that measuring changes in health-related quality of life (HRQoL) over time would have obvious relevance to older adults with MCCs.

In this study, we examined two well-validated HRQoL instruments to determine if they could be used as patient reported outcome (PRO) based PMs. With two datasets, we examined the feasibility of using these PROs as PMs by examining their reliability as PMs in an older population with MCCs. The purpose of this effort was to inform expectations about the sample size that will be needed to develop and validate PRO-based PMs.

METHODS

Data for this study came from two sources: the Medicare Health Outcomes Survey (HOS), a questionnaire administered by the Centers for Medicare and Medicaid Services (CMS) to track the health status of individuals enrolled in Medicare Advantage Organizations (MAOs), with which we measured changes in HRQoL over a two-year period using the VR-12 ("Phase 1"). We also collected primary data from primary care patients enrolled with Kaiser Permanente of Colorado (KPCO), using the PROMIS-29 and concurrent measurement of the VR-36 ("Phase 2").

We collected data from participants at baseline and at follow-up, which was two years later for Phase one and six months later for Phase 2. Analyses include data only from participants who contributed data at both time points. We used summary scores capturing physical and mental HRQoL to construct binary and continuous candidate PMs. The binary PMs

indicated whether the respondent had a stable or improved score at follow-up, and the continuous PMs measured the mean score change over time.

We examined the reliability that could be obtained with these PMs under these conditions (i.e., with these sample sizes and these measures of statistical dispersion). It is important to distinguish the use of the term “reliability” here from how it is used in other contexts. In the context of validating a multi-item instrument, “reliability” refers to measures of how closely the items in the instrument relate to each other, such as a Cronbach’s coefficient alpha.⁴ In contrast, here, in the context of performance measurement, “reliability” refers to efforts to distinguish between the performance of two entities (providers, clinics, or health plans). In the context of performance measurement, as we will explain below, reliability is a function of the intraclass correlation (ICC) and the sample size, or the number of patients per entity being profiled. Following the work of others,^{5–8} we considered a minimum reliability of 0.7 to be necessary for an acceptable PM. When the reliability as a PM was less than 0.7, we calculated the minimum sample size that would have been needed to achieve 0.7 reliability.

Phase 1 Dataset: Medicare Health Outcomes Study

Since 1998, the HOS has been administered annually to a random sample of enrollees of each MAO, with a follow-up survey administered two years later. Risk-adjusted changes in HRQoL, as reflected by the PCS and MCS, have been used to rank MAOs and contribute to the providers’ Star Rating assigned by CMS. The analysis for this study focused on cohort 15, whose baseline collection was in 2012 and follow-up was in 2014. We also analyzed cohorts 9–14; results were similar.

To be included in our analysis, a patient needed to be age 65 or older at baseline, a member of the same MAO for both survey time points, and provide sufficiently complete responses to calculate PCS and MCS for both time points. In addition, patients needed to have at least 2 of 13 pre-specified conditions, based on patient self-report. Twelve of these 13 conditions were ascertained based on a positive response to a question on the baseline survey, asking if the respondent “had ever been told by a health professional that you have...”. These 12 conditions were: arthritis, cancer, chronic lung disease, congestive heart failure, diabetes, hypertension, inflammatory bowel disease, ischemic heart disease, osteoporosis, other heart problems, sciatica, and stroke. The language used to describe these conditions in the survey is found in eTable 1. For the 13th condition, depression, ascertainment was based on answering in the affirmative to at least one of three depression screening questions (Supplemental Digital Content 1, methods details).

Phase 2 Dataset: Primary Data Collection from KPCO

Participants for Phase 2 were recruited from KPCO, a not-for-profit integrated delivery system. We obtained institutional review board approvals from both KPCO and the RAND Corporation. We identified 3,749 participants who met the eligibility criteria, which were similar to those for Phase 1, except that chronic conditions were defined using ICD-10 codes (Supplemental Digital Content 1, methods details). Subjects were sent a letter with an option to opt out by returning a postcard. Those who did not opt out were sent a link to a web-based survey by email. Respondents to the baseline survey were sent a link to an identical follow-

up survey six months later. To be included, individuals also needed to respond to both the baseline and follow-up surveys with sufficiently complete answers to calculate summary scores. Details of survey data collection are described in Supplemental Digital Content 2.

Measurement Scales

The Veterans RAND 12-Item Short Form (VR-12), used to measure HRQoL in Phase 1, is a general HRQoL instrument⁹ with eight domains and two summary scales: the Physical Component Score, or PCS, and the Mental Component Score, or MCS.^{10,11} The VR-12 has been validated for use with both Veteran and non-Veteran populations.⁹ It was originally developed among Veterans and has continued to be used in Veteran populations with a diverse set of conditions including spinal cord injury,¹² posttraumatic stress disorder,¹³ coronary artery disease,¹⁴ and multiple sclerosis.¹⁵

The Patient-Reported Outcomes Measurement Information System 29-Item Profile Measure (PROMIS-29), used to measure HRQoL in Phase 2, is a relatively new HRQoL instrument that was developed using modern measurement theory and was calibrated and scored based on contemporary samples (post-2000). The instrument produces eight scores, which include four items from each of the seven PROMIS domains (anxiety, depression, fatigue, pain interference, physical functioning, sleep disturbance, and ability to participate in social roles), as well as a pain intensity item. PROMIS-29 has been tested and validated in a variety of patient populations. In addition to a representative sample of the U.S. population,¹⁶ the instrument has successfully measured HRQoL in chiropractic patients¹⁷ and in people with neuroendocrine tumors,^{18,19} systemic sclerosis,²⁰ irritable bowel syndrome,²¹ rheumatoid arthritis and osteoarthritis,²² and systemic lupus erythematosus,²³ and HIV.²⁴

We used two summary scales for the PROMIS-29, which we call the Physical Health Score (PHS) and the Mental Health Score (MHS). These summary scores were intended to measure the physical and mental domains of HRQoL. The use of these two scores was meant to be analogous to the VR-12 PCS and MCS, and to obviate the need to examine a large number of PMs based on the eight scores. The process for developing the PHS and MHS is described in Supplemental Digital Content 3, methods supplement, and has been described elsewhere by Hays, et al.²⁵ Concurrently, we used the VR-36 as a HRQoL measure in Phase 2. The VR-36 is an expanded version of the VR-12 and produces the same scores and summary scores.

Construction of Candidate Performance Measures

We created two groups of candidate PMs based on the comparison between baseline and follow-up survey scores (two years for Phase 1, and six months for Phase 2), for responses with sufficient information to calculate HRQoL scores. The first group of candidate PMs was based on having a stable or improved score for an aspect of HRQoL; the second was based on the mean change in an aspect of HRQoL. In Phase 1, the “stable or improved” score was defined based on the score not having decreased by more than a standard deviation (SD; see Table 1). In Phase 2, the “stable or improved” score was defined using various cutoffs: not having decreased at all, not having decreased by more than ¼ SD, by

more than ½ SD, or by more than a full SD. These effect sizes correspond to a generally accepted framework defining them as small, moderate, or large to very large effect sizes.²⁶

Calculating Reliability of Candidate Performance Measures

We formally evaluated the eight candidate PRO-based PMs based on their reliability as PMs. The term “reliability” is generally understood to refer to the proportion of variation in the PM attributable to systematic differences across the measured entities, rather than to random error. This is also sometimes called the “signal to noise ratio.” A more reliable PM will capture more signal and less noise.^{5–8} We measured reliability of PMs at two organizational levels: health plans in Phase 1, and ambulatory clinics in Phase 2.

To measure PM reliability, we used a widely accepted approach based on calculation of intraclass correlation (ICC), which is independent of sample size. The ICCs for continuous outcomes were calculated using random effects models to determine the proportion of the variance explained by the clinic. ICCs for continuous outcomes were calculated using random effects models within the lmer package in the R statistics program (R Foundation for Statistical Computing, Vienna, Austria) to determine the proportion of the variance explained by the clinic. For binary outcomes, we used the aod package in R, with the function iccbin(). Other procedures were applied, including those implemented in the ICC and psych packages in R to assess sensitivity to these methodological choices; these analyses yielded similar results and are not presented here.

After calculating ICC, PM reliability (r) can be calculated using an adaptation of the Spearman-Brown Prophecy formula,^{27,28} where n is defined here as the mean number of respondents per entity being profiled, instead of the number of items or tests as in the original formula:

$$r = \frac{(ICC)(n)}{1 + (ICC)(n-1)}$$

This equation implies that, for a given ICC, reliability of a PM can be increased by increasing n . The equation can also be rearranged to calculate how large an n is needed to achieve a specific reliability as a PM for a given ICC:

$$n = \frac{r(1-ICC)}{(ICC)(1-r)}$$

A PM must have sufficient reliability as a PM to demonstrate differences between entities that are due to actual differences between entities, rather than chance alone. Increasing the sample size can improve reliability of a PM; here, we examine the reliability achieved with these candidate PMs and these sample sizes. While there is no firm minimum reliability for a PM, most previous studies consider a minimum reliability of 0.7 to 0.8 to be valid to use a PM for comparisons among providers.^{5–8} Accordingly, we calculated three statistics for each of the candidate PMs we examined: 1) ICC; 2) reliability as a PM estimated using the available sample; and 3) the minimum number of respondents per site needed to achieve a

reliability of 0.7 for that PM. Analyses were conducted using SAS, version 9.4 (SAS Corporation) and R, version 3.2.5 (R Foundation).

RESULTS

Phase 1 Analyses

The dataset for Phase 1 consisted of 79,972 HOS respondents, age 65 or older and having at least two MCCs, with sufficiently complete responses to calculate the PCS and MCS at baseline and two years later. Respondents were distributed among 466 MAOs, with an average of 160 respondents per MAO. The response rate for the baseline survey was 51.5% and the response rate for the follow-up survey was 72.2%. Table 2 shows the proportion of these individuals who had each of the 13 chronic conditions and who had 2, 3, 4, or 5+ conditions, which helps characterize respondents' burden of chronic disease.

Overall, 68% of individuals had stable or improved PCS at follow-up (mean change -0.38 on a 100-point scale, SD 0.84), and 74% had stable or improved MCS (mean change 0.36, SD 0.90). Table 3 shows the performance characteristics of PM1a, PM1b, PM2a, and PM2b among cohort 15 of the HOS, adjusting for age only. PM1b achieved the highest reliability (0.315), which falls short of our prespecified minimum reliability of 0.7 for a PM. To achieve a reliability of 0.7 for PM1b, we would have needed at least 882 respondents per MAO; the other candidate PMs, with lower reliabilities as PMs, would have required much larger sample sizes.

We also performed several analyses to examine the sensitivity of these results to our methodological choices (Table, Supplemental Digital Content 4). After adding risk adjustment, all four PMs required a sample size of at least 2,000 respondents per MAO to achieve acceptable reliability as PMs, whereas one PM required only 882 without risk adjustment. We also re-examined reliability for these PMs after deleting proxy responses. Including proxy responses, there had been an average of 160 respondents per MAO; after excluding proxy responses, there was an average of 131 respondents per MAO. Excluding proxy responses generally resulted in lower reliability as PMs. Finally, we distinguished between the worst-performing quartile of Medicare Advantage Organizations and the other three quartiles, which worsened reliability for PM2a and improved it for PM2b, such that a sample size of only 662 per MAO would have sufficed to generate a reliability of 0.7 for PM2b.

Phase 2 Analyses

The dataset for Phase 2 consisted of 337 KPCO members, distributed among 13 KPCO-affiliated ambulatory care clinics, with sufficiently complete responses to calculate PHS and MHS at a six-month follow-up interval. The following is a description of how the sample of 337 was achieved. At baseline, 1996 patients were invited to participate in the survey. Of these, 779 responded (39%), and 490 had sufficiently complete responses for analysis (25%). By design, not all of these 490 were invited to participate in the follow-up survey. 401 of them were chosen at random and invited to participate, of whom 352 responded

(88%), and 337 had sufficiently complete responses for analysis (84%). It is these 337 upon whom the Phase 2 analyses were based.

Table 4 shows the overall sample characteristics of these individuals. A majority of individuals were age 80 or older (due to oversampling), and most were White, non-Hispanic. There was a high burden of chronic disease, roughly comparable to the HOS population seen in Phase 1.

Overall, 53% of individuals had stable or improved PCS at follow-up (mean change -0.23 on a 100-point scale, SD 5.08), and 53% had stable or improved MCS (mean change -0.10 , SD 4.74). Table 5 shows the performance characteristics of PM1c, PM1d, PM2c, and PM2d among these 13 KPCO clinics, without adjustment for covariates. The version of PM1c which allowed up to $\frac{1}{4}$ standard deviation of decrease to define stability achieved a reliability as a PM of 0.16. To achieve a reliability of 0.7 for this PM, we would have needed at least 341 respondents per clinic. The only other PM that achieved a measurable reliability as a PM was PM2c (reliability = 0.03). This measure would have required a sample size of 2,317 per clinic to achieve a reliability of 0.7 as a PM. We also examined PM1a, PM1b, PM2a, and PM2b for the KPCO cohort, using VR-36 based measures and defining PM1a and PM1b as “equal to or greater than the previous value”. We repeated the reliability analyses (results not shown). None of these PMs achieved measurable ICC or reliability as a PM in this sample.

DISCUSSION

We examined eight candidate PMs, which focus on change in HRQoL over time, for profiling MAOs (Phase 1) and ambulatory clinics (Phase 2), using a population of community-dwelling older adults with multiple chronic conditions. In Phase 1, with a follow-up interval of two years, the highest reliability as a PM was achieved by PM1b (0.315). In Phase 2, with a follow-up interval of only six months, the highest reliability as a PM was achieved by a variant of PM1c (0.16). Our results imply that larger sample sizes would have been needed for acceptable reliability of PMs, at least with this population.

One way to increase the reliability of PMs based on changes in HRQoL is to increase the follow-up interval, and indeed our results imply that a two-year follow-up interval allows more opportunity for HRQoL to change than six months. There is a natural limit to follow-up intervals, however, because longer follow-up intervals erode the immediacy of PMs and could lead to issues with loss to follow-up and attribution. Another way to increase reliability of PMs would be to select a sample with higher comorbidity and thus greater propensity to decline in HRQoL over time. Our analyses used a group of patients with a high propensity for decline, namely, patients age 65+ with 2 or more of 13 chronic conditions. Limiting the inclusion criteria to patients with even greater illness burden (e.g., 3 or more chronic conditions) might be expected to increase the signal further.

Another approach might be to study PRO-based PMs in a population with a specific illness that predicts a high likelihood of decline over time, and ideally an illness for which effective treatments are available. A possible example would be a population of patients with severe

congestive heart failure, whose decline over time could be measured either with the general HRQoL measures we used or potentially a disease-specific measure.²⁹ It is worth noting that PRO-based PMs have shown particular promise in the setting of joint replacement, a procedure usually expected to produce a rapid and marked improvement in HRQoL.³⁰ Thus, while we found that PRO-based PMs would require infeasibly large sample sizes in a population of older adults with MCCs, this does not necessarily mean that PRO-based PMs would be unworkable in all populations.

One important consideration for PRO-based PMs is that they should be measured as close as possible to the person or entity with the ability to improve the patient's HRQoL or slow its decline. PRO-based PMs could be used at three levels of the healthcare system: physician (closest to the patient), practice (farther), and payer or plan (farthest). In our study, the smallest sample size that would have been required to produce an acceptable reliability as a PM was found as part of the clinic level analysis (341 patients per clinic), rather than at the health plan level (at least 880 patients per clinic). It is possible that measuring even closer to the patient (physician-level) could increase the ICC above what was seen in this study – possibly enough to compensate for the smaller sample size. This would need to be examined directly. For now, it is noteworthy that none of the PMs studied here produced acceptable reliability as a PM with the sample sizes we had, in this population. Future efforts to validate PRO-based PMs in a similar population should use sample sizes at least as large as those estimated in this study as the minimum that would have been necessary to produce acceptable reliability as a PM.

It is important to emphasize that this study does not in any way challenge the validity or reliability of the VR-12, the PROMIS-29, or their summary scores. Rather, what we examined here was the reliability of their summary scores when operationalized as PMs at the clinic or health plan level. The VR-12 and PROMIS-29 have been thoroughly validated as measures of HRQoL, and none of our findings call that into question at all.

This study has important methodological strengths. We used two different samples. Our results were obtained using two different HRQoL instruments, the VR-12/VR-36 and the PROMIS-29. Therefore, the results that we obtained do not rely on a single dataset or a single HRQoL instrument. The rigorous prospective and longitudinal data collection for Phase 2, performed in the setting of ambulatory care clinics, is an added strength.

We also acknowledge important limitations. First, as we have noted above, we only examined the use of changes in HRQoL over time in this specific population; namely, community-dwelling older adults with multiple chronic conditions. Our results should not be taken to preclude the use of PRO-based PMs in other populations that may provide more of a signal in terms of changes in HRQoL over time. Second, the sample sizes that we estimated would have been necessary for profiling performance in this population at the level of health plans or clinics would not be impossible to achieve. While collecting these sample sizes would require resources, it is ultimately up to policymakers and health plan managers to decide whether the need to measure performance would justify the resources required. Finally, while these datasets offered unique opportunities to examine the questions we did, we did not have the data necessary to perform provider-level analyses, or to examine

similar PMs in certain sub-populations such as recipients of joint replacement surgery. These analyses would be important priorities for future studies.

In conclusion, we used two separate datasets, collected under different conditions, to examine the reliability as PMs of PRO-based PMs as well as the sample sizes needed to produce acceptable reliability. Our results suggest that developing and validating performance measures based on change in HRQoL over time, using a two-year follow-up period, would require at least 882 respondents per health plan being profiled, to achieve a reliability of 0.7 as a PM. **Similarly, measuring at the clinic level with a six-month follow-up period would have required at least 341 respondents per clinic being profiled.** Shorter follow-up, or adding risk adjustment, would require even more respondents. These estimates should be kept in mind when planning for the development and validation of PRO-based PMs in similar settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

Funding: Funded by contract #HHSN271201500064C NIH NIA (PI: Edelen).

References

1. Frankel BA, Bishop TF. A Cross-Sectional Assessment of the Quality of Physician Quality Reporting System Measures. *Journal of general internal medicine*. 2016;31:840–845. [PubMed: 27197975]
2. MacLeod S, Schwebke K, Hawkins K, et al. The Need for Comprehensive Health Care Quality Measures for Older Adults. *Population health management*. 2017.
3. Patient Reported Outcomes (PROs) in Performance Measurement. Washington, DC: National Quality Forum;2013.
4. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334.
5. Hargraves JL, Hays RD, Cleary PD. Psychometric properties of the Consumer Assessment of Health Plans (CAHPS) 2.0 Adult Core Survey. *Health Serv Res*. 2003;38:1509–1527. [PubMed: 14727785]
6. Hofer TP, Hayward RA, Greenfield S, et al. The Unreliability of Individual Physician “Report Cards” for Assessing the Costs and Quality of Care of a Chronic Disease. *JAMA*. 1999;281:2098–2105. [PubMed: 10367820]
7. Price RA, Stucky B, Parast L, et al. Development of valid and reliable measures of patient and family experiences of hospice care for public reporting. *J Palliative Med*. doi: 10.1089/jpm.2017.0594. [Epub ahead of print].
8. Sequist TD, Schneider EC, Li A, et al. Reliability of medical group and physician performance measurement in the primary care setting. *Med Care*. 2011;49:126–131. [PubMed: 20421826]
9. Kazis LE, Miller DR, Clark JA, et al. Improving the response choices on the veterans SF-36 health survey role functioning scales: results from the Veterans Health Study. *J Ambul Care Manage*. 2004;27:263–280. [PubMed: 15287216]
10. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical care*. 1992;30:473–483. [PubMed: 1593914]
11. Tarlov AR, Ware JE, Jr, Greenfield S, et al. The medical outcomes study: An application of methods for monitoring the results of medical care. *JAMA*. 1989;262:925–930. [PubMed: 2754793]

12. Ames H, Wilson C, Barnett SD, et al. Does Functional Motor Incomplete (AIS D) Spinal Cord Injury Confer Unanticipated Challenges? *Rehabilitation psychology*. 2017.
13. Goldberg J, Magruder KM, Forsberg CW, et al. The association of PTSD with physical and mental health functioning and disability (VA Cooperative Study #569: the course and consequences of posttraumatic stress disorder in Vietnam-era veteran twins). *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*. 2014;23:1579–1591.
14. Bishawi M, Shroyer AL, Rumsfeld JS, et al. Changes in health-related quality of life in off-pump versus on-pump cardiac surgery: Veterans Affairs Randomized On/Off Bypass trial. *The Annals of thoracic surgery*. 2013;95:1946–1951. [PubMed: 23453761]
15. Turner AP, Kivlahan DR, Haselkorn JK. Exercise and quality of life among people with multiple sclerosis: looking beyond physical functioning to mental health and participation in life. *Archives of physical medicine and rehabilitation*. 2009;90:420–428. [PubMed: 19254606]
16. Craig BM, Reeve BB, Brown PM, et al. US valuation of health outcomes measured using the PROMIS-29. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*. 2014;17:846–853. [PubMed: 25498780]
17. Alcantara J, Ohm J, Alcantara J. The use of PROMIS and the RAND VSQ9 in chiropractic patients receiving care with the Webster Technique. *Complementary therapies in clinical practice*. 2016;23:110–116. [PubMed: 25990687]
18. Beaumont JL, Cella D, Phan AT, et al. Comparison of health-related quality of life in patients with neuroendocrine tumors with quality of life in the general US population. *Pancreas*. 2012;41:461–466. [PubMed: 22422138]
19. Pearman TP, Beaumont JL, Cella D, et al. Health-related quality of life in patients with neuroendocrine tumors: an investigation of treatment type, disease status, and symptom burden. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*. 2016;24:3695–3703. [PubMed: 27029477]
20. Hinchcliff M, Beaumont JL, Thavarajah K, et al. Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item Health Profile and Functional Assessment of Chronic Illness Therapy-Dyspnea short form. *Arthritis care & research*. 2011;63:1620–1628. [PubMed: 22034123]
21. IsHak WW, Pan D, Steiner AJ, et al. Patient-Reported Outcomes of Quality of Life, Functioning, and GI/Psychiatric Symptom Severity in Patients with Inflammatory Bowel Disease (IBD). *Inflammatory bowel diseases*. 2017;23:798–803. [PubMed: 28301432]
22. Katz P, Pedro S, Michaud K. Performance of the PROMIS 29-Item Profile in Rheumatoid Arthritis, Osteoarthritis, Fibromyalgia, and Systemic Lupus Erythematosus. *Arthritis care & research*. 2016.
23. Lai JS, Beaumont JL, Jensen SE, et al. An evaluation of health-related quality of life in patients with systemic lupus erythematosus using PROMIS and Neuro-QoL. *Clinical rheumatology*. 2017;36(3):555–562. [PubMed: 27848056]
24. Schnall R, Liu J, Cho H, et al. A Health-Related Quality-of-Life Measure for Use in Patients with HIV: A Validation Study. *AIDS patient care and STDs*. 2017;31(2):43–48. [PubMed: 28051875]
25. Hays RD, Spritzer KL, Schalet BD, et al. PROMIS-29 v.2.0 Profile physical and mental health summary scores. *Quality of Life Research*. Epub ahead of print. 10.1007/s11136-018-1842-3.
26. Cohen J *Statistical Power Analysis for the Behavioral Sciences*. 2 ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
27. Brown W Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 1904–1920. 1910;3:296–322.
28. Spearman C Correlation calculated from faulty data. *British Journal of Psychology*, 1904–1920. 1910;3:271–295.
29. Green CP, Porter CB, Bresnahan DR, et al. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. *J Am Coll Cardiol*. 2000;35:1245–1255. [PubMed: 10758967]
30. Hung M, Saltzman CL, Greene T, et al. Evaluating instrument responsiveness in joint function: The HOOS JR, the KOOS JR, and the PROMIS PF CAT. *J Orthop Res* 2018;36:1178–1184. [PubMed: 28921658]

Table 1:

Definitions and specifications of the eight candidate performance measures.

Measure Label	Description	Binary/Continuous	VR-12/PROMIS-29	Summary Measure	Physical/Mental HRQoL	Phase
PM1a	Stable or improved PCS	Binary	VR-12 or VR-36 *	PCS	Physical	1, 2
PM1b	Stable or improved MCS	Binary	VR-12 or VR-36	MCS	Mental	1, 2
PM1c	Stable or improved PHS	Binary	PROMIS-29	PHS	Physical	2
PM1d	Stable or improved MHS	Binary	PROMIS-29	MHS	Mental	2
PM2a	Mean change in PCS	Continuous	VR-12 or VR-36	PCS	Physical	1, 2
PM2b	Mean change in MCS	Continuous	VR-12 or VR-36	MCS	Mental	1, 2
PM2c	Mean change in PHS	Continuous	PROMIS-29	PHS	Physical	2
PM2d	Mean change in MHS	Continuous	PROMIS-29	MHS	Mental	2

* VR-12 in Phase 1, VR-36 in Phase 2. Both instruments produce equivalent summary measures (i.e., the PCS and MCS).

PM: Performance measure

PROMIS-29: Patient-Reported Outcomes Measurement Information System, 29-Item Profile Measure

VR-12: Veterans RAND 12-Item Short Form

VR-36: Veterans RAND 36-Item Survey

PCS: the physical health summary score of the VR-12 or the VR-36

MCS: the mental health summary score of the VR-12 or the VR-36

PHS: the physical health summary score of the PROMIS-29

MHS: the mental health summary score of the PROMIS-29

Table 2:

Characteristics of respondents included in the Health Outcomes Survey, cohort 15 (n = 79,972).

Variable	Percent
Age (Mean, SD)	74.7 (6.8)
Age Categories	
65–69	28%
70–74	27%
75–79	21%
80–84	15%
85+	10%
Sex	
Male	39%
Female	61%
Race/Ethnicity	
White non-Hispanic	74%
Hispanic	10%
Non-White, non-Hispanic	12%
Missing	5%
Percentage with Chronic Condition	
Arthritis	66%
Cancer	7%
Chronic Lung Disease	19%
Congestive Heart Failure	10%
Depression	44%
Diabetes	32%
Hypertension	78%
Inflammatory Bowel Disease	6%
Ischemic Heart Disease	21%
Osteoporosis	25%
Other Heart Problems	27%
Sciatica	29%
Stroke	9%
Number of Chronic Conditions	
2	28%
3	26%
4	20%
5+	27%

Table 3:

Performance measures in Health Outcomes Survey, cohort 15. There were 449 Medicare Advantage Organizations, with a mean of 160 respondents per Medicare Advantage Organization.

Performance Measure	Intraclass Correlation (ICC)	Reliability as PM	Number Needed for Reliability of 0.7 (as PM)
PM1a	0.000973	0.144	2399
PM1b	0.002644	0.315	882
PM2a	0.001615	0.219	1445
PM2b	0.000801	0.122	2912

PM: Performance Measure

PM1a: Stable or improved physical component subscale (PCS), the physical health summary score of the Veterans RAND 12-Item Short Form (VR-12)

PM1b: Stable or improved mental component subscale (MCS), the mental health summary score of the Veterans RAND 12-Item Short Form (VR-12)

PM2a: Mean change in physical component subscale (PCS), the physical health summary score of the Veterans RAND 12-Item Short Form (VR-12)

PM2b: Mean change in mental component subscale (MCS), the mental health summary score of the Veterans RAND 12-Item Short Form (VR-12)

Table 4:

Characteristics of respondents included in primary data collection from Kaiser Permanente Colorado (n = 337).

	Percentage (or Mean, SD where noted)
Mean Age, Years (SD)	79.0 (7.2)
Age Groups	
65–69	14%
70–74	20%
75–79	9%
80–84	35%
85+	21%
Sex	
Male	50%
Female	50%
Race/Ethnicity	
White/Non-Hispanic	93%
Hispanic	4%
Non-White/Non-Hispanic	2%
Missing Race/Non-Hispanic	<2% *
Total Number of Chronic Conditions	
2	44%
3	31%
4	14%
5+	11%
Presence of a Specific Chronic Condition	
Arthritis	23%
Cancer	9%
Chronic Lung Disease	38%
Congestive Heart Failure	10%
Depression	22%
Diabetes	31%
Hypertension	80%
Inflammatory Bowel Disease	<2% *
Ischemic Heart Disease	27%
Osteoporosis	18%
Other Heart Problems	31%
Sciatica	6%
Stroke	2%
Summary PROMIS-29 Scores (Mean, SD)	

	Percentage (or Mean, SD where noted)
PHS	43.8 (8.1)
MHS	50.4 (7.6)
Summary VR-36 Scores (Mean, SD)	
PCS	39.7 (10.5)
MCS	55.9 (9.2)

*
Cell suppressed due to small numbers.

PROMIS-29: Patient-Reported Outcomes Measurement Information System, 29-Item Profile Measure

VR-36: Veterans RAND 36-Item Survey

PCS: the physical health summary score of the VR-36

MCS: the mental health summary score of the VR-36

PHS: the physical health summary score of the PROMIS-29

MHS: the mental health summary score of the PROMIS-29

Table 5:

Candidate performance measures drawn from 13 ambulatory care clinics within Kaiser Permanente Colorado (mean respondents per clinic: 27).

Performance Measure (PM)	Intraclass Correlation (ICC)	Reliability as PM	Number Needed for Reliability of 0.7 (as PM)
PM1c (Unchanged or improved)	0	0	n/a
PM1c (Not lower by more than ¼ SD)	0.007	0.16	341
PM1c (Not lower by more than ½ SD)	0	0	n/a
PM1c (Not lower by more than 1 SD)	0	0	n/a
PM1d (Unchanged or improved)	0	0	n/a
PM1d (Not lower by more than ¼ SD)	0	0	n/a
PM1d (Not lower by more than ½ SD)	0	0	n/a
PM1d (Not lower by more than 1 SD)	0	0	n/a
PM2c	0.001	0.03	2317
PM2d	0	0	n/a

n/a: Not applicable

PM1c: Stable or improved physical health score (PHS), the physical health summary score of the Patient-Reported Outcomes Measurement Information System 29-Item Profile Measure (PROMIS-29)

PM1d: Stable or improved mental health score (MHS), the mental health summary score of the Patient-Reported Outcomes Measurement Information System 29-Item Profile Measure (PROMIS-29)

PM2c: Mean change in physical health score (PHS), the physical health summary score of the Patient-Reported Outcomes Measurement Information System 29-Item Profile Measure (PROMIS-29)

PM2d: Mean change in mental health score (MHS), the mental health summary score of the Patient-Reported Outcomes Measurement Information System 29-Item Profile Measure (PROMIS-29)