Research Report

Patient-Reported Outcome–Based Performance Measures for Older Adults with Multiple Chronic Conditions

Maria Orlando Edelen, Adam J. Rose, Elizabeth Bayliss, Lesley Baseman, Emily Butcher, Rosa-Elena Garcia, David Tabano, Brian D. Stucky

For more information on this publication, visit www.rand.org/t/RR2176

This study was conducted in response to contract HHSN271201500064C NIH NIA (AG).

Published by the RAND Corporation, Santa Monica, Calif.

Preface

Ensuring patients' survival and curing disease are no longer the sole goals of patient care; functional status and well-being must also be optimized. Standardized assessments through patient-reported outcome (PRO) performance measures (PMs) allow evaluators to see how well these aspects of care are being delivered and compare the performance of health care systems and different provider groups.

To meaningfully assess health-related quality of life, PMs must align with patients' goals. Most PMs in current use assess technical quality of care, the benefit of which can take years to realize. These sorts of measures might not be particularly salient to older adults (especially age 80 and older) with multiple chronic conditions (MCC) because their expectations for long-term survival are limited. The desire of such patients to maximize quality of life for the remainder of their life spans is obviously undiminished. However, no PMs have been formally developed or validated for use in this population. Therefore, policymakers and researchers who focus on quality of health care are especially interested in developing PRO-based PMs to measure the care delivered to older adults with MCC.

This project aimed to address this gap in validated PRO-based PMs through the completion of two aims. The first aim was to determine whether the Patient-Reported Outcomes Measurement Information System 29-item (PROMIS-29) profile instrument is valid for use in this population. The second aim was to study how well PMs based on PROs, specifically the Veterans RAND 36 Item Health Survey (VR-36) and the PROMIS-29, measure functional status and well-being in the same population. This report summarizes our analyses and provides recommendations for future efforts to develop and validate PRO-based PMs in similar populations.

This study was conducted in response to a contract between the National Institutes of Health and the RAND Corporation (prime)/Kaiser Permanente Colorado and Boston University School of Public Health (subcontractors) for the project titled "Outcome Performance Measure Development for Persons with Multiple Chronic Conditions (MCC)." Results herein are reported for phase 2, year 2 of the project.

Contents

Tables

Summary

As efforts to measure health care quality have become more sophisticated, the goals of patient care have extended beyond ensuring survival and curing disease to helping patients optimize their functional status and well-being. Assessments that use patient-reported outcome (PRO) performance measures (PMs) can measure how well these aspects of care are being delivered and compare the performance of health care systems and different provider groups. Most PMs focus on technical quality of care or such outcomes as survival. For older adults, especially those over the age of 80 with multiple chronic conditions (MCC), it might be equally important or even more important to have a good quality of life. Therefore, policymakers and researchers have been particularly interested in designing PMs that reflect the goals of these patients. To date, no PRO-based PMs have been formally developed or validated specifically for use in older adults with MCC.

We tested PMs that were based on two prominent instruments for assessing health-related quality of life (HRQoL): the Veterans RAND 36 Item Health Survey (VR-36) (which, despite its name, is also widely used in nonveteran populations) and the Patient-Reported Outcomes Measurement Information System 29-item (PROMIS-29) profile instrument. The PROMIS-29 is in widespread use but has undergone limited validation in a geriatric population with MCC. There were two main aims of the study: first, to validate the PROMIS-29 in this population, and second, to develop a better understanding of the practical use of PRO-based PMs in a geriatric population. To this end, we assessed the performance of PMs based on serial administration of the VR-36 or PROMIS-29, specifically in the MCC population that we studied.

We surveyed members of Kaiser Permanente Colorado (KPCO), a not-for-profit integrated delivery system. All participants were age 65 or older, and we oversampled those age 80 and older. Each participant had at least two of 13 specific chronic conditions: arthritis, cancer, chronic lung disease, congestive heart failure, depression, diabetes, hypertension, inflammatory bowel disease, ischemic heart disease, osteoporosis, other heart problems, sciatica, or stroke. We identified participants meeting eligibility criteria using KPCO clinical data, and we randomized participants to respond to the survey via the web, telephone, or mail mode. We randomly selected a subset of those who had not responded to the web mode after three weeks to complete the baseline survey by a different mode (mail or telephone). Six months later, we randomly selected 400 baseline web respondents for a second web survey.

PROMIS-29 scores were distributed in expected ways based on such characteristics as age, sex, and number of chronic conditions. After an exhaustive search, we found no evidence of differential item functioning in the sample. Case mix–adjusted PROMIS-29

scores demonstrated convergent construct validity with utilization of primary care (a marker for illness burden) and with scores from the VR-36.

We constructed eight PRO-based PMs, based on four HRQoL scores: the two summary scores of the VR-36 (physical component score and mental component score) and two summary scales of the PROMIS-29 (Physical Health and Mental Health) (see HealthMeasures, undated). Each of these four scores was developed into two different PMs: one for the proportion of patients alive after six months with a stable or improved score, and the other for the mean change in the score at each site of care being profiled. These eight PMs generally performed poorly in terms of distinguishing among the 13 KPCO ambulatory clinics that we profiled. These clinics averaged 26 respondents per clinic. It is generally agreed that PMs must have a reliability of at least 0.7 to be acceptable for use; none of the PMs that we examined achieved a reliability above 0.13. According to the data that we collected, most PMs did not generate measurable reliability under these conditions. A few PMs would have required several thousand observations per entity, which is theoretically possible but still infeasible.

We conclude that scores on the PROMIS-29 are valid for use in an older population with MCC. Using data collected at baseline and a six-month follow-up, PMs based on the VR-36 or PROMIS-29 would have required infeasible or impossible sample sizes to achieve acceptable reliability. Efforts to develop PRO-based PMs for general use should continue because this is such an important goal. However, our findings do not support the general use of these PMs without further refinement and testing. Our results suggest that a major limitation of the approach taken here was the small amount of change in HRQoL that can be expected over a six-month interval. Future efforts to develop PRO-based PMs should focus on changes in HRQoL over longer follow-up intervals.

Three main policy-relevant implications emerge from this work:

1. It is feasible to collect HRQoL data among older adults with MCC in an integrated health system.
2. PROMIS-29 is valid for use in older adults with MCC.
3. Future efforts to develop PRO-based PMs require longer follow-up intervals and larger samples than we used.

Acknowledgments

We thank the RAND Survey Research Group staff who so ably coordinated the survey distribution and successfully recruited and interviewed survey respondents. We also appreciate the valuable insights we received from Karon F. Cook and Justin W. Timbie during RAND's quality review process. We thank James Rothendler and Lewis Kazis of the Boston University School of Public Health (the lead investigators for phase 1 of this contract) for their input on the research design and on this report. Finally, we thank our contracting officer, Marcel Salive, for his continued advice and support throughout the project.

Abbreviations

AOR	adjusted odds ratio
DIF	differential item function
HOS	Medicare Health Outcomes Survey
HRQoL	health-related quality of life
ICC	intraclass correlation coefficient
ICD-10	International Classification of Diseases, Tenth Revision, Clinical Modification
KPCO	Kaiser Permanente Colorado
MAO	Medicare Advantage Organization
MCC	multiple chronic conditions
MCS	mental component score
NQF	National Quality Forum
PCS	physical component score
PM	performance measure
PRO	patient-reported outcome
PROMIS-29	Patient-Reported Outcomes Measurement Information System 29-Item Survey
SD	standard deviation
SES	socioeconomic status
SF-36	36-Item Short-Form
SRG	Survey Research Group
VDW	virtual data warehouse
VR-36	Veterans RAND 36 Item Health Survey

Chapter One. Background

Adults age 65 and older with multiple chronic conditions (MCC) often have intensive health care needs that require complex care coordination. The specific needs of this population are often not reflected in current performance measures (PMs). Further, many commonly used PMs are based either on outcomes that might not necessarily be important to people with MCC or on processes of care that might affect outcomes only in the long run. For example, many PMs assess processes of care that affect health outcomes over the course of years or even decades. The relevance of such measures to an older adult with MCC might be questionable. In contrast, health outcomes based on patient reports, known as patient-reported outcomes (PROs), are inherently important to everyone, including those with MCC. The most commonly measured PROs relate to health-related quality of life (HRQoL), a multifaceted construct that is usually assumed to contain at least two components: physical health and mental health. The importance of HRQoL does not wane with age, nor is it less relevant in light of limited life expectancy or the presence of MCC.

Given the inherent importance of PROs for older adults with MCC, using them as PMs for this population is an attractive idea. However, we need to know more about the operating characteristics of PMs based on PROs before they can be used for performance profiling. The work described in this report is part of a larger, two-part contract with the National Institute on Aging to develop and validate PRO-based PMs for use with older adults with MCC. The first phase of the contract, described in more detail in the next paragraph, used a retrospective database to examine the suitability of PMs that are based on change in HRQoL in older adults with MCC (Kazis, Rogers, et al., 2017). This second phase of the contract used prospectively collected data to achieve two main aims: (1) to assess the performance of two widely used HRQoL instruments in a population of older adults with MCC and (2) to evaluate the psychometric properties of PMs that are based on PROs.

For the first aim, we looked at the performance of the Patient-Reported Outcomes Measurement Information System 29-Item (PROMIS-29) health survey in this population of older adults with MCC, a population for whom the PROMIS-29 has received only limited previous validation. Analyses included (1) examining how PROMIS-29 scores vary across key patient characteristics, such as MCC diagnoses and health care utilization; and (2) formally testing for differential item functioning (DIF) of PROMIS-29 items in this population, especially according to mode of administration (mail, phone, or web).

For the second aim, we examined the psychometric properties of PMs that are based on the PROMIS-29 and the Veterans RAND 36-Item Health Survey (VR-36) in this population, including the ability of such PMs to meaningfully profile the performance of ambulatory clinics over time. The long-term goal of this work, which is beyond the immediate scope of

this contract, was to provide documentation for an eventual application for National Quality Forum (NQF) endorsement of PMs based on the PROMIS-29 or the VR-36, should the reliability of such PMs prove sufficient to support an application.

Potential Uses of Performance Measures

Different types of PMs exist, each with different uses. Here, we briefly explain which PMs we considered in this study and how they are being used. PMs can be categorized based on their *place in the structure–process–outcome model* for quality of health care (Donabedian, 1988). This model posits that the setting and resources available form the context within which care is delivered (*structure*), which, in turn, can influence providers' ability to perform specific actions on the patient's behalf, particularly those that are concordant with the best available evidence (*process*). The effect of better process of care is measured in the form of improved *outcomes* of care. Outcomes themselves can come in different forms and can include both *definitive outcomes* (such as mortality or amputation) that are intrinsically important to patients, as well as *intermediate outcomes* (such as blood pressure control), whose importance is due to a demonstrated link to definitive outcomes. This study considered PROs, which we measure in the form of HRQoL. PROs reflect a patient's own report about his or her HRQoL, and they are a form of definitive outcome because they intrinsically matter to the patient.

PMs can also be categorized based on the *level at which they are intended to function*. PMs might profile the delivery of health care at different levels of the system, including the individual health care provider, clinic or hospital, and health plan. In the present study, we used PMs to profile performance among 13 ambulatory clinics that were part of Kaiser Permanente Colorado (KPCO). The level at which measurement is taking place has implications not only for sample size and power but also for the locus of control (i.e., an individual provider might be assumed to have more control over the outcomes of his patients than exists at the clinic or particularly the health plan level).

Lastly, another way to categorize PMs is by their *intended use*. The most limited use for a PM is to use it for monitoring the adequacy of care for individual patients and possibly to alert their providers if care is not concordant with what is considered best practice (this is most relevant for process-of-care measures). A more expansive use for a PM, suitable for either process or outcome measures, would be to use it to compare the performance of different entities, as we did in this study, comparing PROs among 13 ambulatory clinics. An even more expansive use would be to make the results of such a comparison publicly available, often in hopes of influencing patients or consumers to choose better-performing providers, clinics, or health plans. The most expansive use of a PM would be to attach tangible rewards and penalties to these rankings, such as performance payments. As a rule of thumb, the more expansive the intended use of a PM, the more that PM should be thoroughly

vetted before it is introduced, in part because of the potential for unintended consequences from a poorly designed or poorly vetted PM.

Previous Testing and Use of the Veterans RAND 36 Item Health Survey

The VR-36 is a widely used measure of HRQoL both for veteran and for nonveteran populations. It assesses quality of life across eight domains: (1) limitations in physical activities because of health problems, (2) limitations in social activities because of physical or emotional problems, (3) limitations in usual-role activities because of physical health problems, (4) bodily pain, (5) general mental health (psychological distress and well-being), (6) limitations in usual-role activities because of emotional problems, (7) vitality (energy and fatigue), and (8) general health perceptions. The VR-36 also has two summary scales, both of which draw on all eight domains to some extent: the physical component score (PCS) and the mental component score (MCS) (Ware and Sherbourne, 1992; Tarlov et al., 1989). These are generally felt to represent physical and mental HRQoL, respectively. The VR-36 is a modification of the 36-Item Short Form Survey (SF-36), which was developed in response to findings from the Medical Outcomes Study (Ware and Sherbourne, 1992; Tarlov et al., 1989). To reduce ceiling and floor effects and to increase the items' explanatory power, the response scale for the two role functioning items was changed to a five-point scale instead of a dichotomized scale. This change in the scale is the only difference between the two instruments but an important one (Kazis, Miller, Clark, Skinner, Lee, Ren, et al., 2004).

The VR-36 has undergone rigorous psychometric testing that has proven that it is a reliable and valid scale for use with veteran and nonveteran populations (Kazis, Miller, Clark, Skinner, Lee, Ren, et al., 2004). It was originally developed for veterans and has continued to be used in veteran populations with a diverse set of conditions, including spinal cord injury, posttraumatic stress disorder (Goldberg et al., 2014), coronary disease (Bishawi et al., 2013), and multiple sclerosis (Turner, Kivlahan, and Haselkorn, 2009). It has also been included as part of the Healthcare Effectiveness Data and Information Set (HEDIS) performance measures since 2006 (National Committee for Quality Assurance, 2006).

Previous Testing and Use of the Patient-Reported Outcomes Measurement Information System 29-Item Survey

The PROMIS-29 is also a widely used and well-validated assessment of HRQoL. It is meant to be an efficient means of assessing a broad range of HRQoL, providing a comprehensive assessment of a patient's overall quality of life. By covering a broad spectrum of health issues, the PROMIS-29 was designed to be used with adults with any disease or condition. The instrument includes four items from each of the seven Patient-Reported Outcomes Measurement Information System (PROMIS) domains (Anxiety, Depression,

Fatigue, Pain Interference, Physical Function, Sleep Disturbance, and Ability to Participate in Social Roles and Activities), as well as an additional Pain Intensity item. Currently, the PROMIS-29 yields eight distinct scores, one for each domain represented in the profile.

The PROMIS-29 has been tested and validated in a variety of patient populations. In addition to deployment with a representative sample of the U.S. population (Craig et al., 2014), the instrument has successfully measured HRQoL in chiropractic patients (Alcantara, Ohm, and Alcantara, 2016) and people with neuroendocrine tumors (Beaumont et al., 2012; Pearman et al., 2016), systemic sclerosis (Hinchcliff et al., 2011), irritable bowel syndrome (IsHak et al., 2017), rheumatoid arthritis and osteoarthritis (Katz, Pedro, and Michaud, 2017), systemic lupus erythematosus (Lai et al., 2017), and human immunodeficiency virus (HIV) (Schnall et al., 2017), among other populations. In prior testing, the PROMIS-29 has also been administered in a variety of settings (e.g., at the patient's home or in a medical clinic) and by a variety of methods (e.g., online or via paper and pencil).

Comparison of the Veterans RAND 36 Item Health Survey and the Patient-Reported Outcomes Measurement Information System 29-Item Survey

Because certain parts of this report compare scores on the VR-36 and the PROMIS-29, we briefly contrast their salient features here. The VR-36 is based on the SF-36, which was developed in the 1980s as part of the Medical Outcomes Study (Ware and Sherbourne, 1992; Tarlov et al., 1989). The PROMIS-29 can be thought of as a "modern" version of the SF-36 or the VR-36. It was developed using modern measurement theory, known as item response theory, and was calibrated and scored based on more-contemporary samples (post-2000). Because both instruments are measuring a common construct (namely, HRQoL), there is considerable overlap. However, by design, the PROMIS-29 contains some new content, such as a scale about sleep disturbance, which is not directly measured in the SF-36 or the VR-36. The PROMIS-29 is also distinguished by having different proportions of individual survey items making up each of its dimensions and having been calibrated using a more modern calibration method. Crosswalks exist to compare or coscore the two instruments but not specifically for a population of older adults with MCC.

Summary of Phase 1

We completed phase 1 of the contract (Kazis, Rogers, et al., 2017) with existing data, whereas we completed phase 2 (the focus of the present report) with prospectively collected data. For context, we summarize phase 1 activities and results here.

Phase 1 analyses used data from the Medicare Health Outcomes Survey (HOS), which collected the VR-12 (a briefer version of the VR-36) between 2006 and 2014 (Dalewitz, Khan,

and Hershey, 2000; Health Services Advisory Group, 2017). We examined the ability of PRO-based PMs to profile Medicare Advantage plans. We constructed four plan-level PMs based on the PCSs and MCSs according to the percentage of people alive at two years with stable or improved (1) PCSs or (2) MCSs and the average change over two years in (3) PCSs or (4) MCSs. We evaluated the ability of these four PMs to profile Medicare Advantage Organizations (MAOs), including their ability to support meaningful and valid comparisons between organizations. We used the metrics of intraclass correlation coefficient (ICC) and reliability at the group level to assess the PMs' potential to distinguish levels of performance.

We found that, in the context of differentiating MAOs, PMs based on the MCS of the VR-12 performed better than PMs based on the PCS. However, for the four basic PMs applied to the MAO samples, reliabilities for both PCS- and MCS-based PMs were low. Therefore, we explored several options for improving the ability to differentiate MAO performance, including examining how specific subsets of the population and threshold effects might distinguish low-performing MAOs from others. We also examined the effect of maximizing sample sizes by combining Medicare HOS cohorts. Despite these maneuvers, none of the PMs we examined achieved the criterion threshold of 0.7 that is considered sufficient reliability for a PM (Sequist et al., 2011), although some came close, particularly measures based on the MCS.

Major Activities, Phase 2

In contrast to phase 1, which used existing data, in phase 2, we collected primary data from a cohort of older adults with MCC. The survey instrument included both the VR-36 and the PROMIS-29, allowing direct comparison and cross-validation between the two measures. The major activities of this phase of the contract were as follows: First, we developed a survey that included the VR-36, the PROMIS-29, a question about financial difficulty due to the expense of health care, and a question about proxy response ("Did someone help you to complete this survey?"). Second, we identified eligible participants from KPCO, a not-for-profit integrated delivery system. Third, we collected a first round of survey data from these participants using three different survey modalities (web, mail, and telephone). Fourth, we collected a second round of survey data approximately six months later from participants who had responded to the first round's web-based survey. Fifth, we performed analyses of the data collected to generate basic descriptive statistics for the VR-36 and the PROMIS-29 and assess PROMIS-29 items for DIF, especially according to survey mode, in this population. Sixth, we used longitudinal data from the subset of participants who responded to the survey at both time points to examine the properties of the VR-36 and the PROMIS-29 when used as PMs in this population.

Organization of This Report

The remainder of this report is organized as follows:

- Chapter Two presents the data collection methods.
- Chapter Three presents the analytic approach and results for analyses related to the validation of items from the PROMIS-29, especially tests for DIF in this population.
- Chapter Four presents the analytic approach and results for analyses related to the operating characteristics of VR-36 and PROMIS-29 scales when used as PMs in this population of community-dwelling older adults with MCC.
- Chapter Five includes a discussion of the results, a summary and conclusion, and policy implications of the findings.

Chapter Two. Data Collection Methods

Setting

We recruited participants from KPCO, which directly provides primary and specialty care, including both ambulatory and hospital-based care. Care received outside the integrated delivery system, such as emergency care, is documented through claims made to KPCO. As a result, clinical data for KPCO members encompass essentially all of the care that they receive. Data on KPCO study participants were derived from the virtual data warehouse (VDW), a standardized quality-controlled secondary observational database that incorporates data from the KPCO electronic health record, pharmacy fills, claims, demographic, membership, and administrative systems. Participants were assigned to primary care physicians at one of 13 KPCO clinics in the Denver metropolitan area. We obtained approvals from the institutional review boards of both KPCO and the RAND Corporation.

Eligibility and Identification of Participants

A KPCO member was eligible to participate if he or she was age 65 or older, was actively enrolled in KPCO, was assigned to a primary care provider at a KPCO ambulatory clinic (which demonstrates that he or she was not permanently residing at a skilled nursing facility), had been seen for clinical care at least once in the past 12 months, had a valid email address, and had at least two of 13 specific chronic conditions. These conditions were

- arthritis
- cancer
- chronic lung disease
- congestive heart failure
- depression
- diabetes
- hypertension
- inflammatory bowel disease
- ischemic heart disease
- osteoporosis
- other heart problems
- sciatica
- stroke.

Each condition was defined using a list of International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes, which are in Appendix A. Next we describe other study eligibility criteria.

Developing an appropriate list of codes for the present study required some work by our study team. We started with a set of codes developed by Hude Quan et al., which were based on those that Anne Elixhauser originally established for risk adjustment of chronic conditions (Quan et al., 2005). We further modified the Quan set of codes in consultation with clinically trained members of our team and other practicing clinicians at the RAND Corporation.

We modified the Quan set of codes in two main ways. First, we added codes to capture some conditions that the Quan paper had not included. The purpose of risk-adjustment models, such as the Elixhauser/Quan model, is to detect chronic conditions that drive hospitalizations, costs, or mortality. Our purpose here, in contrast, was to detect conditions that would have an important impact on HRQoL. We therefore added codes to the Quan set to capture certain conditions that can affect HRQoL but are not a major cause of hospitalization or death, such as osteoarthritis, osteoporosis, and sciatica.

Second, because the spectrum of diseases that we needed to detect differed from that in the Quan model, we added codes to extend that spectrum for some conditions. Quan was appropriately interested in detecting only the most-severe manifestations of disease because these tend to drive morbidity and mortality. For our study, we were also interested in capturing less severe manifestations of disease because those can affect HRQoL as well. For ischemic heart disease, for example, we added the codes for angina pectoris to the codes that Quan had used to capture more-severe manifestations of disease, such as myocardial infarction. We reasoned that, although angina pectoris is less severe, it could still meaningfully affect HRQoL.

Survey Design

We constructed the survey mainly from two previously validated instruments, the VR-36 and the PROMIS-29. We described the history and previous validation of these instruments earlier. In addition to the VR-36 and the PROMIS-29, we included two questions: one about the extent to which health-related expenses have caused financial problems for the respondent or his or her family and another asking about *proxy response* (a concept that includes a range of options from the proxy assisting with recording through responding entirely on the subject's behalf).

We included the financial problem question because it captured additional information about financial status that could not otherwise be obtained from KPCO automated data or from the HRQoL surveys. We used this as a stratifying variable (see the discussion of question 4 in Chapter Three) and as part of case mix–adjustment models (e.g., for the DIF analyses). We included the proxy response item because, although we recognize the importance of allowing proxy responses in this population, we also felt that it would be important to record which respondents had help from a proxy. Proxy response is known to affect responses to HRQoL instruments in certain ways, so we needed to record it to be able

to adjust for it (Corder, Woodbury, and Manton, 1996; Hays et al., 1995; Pickard and Knight, 2005). We reproduce the full survey instrument as Appendix B. We also administered the same survey six months after the first, without changes.

Development of Patient-Reported Outcomes Measurement Information System 29-Item Summary Scores

In addition to the eight usual scores for the PROMIS-29 (Anxiety, Depression, Fatigue, Pain Intensity, Pain Interference, Physical Function, Sleep Disturbance, and Ability to Participate in Social Roles and Activities), we used two summary scales for the PROMIS-29, which we call Physical Health and Mental Health. The intent of these two scores was to allow the examination of a smaller number of PMs (based on PCS, MCS, PROMIS-29 Physical Health, and PROMIS-29 Mental Health), as opposed to the much larger number of PMs that would have been required if based on the PCS, the MCS, and the eight scores of the PROMIS-29. In addition, we expected that these summary scores would be more stable because they contain contributions from more survey items and would be less subject to type I errors (erroneously finding a difference that is not truly present, in some cases because of a large number of comparisons).

Although the development of the two summary scores for the PROMIS-29 was not formally part of this project, we describe the process here. We conducted our work to develop these scores in close collaboration with PROMIS researchers, and Ron D. Hays, Ph.D., led analyses using data from a general population. We worked iteratively with Hays and his team to ensure that the final factor solution was suitable for our sample.

We conducted exploratory factor analysis (EFA) of the PROMIS-29 scales and the pain intensity items using the promax oblique rotation and maximum likelihood estimation methods. We used eigenvalues, scree plot, and proportion of variance explained as guidelines to determine the number of factors to extract from the eight PROMIS-29 domains. We also tested alternatives of combining some of the variables to create a reduced set of six items. For example, we tried combining the two pain items, and we tried combining depression and anxiety into a single variable that represents emotional distress. Using the rotated factor loading pattern, we examined the interpretability of the EFA results.

When the criteria of the number of factors extracted were satisfied and the results were interpretable, we proceeded to confirmatory factor analysis (CFA) to examine model fit, based on the same loading pattern that we extracted from EFA results. We confirmed the final two-factor CFA model using data from our MCC sample, as well as data from a sample representing the general population. This final solution was a two-factor solution in which the factors were correlated; these factors represented physical and mental HRQoL, respectively. We arrived at the solution after combining the two pain scales and after

combining the anxiety and depression items because we noticed that they were collinear and that the factor solution converged better with them combined.

Once he had confirmed them across samples, Hays generated the scoring coefficients based on the CFA solution from the general population data. This allowed for clearly interpretable profile scores relative to known general population averages for these similar to interpretation of scores for the PCS, MCS, and PROMIS-29 domains. We examined the correlations between these PROMIS-29 summary scores and other measures (to address question 6 in Chapter Three), especially the PCS and MCS, to give a sense of what these PROMIS summary scores represent.

Survey Administration

Using KPCO VDW data (as described above), we identified 3,749 participants who met the eligibility criteria. We randomly assigned participants to different survey modes, with the same survey instrument administered across all modes. Initially, we assigned 2,764 participants to the web survey, 376 to the mail survey, and 372 to the phone survey. We based the number assigned to the mail and phone surveys on previous experience by the RAND Survey Research Group (SRG) with similar surveys, as well as the final number of responses we wanted to gather. We assigned a much larger number to the web survey, with the expectation that the response rate would be lower, and using a conservative assumption for what the response rate might be.

Participants who had not responded to the web-based survey after four weeks were then eligible to be reassigned to complete the survey either via mail or phone. We invited a randomly selected subset of the web nonresponders (384) to complete the survey by mail (web–mail group) and another subset (382) to complete the survey by phone (web–phone group). We anticipated that some of the participants would require help from a proxy, such as a family member, to complete the survey, so we allowed proxy responses for all modes.

KPCO identified the sample through queries of VDW using SAS 9.4. The study eligibility criteria were

- The participant had to have made at least one clinic visit during the prior 12 months.
- The participant had to have at least two chronic conditions as specified above.
- The participant had to be age 65 or older.
- The participant had to have a valid email address.
- The participant did not require a language interpreter, according to KPCO records (i.e., spoke English).

Because the oldest old were a special group of interest for this study, we oversampled participants from the age 80–and-older category. KPCO sent an email invitation to each eligible patient with an information sheet that introduced and explained the purpose of the

survey. At that point, we gave participants two weeks to opt out, after which we presumed them to be eligible for initial contact. However, they could opt out of survey completion at any time. Prior to sending the email invitation, KPCO performed one final check to remove any deceased respondents from the list. The final sample contact file contained the following variables for each participant: name, address, phone number, email address, gender, and age.

KPCO then sent RAND the sample file, which contained a total of 3,749 potential participants. SRG conducted all survey administration and data collection activities, including hosting the web-based survey. We sent an email invitation to each person assigned to the web survey, followed by weekly email reminders for three weeks, if the participant had not completed the survey. We sent reminders to participants who only partially completed the web-based survey, asking them to finish the surveys they had started. Anyone invited to complete the survey by mail (the mail and web–mail groups) received an initial packet that included a personalized invitation letter, information sheet, mail survey, and prepaid business reply envelope. Four and six weeks after the first survey mailing, we sent nonresponders materials identical to those in the first packet. Telephone center staff contacted each respondent invited to complete the survey by phone (the phone and web–phone groups) an average of five times. We made calls on weekdays and weekends in morning, afternoon, and evening, between the hours of 9:00 a.m. and 8:00 p.m. respondent time. We also noted and called back at times that respondents or their family members identified as good times to reach them.

The second wave of data collection began six months after the first. It included participants who had completed the baseline web-based survey and who had agreed to be contacted again for the six-month survey. We sent a total of 401 qualifying participants cover letters with the uniform resource locator and link and an email invitation that included information about the survey. Informed by our experience with the baseline web-based survey, we emailed reminders to nonresponders on days 2, 4, 7, 14, and 21 after the first email. Each follow-up survey participant received $30 for completing the second survey. Table 2.1 reports the dates during which we fielded the survey.

Table 2.1. Dates of Fielding the Survey, by Mode

Survey Mode	Start Date	Mode Reassigned	End Date
Web only	November 11, 2016	N/A	February 4, 2017
Web–phone	November 11, 2016	December 6, 2016	March 6, 2017
Web–mail	November 11, 2016	December 6, 2016	February 4, 2017
Phone only	November 13, 2016	N/A	March 6, 2017
Mail only	November 14, 2016	N/A	February 4, 2017
Web only (second round)	April 14, 2017	N/A	May 11, 2017

NOTE: N/A = not applicable.

Collection of Automated Data

We also queried KPCO VDW for each person who had been sent an invitation to participate. This enabled us to supplement self-report information from the survey with information that was both (1) more detailed than would have been possible for many respondents to report and (2) collected without adding to respondent burden. This separate data source also enabled us to compare survey respondents with nonrespondents on some of their characteristics to assess representativeness. Data obtained from VDW included the following variables: age, sex, race and ethnicity, percentage of residents below poverty in the neighborhood (U.S. census tract), chronic health conditions, and number of home health visits, primary and specialty care visits, and hospitalizations. We based decisions regarding how to divide variables into categories on their univariate distribution.

Age

We assessed participant age as of the day we began the survey. As discussed above, all participants were age 65 or older, with oversampling of those age 80 and above. For analysis, we divided age into the following categories: 65–69, 70–74, 75–79, 80–84, and 85+.

Sex

We characterized participant sex as male or female.

Race and Ethnicity

In VDW, race is characterized as white, black, Asian, Native American, and other. Missingness for race, meaning that the information is not recorded, is approximately 3 percent. Ethnicity is characterized as Hispanic or non-Hispanic. Because a majority of our population was white and non-Hispanic and several other categories were relatively uncommon, we created four merged categories that would each contain sufficient numbers to support analyses: white non-Hispanic, nonwhite and non-Hispanic, Hispanic (any race), and missing-race non-Hispanic.

Percentage Below Poverty in the Census Tract

Although controlling for household income in analyses would be ideal, doing so is often impractical because such data are not always collected. A large body of research has shown, however, that knowing where a person lives and linking to U.S. census data can be almost as good in many cases—a concept known as *area socioeconomic status* (area SES). We assessed area SES using the census tract of residence, a highly specific level of geocoding that tends to be associated with an extremely homogeneous SES. In accordance with best practices, we linked the census tract with the percentage of households below the federal poverty threshold from the American Community Survey 2010–2015 estimates (Krieger et

al., 2005; see also Office of the Assistant Secretary for Planning and Evaluation, undated). The percentage of households below the FPL has been shown to be the best single measure of area SES (Krieger et al., 2005). We divided this variable into the following categories: 0–9.99 percent of households below the threshold, 10–19.99 percent of households, and 20 percent or more.

Chronic Health Conditions

We characterized each participant regarding whether he or she had each of the 13 chronic conditions used to define our sample. The definition for each chronic condition and the International Classification of Diseases code used to define it are the same as discussed earlier under the eligibility criteria (see "Eligibility and Identification of Participants," earlier in this chapter; see also Appendix A).

Number of Chronic Conditions

We characterized each participant based on how many of the 13 chronic conditions he or she had. By definition, every participant had at least two. The levels of this variable were two conditions, three conditions, four conditions, and five or more conditions.

Home Health Encounters

We characterized each participant as having had any home health encounters in the past 12 months or having had no such encounters.

Primary Care Visits

We characterized each participant regarding his or her number of primary care visits within the past 12 months. We divided this variable into the following categories: zero to three, four to six, seven to nine, and ten or more visits.

Specialty Care Visits

We characterized each participant regarding his or her number of specialty care visits within the past 12 months. We divided this variable into the following categories: zero to three, four to six, seven to nine, and ten or more visits.

Hospitalizations

We characterized each participant regarding how many times he or she had been hospitalized in the past 12 months. We divided this variable into the following categories: never, one time, and two or more times.

Chapter Three. Analytic Methods and Results: Validation of Items

Key Questions

Our analyses of the first round of survey data were intended to answer several key questions. We have organized this chapter according to these six key questions:

1. What are the characteristics of the baseline respondents?
2. Does the likelihood of responding to the survey vary by subject characteristics?
3. Do HRQoL scores vary by survey mode, both before and after case mix adjustment?
4. Do HRQoL scores vary according to respondent characteristics, such as demographics and comorbid conditions?
5. Is there DIF for items of the PROMIS-29 across survey modes or respondent-level characteristics?
6. Do we find construct validity linking risk-adjusted PROMIS-29 scores with health outcomes, such as utilization of primary care visits or hospitalizations?

Question 1: What Are the Characteristics of the Baseline Respondents?

A total of 1,359 participants responded to the baseline survey; Table 3.1 shows the response rates by survey mode. The web mode was noteworthy in that 37 percent of those who began to respond to the web-based survey (289 out of 490) ultimately abandoned the effort before completing enough of the survey to support including them in the HRQoL analyses, which we defined by having enough responses to calculate at least one PROMIS-29 or VR-36 scale (e.g., anxiety, vitality). With the phone mode, the interviewer made sure the respondent completed all items; with the mail mode, it was also uncommon for the respondent to return a survey without enough information to support analysis. Ultimately, the response rate with sufficient information was lowest for the web mode (25 percent); this was about 50 percent for the other modes. However, in the second round of the survey, the response rate was 88 percent, among a group of participants who all had completed all of a web survey in the first round, with 84 percent submitting sufficient data for analysis.

Table 3.1. Survey Response, by Mode

Response	Survey Round 1: Web Only	Survey Round 1: Mail Only	Survey Round 1: Web–Mail	Survey Round 1: Phone Only	Survey Round 1: Web–Phone	Survey Round 2: Web Only
Invited	1,996	376	384	609	384	401
Responded	779	219	168	305	176	352
Percentage who responded	39	58	44	50	46	88
Sufficient for analysis	490	219	168	305	176	335
Percentage of responses that were sufficient for analysis	63	100	100	100	100	95
Response rate with sufficient information for analysis, as a percentage	25	58	44	50	46	84

We examined the respondents' demographic and clinical characteristics (Table 3.2). The mean age of the overall sample was 80.7 years. Web respondents were somewhat younger (79.3), and mail and phone respondents were somewhat older (81.4 and 81.6, respectively). A majority of respondents (89 percent) were of non-Hispanic white race and ethnicity. Most respondents had either two (35 percent) or three (31 percent) of the 13 chronic conditions we studied, while 18 percent had four and 16 percent had five or more. Some specific chronic conditions were uncommon, such as inflammatory bowel disease (1 percent), whereas others were comparatively common (hypertension, 82 percent). A considerable number of respondents (46 percent) had received at least some home health care in the preceding 12 months. However, only 13 percent had been hospitalized at least once during that period.

Table 3.2. Demographic and Clinical Characteristics for All Participants Who Completed the Survey, by Mode

Variable	Web: 491 Respondents, Mean Age 79.3 Years (SD 7.2)	Mail: 387 Respondents, Mean Age 81.4 Years (SD 6.4)[a]	Phone: 481 Respondents, Mean Age 81.6 Years (SD 6.6)[a]	All Modes Combined: 1,359 Respondents, Mean Age 80.7 Years (SD 6.8)
Age, in years, as a percentage; *p*-value = 0.58				
65–69	15	8	8	10
70–74	17	11	11	13
75–79	10	12	8	10
80–84	37	37	40	38
85+	22	32	33	28

Variable	Web: 491 Respondents, Mean Age 79.3 Years (SD 7.2)	Mail: 387 Respondents, Mean Age 81.4 Years (SD 6.4)[a]	Phone: 481 Respondents, Mean Age 81.6 Years (SD 6.6)[a]	All Modes Combined: 1,359 Respondents, Mean Age 80.7 Years (SD 6.8)
Sex, as a percentage; *p*-value = 0.86				
Male	50	49	45	48
Female	50	51	55	52
Race and ethnicity, as a percentage; *p*-value = 0.78				
White non-Hispanic	92	88	88	89
Hispanic	3	5	3	4
Nonwhite non-Hispanic	2	5	6	4
Missing race non-Hispanic	2	2	4	3
Percentage below poverty in the census tract; *p*-value = 0.79				
0–9.99	60	59	58	59
10–19.99	31	31	33	32
20+	9	10	10	9
Total number of chronic conditions, as a percentage; *p*-value = 0.68				
2	43	33	30	35
3	30	30	33	31
4	15	19	20	18
5+	12	18	18	16
Proportion with specific chronic condition, as a percentage[b]				
Arthritis; *p*-value = 0.46	22	26	27	25
Cancer; *p*-value = 0.78	10	9	9	9
Chronic lung disease; *p*-value = 0.43	37	37	39	38
Congestive heart failure; *p*-value = 0.46	11	18	19	16

Variable	Web: 491 Respondents, Mean Age 79.3 Years (SD 7.2)	Mail: 387 Respondents, Mean Age 81.4 Years (SD 6.4)[a]	Phone: 481 Respondents, Mean Age 81.6 Years (SD 6.6)[a]	All Modes Combined: 1,359 Respondents, Mean Age 80.7 Years (SD 6.8)
Depression; *p*-value = 0.15	22	21	27	23
Diabetes; *p*-value = 0.02	30	35	28	31
Hypertension; *p*-value = 0.32	81	84	82	82
Inflammatory bowel disease; *p*-value = 0.23	1	1	2	1
Ischemic heart disease; *p*-value = 0.58	28	32	27	29
Osteoporosis; *p*-value = 0.74	18	22	27	23
Other heart problems; *p*-value = 0.17	33	36	35	35
Sciatica; *p*-value = 0.58	6	4	5	5
Stroke; *p*-value = 0.89	5	6	7	6
Any home health encounters in the past 12 months, as a percentage; *p*-value = 0.03				
Yes	40	53	48	46
No	60	47	52	54
Number of primary care visits in the past 12 months, as a percentage; *p*-value = 0.69				
0–3	46	42	40	43
4–6	34	36	35	35
7–9	9	12	14	12
10+	10	9	11	10
Number of specialty care visits in the past 12 months, as a percentage; *p*-value = 0.65				
0–3	53	58	57	56
4–6	24	21	20	22
7–9	11	11	12	11

Variable	Web: 491 Respondents, Mean Age 79.3 Years (SD 7.2)	Mail: 387 Respondents, Mean Age 81.4 Years (SD 6.4)[a]	Phone: 481 Respondents, Mean Age 81.6 Years (SD 6.6)[a]	All Modes Combined: 1,359 Respondents, Mean Age 80.7 Years (SD 6.8)
10+	12	11	11	11
Number of hospitalizations in the past 12 months, as a percentage; *p*-value = 0.04				
0	89	86	85	87
1	9	10	12	10
2+	2	5	3	3

NOTE: SD = standard deviation. The *p*-value for mean age is 0.26.
[a] Includes participants initially assigned to the web mode and then reassigned to another mode.
[b] Conditions do not sum to 100 percent.

Question 2: Does the Likelihood of Responding to the Survey Vary by Subject Characteristics?

Table 3.3 shows the adjusted odds ratio (AOR) for responding to the survey associated with survey mode and respondent characteristics, comparing the 1,359 participants who responded to the survey with the 2,390 who did not. Survey mode was a fairly strong predictor of the likelihood of survey response, with the highest odds of completion for those assigned to the mail survey mode and the lowest for the web mode (AOR = 0.22, compared with mail mode). Age was not a strong predictor of responding to the survey and was not statistically significant. Female sex predicted a lower likelihood of responding to the survey (AOR = 0.77, $p = 0.002$). A participant whose race or ethnicity was other than white non-Hispanic was less likely to respond: Hispanics and nonwhites had AORs of 0.47 and 0.58, respectively, compared with white non-Hispanics. Residents of the highest-poverty census tracts had slightly lower odds of responding to the survey than residents of the wealthiest census tracts (AOR = 0.80).

Table 3.3. Adjusted Odds Ratios for Responding to the Initial Survey

Variable	AOR	95% CI	*p*-Value
Survey mode			<0.001
Mail	Reference		
Phone	0.70	0.54–0.92	
Web	0.22	0.18–0.28	
Web–mail	0.58	0.43–0.78	
Web–phone	0.60	0.44–0.80	

Variable	AOR	95% CI	*p*-Value
Age, in years			0.24
65–69	Reference		
70–74	1.12	0.83–1.52	
75–79	1.14	0.82–1.58	
80–84	0.96	0.75–1.25	
85+	0.88	0.67–1.16	
Sex			0.002
Male	Reference		
Female	0.77	0.66–0.91	
Race and ethnicity			<0.001
White non-Hispanic	Reference		
Hispanic	0.47	0.33–0.66	
Nonwhite non-Hispanic	0.58	0.42–0.80	
Missing race non-Hispanic	1.22	0.77–1.96	
Percentage below poverty in the census tract			0.04
0–9.99	Reference		
10–19.99	1.10	0.94–1.30	
20+	0.80	0.63–1.02	
Total number of chronic conditions			0.76
2	Reference		
3	1.11	0.85–1.46	
4	1.26	0.79–2.00	
5+	1.52	0.70–3.30	
Presence of a specific chronic condition			
Arthritis	1.05	0.80–1.37	0.73
Cancer	1.06	0.76–1.47	0.73
Chronic lung disease	1.03	0.80–1.32	0.84
Congestive heart failure	0.68	0.51–0.90	0.007
Depression	0.72	0.55–0.93	0.01
Diabetes	0.81	0.62–1.04	0.098
Hypertension	0.84	0.63–1.11	0.22
Inflammatory bowel disease	0.78	0.41–1.47	0.44
Ischemic heart disease	0.84	0.65–1.09	0.19
Osteoporosis	0.78	0.59–1.02	0.07
Other heart problems	1.01	0.78–1.32	0.93
Sciatica	0.90	0.62–1.30	0.57
Stroke	0.89	0.62–1.26	0.50
Any home health encounters in the past 12 months?			0.81
No	Reference		

Variable	AOR	95% CI	*p*-Value
Yes	0.98	0.84–1.15	
Number of primary care visits in the past 12 months			0.02
0–3	Reference		
4–6	1.30	1.1–1.53	
7–9	1.12	0.88–1.43	
10+	1.21	0.93–1.58	
Number of specialty care visits in the past 12 months			0.008
0–3	Reference		
4–6	1.2	1.01–1.45	
7–9	1.41	1.11–1.81	
10+	1.35	1.05–1.74	
Number of hospitalizations in the past 12 months			0.003
0	Reference		
1	0.70	0.55–0.89	
2+	0.63	0.42–0.94	

NOTE: n = 3,749, of whom 1,359 responded and 2,390 did not. CI = confidence interval.

Total number of chronic conditions was not a significant predictor of survey response in this multivariable analysis. The presence or absence of a specific chronic condition was not usually predictive of survey response, with some exceptions, including congestive heart failure and depression, both of which were associated with lower odds of survey completion than participants without these conditions (AOR = 0.68 and 0.72, $p = 0.007$ and 0.01, respectively). Participants with one or more home health encounters in the past 12 months and those without were similarly likely to respond. Participants with the fewest primary care visits and the fewest specialty care visits (zero to three visits) in the past 12 months were somewhat less likely to respond than those with more visits, although the effect was small. Participants who had been hospitalized within the past 12 months were less likely to respond to the survey than those who had not.

In summary, survey mode appeared to be the strongest predictor of response, with mail and, to a lesser extent, phone modes having higher response rates than web mode. Age was not an important predictor of survey response among this population of participants, all of whom were age 65 or older. In general, sicker participants were somewhat less likely to respond, although the effect was small. Nonwhite participants were considerably less likely to respond, as were residents of census tracts with high rates of poverty.

Question 3: Do Health-Related Quality-of-Life Scores Vary by Survey Mode, Both Before and After Case Mix Adjustment?

We began by examining unadjusted HRQoL scores by survey mode (Table 3.4). There were a few statistically significant differences across modes, but most were relatively small

in terms of absolute magnitude. On VR-36 scores, the three modes showed statistically significant differences on PCS, physical functioning, social functioning, mental health index, and vitality, based on an analysis-of-variance (ANOVA) test of the null hypothesis that all three modes would have equal scores. In general, the magnitudes of these differences were fairly small. For example, on the PCS, the population norm is defined by a mean of 50 and an SD of 10. Here, the difference among groups is 2 points or less, or less than 0.25 of an SD, which is generally considered a small effect size (Cohen, 1969). A larger difference was seen across modes with physical functioning, in which mail respondents were almost a full SD below web respondents; phone respondents were in between.

Table 3.4. Comparison of Unadjusted Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Survey Mode

Variable	Web: 491 Respondents (SD)	Mail: 387 Respondents[a] (SD)	Phone: 481 Respondents[a] (SD)	*p*-Value Generated from ANOVA
VR-36				
Standardized PCS	39 (10.94)	37 (11.5)	38 (12.05)	0.03
Standardized MCS	55 (9.92)	55 (10.36)	56 (9.51)	0.10
General health	63 (20.55)	60 (20.04)	61 (19.84)	0.12
Role—physical	72 (26.01)	69 (28.27)	67 (28.84)	0.02
Physical functioning	58 (28.42)	49 (29.38)	54 (30.75)	<0.001
Pain index	61 (23.87)	63 (24.76)	62 (26.2)	0.31
Social functioning	81 (23.92)	77 (26.08)	76 (27.12)	0.02
Mental health	78 (17.22)	77 (17.86)	80 (15.89)	0.002
Role—emotional	89 (19.01)	88 (20.42)	89 (20.55)	0.40
Vitality	52 (21.27)	50 (22.2)	54 (22.79)	0.03
PROMIS-29				
Physical Health	43 (8.64)	42 (9.25)	42 (9.63)	0.10
Mental Health	50 (8.00)	50 (8.42)	51 (7.74)	0.02
Anxiety (–)	50 (8.32)	50 (8.94)	49 (8.68)	0.02

Variable	Web: 491 Respondents (SD)	Mail: 387 Respondents[a] (SD)	Phone: 481 Respondents[a] (SD)	*p*-Value Generated from ANOVA
Depression (–)	49 (8.36)	51 (8.56)	49 (8.15)	0.003
Fatigue (–)	51 (9.06)	51 (9.22)	51 (8.85)	0.65
Pain Interference (–)	53 (9.12)	54 (9.4)	54 (9.55)	0.75
Physical Function (+)	42 (8.37)	41 (9.15)	40 (9.64)	0.002
Sleep Disturbance (–)	47 (8.53)	47 (9.19)	47 (8.48)	0.79
Social Roles (+)	50 (8.99)	49 (9.14)	50 (9.69)	0.06
Pain Intensity (score 0–10) (–)	3 (2.32)	3 (2.36)	3 (2.51)	0.20

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.
[a] Includes participants initially assigned to the web mode and then reassigned to another mode.

On PROMIS-29 scores, between-group differences by survey mode were also generally small but sometimes statistically significant. Groups differed at the $p < 0.05$ level on Anxiety, Depression, and Physical Function. The magnitudes of these differences were less than 0.25 of an SD, generally regarded as a small effect size (Cohen, 1969).

We then compared HRQoL scores across survey modes after adjusting for case mix (Table 3.5). Our case mix–adjustment model used all the variables in Table 3.2. This allowed us to consider whether HRQoL scores differed across survey modes even after accounting for differences in such factors as age and illness burden. Risk adjustment did not appreciably change most of the comparisons across modes. Although several of the small between-mode differences were somewhat attenuated after adjustment, all remained statistically significant because of the large sample size. In general, risk adjustment seemed to affect cross-mode comparisons more for the VR-36 than for the PROMIS-29. The biggest change due to risk adjustment was the mode comparison on VR-36 physical functioning scores. Before adjustment, there was a 9-point difference between the web and mail modes; after adjustment, the difference narrowed to 5 points, changing the effect size classification from large to moderate (Cohen, 1969).

Table 3.5. Comparison of Risk-Adjusted Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Survey Mode

Variable	Web: 491 Respondents	Mail: 387 Respondents[a]	Phone: 481 Respondents[a]	*p*-Value Generated from ANOVA
VR-36				
Standardized PCS	39 (5.29)	37 (5.45)	37 (5.41)	<0.001
Standardized MCS	56 (3.73)	56 (3.58)	55 (3.71)	0.02
General health	63 (7.75)	61 (7.89)	61 (7.59)	<0.001
Role—physical	72 (11.32)	68 (12.14)	68 (11.87)	<0.001
Physical functioning	58 (14.76)	53 (14.85)	52 (15.21)	<0.001
Pain index	63 (9.67)	62 (10.15)	61 (10.37)	0.006
Social functioning	80 (9.73)	78 (10.23)	77 (10.07)	<0.001
Mental health	79 (6.32)	79 (5.92)	78 (6.39)	0.008
Role—emotional	90 (7.16)	88 (7.55)	88 (7.25)	<0.001
Vitality	54 (8.28)	52 (8.36)	51 (8.44)	<0.001
PROMIS-29				
Physical Health	43 (4.64)	42 (4.76)	41 (4.81)	<0.001
Mental Health	51 (3.42)	50 (3.49)	50 (3.49)	<0.001
Anxiety (–)	49 (2.82)	50 (2.70)	50 (2.82)	<0.001
Depression (–)	49 (3.23)	50 (3.25)	50 (3.42)	<0.001
Fatigue (–)	50 (3.64)	51 (3.57)	51 (3.57)	<0.001
Pain Interference (–)	53 (3.50)	54 (3.72)	54 (3.79)	<0.001
Physical Function (+)	42 (4.56)	41 (4.62)	40 (4.69)	<0.001
Sleep Disturbance (–)	47 (1.88)	47 (1.82)	47 (1.84)	0.04
Social Roles (+)	51 (3.84)	50 (4.07)	49 (4.00)	<0.001

Variable	Web: 491 Respondents	Mail: 387 Respondents[a]	Phone: 481 Respondents[a]	*p*-Value Generated from ANOVA
Pain Intensity (score 0–10) (–)	3 (0.87)	3 (0.92)	3 (0.97)	0.002

NOTE: All results are adjusted for age, sex, race and ethnicity, area SES, receipt of home health care, total number of chronic conditions, presence or absence of each of 13 chronic conditions, number of primary care visits in the previous 12 months, number of specialty care visits in the previous 12 months, and number of hospitalizations in the previous 12 months. Social Roles = Ability to Participate in Social Roles and Activities. For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL.
[a] Includes participants initially assigned to the web mode and then reassigned to another mode.

Question 4: Do Health-Related Quality-of-Life Scores Vary by Respondent Characteristics, Such as Demographics and Comorbid Conditions?

We started to answer this question by comparing mean HRQoL scores by age group (Table 3.6). Differences in some HRQoL dimensions were on the order of 0.5 SD or more, which is generally agreed to constitute a moderate to large effect size (Cohen, 1969). We saw differences in aspects of physical function, such as the VR-36 PCS, the VR-36 physical functioning and role—physical scores, and the PROMIS-29 Physical Health function score (all $p < 0.001$ for ANOVA). Other aspects of HRQoL had differences among groups, but the magnitudes of differences were considerably less. In particular, aspects of mental health did not appear to decline with increasing age, as seen with the VR-36 MCS ($p = 0.10$). A decline in physical function with increasing age, without a corresponding decline in mental health, corresponds with what might be expected, based on previous work (Selim et al., 2004).

Table 3.6. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Age Group

Variable	Age 65–69: 170 Respondents (SD)	Age 70–74: 223 Respondents (SD)	Age 75–79: 157 Respondents (SD)	Age 80–84: 630 Respondents (SD)	Age 85+: 467 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36						
Standardized PCS	40 (11.54)	42 (11.50)	39 (11.91)	38 (10.96)	35 (11.25)	<0.001
Standardized MCS	54 (9.51)	56 (8.35)	56 (10.44)	56 (9.57)	55 (10.87)	0.10
General health	60 (20.83)	67 (19.73)	62 (19.60)	62 (19.33)	59 21.01	<0.001
Role—physical	75 (24.18)	77 (25.68)	73 (27.37)	71 (25.97)	61 (30.00)	<0.001
Physical functioning	65 (26.99)	68 (27.85)	58 (29.17)	54 (28.04)	42 (29.26)	<0.001
Pain index	60 (24.16)	63 (24.37)	62 (24.65)	63 (24.51)	60 (26.17)	0.29
Social functioning	80 (22.07)	81 (23.95)	82 (24.60)	80 (24.59)	73 (28.67)	<0.001
Mental health	76 (16.27)	80 (16.44)	79 (17.36)	79 (16.64)	77 (17.78)	0.31
Role—emotional	89 (15.94)	92 (15.04)	90 (18.62)	90 (19.59)	85 (23.45)	<0.001
Vitality	51 (21.90)	56 (22.67)	55 (23.35)	53 (20.91)	48 (22.4)	<0.001
PROMIS-29						
Physical Health	45 (8.82)	46 (9.19)	44 (8.81)	42 (8.50)	39 (9.02)	<0.001
Mental Health	50 (8.01)	52 (7.95)	51 (8.19)	50 (7.56)	48 (8.45)	<0.001
Anxiety (–)	50 (8.43)	49 (8.29)	49 (8.45)	50 (8.49)	51 (9.10)	0.16

Depression (–)	49 (7.99)	48 (7.38)	48 (8.39)	49 (8.29)	51 (8.77)	<0.001
Fatigue (–)	51 (8.90)	49 (9.23)	50 (8.99)	51 (8.31)	52 (9.73)	0.001
Pain Interference (–)	53 (9.57)	53 (9.22)	53 (9.33)	53 (8.99)	54 (9.80)	0.41
Physical Function (+)	45 (8.47)	45 (8.93)	43 (8.46)	41 (8.34)	37 (8.83)	<0.001
Sleep Disturbance (–)	49 (8.57)	47 (9.01)	47 (8.45)	47 (8.55)	48 (8.83)	0.29
Social Roles (+)	52 (8.97)	52 (9.40)	51 (9.19)	50 (8.55)	47 (9.60)	
Pain Intensity (score 0–10) (–)	3 (2.43)	3 (2.36)	3 (2.54)	3 (2.34)	3 (2.45)	0.75

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

HRQoL scores varied by the respondent's sex (Table 3.7). Female respondents scored lower on some aspects of HRQoL, such as physical function, as exemplified by the PCS (4 points lower, $p < 0.001$). On the PROMIS instrument, females scored lower on almost every aspect of HRQoL, including aspects reflecting both physical and mental well-being, although some of the differences were relatively small.

Table 3.7. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Sex

Variable	Male: 775 Respondents (SD)	Female: 872 Respondents (SD)	*p*-Value Generated from t-Test
VR-36			
Standardized PCS	40 (11.10)	36 (11.62)	<0.001
Standardized MCS	56 (9.93)	55 (9.83)	0.27
General health	62 (20.53)	61 (19.87)	0.54
Role—physical	72 (27.66)	67 (27.53)	0.003
Physical functioning	59 (29.13)	50 (29.52)	<0.001
Pain index	67 (23.56)	57 (25.31)	<0.001
Social functioning	81 (24.86)	76 (26.25)	<0.001
Mental health	80 (17.02)	77 (16.84)	0.003
Role—emotional	89 (20.64)	88 (19.26)	0.34
Vitality	55 (21.86)	50 (22.02)	<0.001
PROMIS-29			
Physical Health	44 (9.09)	41 (8.96)	<0.001
Mental Health	51 (7.76)	49 (8.11)	<0.001
Anxiety (–)	49 (8.47)	51 (8.70)	<0.001
Depression (–)	49 (8.25)	50 (8.43)	0.002
Fatigue (–)	50 (8.85)	52 (9.04)	<0.001
Pain Interference (–)	52 (8.62)	55 (9.77)	<0.001

Variable	Male: 775 Respondents (SD)	Female: 872 Respondents (SD)	*p*-Value Generated from t-Test
Physical Function (+)	43 (9.07)	40 (8.82)	<0.001
Sleep Disturbance (–)	46 (8.34)	48 (8.92)	<0.001
Social Roles (+)	51 (9.10)	49 (9.30)	<0.001
Pain Intensity (score 0–10) (–)	3 (2.23)	4 (2.47)	<0.001

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

We also compared HRQoL across categories of race and ethnicity (Table 3.8). Although small numbers of participants in the nonwhite groups somewhat limited statistical comparisons, we observed that Hispanic respondents reported generally worse mental HRQoL, with lower MCS, higher levels of anxiety and depression on the PROMIS-29, and higher levels of fatigue and pain. Generally, HRQoL for nonwhite non-Hispanic respondents was similar to that for white non-Hispanic respondents.

Table 3.8. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Race and Ethnicity

Variable	White Non-Hispanic: 1,455 Respondents (SD)	White Hispanic: 60 Respondents (SD)	Nonwhite Non-Hispanic: 81 Respondents (SD)	Missing Race Non-Hispanic: 42 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36					
Standardized PCS	38 (11.43)	35 (11.72)	38 (12.04)	39 (13.07)	0.41
Standardized MCS	55 (9.67)	52 (12.74)	53 (11.56)	59 (6.80)	0.003
General health	62 (20.02)	52 (21.48)	60 (20.47)	66 (20.77)	0.006
Role—physical	70 (27.36)	60 (32.25)	67 (29.48)	69 (26.51)	0.08
Physical functioning	54 (29.47)	50 (31.09)	53 (31.98)	59 (31.28)	0.57
Pain index	62 (24.53)	54 (29.07)	59 (27.56)	63 (26.77)	0.16
Social functioning	79 (25.26)	67 (33.26)	76 (27.73)	84 (20.99)	0.006
Mental health	79 (16.69)	73 (21.33)	75 (19.04)	83 (14.52)	0.06
Role—emotional	90 (18.88)	78 (30.11)	82 (25.58)	91 (19.52)	<0.001

Variable	White Non-Hispanic: 1,455 Respondents (SD)	White Hispanic: 60 Respondents (SD)	Nonwhite Non-Hispanic: 81 Respondents (SD)	Missing Race Non-Hispanic: 42 Respondents (SD)	*p*-Value Generated from ANOVA
Vitality	52 (21.90)	47 (23.83)	56 (23.39)	61 (21.51)	0.02
PROMIS-29					
Physical Health	42 (9.10)	39 (9.89)	42 (10.02)	44 (9.14)	0.009
Mental Health	50 (7.87)	46 (10.48)	50 (8.71)	51 (7.50)	0.08
Anxiety (–)	50 (8.50)	52 (10.77)	51 (9.51)	49 (8.23)	0.12
Depression (–)	49 (8.26)	53 (10.06)	50 (9.06)	50 (7.62)	0.03
Fatigue (–)	51 (8.98)	54 (9.41)	49 (9.00)	50 (9.37)	0.02
Pain Interference (–)	53 (9.19)	58 (11.07)	55 (9.63)	53 (9.92)	0.003
Physical Function (+)	41 (9.01)	39 (9.07)	42 (10.03)	43 (9.26)	0.08
Sleep Disturbance (–)	47 (8.52)	51 (9.93)	48 (10.12)	47 (8.85)	0.02
Social Roles (+)	50 (9.18)	47 (11.28)	51 (9.95)	50 (8.21)	0.13
Pain Intensity (score 0–10) (–)	3 (2.35)	4 (2.88)	4 (2.65)	3 (2.44)	0.02

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

We compared the HRQoL of respondents based on the percentage of households in their census tract of residence with incomes below the federal poverty threshold (Table 3.9). Relatively few respondents lived in areas with high levels of poverty. Respondents who lived in poorer census tracts had modestly lower scores on some aspects of HRQoL, although most differences were not statistically significant. The largest difference was seen on the VR-36 physical functioning score, for which reporting scores among respondents living in census tracts with more than 10 percent of households in poverty were approximately 0.5 SD lower than those living in census tracts with less than 10 percent in poverty.

Table 3.9. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Percentage of Households in Poverty in the Census Tract of Residence

Variable	0–9.99%: 971 Respondents (SD)	10–19.99%: 519 Respondents (SD)	20%+: 157 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36				
Standardized PCS	39 (11.49)	37 (11.34)	38 (12/12)	0.09
Standardized MCS	56 (9.71)	55 (10.31)	55 (9.46)	0.64
General health	62 (20.68)	60 (19.72)	63 (18.54)	0.34
Role—physical	71 (27.45)	67 (27.53)	69 (29.04)	0.04
Physical functioning	56 (29.2)	51 (29.69)	50 (31.55)	0.004
Pain index	63 (24.52)	60 (26.74)	61 (24.78)	0.24
Social functioning	79 (24.99)	77 (27.02)	78 (25.41)	0.25
Mental health	79 (17.02)	77 (17.11)	78 (16.25)	0.42
Role—emotional	90 (18.78)	87 (21.24)	87 (21.70)	0.01
Vitality	53 (22.10)	51 (21.87)	52 (22.84)	0.60
PROMIS-29				
Physical Health	43 (9.12)	41 (9.22)	42 (9.26)	0.11
Mental Health	51 (8.04)	50 (8.08)	49 (7.76)	0.02
Anxiety (–)	49 (8.66)	50 (8.65)	50 (8.45)	0.27
Depression (–)	49 (8.23)	50 (8.52)	51 (8.66)	0.08
Fatigue (–)	51 (9.01)	52 (9.21)	52 (8.38)	0.10
Pain Interference (–)	53 (9.12)	54 (9.68)	53 (9.53)	0.27
Physical Function (+)	42 (8.99)	40 (9.16)	41 (9.11)	0.01
Sleep Disturbance (–)	47 (8.58)	47 (8.95)	48 (8.53)	0.72
Social Roles (+)	50 (9.41)	49 (9.19)	50 (8.70)	0.09

Variable	0–9.99%: 971 Respondents (SD)	10–19.99%: 519 Respondents (SD)	20%+: 157 Respondents (SD)	*p*-Value Generated from ANOVA
Pain Intensity (score 0–10) (–)	3 (2.39)	3 (2.4)	3 (2.5)	0.84

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

We also examined the relationship between the number of chronic conditions and HRQoL scores (Table 3.10). As stated previously, by definition, each participant had at least two of the 13 chronic conditions we examined as a condition of eligibility; we compared those who had two, three, four, or at least five of those conditions. Unsurprisingly, every aspect of HRQoL was worse among those with more chronic conditions. We saw a stronger effect for aspects of physical health than mental health. We observed this across both the VR-36 and the PROMIS-29. The difference in PCS, for example, between those with the fewest and the most chronic conditions—a 10-point difference—is equivalent to a full SD and thus a large effect size (Cohen, 1969). The notable exception was the sleep scale on the PROMIS-29, for which having more chronic conditions was associated with only a small increase in Sleep Disturbance.

Table 3.10. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Total Number of Chronic Conditions

Variable	2: 580 Respondents (SD)	3: 523 Respondents (SD)	4: 290 Respondents (SD)	5+: 254 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36					
Standardized PCS	41 (10.78)	39 (11.08)	36 (11.56)	31 (10.58)	<0.001
Standardized MCS	57 (8.49)	55 (10.35)	55 (9.99)	53 (11.23)	<0.001
General health	67 (19.11)	62 (20.19)	59 (19.32)	52 (19.82)	<0.001
Role—physical	77 (24.69)	71 (26.74)	67 (28.67)	54 (27.95)	<0.001
Physical functioning	63 (28.16)	56 (28.48)	48 (30.11)	38 (27.22)	<0.001
Pain index	67 (23.47)	63 (24.34)	60 (25.20)	50 (25.14)	<0.001
Social functioning	84 (22.68)	79 (25.30)	76 (26.53)	67 (28.03)	<0.001
Mental health	81 (15.80)	78 (17.10)	77 (16.97)	74 (18.14)	<0.001
Role—emotional	92 (16.35)	88 (20.80)	87 (21.32)	84 (22.33)	<0.001

Variable	2: 580 Respondents (SD)	3: 523 Respondents (SD)	4: 290 Respondents (SD)	5+: 254 Respondents (SD)	*p*-Value Generated from ANOVA
Vitality	57 (21.41)	53 (21.76)	49 (22.09)	43 (21.00)	<0.001
PROMIS-29					
Physical Health	45 (8.81)	42 (9.20)	41 (8.64)	37 (7.97)	<0.001
Mental Health	52 (7.47)	50 (8.12)	49 (7.80)	46 (7.60)	<0.001
Anxiety (–)	49 (7.92)	50 (8.83)	50 (8.81)	52 (9.22)	<0.001
Depression (–)	48 (7.73)	49 (8.23)	50 (8.46)	53 (8.98)	<0.001
Fatigue (–)	49 (8.53)	51 (9.19)	53 (8.88)	55 (8.39)	<0.001
Pain Interference (–)	51 (8.90)	53 (8.94)	55 (9.53)	57 (9.46)	<0.001
Physical Function (+)	44 (8.66)	41 (9.10)	40 (8.75)	36 (7.77)	<0.001
Sleep Disturbance (–)	47 (8.15)	47 (9.04)	48 (9.14)	49 (8.47)	0.02
Social Roles (+)	53 (8.70)	50 (9.24)	48 (9.13)	45 (8.56)	<0.001
Pain Intensity (score 0–10) (–)	3 (2.28)	3 (2.36)	3 (2.46)	4 (2.40)	<0.001

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

We also compared HRQoL scores based on the utilization of primary care visits in the previous 12 months (Table 3.11). We found that aspects of physical HRQoL were lower among those with more primary care visits, whereas there was no discernible effect on mental HRQoL. For example, PCS scores were 8 points lower among those with ten or more primary care visits than for those with one to three visits, a difference of almost a full SD. On the PROMIS-29, the dimensions of Pain Intensity, Pain Interference, and Fatigue were statistically significantly different across categories of primary care utilization, with high utilizers of primary care reporting more pain, more pain interference, and more fatigue.

Table 3.11. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Number of Primary Care Visits in the Previous 12 Months

Variable	0–3: 720 Respondents (SD)	4–6: 566 Respondents (SD)	7–9: 200 Respondents (SD)	10+: 161 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36					
Standardized PCS	40 (11.56)	38 (11.15)	35 (10.90)	32 (10.66)	<0.001
Standardized MCS	56 (9.41)	56 (9.42)	54 (11.81)	53 (10.76)	0.56
General health	64 (20.25)	62 (19.3)	57 (21.43)	54 (18.62)	<0.001
Role—physical	74 (27.12)	70 (26.16)	65 (28.63)	57 (29.88)	0.002
Physical functioning	57 (30.10)	56 (28.70)	49 (28.85)	40 (27.74)	0.20
Pain index	67 (23.82)	61 (24.00)	55 (25.46)	50 (26.85)	0.02
Social functioning	82 (24.25)	78 (25.26)	74 (28.33)	68 (26.86)	0.80
Mental health	80 (16.19)	79 (16.29)	77 (19.71)	73 (18.34)	0.97
Role—emotional	90 (18.90)	89 (18.92)	86 (23.65)	85 (22.22)	0.69
Vitality	54 (22.12)	54 (21.05)	47 (24.12)	45 (20.80)	0.04
PROMIS-29					
Physical Health	43 (9.53)	43 (8.79)	40 (8.24)	37 (8.09)	<0.001
Mental Health	51 (7.76)	50 (7.80)	48 (8.46)	46 (8.18)	<0.001
Anxiety (–)	49 (8.40)	50 (8.42)	50 (9.73)	52 (8.67)	0.83
Depression (–)	49 (7.96)	49 (8.28)	50 (8.67)	53 (9.27)	0.47
Fatigue (–)	50 (8.89)	51 (8.62)	53 (9.79)	55 (8.68)	0.04
Pain Interference (–)	52 (8.98)	53 (9.12)	56 (9.03)	58 (9.77)	0.004
Physical Function (+)	42 (9.52)	42 (8.73)	39 (8.04)	37 (7.84)	0.10
Sleep Disturbance (–)	47 (8.24)	47 (8.66)	48 (9.52)	49 (9.47)	0.23
Social Roles (+)	51 (9.27)	50 (9.06)	48 (9.22)	46 (9.10)	0.06

Variable	0–3: 720 Respondents (SD)	4–6: 566 Respondents (SD)	7–9: 200 Respondents (SD)	10+: 161 Respondents (SD)	*p*-Value Generated from ANOVA
Pain Intensity (score 0–10) (–)	3 (2.30)	3 (2.38)	4 (2.42)	4 (2.44)	0.02

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

We also examined the relationship between specialty care utilization and HRQoL scores (Table 3.12). Again, aspects of physical HRQoL were more strongly affected than mental HRQoL, although the magnitudes of the effects were less than we saw with primary care utilization. On the VR-36, this was most notably evidenced by a lower PCS, in which the magnitude of the difference (3 points) was commensurate with a small effect (Cohen, 1969). On the PROMIS-29, we saw between-group differences on both Physical Health and Mental Health, although those differences were also small.

Table 3.12. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Number of Specialty Care Visits in the Previous 12 Months

Variable	0–3: 927 Respondents (SD)	4–6: 355 Respondents (SD)	7–9: 181 Respondents (SD)	10+: 184 Respondents (SD)	*p*-Value Generated from ANOVA
Number of respondents	927	355	181	184	
VR-36					
Standardized PCS	39 (11.51)	37 (11.14)	36 (12.32)	36 (10.81)	<0.001
Standardized MCS	55 (9.87)	56 (10.40)	56 (9.34)	56 (9.41)	0.17
General health	63 (20.29)	62 (18.99)	57 (20.81)	57 (20.08)	<0.001
Role—physical	72 (27.08)	68 (28.21)	65 (28.70)	65 (27.31)	<0.001
Physical functioning	55 (30.42)	55 (27.93)	52 (30.30)	51 (28.49)	<0.001
Pain index	64 (24.69)	60 (24.71)	59 (25.73)	59 (25.22)	<0.001
Social functioning	79 (26.17)	79 (25.47)	78 (25.66)	77 (24.01)	<0.001
Mental health	78 (16.66)	79 (17.63)	79 (17.16)	78 (17.17)	0.03
Role—emotional	88 (20.12)	89 (20.06)	90 (17.36)	89 (21.22)	0.054
Vitality	53 (22.04)	51 (21.75)	51 (22.65)	49 (22.13)	<0.001

Variable	0–3: 927 Respondents (SD)	4–6: 355 Respondents (SD)	7–9: 181 Respondents (SD)	10+: 184 Respondents (SD)	*p*-Value Generated from ANOVA
PROMIS-29					
Physical Health	43 (9.47)	42 (8.63)	41 (9.39)	41 (8.42)	0.11
Mental Health	51 (8.08)	50 (7.96)	49 (8.32)	49 (7.56)	0.08
Anxiety (–)	50 (8.48)	50 (8.82)	50 (8.94)	49 (8.78)	0.04
Depression (–)	50 (8.37)	49 (8.49)	49 (8.60)	49 (8.60)	<0.001
Fatigue (–)	50 (9.02)	51 (8.99)	51 (9.15)	53 (8.85)	<0.001
Pain Interference (–)	53 (9.37)	54 (8.93)	55 (9.60)	55 (9.45)	0.01
Physical Function (+)	42 (9.41)	41 (8.46)	40 (9.12)	40 (8.40)	<0.001
Sleep Disturbance (–)	47 (8.72)	48 (8.49)	48 (8.55)	47 (8.99)	0.35
Social Roles (+)	51 (9.48)	50 (9.12)	50 (8.84)	49 (8.98)	<0.001
Pain Intensity (score 0–10) (–)	3 (2.42)	3 (2.38)	4 (2.44)	3 (2.23)	0.08

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

We also examined the HRQoL of respondents who had been hospitalized in the previous 12 months in comparison with those who had not (Table 3.13). Physical function appeared to be poorer among those who had been hospitalized. However, the relatively small number of respondents who had been hospitalized meant that these comparisons had limited statistical power to show a difference.

Table 3.13. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Number of Hospitalizations in the Previous 12 Months

Variable	0: 1,429 Respondents (SD)	1: 167 Respondents (SD)	2+: 51 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36				
Standardized PCS	39 (11.33)	34 (11.55)	31 (11.26)	0.09
Standardized MCS	56 (9.71)	54 (10.93)	53 10.92)	0.64

Variable	0: 1,429 Respondents (SD)	1: 167 Respondents (SD)	2+: 51 Respondents (SD)	*p*-Value Generated from ANOVA
General health	63 (19.90)	55 (20.21)	53 (22.25)	0.34
Role—physical	72 (26.81)	59 (29.56)	45 (26.86)	0.04
Physical functioning	56 (29.13)	45 (30.19)	34 (30.57)	0.004
Pain index	63 (24.17)	55 (29.03)	56 (28.07)	0.24
Social functioning	80 (25.00)	71 (28.25)	62 (26.92)	0.25
Mental health	79 (16.51)	76 (19.69)	74 (19.06)	0.42
Role—emotional	89 (19.46)	86 (22.43)	84 (23.3)	0.01
Vitality	53 (21.84)	46 (22.76)	43 (22.04)	0.60
PROMIS-29				
Physical Health	43 (9.04)	39 (8.93)	36 (9.34)	<0.001
Mental Health	51 (7.84)	48 (8.83)	46 (8.41)	<0.001
Anxiety (–)	50 (8.53)	51 (9.25)	52 (8.98)	0.27
Depression (–)	49 (8.15)	52 (9.71)	52 (8.44)	0.08
Fatigue (–)	50 (8.88)	54 (9.60)	54 (8.58)	0.10
Pain Interference (–)	53 (9.10)	55 (10.72)	56 (10.41)	0.27
Physical Function (+)	42 (8.98)	38 (8.38)	35 (9.22)	0.01
Sleep Disturbance (–)	47 (8.55)	48 (9.55)	47 (9.51)	0.72
Social Roles (+)	51 (9.16)	47 (9.30)	44 (8.69)	0.09
Pain Intensity (score 0–10) (–)	3 (2.37)	4 (2.58)	3 (2.49)	0.84

NOTE: Social Roles = Ability to Participate in Social Roles and Activities.

We examined HRQoL scores based on whether the respondent had received some home health care in the previous year or had received none (Table 3.14). Respondents who had received home care scored lower on every aspect of HRQoL except sleep disturbance, with larger effect sizes for physical aspects of HRQoL such as the PCS (almost a full standard deviation), vitality, physical functioning, and pain.

Table 3.14. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Receipt of Home Health Care

Variable	No Home Health Care: 901 Respondents (SD)	Some Home Health Care: 746 Respondents (SD)	*p*-Value Generated from t-Test
VR-36			
Standardized PCS	41 (10.70)	34 (11.43)	<0.001
Standardized MCS	56 (9.02)	54 (10.77)	0.002
General health	66 (18.28)	57 (21.16)	<0.001
Role—physical	76 (24.52)	62 (29.33)	<0.001
Physical functioning	63 (27.65)	44 (29.00)	<0.001
Pain index	64 (23.60)	59 (26.20)	<0.001
Social functioning	83 (22.87)	73 (27.70)	<0.001
Mental health	80 (15.78)	77 (18.18)	<0.001
Role—emotional	91 (17.12)	86 (22.44)	<0.001
Vitality	56 (20.82)	47 (22.71)	<0.001
PROMIS-29			
Physical Health	45 (8.68)	39 (8.86)	<0.001
Mental Health	52 (7.36)	48 (8.40)	<0.001
Anxiety (–)	49 (8.24)	51 (9.01)	<0.001
Depression (–)	48 (7.71)	51 (8.91)	<0.001
Fatigue (–)	49 (8.35)	53 (9.32)	<0.001
Pain Interference (–)	52 (8.74)	55 (9.80)	<0.001
Physical Function (+)	44 (8.72)	38 (8.58)	<0.001
Sleep Disturbance (–)	47 (8.43)	48 (8.97)	0.10
Social Roles (+)	52 (8.78)	48 (9.31)	<0.001

Variable	No Home Health Care: 901 Respondents (SD)	Some Home Health Care: 746 Respondents (SD)	*p*-Value Generated from t-Test
Pain Intensity (score 0–10) (–)	3 (2.30)	3 (2.48)	<0.001

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

Although the majority of the survey instrument consisted of the VR-36 and the PROMIS-29, we posed two additional questions. The first asked, "To what degree have your health-related expenses caused financial problems for you and your family?" There were four response options: not at all, a little, somewhat, and a lot. Table 3.15 shows differences in respondent HRQoL based on those four options. HRQoL scores decreased with increasing levels of this perceived financial insecurity variable. We observed a clear gradient for every aspect of HRQoL, including sleep quality, which was not predicted by most other respondent-level characteristics. In Tables 3.6–3.14, many aspects of illness burden or age were associated with a decrease in physical but not mental aspects of HRQoL. In contrast, the responses to the question about financial strain from health care–related expenses predicted similar effects on both physical and mental HRQoL. These effects were larger than any other stratifying variable. For example, the difference between the highest and lowest level of this variable for the PCS and the MCS was more than a full SD. It is noteworthy that the effect of this question diverged from the rather small effect we saw earlier for area SES. This result implies that the effect of perceived financial insecurity on HRQoL might be much greater than the effect of census-based classification of poverty, based on the neighborhood of residence.

Table 3.15. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Perceived Difficulty in Paying for Health Care

Variable	Not at All: 831 Respondents (SD)	A Little: 242 Respondents (SD)	Somewhat: 178 Respondents (SD)	A Lot: 77 Respondents (SD)	*p*-Value Generated from ANOVA
VR-36					
Standardized PCS	41 (10.89)	36 (11.29)	33 (10.86)	29 (9.21)	<0.001
Standardized MCS	57 (8.34)	54 (10.02)	52 (9.89)	46 (14.34)	<0.001
General health	67 (18.25)	56 (19.67)	54 (19.15)	43 (21.97)	<0.001
Role—physical	76 (24.65)	66 (26.97)	57 (29.11)	41 (28.82)	<0.001
Physical functioning	61 (27.99)	49 (28.97)	42 (28.02)	28 (25.98)	<0.001

Variable	Not at All: 831 Respondents (SD)	A Little: 242 Respondents (SD)	Somewhat: 178 Respondents (SD)	A Lot: 77 Respondents (SD)	*p*-Value Generated from ANOVA
Pain index	67 (23.64)	59 (23.24)	51 (24.9)	42 (24.43)	<0.001
Social functioning	85 (21.88)	75 (26.62)	66 (26.02)	51 (30.4)	<0.001
Mental health	82 (14.98)	76 (16.18)	72 (16.82)	63 (21.37)	<0.001
Role—emotional	93 (16.19)	87 (19.64)	84 (21.58)	67 (30.44)	<0.001
Vitality	57 (21.13)	49 (20.51)	43 (20.97)	36 (22.4)	<0.001
PROMIS-29					
Physical Health	45 (8.65)	40 (8.63)	38 (8.22)	33 (7.52)	<0.001
Mental Health	52 (7.26)	48 (7.23)	46 (7.14)	42 (8.67)	<0.001
Anxiety (–)	48 (7.9)	52 (8.14)	53 (8.35)	57 (10)	<0.001
Depression (–)	48 (7.56)	51 (8.04)	53 (8.32)	56 (10.18)	<0.001
Fatigue (–)	49 (8.58)	52 (7.95)	55 (8.43)	59 (9.89)	<0.001
Pain Interference (–)	51 (8.87)	55 (8.57)	58 (8.5)	61 (9.51)	<0.001
Physical Function (+)	43 (8.61)	39 (8.46)	37 (8.56)	33 (7.37)	<0.001
Sleep Disturbance (–)	46 (8.33)	49 (8.4)	50 (8.56)	51 (9.8)	<0.001
Social Roles (+)	52 (8.87)	48 (8.22)	45 (8.13)	42 (9.11)	<0.001
Pain Intensity (score 0–10) (–)	3 (2.26)	3 (2.19)	4 (2.38)	5 (2.54)	<0.001

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

Another question that we added to the survey concerned proxy response: "Did someone help you complete this survey?" Previous literature has shown that proxy respondents might give answers that differ from respondents' self-reports (Corder, Woodbury, and Manton, 1996; Hays et al., 1995; Pickard and Knight, 2005). Here, we cannot be certain whether the proxy answered the questions on the respondent's behalf, wrote down the answers that the respondent gave, helped them by explaining some of the more-difficult questions, or provided some other level of help. We found that respondents who used proxies had poorer

HRQoL in every aspect except sleep quality (Table 3.16). The magnitude of these effects was in the moderate to large range; for example, between a half and a full SD on both MCS and PCS.

Table 3.16. Comparison of Mean Health-Related Quality-of-Life Scores (with Standard Deviation), by Respondents' Use of Proxy Assistance

Variable	No: 1,191 Respondents (SD)	Yes: 143 Respondents (SD)	*p*-Value Generated from t-Test
VR-36			
Standardized PCS	39 (11.25)	30 (10.92)	<0.001
Standardized MCS	56 (9.04)	50 (12.91)	<0.001
General health	63 (19.54)	49 (21.48)	<0.001
Role—physical	72 (25.84)	49 (33.59)	<0.001
Physical functioning	58 (28.27)	29 (29.18)	<0.001
Pain index	63 (24.44)	53 (27.28)	<0.001
Social functioning	81 (23.79)	59 (32.94)	<0.001
Mental health	80 (15.82)	68 (20.41)	<0.001
Role—emotional	91 (17.36)	75 (30.1)	<0.001
Vitality	54 (21.35)	39 (23.23)	<0.001
PROMIS-29			
Physical Health	43 (8.71)	34 (9.1)	<0.001
Mental Health	51 (7.61)	45 (9.59)	<0.001
Anxiety (–)	49 (8.33)	53 (10.04)	<0.001
Depression (–)	49 (7.97)	55 (9.69)	<0.001
Fatigue (–)	50 (8.58)	56 (10.79)	<0.001
Pain Interference (–)	53 (9.06)	57 (11.09)	<0.001
Physical Function (+)	42 (8.58)	33 (8.85)	<0.001

Variable	No: 1,191 Respondents (SD)	Yes: 143 Respondents (SD)	*p*-Value Generated from t-Test
Sleep Disturbance (–)	47 (8.62)	48 (9.35)	0.24
Social Roles (+)	51 (8.72)	43 (10.81)	<0.001
Pain Intensity (score 0–10) (–)	3 (2.37)	4 (2.62)	0.02

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

Question 5: Is There Differential Item Functioning for Items of the Patient-Reported Outcomes Measurement Information System 29-Item Survey Across Survey Mode or Respondent-Level Characteristics?

DIF refers to a systematic tendency for respondents of a certain type to score higher or lower than expected on a particular survey item. For example, women might systematically score higher than men on an item related to wrist pain, *beyond what would be predicted from their levels of the symptom as estimated by other items of the measure*. It is important to detect DIF because its presence implies some bias in scoring. Thus, if it is detected, survey creators might want to adjust the scoring of items to reduce it or at least be aware of the items that display DIF, for whom, and in what direction (i.e., higher or lower scores). No previous effort has assessed PROMIS-29 items for DIF among older adults with two or more chronic health conditions.

There are several different ways to test for the presence of DIF. After considering some well-regarded methods, we chose to test for DIF using the method that Gelin and Zumbo used because of its wide acceptance and its suitability for performing the large number of DIF analyses we needed to perform in an efficient manner (Gelin and Zumbo, 2003). The Gelin and Zumbo method is similar to the logistic regression approach that Swaminathan and Rogers introduced, but it uses ordinal rather than binary logistic regression to accommodate items with more than two response levels (Swaminathan and Rogers, 1990). We corrected for multiple comparisons using a Benjamini–Hochberg correction (Benjamini and Hochberg, 1995). We tested every item in the PROMIS-29 for DIF against every respondent-level characteristic, including survey mode, by estimating a series of nested models. These models predicted each item response as the outcome. The first model specified a main effect for the case mix–adjusted score on the PROMIS-29 scale score from which the item came (model 1). The next added a main effect for the grouping variable indicator (model 2). And the last added an interaction effect (score × group; model 3). For each set of analyses, we omitted the stratifying variable from the case mix model (e.g., we did not adjust for sex in the

case mix model when testing for the presence of DIF by sex). When a respondent-level characteristic had more than two levels, we created a binary version of the variable based on the univariate distribution of the data, with the exception of survey mode, in which we tested all levels of the variable against one another.

Therefore, the variables were as follows:

- web mode versus mail mode
- web mode versus phone mode
- phone mode versus mail mode
- male sex versus female sex
- age 65–79 versus age 80+
- white non-Hispanic race and ethnicity versus all others
- 0- to 9.99-percent poverty in the census tract of residence versus 10 percent or more
- two chronic conditions versus three or more
- any home health care versus none
- any hospitalizations versus none
- zero to three primary care visits versus four or more
- zero to three specialty care visits versus four or more
- proxy assistance with survey completion versus not
- presence versus absence of each of 13 chronic medical conditions
  - arthritis
  - cancer
  - chronic lung disease
  - congestive heart failure
  - depression
  - diabetes
  - hypertension
  - inflammatory bowel disease
  - ischemic heart disease
  - osteoporosis
  - other heart problems
  - sciatica
  - stroke.

Of note is the fact that we did not test for DIF for the perceived financial insecurity question, "To what degree have your health-related expenses caused financial problems for you and your family?" (not at all, a little, somewhat, or a lot). Because this question requires a subjective assessment of one's level of financial security, it is likely to be highly correlated with aspects of mental HRQoL, especially anxiety. DIF analysis is not meaningful when the stratifying variable is highly related to an aspect of the construct that is being evaluated for DIF (in this case, HRQoL). We also did not examine DIF for the Pain Intensity item because it constitutes an entire scale by itself and thus is not subject to DIF in the usual sense of

comparing the response on one item with those for the other items used to calculate its scale. Thus, we examined DIF for only 28 items from the PROMIS-29.

We therefore tested 26 binary variables against 28 items from the PROMIS-29, for a total of 728 analyses to evaluate DIF. We applied the Benjamini–Hochberg correction for multiple testing once for each of the 26 binary variables we tested; thus, we applied it for groups of 28 analyses, as opposed to once across all 728 analyses. We determined statistical significance based on a comparison of three nested models—namely, model 1 (case mix–adjusted scale score only), model 2 (scale score and group indicator), and model 3 (scale score, group indicator, and the interaction between these two predictors). In addition to considering statistical significance, we considered the magnitude of the effect size for DIF, in part because the ability to demonstrate a statistically significant effect is highly related to sample size and not necessarily to clinical significance. Per the recommendations of Gelin and Zumbo, an effect can be considered negligible in magnitude if the increase in R^2 between two nested models is less than 0.035 (Gelin and Zumbo, 2003). An increase of 0.035 to 0.070 is considered a moderate-size effect, while an increase of 0.70 or greater is considered a large effect.

We therefore first evaluated the overall presence of DIF by (1) examining the statistical significance of the 2-degree-of-freedom change in model chi-square from model 1 to model 3 and (2) evaluating the magnitude of change in R^2 from model 1 to model 3 (Gelin and Zumbo, 2003). If the 2-degree-of-freedom change was statistically significant at the corrected 0.01 level or less, *and* the difference in R^2 was larger than 0.035, the item was considered to have DIF. For these items, we then examined each of the 1-degree-of-freedom changes (from model 1 to model 2 and from model 2 to model 3) to determine whether the DIF was uniform (significance in difference between model 1 and model 2) or nonuniform (significance in difference between model 2 and model 3) and evaluated the magnitude of that effect.

Of the 728 item–characteristic pairs that we tested for DIF, 56 demonstrated statistical significance at the Benjamini–Hochberg–corrected 0.01 level. (Appendix C contains the full results of the DIF analyses.) However, of these, only one item pair had a moderate-size effect: the combination of Physical Function question 4 and proxy response to the survey (change in R^2: 0.063). The other 55 item pairs all had negligible effect sizes, ranging from 0.007 (the smallest) to 0.033 (the largest but still within the "negligible" range).

The text of Physical Function question 4 read, "Are you able to run errands and shop?" The response is a five-point Likert scale from "without any difficulty" to "unable to do." This showed DIF for respondents based on proxy response. A closer examination of the 1-degree-of-freedom model chi-square changes showed that the observed DIF was uniform—that is, because of the addition of the main effect for group when going from model 1 to model 2 [chi-square change 113.5; $p < 0.001$]. Specifically, those with proxy response had higher scores (better HRQoL) on this question than would have been predicted based on their other

personal characteristics (i.e., the case mix model) and their responses to the other Physical Function items. In other words, those with proxy response had ORs of 6.8 (95-percent CI 4.8–9.6) to have a higher score on this item than those who did not.

Question 6: Do We Find Construct Validity Linking Risk-Adjusted Patient-Reported Outcomes Measurement Information System 29-Item Survey Scores with Health Outcomes, Such as Utilization of Primary Care Visits or Other Prominent Measures of Health-Related Quality of Life?

Validity of items can be established in three main ways. *Content validity* establishes that the content area that an item covers matches that of the construct it is meant to measure, neither excluding important content nor including extraneous content. *Criterion validity* establishes a direct link between the items and the construct to be measured via a direct consequential relationship. For example, the criterion validity of an item intended to measure medication adherence might best be demonstrated through direct correlation with blood levels of the medication in question. Although *construct validity* establishes more-indirect links between the items and the construct than criterion validity does, it is preferable that the links be as direct as possible. Here, although there might not be a single link that conclusively establishes validity on its own, we relied on a combination of several inferential links to establish validity.

We aimed to demonstrate the construct validity of PROMIS-29 scores using the number of primary care visits, which is a measure of health care utilization intensity, and correlations with VR-36 summary scores (MCS and PCS), which we measured concurrently. We used the number of primary care visits over other possible variables, such as the number of hospitalizations. Although the number of hospitalizations might be even more strongly tied to illness burden, relatively few patients had any hospitalizations, rendering this a fairly coarse measure of illness burden. It is also worth noting that some of the potential disadvantages of using the number of primary care visits might be minimized in this study sample. In many contexts, primary care utilization is determined not only by illness burden but also by access to care, which might be related to SES. Here, we consider all KPCO patients to have similar access to care under that system, reducing that concern. An additional demonstration of construct validity can be found in Tables 3.6 through 3.16, in which we analyzed PROMIS-29 scores according to various subject characteristics, including the number of chronic medical conditions.

Analysis A: Comparing Risk-Adjusted Patient-Reported Outcomes Measurement Information System 29-Item Survey Scores at Different Levels of Primary Care Utilization

We began by examining risk-adjusted PROMIS-29 scores across categories of primary care visits in the previous 12 months. Our risk-adjustment model used the full set of respondent-level variables in Table 3.2 to further support the idea that, even after controlling for demographics and illness burden, the PROMIS-29 is still measuring aspects of HRQoL that other variables do not capture. Because these analyses are risk-adjusted, PROMIS scores are represented as being higher or lower than would be predicted for such respondents, based on the case mix model (a concept also known as an "observed minus predicted" value). For example, an Anxiety score of –1.0 would mean that respondents in that category scored 1 point below what the case mix model would predict. An ANOVA test for multiple groups, or a t-test for two groups, tested the null hypothesis that group scores did not differ.

We found that respondents with the fewest primary care visits (zero to three) had better risk-adjusted HRQoL on all eight PROMIS scales—and both PROMIS summary scales—than respondents with more primary care visits (Table 3.17). The difference between the groups was small, less than 0.25 SD. Nevertheless, these results are supportive of the construct validity of the PROMIS-29 for measuring HRQoL in this older population of patients with MCC, beyond what can be predicted by patient characteristics, such as demographics and comorbid conditions.

Table 3.17. Comparison of Risk-Adjusted Patient-Reported Outcomes Measurement Information System 29-Item Survey Scores, by Number of Primary Care Visits in the Previous 12 Months

Scale	0–3: 584 Respondents	4–6: 478 Respondents	7–9: 160 Respondents	10+: 137 Respondents	*p*-Value, Test for Linear Trend
Physical Health	–0.12	0.66	–0.75	–0.96	<0.001
Mental Health	0.09	0.43	–0.65	–1.24	<0.001
Anxiety (–)	–0.25	–0.09	0.23	1.11	<0.001
Depression (–)	–0.11	–0.41	0.23	1.62	<0.001
Fatigue (–)	–0.25	–0.45	0.87	1.58	<0.001
Pain Intensity (–)	–0.11	–0.10	0.36	0.42	<0.001
Pain Interference (–)	–0.69	–0.13	1.23	1.92	<0.001
Physical Function (+)	–0.24	0.72	–0.62	–0.78	<0.001
Sleep Disturbance (–)	–0.00	–0.23	0.22	0.58	<0.001
Social Roles (+)	–0.13	0.46	–0.54	–0.44	<0.001

NOTE: (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

Analysis B: Correlation Between Patient-Reported Outcomes Measurement Information System 29-Item Survey Scores and Another Well-Established Health-Related Quality-of-Life Instrument

We assessed the correlation between PROMIS-29 scores and summary scores from the VR-36, which were measured concurrently. We also examined correlations between the two PROMIS-29 summary scores and the two VR-36 summary scores and calculated a Pearson correlation among these scores (Table 3.18). Four PROMIS-29 scores showed strong associations with PCS, as measured by a correlation of 0.5 or higher: Physical Function, Ability to Participate in Social Roles and Activities, Pain Interference, and Pain Intensity (Cohen, 1969). Scores on one PROMIS-29 domain—Fatigue—were associated strongly with both PCS and MCS. Scores on two other scales, Anxiety and Depression, were strongly associated with MCS scores. Finally, scores on the PROMIS-29 Sleep Disturbance scale were only weakly and negatively related to the scores of both the PCS and the MCS. These findings are generally consistent with what is known about the relationship between the PCS and MCS and among scores on the eight scales of the PROMIS-29. All the correlations in the table are statistically significant and in the expected direction, suggesting that poorer health, as measured by any part of the VR-36, is associated with poorer health as measured by any part of the PROMIS-29. In general, therefore, these results are supportive of the construct

validity of the PROMIS-29 as a measurement of HRQoL in the older population of adults with MCC.

Table 3.18. Pearson Correlation Between the Eight Patient-Reported Outcomes Measurement Information System 29-Item Survey Scales and the Two Summary Scores of the Veterans RAND 36 Item Health Survey, Measured Concurrently (n = 1,198)

PROMIS-29 Scale	PCS (VR-36)	MCS (VR-36)
Physical Function (+)	0.81	0.28
Social Roles (+)	0.72	0.49
Pain Interference (–)	–0.70	–0.33
Pain Intensity (–)	–0.61	–0.25
Fatigue (–)	–0.63	–0.54
Anxiety (–)	–0.30	–0.66
Depression (–)	–0.40	–0.66
Sleep Disturbance (–)	–0.35	–0.38

NOTE: All correlations are statistically significant at the $p < 0.001$ level. (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

The VR-36 PCS and MCS are established summary scores that represent physical and mental HRQoL, respectively. We compared these scores with those from our two prototype PROMIS-29 summary scores, Physical Health and Mental Health (Table 3.19). As we know from previous work, the correlation between PCS and MCS scores is low; here, it was 0.13 (Rogers et al., 2004). The two summary scores of the PROMIS-29 were much more highly correlated (r = 0.74). Both PROMIS summary scores were also highly correlated with the PCS (0.85 for PROMIS-29 Physical Health and 0.70 for PROMIS-29 Mental Health). MCS was highly correlated with PROMIS-29 Mental Health (0.64) and much less correlated with PROMIS-29 Physical Health (0.32). This tells us that both summary scores for the PROMIS-29 measure a construct more closely related to the PCS than to the MCS. The implication of this finding is that the PROMIS-29 summary scores are not the same as the PCS and MCS and should not be interpreted as if they are.

Table 3.19. Pearson Correlation Between the Two Prototype Patient-Reported Outcomes Measurement Information System 29-Item Survey Summary Scores and the Two Summary Scores of the Veterans RAND 36 Item Health Survey, Measured Concurrently (n = 1,200)

Variable	PROMIS-29 Physical Health	PROMIS-29 Mental Health	PCS	MCS
PROMIS-29 Physical Health	—	—	—	—
PROMIS-29 Mental Health	0.74	—	—	—
PCS	0.85	0.70	—	—
MCS	0.32	0.64	0.13	—

NOTE: All correlations are statistically significant at the $p < 0.001$ level.

Chapter Four. Analytic Methods and Results: Utility of Items as Performance Measures

Tasks

Our analyses of the second round of survey data, collected at month 6, can be divided into the following four discrete tasks:

1. Examine the response rate to the second-round survey and the representativeness of respondents, compared with those who did not respond, as a means of assessing threats to validity.
2. Construct PRO-based PMs, using data from the first and second rounds of survey data. Define concepts, such as numerator, denominator, and any exclusion criteria.
3. Perform reliability testing of these PMs, examining their ability to reliably distinguish performance across several KPCO ambulatory clinics.
4. Using the collected data and simulations, assess the potential for the PMs to demonstrate differences among ambulatory clinics or health systems. Propose suggestions for the future collection of PRO-based PM data.

Task 1: Examine the Response to the Second Round of the Survey and Implications for Threats to Validity

A total of 490 people responded to the first web survey. Of those, we now know that six (1.2 percent) had died prior to the date the second web survey began (April 14, 2017). The second round of the survey was sent out six months after the first round, to 401 randomly selected participants who had responded to the web mode survey on the first round (one of whom we know now, but did not know at the time, was already deceased). Of those, 352 responded to the second round, for a response rate of 88 percent. Among responders, 337, or 84 percent, submitted sufficient data to calculate scores. Table 4.1 compares the characteristics of the 337 participants who provided sufficient data in the second round with the characteristics of the 64 participants who did not.

Table 4.1. Comparison of Characteristics Between 337 Participants Who Responded to the Second-Round Survey with Sufficient Data and 64 Who Did Not

Variable	Respondents (n = 337) (SD)	Nonrespondents (n = 64) (SD)	*p*-Value
Age			
Mean	79.0 (7.18)	80.3 (6.71)	0.18
Age, in years			0.26
65–69	14%	11%	
70–74	20%	14%	
75–79	9%	8%	
80–84	35%	49%	
85+	21%	17%	
Sex			0.80
Male	50%	49%	
Female	50%	51%	
Race and Ethnicity			0.03
White non-Hispanic	93%	86%	
Hispanic	4%	3%	
Nonwhite non-Hispanic	2%	5%	
Missing race non-Hispanic	1%	6%	
Percentage below poverty in the census tract			0.39
0–9.99	60%	57%	
10–19.99	31%	30%	
20+	8%	13%	
Total number of chronic conditions			0.43
2	44%	33%	
3	31%	32%	
4	14%	21%	
5+	11%	14%	
Presence of a specific chronic condition			
Arthritis	23%	24%	0.92
Cancer	9%	6%	0.40
Chronic lung disease	38%	35%	0.65
Congestive heart failure	10%	11%	0.72
Depression	22%	24%	0.84
Diabetes	31%	32%	0.88
Hypertension	80%	84%	0.40

Variable	Respondents (n = 337) (SD)	Nonrespondents (n = 64) (SD)	*p*-Value
Inflammatory bowel disease	1%	2%	0.80
Ischemic heart disease	27%	27%	0.90
Osteoporosis	18%	21%	0.68
Other heart problems	31%	38%	0.23
Sciatica	6%	13%	0.08
Stroke	2%	10%	0.01
Any home health encounters in the previous 12 months?			0.13
No	36%	48%	
Yes	64%	52%	
Number of primary care visits in previous 12 months			0.19
0–3	48%	38%	
4–6	34%	37%	
7–9	8%	16%	
10+	11%	10%	
Number of specialty care visits in the previous 12 months			0.08
0–3	52%	65%	
4–6	26%	13%	
7–9	11%	13%	
10+	12%	10%	
Number of hospitalizations in the previous 12 months			0.23
0	90%	84%	
1	8%	11%	
2+	2%	5%	
Endorsed difficulty paying for health care on the first survey?			0.74
Not at all	68%	71%	
A little	1%	13%	
Somewhat	9%	11%	
A lot	6%	5%	
Did a person help the respondent complete the first survey?			0.004
Yes	4%	13%	
No	96%	87%	

Variable	Respondents (n = 337) (SD)	Nonrespondents (n = 64) (SD)	*p*-Value
Mean PROMIS-29 summary scores on the first survey (SDs in parentheses)			
Physical Health	43.8 (8.13)	42.2 (9.17)	0.18
Mental Health	50.4 (7.55)	48.6 (8.93)	0.10
Mean VR-36 scores on the first survey (SDs in parentheses)			
PCS	39.7 (10.49)	37.9 (11.17)	0.24
MCS	55.9 (9.18)	53.6 (10.36)	0.09

NOTE: Because of rounding, some distributions do not sum to 100.

Table 4.2 shows that the 337 participants who responded to the second-round survey are largely similar to the 64 who did not respond, although the sample size limits the statistical power to show small differences. Several variables were significantly different between groups or otherwise bear comment. First, those of white non-Hispanic race and ethnicity were more likely to respond (93 percent of responders were non-Hispanic white, versus 86 percent of nonresponders, $p = 0.03$ for the race and ethnicity category). Second, only 4 percent of second-round responders had received assistance completing the survey on the first administration, compared with 13 percent of nonresponders ($p = 0.004$). Third, HRQoL scores recorded in the first-round survey were somewhat higher among second-round responders, a difference of approximately 0.25 SD (small effect). These differences did not achieve statistical significance at the 95-percent confidence level; for example, MCS was 55.9 among responders and 53.6 among nonresponders, $p = 0.09$. In summary, second-round responders were more likely than second-round nonresponders to be of white non-Hispanic race and ethnicity and not to require assistance to complete the survey. There was a notable but not statistically significant finding of higher HRQoL in those who responded. We judged these small differences not to constitute a threat to the validity of comparing HRQoL between the first- and second-round surveys as a means of generating PRO-based PMs.

Table 4.2. Changes in Health-Related Quality-of-Life Scores from the First- to the Second-Round Surveys, Among 337 Participants Who Responded to Both Surveys

Variable	First-Round Survey, Mean (SD)	Second-Round Survey, Mean (SD)	*p*-Value
VR-36			
Standardized PCS	39.7 (10.49)	38.9 (10.82)	0.07
Standardized MCS	55.9 (9.18)	56.3 (9.26)	0.40
General health	64.4 (19.58)	63.2 (19.79)	0.14
Role—physical	74.5 (24.21)	72.9 (24.38)	0.13
Physical functioning	60.9 (26.59)	58.9 (27.03)	0.02
Pain index	61.7 (23.1)	61.5 (23.44)	0.90
Social functioning	83.2 (21.99)	83.1 (23.32)	0.97
Mental health	79.2 (16.33)	79.8 (15.93)	0.36
Role—emotional	91.6 (15.74)	90.7 (18.03)	0.26
Vitality	52.9 (20.49)	52.8 (21.29)	0.91
PROMIS-29			
Physical Health	43.8 (8.13)	43.5 (8.15)	0.40
Mental Health	50.4 (7.55)	50.3 (7.65)	0.69
Anxiety (–)	49.8 (8.1)	49.2 (7.97)	0.21
Depression (–)	48.5 (7.79)	48.2 (7.46)	0.45
Fatigue (–)	51.0 (8.68)	51.1 (8.79)	0.40
Pain Interference (–)	53.2 (8.93)	52.7 (8.6)	0.23
Physical Function (+)	42.9 (7.77)	42.7 (7.81)	0.61
Sleep Disturbance (–)	47.2 (8.41)	47.3 (8.46)	0.83
Social Roles (+)	51.0 (8.61)	50.8 (9.15)	0.60

Variable	First-Round Survey, Mean (SD)	Second-Round Survey, Mean (SD)	*p*-Value
Pain Intensity (score 0–10) (–)	2.99 (2.25)	2.95 (2.2)	0.83

NOTE: For the VR-36, higher scores are always indicative of better HRQoL. For the PROMIS-29, (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

Changes in Health-Related Quality of Life over Time

It is important to understand how HRQoL scores changed over time in our population, because it is these changes on which PRO-based PMs will be based. We therefore compared HRQoL scores for the 337 participants who responded to both the first- and second-round surveys to see whether those scores changed over time (Table 4.2).

The results in Table 4.2 demonstrate that there was little change in mean HRQoL scores between the first and second survey administrations, which were performed six months apart (a decision dictated by the duration of the contract and limitations of the budget). Although the general pattern of HRQoL, especially physical HRQoL, is to decline with increasing age, the decline that would be expected over such a short period is minimal (Crosby, Kolotkin, and Williams, 2003). This emphasizes a limitation of our study design—namely, that measuring changes in HRQoL over a short interval, which limits the potential for HRQoL to change meaningfully, also limits the ability to use such changes to profile performance (Crosby, Kolotkin, and Williams, 2003). The only score that was statistically significantly lower at the second assessment was the physical functioning score of the VR-36, which declined by 2 points. No other score, including the Physical Function score of the PROMIS-29, showed a statistically significant change. Although overall population mean scores did not change, some people did experience score changes in some cases. This will be the topic that the PMs will address.

Task 2: Construct Patient-Reported Outcome–Based Performance Measures

Our PRO-based PMs were based on those that we had examined in the phase 1 report, using data from the Medicare HOS (see Kazis, Rogers, et al., 2017, p. 8, Table 1.5). The first group of PMs (group 1) was based on the status of being alive with a stable or improved score for some aspect of HRQoL. The second group of PMs (group 2) was based on the mean change in an aspect of HRQoL. From these two concepts, we created a total of eight candidate PMs: four based on the two summary scores of the VR-36 (the PCS and MCS) and four based on the two summary scores for the PROMIS-29 (Physical Health and Mental Health). An additional 16 potential PMs were based on the eight domain scores of the PROMIS-29 (Anxiety, Depression, Fatigue, Pain Intensity, Pain Interference, Physical Function, Sleep Disturbance, and Ability to Participate in Social Roles and Activities).

Results for these additional 16 PMs were generally similar to the main results and are provided in Table D.1 in Appendix D.

The numerator and denominator for the PMs consisted of the population who had responded to the survey in both the first and the second administrations (the 337 participants addressed in Tables 4.1 and 4.2). This requirement to respond to surveys at multiple time points as a criterion for inclusion in such a PM should be considered a potential limitation of such measures.

Table 4.3 sets forth the eight candidate PMs. For group 1, which are binary PMs, the percentage of respondents meeting the measure is given. For group 2, which contains continuous PMs, the mean change across the population is given, with the SD.

Table 4.3. Population-Level Performance on Each Performance Measure (total n = 337)

Group and Measure	Population Performance
Group 1: Percentage alive with stable or improved score	
VR-36–based measures	
PCS	48%
MCS	54%
PROMIS-29–based measures	
Physical Health	53%
Mental Health	53%
Group 2: Mean change in score (with SD)	
VR-36–based measures	
PCS	–0.71 (7.07)
MCS	0.36 (7.68)
PROMIS-29–based measures	
Physical Health	–0.23 (5.08)
Mental Health	–0.10 (4.74)

The upper part of the table shows that, for the PMs based on being alive with stable or improved HRQoL, approximately half of participants fulfilled each measure. In the lower part of the table, we show PMs based on the mean change in each of the four HRQoL scores. In general, these mean changes can be described as very small, especially in comparison to the SDs. This implies that, although some members of the population might experience changes in either direction, the overall net change of the population average is quite small by comparison.

We drew our participants from 13 distinct ambulatory clinics within the KPCO system, in which they received their primary care. These 13 clinics, here identified as clinic A through

clinic L, are the subject of Tables 4.4 and 4.5. The first table (Table 4.4) examines the four binary PMs based on being alive with stable or improved HRQoL at six months:

- PM 1a (alive with stable or improved PCS)
- PM 1b (alive with stable or improved MCS)
- PM 1c (alive with stable or improved PROMIS-29 Physical Health)
- PM 1d (alive with stable or improved PROMIS-29 Mental Health).

Table 4.4. Ambulatory Clinic–Level Performance on Four Binary (0/1) Performance Measures, with 95-Percent Confidence Intervals (total n = 337)

Clinic	Clinic n	PM 1a, PCS as a Percentage (*p*-value = 0.94) (95% CI)	PM 1b, MCS as a Percentage (*p*-value = 0.81) (95% CI)	PM 1c, PROMIS-29 Physical Health as a Percentage (*p*-value = 0.77) (95% CI)	PM 1d, PROMIS-29 Mental Health as a Percentage (*p*-value = 0.65) (95% CI)
A	25	52 (31–73)	56 (35–77)	50 (28–72)	58 (37–80)
B	22	41 (19–63)	50 (27–73)	61 (39–82)	61 (39–82)
C	42	52 (37–68)	50 (34–66)	63 (48–79)	44 (28–60)
D	17	53 (26–79)	65 (39–90)	65 (39–90)	53 (26–79)
E	37	46 (29–63)	62 (46–79)	53 (36–69)	45 (29–61)
F	19	53 (28–77)	53 (28–77)	45 (21–69)	55 (31–79)
G	27	63 (44–82)	52 (32–72)	44 (24–64)	67 (48–86)
H	17	53 (26–79)	29 (5–54)	44 (19–70)	61 (36–86)
I	23	43 (22–65)	57 (35–78)	55 (32–77)	68 (47–89)
J	38	42 (26–59)	55 (39–72)	49 (32–66)	51 (34–68)
K	27	44 (24–64)	59 (39–79)	44 (24–64)	44 (24–64)
L	32	44 (26–62)	50 (32–68)	63 (45–80)	50 (32–68)

NOTE: For each PM, the definition is "alive at 6 months with stable or improved . . ." with the HRQoL measure that was stable or improved described in the column heading. *p*-value tests whether the performance of any of the clinics differs from the overall mean across all 13 clinics: F-test with clinic-level fixed effects.

Many of the clinics have fairly small samples, a fact that affects the precision of the estimates and therefore limits the ability to distinguish performance among clinics. Performance on some measures tends to be higher and lower overall on others. However, none of the clinics differed from the group mean performance on any of the PMs at the 0.05 level of significance.

Table 4.5 examines clinic-level performance of the 13 ambulatory clinics using the four candidate PMs that are continuous:

- PM 2a (mean change in PCS)
- PM 2b (mean change in MCS)
- PM 2c (mean change in PROMIS-29 Physical Health)
- PM 2d (mean change in PROMIS-29 Mental Health).

Table 4.5. Ambulatory Clinic–Level Performance on Four Continuous Performance Measures, with Standard Deviation (total n = 337)

Clinic	Clinic n	PM 2a, PCS Mean (*p*-value = 0.67) (SD)	PM 2b, MCS Mean (*p*-value = 0.72) (SD)	PM 2c, PROMIS-29 Physical Health Mean (*p*-value = 0.53) (SD)	PM 2d, PROMIS-29 Mental Health Mean (*p*-value = 0.84) (SD)
A	25	–1.59 (7.91)	–0.36 (5.75)	–1.18 (5.10)	0.76 (4.98)
B	22	–0.83 (4.37)	0.33 (5.88)	0.52 (3.52)	0.83 (3.42)
C	42	0.52 (8.36)	–0.77 (7.72)	1.02 (5.23)	–0.38 (4.67)
D	17	0.19 (7.00)	–0.05 (4.62)	0.99 (4.06)	0.31 (4.87)
E	37	–2.63 (10.59)	1.29 (10.73)	–1.20 (6.85)	–1.51 (5.31)
F	19	–1.34 (4.87)	0.13 (9.95)	–0.96 (5.06)	0.08 (4.78)
G	27	0.46 (5.36)	–0.32 (7.84)	–0.07 (3.83)	0.5 (5.28)
H	17	0.90 (5.53)	–1.05 (5.60)	–0.72 (2.69)	–0.25 (4.57)
I	23	–1.79 (5.91)	1.64 (7.88)	–1.46 (5.43)	0.58 (4.96)
J	38	–0.78 (6.30)	0.78 (7.38)	–0.58 (5.16)	–0.11 (3.86)
K	27	–1.85 (7.30)	3.30 (8.64)	–0.47 (5.13)	–0.30 (4.86)
L	32	0.49 (5.54)	–0.86 (5.85)	0.92 (5.31)	–0.39 (5.25)

NOTE: For each PM, the definition is "mean change in . . .at six months" with the applicable HRQoL measure described in the column heading. *p*-value tests whether the performance of any of the clinics differs from the overall mean across all 13 clinics: F-test with clinic-level fixed effects.

The mean change between the first and second survey administrations on all of the HRQoL scores was very small. However, the SDs were large, indicating that some people might have changed by a considerably greater margin in both the upward and downward directions. None of the clinics' performance on any of the PMs differed from the group mean performance at the 0.05 level of statistical significance.

Task 3: Perform Reliability Testing of Performance Measures, and Task 4: Assess the Potential of These Performance Measures to Distinguish Between Providers, Practices, or Health Plans

Having specified eight potential PRO-based PMs, we proceeded to formally evaluate their reliability and their ability to distinguish performance across ambulatory clinic sites. The data collected for phase 2 (the current analyses) consisted of 13 ambulatory clinics at KPCO. Using these clinics as our cohort, we examined the ability of these 24 PMs to distinguish among clinics' performance. The performance of PMs can be measured in several ways. Here, we used a widely accepted approach based on calculation of ICC and calculation of reliability coefficients. This approach also matches the analytic approach that we used for the phase 1 analyses.

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 (NQF, 2013).

Our approach to measuring reliability began with the calculation of ICC, which is independent of sample size. Here, we calculated the ICCs for continuous outcomes using random-effect models within the immer package in the R statistics program to determine the proportion of the variance that the clinic explains. For binary outcomes, we used the aod package in R, with the function iccbin(). We applied other procedures, 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.

Once ICC has been calculated, reliability (r) can then be calculated from ICC using the Spearman–Brown prediction formula, where n is the sample size; in our case, this is the mean number of respondents per ambulatory clinic, measured across all the clinics (Brown, 1910; Spearman, 1910):

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

This equation implies that, for a given ICC, reliability 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 for a given ICC:

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

A PM must have sufficient reliability to demonstrate differences between providers that are due to actual differences in performance rather than chance alone. Increasing the sample size can improve reliability; here, we examine the reliability that was achieved with these PMs and these sample sizes. Although there is no firm minimum reliability that is considered

sufficient for a PM, the general consensus of most previous studies is that a minimum reliability of 0.7 to 0.8 is needed to allow valid comparisons among providers (Sequist et al., 2011; Adams, 2009). A minimum reliability level of 0.7 is noted in the NQF-commissioned paper on PRO PMs (NQF, 2013).

Accordingly, for each of the eight performance measures we examined, we present three calculated statistics: (1) ICC, (2) reliability estimated using the available sample, and (3) calculated minimum number of respondents per site that would be necessary to achieve a reliability of 0.7. Table 4.6 presents these statistics for our entire data set, examining the ability of these eight PMs to distinguish performance among the 13 KPCO ambulatory clinics.

Table 4.6. Intraclass Correlation and Reliability, by Measure (total n = 337)

Group and Measure	Mean Patients per Site (SD)	ICC	Reliability	Minimum n for Reliability ≥ 0.7
1: Alive with stable or improved score				
1a: PCS	27.2 (8.42)	<0.001	0	N/A
1b: MCS	27.2 (8.42)	0	0	N/A
1c: PROMIS-29 Physical Health	27.3 (8.36)	<0.001	0	N/A
1d: PROMIS-29 Mental Health	27.3 (8.36)	<0.001	0	N/A
2: Mean change in score				
2a: PCS	27.2 (8.42)	0	0	N/A
2b: MCS	27.2 (8.42)	0	0	N/A
2c: PROMIS-29 Physical Health	27.3 (8.36)	0.001	0.03	2,317
2d: PROMIS-29 Mental Health	27.3 (8.36)	0	0	N/A

NOTE: "Minimum n for Reliability ≥ 0.7" indicates the minimum mean number of respondents per profiled entity that is required to ensure reliability of at least 0.7.

The sample size per site of care, after excluding respondents whose data were too incomplete to calculate PMs, was about 26. Most ICCs were 0, meaning that reliability was also 0. Only PM 2c (Mean Change in PROMIS-29 Physical Health) had an ICC of 0.001, meaning that site of care explained approximately 0.1 percent of the variation on that measure. It was not possible to calculate reliability for the other seven PMs or to estimate the minimum number required to achieve a reliability of 0.7. For PM 2c, to achieve a reliability of 0.7, it would have been necessary to collect 2,317 responses for each site of care to be

profiled. Such a requirement would greatly limit the ability to use such a measure at the clinic level and might restrict its use in comparing health plans or other large entities.

Table 4.7 shows a similar calculation after removing all respondents who used proxies to help answer the survey at baseline, at six months, or both. Proxy response could theoretically reduce reliability if such responses do not accurately reflect the actual health of enrollees. Using the signal-to-noise analogy described above, random misallocation would lead to increased noise, making it more difficult to detect any underlying signal.

Table 4.7. Intraclass Correlation and Reliability, by Measure, After Deleting Proxy Responses (total n = 212)

Group and Measure	Mean Patients per Site (SD)	ICC	Reliability	Minimum n for Reliability ≥ 0.7
1: Alive with stable or improved score				
1a: PCS	16.8 (5.98)	0	0	N/A
1b: MCS	16.8 (5.98)	<0.001	0.01	7,218
1c: PROMIS-29 Physical Health	17.1 (6.16)	0	0	N/A
1d: PROMIS-29 Mental Health	17.1 (6.16)	0	0	N/A
2: Mean change in score				
2a: PCS	16.8 (5.98)	0	0	N/A
2b: MCS	16.8 (5.98)	0	0	N/A
2c: PROMIS-29 Physical Health	17.1 (6.16)	0	0	N/A
2d: PROMIS-29 Mental Health	17.1 (6.16)	0.01	0.13	266

NOTE: "Minimum n for Reliability ≥ 0.7" indicates the minimum mean number of respondents per profiled entity that is required to ensure reliability of at least 0.7.

Removing proxy responses decreased the sample size by about one-third, leaving an analyzed sample size per site of about 16. However, removing proxy responses improved estimated ICC for some measures. There were now two measures (PM 1b and PM 2d, MCS and PROMIS-29 Mental Health, respectively) for which it was possible to calculate a minimum number needed to sample to achieve reliability of 0.7 or better. For PM 2d, the number of respondents needed per site of care would be 266 or more. Although this is approximately ten times as large as the 26 patients per site that we had in our study, it might be possible to collect this number, especially at large sites of care or within health plans.

Chapter Five. Discussion and Policy Implications

This chapter begins with a discussion of key limitations of this work that should be considered when interpreting this report. It then incorporates the results from Chapters Three and Four. We have organized the discussion by key question from Chapter Three (each of the six key questions) and by task from Chapter Four (we discuss tasks 1 and 4; tasks 2 and 3 are instrumental to task 4 and are not discussed). The chapter concludes with a consideration of the bigger-picture implications of this work on future policy efforts and a discussion of the three main policy-relevant recommendations to emerge from this report:

- It is feasible to collect HRQoL data among older adults with MCC in an integrated health system.
- The PROMIS-29 is valid for use in older adults with MCC.
- Future efforts to develop PRO-based PMs require longer follow-up intervals and larger sample sizes than we used.

Key Limitations

Limitation 1: Six-Month Follow-Up Interval

Perhaps the most important limitation to consider in interpreting this report is the six-month interval between the first and second web-based surveys, an interval that was dictated by the terms of the contract. Meaningful changes in HRQoL are not typically expected over a six-month period; indeed, we saw very little overall population change on any aspect of HRQoL during the course of our study (see Table 4.2 in Chapter Four) (Crosby, Kolotkin, and Williams, 2003). The results from phase 1 of this contract might provide some insight: Those data reflected used a two-year follow-up interval, but none of these PMs tested achieved a reliability of 0.7. The highest reliability achieved was 0.60, and much lower reliabilities were seen with most PMs, with an average sample size of 546 respondents per entity being profiled. Although it remains unclear what follow-up interval would best support PRO-based PMs, our phase 1 results suggest that even two years is unlikely to be a sufficiently long follow-up interval.

Limitation 2: Limited Sample Size

The sample size (1,359 respondents for round 1 and 337 respondents for round 2) matches the recruitment targets that were set forth in the contract and is dictated in large part by budgetary limitations and the number of eligible patients available at KPCO. At first glance, the number of respondents might look sufficient, and, in fact, the numbers for validation of the PROMIS-29 in the MCC population were adequate. However, a sample of

337 total respondents was likely insufficient for the purpose of profiling 13 ambulatory clinics, each of which had an average of 26 respondents. In our calculations, we found that the PM with the highest ICC would have required at least 266 respondents per clinic. This number would likely have been lower had we used a follow-up interval longer than six months because it would have increased the signal-to-noise ratio. However, it seems clear that future efforts to test PRO-based PMs should aim to have more than 26 patients per entity being profiled. In our phase 1 report, we achieved a maximum reliability of 0.60 (falling short of the minimum acceptable reliability of 0.7), despite having a sample size of approximately 500 respondents per entity profiled, as well as a two-year follow-up interval. This implies that future efforts to develop valid PRO-based PMs should use a sample larger than 500 patients per entity being profiled, a longer follow-up interval than two years, or both.

Limitation 3: Response Rate

Survey response rate varied by mode. Considering only those who submitted sufficient data to support analyses, the response rates were 25 percent for the web mode, 58 percent for mail only, 44 percent for web–mail, 50 percent for phone only, and 46 percent for web–phone. A response rate of around 50 percent (or lower) is typical for many surveys but can be associated with bias if responders differ systematically from nonresponders (e.g., if healthier participants are more likely to respond than those who are sicker). Biased response, when present, can lead to samples that are less than fully representative of the population being studied (Krosnick, 1999; Rogelberg and Stanton, 2007). Although web administration holds great promise for making surveys such as this less expensive and more feasible, a 25-percent response rate for the web mode further accentuates these concerns. Nevertheless, it should be noted that we found only minimal differences between responders and nonresponders (see Table 3.3 in Chapter Three) and, in many cases, minimal differences in respondent characteristics among survey modes (see Tables 3.2, 3.4, and 3.5, also in Chapter Three). These findings are reassuring regarding the magnitude of any bias that might have been introduced by nonresponse and differential response based on mode.

Limitation 4: Relative Homogeneity of Sample

It is also worth noting that the 13 KPCO clinics that we studied, all of which are located in a single state (Colorado), are likely to share many characteristics in common, from the clinical protocols that they use to the populations that they treat. Although we acknowledge that there might also be important differences among these clinics, they likely have more in common than 13 unaffiliated ambulatory clinics located in different states might. This relative homogeneity could have contributed to the relatively low ICCs that we found in our PM analyses (see Tables 4.6 and 4.7 in Chapter Four). Although the common data structure of KPCO undoubtedly aided our data collection efforts, future efforts to validate PRO-based

PMs might wish to use a more heterogeneous sample, both in terms of heterogeneity among sites and in terms of heterogeneity among populations treated.

Discussion of Item Validation

Questions 1 and 2: What Are the Characteristics of Baseline Respondents and Their Impact on the Likelihood of Response to the Survey?

The study population for our primary data collection from KPCO is unlike any previous study of HRQoL. Because of oversampling of the oldest old (age 80+), our sample had a mean age of 80.7 years, although the mean age of web respondents was somewhat lower (79.3 years). By design, every respondent had at least two of 13 predefined conditions, and 16 percent had five or more. In addition to their chronic diseases, 13 percent had been hospitalized in the previous year and 46 percent had received at least some home health care. On PCS and the analogous PROMIS-29 Physical Health summary measure, respondents scored approximately 1 SD below the general population; however, they scored at or somewhat higher than the population mean for overall mental well-being. A picture emerges of a unique population of older patients who receive primary care through an integrated delivery system in Colorado—each of whom has at least a moderate burden of chronic disease but whose emotional well-being is at or above population norms.

To our knowledge, no previous study has collected data of similar sample size from people with MCC nor from participants of such advanced age or in the context of ambulatory care clinics rather than a clinical trial. The present study demonstrated the feasibility of data collection in such a sample. It also showed that data can feasibly be collected through more than one survey modality. It is particularly noteworthy that the web-based modality worked as well as it did in this older population. Although the response rate was lower than that of patients assigned to the other modalities (25 percent versus about 50 percent for the other modalities), the web modality has substantial advantages. It is easier and less expensive to use the web modality, potentially allowing larger numbers of people to be invited to participate. Also noteworthy is the 84-percent response rate on the second web survey among those who had already completed the first web survey. This implies that serial, longitudinal web-based surveys are feasible, even with this older, sicker population. Representativeness of these respondents relative to the studied population remains a concern, however.

We also investigated response likelihood as a function of subject characteristics. In general, the results were reassuring, in that response was not highly conditional on particular characteristics. Age was not a significant predictor of survey response, nor was the number of chronic conditions. There were indications that subjects with particularly severe health conditions were somewhat less likely to respond. For example, congestive heart failure (AOR 0.68) and depression (AOR 0.72) were both statistically significantly associated with lower

odds of survey response. Similarly, someone who had been hospitalized once (AOR 0.70) or twice (AOR 0.63) in the past year was less likely to respond. However, these differences were not particularly large. The most worrisome result, with regard to representativeness, pertained to race and ethnicity: nonwhite non-Hispanics (AOR 0.58) and, especially, Hispanics (AOR 0.47) were among the least likely to respond. Overall, the results imply that a fairly representative subset of subjects responded to the survey, with somewhat lower response rates by those with particularly severe illness and especially by nonwhite or Hispanic subjects.

Questions 3 and 4: What Effect Do Survey Mode and Respondent Characteristics Have on Health-Related Quality-of-Life Scores?

Generally, we found that web respondents were somewhat healthier, in terms of both physical and mental health, than those who responded to the survey via mail or phone. These differences, on the order of one-quarter SD, were statistically significant but of a small effect size (Cohen, 1969). Better HRQoL on the part of web respondents could have been due to their slightly younger age, but these differences persisted after case mix adjustment, which implies that age is not the explanatory factor. It seems that healthier seniors are more comfortable using computers and thus more likely to reply to a web survey modality than their less healthy peers, which, in turn, might correlate with higher income, wealth, or education. We did not collect data to directly examine this supposition, but it could be examined in future studies of older survey respondents.

We also examined the relationship between other respondent characteristics and HRQoL scores. The performance of the VR-36 and the PROMIS-29 has been examined in the past for population subgroups, but never with respondents of such advanced age. We observed a predictable decline in scores on measures of physical HRQoL with advancing age; declines in mental HRQoL scores were observed but were much less pronounced. Female respondents reported lower HRQoL, with a larger decrement in physical than mental HRQoL scores. Nonwhite, especially nonwhite Hispanic, respondents reported lower HRQoL, a result that mirrors findings in younger populations (Zack, 2013). Having more chronic conditions, having been hospitalized, and having received home health care were each associated with lower HRQoL, with a much larger decline in physical than mental HRQoL scores. We anticipated all of these findings: that older and sicker patients would record lower HRQoL, especially in the physical domain. Therefore, these findings are generally reassuring, implying that the VR-36 and the PROMIS-29 function appropriately in this older, sicker patient sample.

Question 5: Is There Differential Item Functioning for Items of the Patient-Reported Outcomes Measurement Information System 29-Item Survey?

We undertook an extensive search for DIF, performing a total of 754 analyses across 26 binary stratifying variables and 29 survey items. Only one variable–item pair met our preset criteria for a statistically significant effect of nontrivial magnitude—namely, proxy response to the survey and Physical Function question 4 ("Are you able to run errands and shop?"). The interpretation of this item's functioning is that proxy respondents systematically judged their respondents' ability to run errands and shop as higher than would be expected based on case mix adjustment and the responses to the other Physical Function items. It might be that proxy respondents answered that shopping and errands are not a problem because the study subject was having these errands done for him or her. Overall, these findings support the conclusion that the PROMIS-29 items do not have relevant DIF when used in a population of older patients with MCC. It is particularly noteworthy that we found no evidence of DIF according to survey administration mode.

Question 6: Do the Risk-Adjusted Patient-Reported Outcomes Measurement Information System 29-Item Survey Scores Have Construct Validity?

The construct validity of the PROMIS-29 has already been extensively evaluated in many populations (Alcantara, Ohm, and Alcantara, 2016; Beaumont et al., 2012; Craig et al., 2014; Hinchcliff et al., 2011; IsHak et al., 2017; Katz, Pedro, and Michaud, 2017; Lai et al., 2017; Pearman et al., 2016; Schnall et al., 2017), but never in a population of such advanced age and high illness burden. We examined criterion validity for PROMIS-29 measures in two ways. First, we compared risk-adjusted PROMIS-29 scores at different levels of primary care utilization. Our hypothesis was that high utilizers of primary care would have lower risk-adjusted HRQoL. This would support the idea that, even after controlling for demographics and illness burden, the PROMIS-29 measures aspects of HRQoL that are not captured by the other variables. Our findings generally supported this hypothesis, in that all variables showed a linear trend in the expected direction (lower HRQoL among patients with higher utilization of primary care visits). The statistical power of these analyses was limited. We examined the PROMIS-29's ability to capture only those aspects of HRQoL not captured by measured covariates, such as demographics and chronic health conditions. The risk-adjustment models for the PROMIS-29 summary scores had R^2 values of 0.30 for PROMIS-29 Mental Health and 0.37 for PROMIS-29 Physical Health scores, indicating that measured patient-level variables accounted for a large proportion of the variation in those measures. Because the covariates are strongly correlated with physical and mental HRQoL, the measures' ability to account for additional variation is noteworthy.

Second, we examined correlations between the eight PROMIS-29 scales and the two PROMIS-29 summary scales and the VR-36 PCS and MCS, summary scales of another well-

established HRQoL measure. Although the VR-36 has not previously been administered in a sample with so many participants over the age of 80, it has been used in samples of patients whose illness burdens are similar to or even higher than those in the current sample (Kazis, Miller, Clark, Skinner, Lee, Rogers, et al., 1998). Because the validity of the VR-36 has been established among those with an extremely high burden of chronic illness, it is a fitting vehicle for evaluating the construct validity of the PROMIS-29 in our sample. We used a traditional cutoff of a correlation of 0.4 or higher as constituting a "moderate" strength of correlation. We found that four of the PROMIS-29 scales (Physical Function, Ability to Participate in Social Roles and Activities, Pain Interference, and Pain Intensity) loaded highly on PCS. Two PROMIS-29 scales (Anxiety and Depression) loaded highly on MCS. One PROMIS-29 scale (Fatigue) loaded highly on both PCS and MCS, and one (Sleep Disturbance) did not load highly on either PCS or MCS and is measuring an aspect of HRQoL not substantially measured by the VR-36. The two summary scales of the PROMIS-29 are highly intercorrelated, and each loads more heavily on PCS than on MCS.

In general, these results support the validity of scores of both instruments for measuring HRQoL. All of the correlations were in the expected direction and statistically significant. However, these results also indicate that the PROMIS-29 is measuring a somewhat different aspect of HRQoL than the VR-36—both by the inclusion of Sleep Disturbance (a construct not directly measured by the VR-36) and because both of the PROMIS-29 summary scores correlate more strongly with the PCS than with the MCS.

Discussion of Utility of Items as Performance Measures

Task 1: Examine the Response Rate in the Second-Round Survey to Assess Threats to Validity

The response rate to the second-round survey was 84 percent, among 401 subjects, all of whom had responded to the first-round web survey. We compared demographics and comorbid conditions between responders and nonresponders to the second-round survey. Groups that were statistically significantly less likely to respond to the second-round survey included those of nonwhite race, those with stroke, and those who had recorded proxy responses to the first survey. A comparison of VR-36 summary scores showed that, on average, nonresponders scored approximately 2 points lower on the PCS and the MCS in the first-round survey, differences that were not statistically significant. This difference is somewhat less than a quarter of an SD and thus constitutes a small effect (Cohen, 1969). In summary, we can conclude that the response rate to the second-round survey was quite high, arguing both for the feasibility of multiround longitudinal web surveys in a population such as this and for the representativeness of the reduced number who respond to a second survey.

Task 4: Assess Potential for Performance Measures to Demonstrate Differences; Propose Suggestions for Future Collection of Patient-Reported Outcome Performance Measures

We examined the potential of eight PMs, which focus on change in HRQoL over time, to profile sites of care:

- PM 1a (alive with stable or improved PCS)
- PM 1b (alive with stable or improved MCS)
- PM 1c (alive with stable or improved PROMIS-29 Physical Health)
- PM 1d (alive with stable or improved PROMIS-29 Mental Health)
- PM 2a (mean change in PCS)
- PM 2b (mean change in MCS)
- PM 2c (mean change in PROMIS-29 Physical Health)
- PM 2d (mean change in PROMIS-29 Mental Health).

In the base case analysis (which includes all respondents with sufficiently complete data to compute scores), the ICC was 0 for most PMs, meaning that reliability could not be calculated. Only one measure (PM 2c, mean change in PROMIS-29 Physical Health) achieved an ICC of 0.001, which means that site of care accounts for 0.1 percent of the variation on this PM. In previous studies using outcome measures, such as hemoglobin A1c or measures of anticoagulation control, PMs have achieved ICCs of 0.03 or 0.04 (Hofer et al., 1999; Rose et al., 2011).

Given the ICC obtained for PM 2c, we calculated that it would require 2,317 respondents per entity being profiled to obtain a reliability of 0.7. This number exceeds the number of patients in the panels of individual clinicians. Although this is not applicable to most medical practices, such a number could be achieved in order to compare health plans or other large entities. However, requiring this many responses raises other issues, including the considerable cost of collecting 2,317 responses per entity. If 100 entities were being compared, this would require the cost and effort of collecting 231,700 surveys.

A sensitivity analysis performed after removing proxy responses did not produce meaningfully different results. It is worth noting that the one PM with a feasible sample size requirement in the base case analysis (PM 2c) differed from the two PMs that had feasible sample size requirements in the no-proxy analysis (PM 1b and PM 2d). Thus, we do not have a strong and consistent signal in favor of any one PM being feasible for measurement with achievable sample sizes. It is also worth noting that we did not adjust these PMs for case mix at the 13 ambulatory clinics being profiled. To the extent that populations differ among these clinics, the ICCs might have been even lower after case mix adjustment.

Policy Implications

1. *It Is Feasible to Collect Health-Related Quality-of-Life Data Among Older Adults with Multiple Chronic Conditions in an Integrated Health System*

One of the policy-relevant goals of this project was to establish the feasibility of HRQoL measurement among this population, which oversampled the oldest old and consisted entirely of patients with at least two of 13 particular chronic medical conditions. In this project, we demonstrated that it is feasible to collect HRQoL data from this population, in large part because of the strong relationships and data collection infrastructure that exist at KPCO. Because of their existing long-standing relationships with their members, their highly curated address and contact files, and their strong data infrastructure, our KPCO colleagues were able to work well with SRG to achieve a high response rate in this population. It is noteworthy that this older population had such a high response rate to a web-based survey and that the response rate remained high (84 percent) in the second round of the web survey. Our data also suggest that web respondents were only slightly younger and healthier than those who responded to other survey modalities, which we did not judge to be a serious threat either to validity or to generalizability. This study, therefore, can serve as a model for how to feasibly collect HRQoL data using a web-based survey. Situating this study in the context of KPCO was a strength, in that it enhanced feasibility and enhanced our chances to complete the study successfully. However, it will also be important to extend the work of this study outside the context of an integrated health system. This will help demonstrate that it is also feasible to collect these sorts of data from older, sicker adults in a more routine setting, without the advantages conferred by an integrated health system, such as KPCO.

2. The Patient-Reported Outcomes Measurement Information System 29-Item Survey Is Valid for Use in Older Adults with Multiple Chronic Conditions

Another policy-relevant goal was to assess the PROMIS-29's validity in this population of older and sicker people. We demonstrated validity in three main ways. First, we did so based on a finding of expected variation with underlying respondent characteristics. For example, older respondents showed greater declines in physical than mental health, as would be expected. Second, we demonstrated validity by assessing for the presence of DIF. After an exhaustive search for DIF, we essentially did not find any. Third, we assessed the construct validity of the PROMIS-29 in this population. We found that the PROMIS-29 was functioning as anticipated and generally capturing HRQoL in the expected way, as stratified by primary care utilization and by comparison with another well-recognized HRQoL instrument (the VR-36). The PROMIS-29 scores performed as expected in this older, sicker population of respondents, supporting its use in this population.

3. Future Efforts to Develop Patient-Reported Outcome–Based Performance Measures Require Longer Follow-Up Intervals and Larger Sample Sizes Than We Used

Lastly, our results shed light on the potential collection and use of PRO-based PMs to compare performance across health care providers, ambulatory clinics, and health plans. We assessed eight main PMs: four based on being alive and with stable or improved HRQoL at six months and four based on mean change in HRQoL over six months, analogous to the measures examined in phase 1 of the study using existing data from the Medicare HOS. We found that most of the measures we studied did not achieve any measurable reliability when deployed under these conditions (e.g., six-month follow-up interval, the sample size used). A few measures could potentially achieve acceptable reliability (0.7 or higher), but only with very large sample sizes. These results are not encouraging for the prospect of using these PRO-based PMs, under these conditions, to profile health care providers, physician practices, or health plans.

However, some caveats should be noted. An interval of only six months might be too short for measuring average changes in HRQoL. Changes in HRQoL take years to be fully realized, and any effect of high-quality care on slowing declines in HRQoL would also occur over years or even decades. Because, on average, HRQoL does not change markedly over a six-month interval, it might be unrealistic to expect any tool to be able to discriminate among sites in this time frame (Crosby, Kolotkin, and Williams, 2003). Therefore, it might be premature to give up on the idea of PRO-based PMs solely because of the results of the present study. Indeed, a key lesson learned from the present effort might be that larger samples—and longer time frames—would be needed to discriminate among sites based on HRQoL.

One option to increase sample size might be to compare larger entities, such as health plans. However, sometimes there is a need to compare entities smaller than health plans, so this might not always be the best solution. Another option would be to increase the follow-up time, which could create issues of feasibility due to cost, or erode the temporal immediacy of the results when they are finally available. Longer follow-up intervals might also accentuate other issues, such as loss to follow-up or patients changing providers, which would complicate the process of attribution.

Although our analyses do incorporate conventional measures of uncertainty in the estimates, it is also important to remember that estimates based on so few observations—26 respondents per site, on average, in our study—will be unstable. Given our phase 1 results, it is likely that a sample size closer to 500 respondents per site of care might be necessary, although this could potentially decrease with a longer follow-up time.

Remaining Questions

As we noted in the phase 1 report (Kazis, Rogers, et al., 2017), questions that remain open include how much of the decline in HRQoL over time is amenable to being slowed by the provision of high-quality health care, and how much of this is under the control of individual providers. Our study does not address these questions directly, but they are worthy of future efforts and indeed important to investigate. It could be that some aspects of HRQoL could be more responsive—that is, easier to slow the decline of functioning and well-being—through high-quality care. But other aspects might prove immutable. It will likely prove challenging to discern which aspects can be improved through better care and which cannot. But being able to focus on improving those aspects of care will surely be worth the effort.

Appendix A. International Classification of Diseases, Tenth Revision, Codes for Chronic Conditions

Table A.1. International Classification of Diseases, Tenth Revision, Codes for Chronic Conditions

Brief Title	Full Plain-English Title	Code
Arthritis	Arthritis of the hip, knee, hand, or wrist	M16.x, M17.x, M18.x, M19.03X, M19.04x, M19.13x, M19.14x, M19.23x, M19.24x
Cancer	Cancer of the colon, rectum, lung, breast, or prostate	C50.x, C61.x, C34.x, C18x–C20x
Chronic lung disease	Emphysema, or asthma, or COPD (chronic obstructive pulmonary disease)	I27.8, I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Congestive heart failure	Congestive heart failure	I09.81, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x
Depression	Depression	F32.x, F33.x
Diabetes	Diabetes, high blood sugar, or sugar in the urine	E08x–E13x
Hypertension	High blood pressure	I10.x, I11.x–I13.x, I15.x
Inflammatory bowel disease	Crohn's disease, ulcerative colitis, or inflammatory bowel disease	K50.x–K51.x
Ischemic heart disease	Angina pectoris, coronary artery disease, myocardial infarction, or heart attack	I20.x–I25.x
Osteoporosis	Osteoporosis, sometimes called thin or brittle bones	M80.x, M81.x
Other heart problems	Other heart conditions, such as problems with heart valves or the rhythm of your heartbeat	I44.1–I44.3, I45.6, I45.9, I47.x–I49.x, R00.0, R00.1, R00.8, T82.1, Z45.0, Z95.0, A52.0, I05.x–I08.x, I09.1, I09.8, I34.x–I39.x, Q23.0–Q23.3, Z95.2–Z95.4
Sciatica	Sciatica (pain or numbness that travels down your leg to below your knee)	M54.3x, M54.4x
Stroke	Stroke	I60.x, I61.x, I62.x, I63.x, I69.x

Draft

The Quality of Life

In Older Adults Study

Baseline Survey

SURVEY INSTRUCTIONS:

- Please mark ONE answer only for each of these questions.
- Your answers to this survey will be kept confidential.
- Please complete the entire survey as carefully as you can.
- Some questions may seem redundant. There are subtle but important differences among the questions, so it is very important that you answer each one.
- If you are not sure how to answer a question or want to provide any comments please make a note in the margin.
- If you have questions about this survey or about the study, please contact Ms. Rosa-Elena Garcia, Associate Survey Director, at 1-800-269-5817 or at rosaeg@rand.org.

Draft

We would like to start with some general health questions.

A1. In general, would you say your health is:

- ☐ Excellent
- ☐ Very Good
- ☐ Good
- ☐ Fair
- ☐ Poor

A2. Compared to one year ago, how would you rate your physical health in general now?

- ☐ Much better
- ☐ Slightly better
- ☐ About the same
- ☐ Slightly worse
- ☐ Much worse

A2B. Compared to one year ago, how would you rate your emotional problems (such as feeling anxious, depressed or irritable) now?

- ☐ Much better
- ☐ Slightly better
- ☐ About the same
- ☐ Slightly worse
- ☐ Much worse

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?

A3. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A4. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A5. Lifting or carrying groceries. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A6. Climbing several flights of stairs. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A7. Climbing one flight of stairs. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A8. Bending, kneeling, or stooping. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A9. Walking more than a mile. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

Draft

A10. Walking <u>several blocks</u>. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A11. Walking <u>one</u> block. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

A12. Bathing or dressing yourself. Does your health now limit you a lot, limit you a little, or not limit you at all?

- ☐ Yes, limited a lot
- ☐ Yes, limited a little
- ☐ No, not limited at all

<u>During the past 4 weeks</u>, have you had any of the following problems with your work or other regular activities, <u>as a result of your physical health</u>?

A13. <u>During the past 4 weeks</u>, have you cut down the <u>amount of time</u> you spent on work or other 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

A14. <u>During the past 4 weeks</u>, have you <u>accomplished less</u> than you would like, 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

A15. <u>During the past 4 weeks</u>, were you limited in the <u>kind</u> of work or other 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

A16. <u>During the past 4 weeks</u>, have you had <u>difficulty</u> performing the work or other activities (for example, it took extra effort), 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

Draft

The next questions ask about the <u>past 4 weeks</u> and any problems you may have had with your work or other regular daily activities <u>as a result of any emotional problems</u> such as feeling depressed or anxious.

A17. <u>During the past 4 weeks</u>, have you cut down on the <u>amount of time</u> you spent on work or other activities, as a result of any emotional problems?

- ☐ 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

A18. <u>During the past 4 weeks</u>, have you <u>accomplished less</u> than you would like, as a result of any emotional problems?

- ☐ 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

A19. <u>During the past 4 weeks</u>, didn't do work or other activities as <u>carefully</u> as usual, as a result of any <u>emotional problems</u>?

- ☐ 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

A20. <u>During the past 4 weeks</u>, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

- ☐ Not at all
- ☐ Slightly
- ☐ Moderately
- ☐ Quite a bit
- ☐ Extremely

A21. How much <u>bodily</u> pain have you had <u>during the past 4 weeks</u>?

- ☐ None
- ☐ Very Mild
- ☐ Mild
- ☐ Moderate
- ☐ Severe
- ☐ Very Severe

A22. <u>During the past 4 weeks</u>, how much did <u>pain</u> interfere with your normal work (including both work outside the home and housework)?

- ☐ Not at all
- ☐ A little bit
- ☐ Moderately
- ☐ Quite a bit
- ☐ Extremely

Draft

The following 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.

A23. How much of the time during the past 4 weeks, did you feel full of pep?

- ☐ 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

A24. How much of the time during the past 4 weeks, have you been a very nervous person?

- ☐ 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

A25. How much of the time during the past 4 weeks, have you felt so down in the dumps that nothing could cheer you up?

- ☐ 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

A26. How much of the time during the past 4 weeks, 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

A27. How much of the time during the past 4 weeks, 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

A28. How much of the time during the past 4 weeks, 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

Draft

A29. How much of the time during the past 4 weeks, did you feel worn out?

- ☐ 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

A30. How much of the time during the past 4 weeks, have you been a happy person?

- ☐ 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

A31. How much of the time during the past 4 weeks, did you feel tired?

- ☐ 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

A32. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?

- ☐ All of the time
- ☐ Most of the time
- ☐ Some of the time
- ☐ A little of the time
- ☐ None of the time

For the next items, please indicate how TRUE or FALSE each of the following statements is for you.

A33. I seem to get sick a little easier than other people.

- ☐ Definitely true
- ☐ Mostly true
- ☐ Don't know
- ☐ Mostly false
- ☐ Definitely false

A34. I am as healthy as anybody I know.

- ☐ Definitely true
- ☐ Mostly true
- ☐ Don't know
- ☐ Mostly false
- ☐ Definitely false

A35. I expect my health to get worse.

- ☐ Definitely true
- ☐ Mostly true
- ☐ Don't know
- ☐ Mostly false
- ☐ Definitely false

A36. My health is excellent.

- ☐ Definitely true
- ☐ Mostly true
- ☐ Don't know
- ☐ Mostly false
- ☐ Definitely false

The following questions are more about your <u>current physical health</u>.

B1. Are you able to do chores such as vacuuming or yard work?

- ☐ Without any difficulty
- ☐ With a little difficulty
- ☐ With some difficulty
- ☐ With much difficulty
- ☐ Unable to do

B2. Are you able to go up and down stairs at a normal pace?

- ☐ Without any difficulty
- ☐ With a little difficulty
- ☐ With some difficulty
- ☐ With much difficulty
- ☐ Unable to do

B3. Are you able to go for a walk of at least 15 minutes?

- ☐ Without any difficulty
- ☐ With a little difficulty
- ☐ With some difficulty
- ☐ With much difficulty
- ☐ Unable to do

B4. Are you able to run errands and shop?

- ☐ Without any difficulty
- ☐ With a little difficulty
- ☐ With some difficulty
- ☐ With much difficulty
- ☐ Unable to do

The following questions are about how you have felt in the <u>past 7 days</u>.

C1. In the <u>past 7 days</u>, I felt fearful.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

C2. In the <u>past 7 days</u>, I found it hard to focus on anything other than my anxiety.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

C3. In the <u>past 7 days</u>, my worries overwhelmed me.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

C4. In the <u>past 7 days</u>, I felt uneasy.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

Draft

D1. In the past 7 days, I felt worthless.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

D2. In the past 7 days, I felt helpless.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

D3. In the past 7 days, I felt depressed.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

D4. In the past 7 days, I felt hopeless.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

E1. In the past 7 days, I felt fatigued.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

E2. In the past 7 days, I had trouble starting things because I was tired.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

E3. In the past 7 days, how run-down did you feel on average?

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

E4. In the past 7 days, how fatigued were you on average?

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

The next questions ask about your sleep in the past 7 days.

F1. In the past 7 days, my sleep quality was…

- ☐ Very poor
- ☐ Poor
- ☐ Fair
- ☐ Good
- ☐ Very good

F2. In the past 7 days, my sleep was refreshing.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

Draft

F3. In the past 7 days, I had a problem with my sleep.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

F4. In the past 7 days, I had difficulty falling asleep.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

The next questions are about activities you may do.

G1. I have trouble doing all of my regular leisure activities with others.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

G2. I have trouble doing all of the family activities that I want to do.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

G3. I have trouble doing all of my usual work (include work at home).

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

G4. I have trouble doing all of the activities with friends that I want to do.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

The next questions are about pain in the past 7 days.

H1. In the past 7 days, how much did pain interfere with your day to day activities?

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

H2. In the past 7 days, how much did pain interfere with work around the home?

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

H3. In the past 7 days, how much did pain interfere with your ability to participate in social activities?

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

H4. In the <u>past 7 days</u>, how much did pain interfere with your household chores?

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

H5. In the <u>past 7 days</u>, how would you rate your pain on average?

- ☐ 0 - No pain
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7
- ☐ 8
- ☐ 9
- ☐ 10 - Worst imaginable pain

The next statements are about support you may have or get.

I1. I have someone who will listen to me when I need to talk.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

I2. I have someone to confide in or talk to about myself or my problems.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

I3. I have someone who makes me feel appreciated.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

I4. I have someone to talk with when I have a bad day.

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

J1. Do you have someone to help you if you are confined to bed?

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

J2. Do you have someone to take you to the doctor if you need it?

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

J3. Do you have someone to help with your daily chores if you are sick?

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

J4. Do you have someone to run errands if you need it?

- ☐ Never
- ☐ Rarely
- ☐ Sometimes
- ☐ Often
- ☐ Always

The next statements ask about how you have been feeling lately.

K1. Lately my life had meaning.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

K2. Lately my life was satisfying.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

K3. Lately I had a sense of balance in my life.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

K4. Lately I felt hopeful.

- ☐ Not at all
- ☐ A little bit
- ☐ Somewhat
- ☐ Quite a bit
- ☐ Very much

L1. How much difficulty do you <u>currently</u> have reading and following complex instructions (e.g., directions for a new medication)?

- ☐ None
- ☐ A little
- ☐ Somewhat
- ☐ A lot
- ☐ Cannot do

L2. How much difficulty do you <u>currently</u> have planning for and keeping appointments that are not part of your weekly routine, (e.g., a therapy or doctor appointment, or a social gathering with friends and family)?

- ☐ None
- ☐ A little
- ☐ Somewhat
- ☐ A lot
- ☐ Cannot do

L3. How much difficulty do you <u>currently</u> have managing your time to do most of your daily activities?

- ☐ None
- ☐ A little
- ☐ Somewhat
- ☐ A lot
- ☐ Cannot do

L4. How much difficulty do you <u>currently</u> have learning new tasks or instructions?

- ☐ None
- ☐ A little
- ☐ Somewhat
- ☐ A lot
- ☐ Cannot do

Draft

M1. To what degree have your health related expenses caused financial problems for you and your family?

- ☐ Not at all
- ☐ A little
- ☐ Somewhat
- ☐ A lot

N1. Did someone help you complete this survey?

- ☐ Yes
- ☐ No, no one helped me complete this survey → **Thank you. Please return the completed survey in the enclosed postage-paid envelope.**

N2. How did that person help you? Mark one or more.

- ☐ Read the questions to me
- ☐ Wrote down the answers I gave
- ☐ Answered the questions for me
- ☐ Translated the questions into my language
- ☐ Helped in some other way → Please describe how that person helped you:

Thank you for completing the survey!
Please return the completed survey in the enclosed postage-paid envelope.

Return to:

Ryan McKay
RAND Survey Research Group
1776 Main Street
PO Box 2138
Santa Monica, CA 90401-2138

Page Intentionally Left Blank


Draft


Draft

Appendix C. Complete Results of Analyses for Differential Item Functioning

This appendix is a spreadsheet accessible on the product page for this report (http://www.rand.org/pubs/research_reports/RR2176).

Appendix D. Intraclass Correlation Coefficients, Reliability, and Number Needed for Alternative Performance Measures Based on the Original Eight Patient-Reported Outcomes Measurement Information System 29-Item Survey Scores

Table D.1. Intraclass Correlation Coefficients and Reliability, by Measure (total n = 337)

Group and Measure	Mean Patients Per Site	ICC	Reliability	Minimum n for Reliability ≥ 0.7
1: Alive with stable or improved score				
Anxiety (–)	26.8	0	0	N/A
Depression (–)	25.7	<0.001	0	N/A
Fatigue (–)	26.0	<0.001	0	N/A
Pain Interference (–)	26.9	0.02	0.31	137
Physical Function (+)	26.2	<0.001	0	N/A
Sleep Disturbance (–)	27.2	0	0	N/A
Social Roles (+)	26.4	0.02	0.40	93
Pain Intensity (score) (–)	27.6	0	0	N/A
2: Mean change in score				
2c: Anxiety (–)	26.8	0.004	0.09	650
2d: Depression (–)	25.7	<0.001	0	N/A
2e: Fatigue (–)	26.0	0	0	N/A
2f: Pain Interference (–)	26.9	<0.001	0	N/A
2g: Physical Function (+)	26.2	<0.001	0	N/A
2h: Sleep Disturbance (–)	27.2	<0.001	0.01	7,107
2i: Social Roles (+)	26.4	0.02	0.35	113
2j: Pain Intensity (score) (–)	27.6	0	0	N/A

NOTE: “Minimum n for Reliability ≥ 0.7” indicates the minimum mean number of respondents per profiled entity that is required to ensure reliability of at least 0.7. (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

Table D.2. Intraclass Correlation Coefficients and Reliability, by Measure, After Deleting Proxy Responses (total n = 212)

Group and Measure	Mean Patients Per Site	ICC	Reliability	Minimum n for Reliability ≥ 0.7
1: Alive with stable or improved score				
Anxiety (–)	16.8	0	0	N/A
Depression (–)	16.0	0	0	N/A
Fatigue (–)	16.0	0.04	0.38	60
Pain Interference (–)	16.8	<0.001	0	N/A
Physical Function (+)	16.2	<0.001	0	N/A
Sleep Disturbance (–)	17.0	0.02	0.28	104
Social Roles (+)	16.4	<0.001	0	N/A
Pain Intensity (score) (–)	17.2	0	0	N/A
2: Mean change in score				
Anxiety (–)	16.8	<0.001	0	N/A
Depression (–)	16.0	<0.001	0	N/A
Fatigue (–)	16.0	0.06	0.49	39
Pain Interference (–)	16.8	0	0	N/A
Physical Function (+)	16.2	0	0	N/A
Sleep Disturbance (–)	17.0	0.03	0.34	78
Social Roles (+)	16.4	<0.001	0	N/A
Pain Intensity (score) (–)	17.2	0	0	N/A

NOTE: “Minimum n for Reliability ≥ 0.7” indicates the minimum mean number of respondents per profiled entity that is required to ensure reliability of at least 0.7. (–) means that a higher score indicates worse HRQoL, and (+) means that a higher score indicates better HRQoL. Social Roles = Ability to Participate in Social Roles and Activities.

References

Adams, John L., *The Reliability of Provider Profiling: A Tutorial*, Santa Monica, Calif.: RAND Corporation, TR-653-NCQA, 2009. As of October 21, 2017: https://www.rand.org/pubs/technical_reports/TR653.html

Alcantara, Joel, Jeanne Ohm, and Junjoe Alcantara, "The Use of PROMIS and the RAND VSQ9 in Chiropractic Patients Receiving Care with the Webster Technique," *Complementary Therapies in Clinical Practice*, Vol. 23, 2016, pp. 110–116. doi:10.1016/j.ctcp.2015.05.003.

Beaumont, Jennifer L., David Cella, Alexandria T. Phan, Seung Choi, Zhimei Liu, and James C. Yao, "Comparison of Health-Related Quality of Life in Patients with Neuroendocrine Tumors with Quality of Life in the General US Population," *Pancreas*, Vol. 41, No. 3, April 2012, pp. 461–466. doi:10.1097/MPA.0b013e3182328045.

Benjamini, Yoav, and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B: *Methodological*, Vol. 57, No. 1, 1995, pp. 289–300.

Bishawi, Muath, Annie Laurie W. Shroyer, John Rumsfeld, John A. Spertus, Janet H. Baltz, Joseph Collins, Jacquelyn A. Quin, G. Hossein Almassi, Frederick L. Grover, and Brack Hattler, "Changes in Health-Related Quality of Life in Off-Pump Versus On-Pump Cardiac Surgery: Veterans Affairs Randomized On/Off Bypass Trial," *Annals of Thoracic Surgery*, Vol. 95, No. 6, February 2013, pp. 1946–1951. doi:10.1016/j.athoracsur.2012.12.014.

Brown, William, "Some Experimental Results in the Correlation of Mental Abilities," *British Journal of Psychology, 1904–1920*, Vol. 3, No. 3, October 1910, pp. 296–322. doi:10.1111/j.2044-8295.1910.tb00207.x.

Centers for Disease Control and Prevention, *ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification*, effective October 1, 2015.

Cohen, Jacob, *Statistical Power Analysis for the Behavioral Sciences*, San Diego: Academic Press, 1969.

Corder, Larry S., Max A. Woodbury, and Kenneth G. Manton, "Proxy Response Patterns Among the Aged: Effects on Estimates of Health Status and Medical Care Utilization from the 1982–1984 Long-Term Care Surveys," *Journal of Clinical Epidemiology*, Vol. 49, No. 2, February 1996, pp. 173–182. doi:10.1016/0895-4356(95)00507-2.

Craig, Benjamin M., Bryce B. Reeve, Paul M. Brown, David Cella, Ron D. Hays, Joseph Lipscomb, A. Simon Pickard, and Dennis A. Revicki, "US Valuation of Health Outcomes Measured Using the PROMIS-29," *Value in Health*, Vol. 17, No. 8, December 1, 2014, pp. 846–853. doi:10.1016/j.jval.2014.09.005.

Crosby, Ross D., Ronette L. Kolotkin, and G. Rhys Williams, "Defining Clinically Meaningful Change in Health-Related Quality of Life," *Journal of Clinical Epidemiology*, Vol. 56, No. 5, May 2003, pp. 395–407. doi:10.1016/S0895-4356(03)00044-1.

Dalewitz, J., N. Khan, and C. O. Hershey, "Barriers to Control of Blood Glucose in Diabetes Mellitus," *American Journal of Medical Quality*, Vol. 15, No. 1, January–February 2000, pp. 16–25.

Donabedian, Avedis, "The Quality of Care: How Can It Be Assessed?" *JAMA*, Vol. 260, No. 12, September 23, 1988, pp. 1743–1748. doi:10.1001/jama.1988.03410120008001.

Gelin, Michaela N., and Bruno D. Zumbo, "Differential Item Functioning Results May Change Depending on How an Item Is Scored: An Illustration with the Center for Epidemiologic Studies Depression Scale," *Educational and Psychological Measurement*, Vol. 63, No. 1, 2003, pp. 65–74.

Goldberg, Jack, Kathryn M. Magruder, Christopher W. Forsberg, Lewis E. Kazis, T. Bedirhan Üstün, Matthew J. Friedman, Brett T. Litz, Viola Vaccarino, Patrick J. Heagerty, Theresa C. Gleason, Grant D. Huang, and Nicholas L. Smith, "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*, Vol. 23, No. 5, June 2014, pp. 1579–1591. doi:10.1007/s11136-013-0585-4.

Hays, R. D., B. G. Vickrey, B. P. Hermann, K. Perrine, J. Cramer, K. Meador, K. Spritzer, and O. Devinsky, "Agreement Between Self Reports and Proxy Reports of Quality of Life in Epilepsy Patients," *Quality of Life Research*, Vol. 4, No. 2, April 1995, pp. 159–168.

Health Services Advisory Group, "Medicare Health Outcomes Survey," last modified August 11, 2017. As of October 21, 2017: http://www.hosonline.org/

HealthMeasures, "Intro to PROMIS®," undated. As of October 31, 2017: http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis

Hinchcliff, M., J. L. Beaumont, K. Thavarajah, J. Varga, A. Chung, S. Podlusky, M. Carns, R. W. Chang, and D. Cella, "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 and Research*, Vol. 63, No. 11, November 2011, pp. 1620–1628. doi:10.1002/acr.20591.

Hofer, Timothy P., Rodney A. Hayward, Sheldon Greenfield, Edward H. Wagner, Sherrie H. Kaplan, and Willard G. Manning, "The Unreliability of Individual Physician 'Report Cards' for Assessing the Costs and Quality of Care of a Chronic Disease," *JAMA*, Vol. 281, No. 22, June 9, 1999, pp. 2098–2105. doi:10.1001/jama.281.22.2098.

IsHak, Waguih William, Dana Pan, Alexander J. Steiner, Edward Feldman, Amy Mann, James Mirocha, Itai Danovitch, and Gil Y. Melmed, "Patient-Reported Outcomes of Quality of Life, Functioning, and GI/Psychiatric Symptom Severity in Patients with Inflammatory Bowel Disease (IBD)," *Inflammatory Bowel Diseases*, Vol. 23, No. 5, 2017, pp. 798–803. doi:10.1097/mib.0000000000001060.

Katz, Patricia, Sofia Pedro, and Kaleb Michaud, "Performance of the Patient-Reported Outcomes Measurement Information System 29-Item Profile in Rheumatoid Arthritis, Osteoarthritis, Fibromyalgia, and Systemic Lupus Erythematosus," *Arthritis Care and Research*, Vol. 69, No. 9, September 2017, pp. 1312–1321. doi:10.1002/acr.23183.

Kazis, L. E., D. R. Miller, J. A. Clark, K. M. Skinner, A. Lee, X. S. Ren, A. Spiro 3rd, W. H. Rogers, and J. E. Ware Jr., "Improving the Response Choices on the Veterans SF-36 Health Survey Role Functioning Scales: Results from the Veterans Health Study," *Journal of Ambulatory Care Management*, Vol. 27, No. 3, July–September 2004, pp. 263–280.

Kazis, L. E., D. R. Miller, J. Clark, K. Skinner, A. Lee, W. Rogers, A. Spiro 3rd, S. Payne, G. Fincke, A. Selim, and M. Linzer, "Health-Related Quality of Life in Patients Served by the Department of Veterans Affairs: Results from the Veterans Health Study," *Archives of Internal Medicine*, Vol. 158, No. 6, March 23, 1998, pp. 626–632.

Kazis, Lewis E., William H. Rogers, James Rothendler, Shirley Qian, Alfredo Selim, Maria Orlando Edelen, Brian Dale Stucky, Adam J. Rose, and Emily Butcher, *Outcome Performance Measure Development for Persons with Multiple Chronic Conditions*, Santa Monica, Calif.: RAND Corporation, RR-1844-NIH, 2017. doi:10.7249/RR1844. As of October 21, 2017: https://www.rand.org/pubs/research_reports/RR1844.html

Krieger, Nancy, Jarvis T. Chen, Pamela D. Waterman, David H. Rehkopf, and S. V. Subramanian, "Painting a Truer Picture of US Socioeconomic and Racial/Ethnic Health Inequalities: The Public Health Disparities Geocoding Project," *American Journal of Public Health*, Vol. 95, No. 2, February 1, 2005, pp. 312–323. doi:10.2105/ajph.2003.032482.

Krosnick, Jon A., "Survey Research," *Annual Review of Psychology*, Vol. 50, No. 1, 1999, pp. 537–567. doi:10.1146/annurev.psych.50.1.537.

Lai, Jin Shei, Jennifer L. Beaumont, Sally E. Jensen, Karen Kaiser, David L. Van Brunt, Amy H. Kao, and Shih Yin Chen, "An Evaluation of Health-Related Quality of Life in Patients with Systemic Lupus Erythematosus Using PROMIS and Neuro-QoL," *Clinical Rheumatology*, Vol. 36, No. 3, March 1, 2017, pp. 555–562. doi:10.1007/s10067-016-3476-6.

National Committee for Quality Assurance, *HEDIS 2006 Specifications for the Medicare Health Outcomes Survey*, Vol. 6, 2006.

National Quality Forum, *Patient Reported Outcomes (PROs) in Performance Measurement*, Washington, D.C., January 10, 2013. As of October 21, 2017: https://www.qualityforum.org/Publications/2012/12/Patient-Reported_Outcomes_in_Performance_Measurement.aspx

NQF—*See* National Quality Forum.

Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, "U.S. Federal Poverty Guidelines Used to Determine Financial Eligibility for Certain Federal Programs," undated. As of October 31, 2017: https://aspe.hhs.gov/poverty-guidelines

Pearman, Timothy P., Jennifer L. Beaumont, David Cella, Maureen P. Neary, and James Yao, "Health-Related Quality of Life in Patients with Neuroendocrine Tumors: An Investigation of Treatment Type, Disease Status, and Symptom Burden," *Supportive Care in Cancer*, Vol. 24, No. 9, 2016, pp. 3695–3703. doi:10.1007/s00520-016-3189-z.

Pickard, A. Simon, and Sara J. Knight, "Proxy Evaluation of Health-Related Quality of Life: A Conceptual Framework for Understanding Multiple Proxy Perspectives," *Medical Care*, Vol. 43, No. 5, May 2005, pp. 493–499.

Quan, Hude, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L. Duncan Saunders, Cynthia A. Beck, Thomas E. Feasby, and William A. Ghali, "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data," *Medical Care*, Vol. 43, No. 11, November 2005, pp. 1130–1139.

Rogelberg, Steven G., and Jeffrey M. Stanton, "Introduction: Understanding and Dealing with Organizational Survey Nonresponse," *Organizational Research Methods*, Vol. 10, No. 2, 2007, pp. 195–209. doi:10.1177/1094428106294693.

Rogers, William H., Lewis E. Kazis, Donald R. Miller, Katherine M. Skinner, Jack A. Clark, Avron Spiro 3rd, and R. Graeme Fincke, "Comparing the Health Status of VA and Non-VA Ambulatory Patients: The Veterans' Health and Medical Outcomes Studies," *Journal of Ambulatory Care Management*, Vol. 27, No. 3, July–September 2004, pp. 249–262.

Rose, Adam J., Elaine M. Hylek, Al Ozonoff, Arlene S. Ash, Joel I. Reisman, and Dan R. Berlowitz, "Risk-Adjusted Percent Time in Therapeutic Range as a Quality Indicator for Outpatient Oral Anticoagulation: Results of the Veterans Affairs Study to Improve Anticoagulation (VARIA)," *Circulation: Cardiovascular Quality and Outcomes*, Vol. 4, No. 1, 2011, pp. 22–29. doi:10.1161/CIRCOUTCOMES.110.957738.

Schnall, Rebecca, Jianfang Liu, Hwayoung Cho, Sabina Hirshfield, Karolynn Siegel, and Susan Olender, "A Health-Related Quality-of-Life Measure for Use in Patients with HIV: A Validation Study," *AIDS Patient Care and STDs*, Vol. 31, No. 2, February 2017, pp. 43–48. doi:10.1089/apc.2016.0252.

Selim, Alfredo J., Dan R. Berlowitz, Graeme Fincke, Zhongxiao Cong, William Rogers, Samuel C. Haffer, Xinhua S. Ren, Austin Lee, Shirley X. Qian, Donald R. Miller, Avron Spiro III, Bernardo J. Selim, and Lewis E. Kazis, "The Health Status of Elderly Veteran Enrollees in the Veterans Health Administration," *Journal of the American Geriatrics Society*, Vol. 52, No. 8, August 2004, pp. 1271–1276. doi:10.1111/j.1532-5415.2004.52355.x.

Sequist, Thomas D., Eric C. Schneider, Angela Li, William H. Rogers, and Dana Gelb Safran, "Reliability of Medical Group and Physician Performance Measurement in the Primary Care Setting," *Medical Care*, Vol. 49, No. 2, February 2011, pp. 126–131. doi:10.1097/MLR.0b013e3181d5690f.

Spearman, C., "Correlation Calculated from Faulty Data," *British Journal of Psychology*, Vol. 3, No. 3, October 1910, pp. 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x.

Swaminathan, Hariharan, and H. Jane Rogers, "Detecting Differential Item Functioning Using Logistic Regression Procedures," *Journal of Educational Measurement*, Vol. 27, No. 4, Winter 1990, pp. 361–370.

Tarlov, Alvin R., John E. Ware Jr., Sheldon Greenfield, Eugene C. Nelson, Edward Perrin, and Michael Zubkoff, "The Medical Outcomes Study: An Application of Methods for Monitoring the Results of Medical Care," *JAMA*, Vol. 262, No. 7, August 18, 1989, pp. 925–930. doi:10.1001/jama.1989.03430070073033.

Turner, Aaron P., Daniel R. Kivlahan, and Jodie K. Haselkorn, "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*, Vol. 90, No. 3, March 2009, pp. 420–428. doi:10.1016/j.apmr.2008.09.558.

Ware, J. E., Jr., and C. D. Sherbourne, "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection," *Medical Care*, Vol. 30, No. 6, June 1992, pp. 473–483.

Zack, Matthew M., "Health-Related Quality of Life: United States, 2006 and 2010," *Morbidity and Mortality Weekly*, Vol. 62, No. 3, Supp., November 22, 2013, pp. 105–111. As of October 21, 2017: https://www.cdc.gov/mmwr/preview/mmwrhtml/su6203a18.htm