Quality of Life Research (2018) 27:2195–2206
https://doi.org/10.1007/s11136-018-1850-3

A new algorithm to build bridges between two patient-reported health outcome instruments: the MOS SF-36® and the VR-12 Health Survey

Alfredo Selim[1] · William Rogers[1] · Shirley Qian[1] · James A. Rothendler[1] · Erin E. Kent[2] · Lewis E. Kazis[3]

Accepted: 4 April 2018 / Published online: 19 April 2018

Abstract

Purpose To develop bridging algorithms to score the Veterans Rand-12 (VR-12) scales for comparability to those of the SF-36® for facilitating multi-cohort studies using data from the National Cancer Institute Surveillance, Epidemiology, and End Results Program (SEER) linked to Medicare Health Outcomes Survey (MHOS), and to provide a model for minimizing non-statistical error in pooled analyses stemming from changes to survey instruments over time.

Methods Observational study of MHOS cohorts 1–12 (1998–2011). We modeled 2-year follow-up SF-36 scale scores from cohorts 1–6 based on baseline SF-36 scores, age, and gender, yielding 100 clusters using Classification and Regression Trees. Within each cluster, we averaged follow-up SF-36 scores. Using the same cluster specifications, expected follow-up SF-36 scores, based on cohorts 1–6, were computed for cohorts 7–8 (where the VR-12 was the follow-up survey). We created a new criterion validity measure, termed "extensibility," calculated from the square root of the mean square difference between expected SF-36 scale averages and observed VR-12 item score from cohorts 7–8, weighted by cluster size. VR-12 items were rescored to minimize this quantity.

Results Extensibility of rescored VR-12 items and scales was considerably improved from the "simple" scoring method for comparability to the SF-36 scales.

Conclusions The algorithms are appropriate across a wide range of potential subsamples within the MHOS and provide robust application for future studies that span the SF-36 and VR-12 eras. It is possible that these surveys in a different setting outside the MHOS, especially in younger age groups, could produce somewhat different results.

Keywords VR-12 · SF-36 · Extensibility

Background

For over two decades, patient-reported outcome surveys have been developed and licensed worldwide for use in population-based quality of life outcomes research [1]. In the United States, the Centers for Medicare & Medicaid Services (CMS) have fielded the Medicare Health Outcomes

SF-36® and SF-12® are registered trademarks of the Medical Outcomes Trust. VR-12© and VR-36© are copyright by the Trustees of Boston University.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-1850-3) contains supplementary material, which is available to authorized users.

✉ Lewis E. Kazis
lek@bu.edu

[1] Center for the Assessment of Pharmaceutical Practices (CAPP), Department of Health Law, Policy and Management, Boston University School of Public Health, Boston, MA, USA

[2] Outcomes Research Branch/Healthcare Delivery Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD, USA

[3] Health Outcomes Unit, Department of Health Law, Policy and Management, Center for the Assessment of Pharmaceutical Practices (CAPP), Boston University School of Public Health, 715 Albany Street, T5W, Boston, MA 02118, USA

Survey (MHOS) to track the functional health and well-being (often referred to as health status or health-related quality of life [HRQoL]) of Medicare beneficiaries enrolled in health plans sponsored by Medicare Advantage organizations (MAOs) [2]. The MHOS has been administered annually since 1998 to a random sample of Medicare beneficiaries with a follow-up sent to those who continue in the plan 2 years later. There are three eras in the types of instruments used to assess HRQoL. For cohorts 1–6 (baseline survey 1998–2003), the Short Form Health Survey-36® version 1.0 (SF-36®) [3] was used for both baseline and follow-up surveys; for cohorts 7 and 8 (baseline surveys in 2004–2005), the SF-36 was used at baseline and the Veterans RAND 12 Item Health Survey (VR-12) [4] at follow-up; and for cohorts 9 until the present (baseline survey 2006-present), the VR-12 was used at both baseline and follow-up [5]. CMS currently uses the MHOS to assess differences in patient-reported outcomes among MAOs, and such differences comprise one component of the Medicare Star Rating system [6].

The MHOS design poses challenges to researchers and policy makers interested in the use of these outcome measures over cohorts that cross the different eras. Such analyses could involve combining observations from different cohorts to improve statistical power or assessing trends over time. This is especially true in oncology. For the purpose of improving outcome surveillance, the National Cancer Institute launched a new data linkage in 2010 that brings together data from the MHOS with patient-level cancer registry data from the Surveillance, Epidemiology, and End Results (SEER) Program [7–10].

The SF-36 includes eight scales: physical functioning (PF), vitality (VT), bodily pain (BP), general health perceptions (GH), role limitations due to physical functioning (RP), role limitations due to emotional functioning (RE), social functioning (SF), and mental health (MH) [11]. The VR-12, a derivative of the Veterans Rand-36/SF-36 questionnaires, contains 12 representative items from these eight scales, with modifications to four items representing "role functioning" (2 of which are related to physical health and 2 related to emotional problems, with yes/no response choices for the SF-36 and 5 Likert response choices for the VR-12) [12]. Scales of the SF-36 comprise 2–10 items compared to 1 or 2 items for the VR-12 scales. In addition, both instruments have a physical and a mental health composite summary score (PCS and MCS) that use all 8 scales in their respective computation [13].

The usual methods of calculating VR-12 scale scores (i.e., by linearly transforming the item response choices to a scale of 0–100) do not precisely align with the corresponding scale scores from the SF-36 due to several factors. One factor is that while there are 1 or 2 questions in common between corresponding SF-36 and VR-12 scales, additional questions that comprise SF-36 but not VR-12 scales may be perceived and responded to in a somewhat different manner than the questions in common [14]. Matching of the VR-12 items and scales to those of the SF-36 is also affected by the context (i.e., instrument length, item positioning) in which the questions appear [15]. In a 12-item format, response patterns can shift in comparison with a response to the same question in an embedded (36-question) format. For example, our preliminary results found that the respondents to the VR-12 report having somewhat more energy on the "energy" item compared with similar respondents to the SF-36, even though the energy item is unchanged. Another factor that creates problems in matching the VR-12 scale scores to those of the SF-36 is the previously noted difference in item response choices for the role items [5, 12].

Additionally, the MHOS has other facets that complicate the analyses designed to match VR-12 items and scales to SF-36 scales. These issues are related to the changes in the cohorts over time between the SF-36 and the VR-12 era surveys. These "sample evolution problems" include issues such as changes in the types of MAOs in the survey; year-to-year differences in administration (e.g., proxy responses, foreign language forms, and the use of telephone administration); and changes in population average disease severity.

Our main goal for this paper was to rescore and thus bridge the VR-12 items and scales to match SF-36 scales. The rescored scales need to be applicable across a broad array of potential MHOS subsamples that could differ by their expected mean values of health status and scale scores. We also developed a rescoring algorithm for SF-36 items so that each item could best represent the full scale. This study was motivated by the needs of researchers using SEER-MHOS data to assess scale scores across cohorts that span the SF-36 and VR-12 eras. It was also designed to have general applicability for other types of studies outside of the oncology area that involve MHOS SF-36 and VR-12 data.

Methods

Data source

We included respondents aged 65 or older in MHOS cohorts 1–12 (1998–2011) who returned/mailed, self-reported surveys in English.

Overview of analytic approach

Our overarching goal was to create algorithms for matching the VR-12 items and scales to SF-36 scales that could be used, without further customization, across a broad array of possible MHOS subsamples. Such subsamples could differ, for example, by respondent characteristics, such as age, gender, number, and type of medical conditions, which in turn, would be reflected by differences in the expected mean values of health status and scale scores. An alternative to such a unified algorithm would be to develop an approach where the rescoring of VR-12 items would be dependent on the characteristics of each subsample. Additionally, it would make it difficult to compare HRQoL scores among different studies.

One of the challenges in creating a unified rescoring algorithm is that a particular SF-36 item, scored in a "simple" manner, does not precisely match with the score of the corresponding SF-36 scale. In this paper, we use the term "simple" scoring to refer to the mostly linearly scaling method described by Ware for the SF-36 [16] or to the purely linear 0–100 scaling for the VR-12. Additionally, the difference in score (both in magnitude and sign in this case) between the item and scale differs by age group [14]. Similar issues would also arise for other respondent characteristics, so that a unified matching algorithm would have to optimize the match between items and scales taking into account a range of respondent characteristics. The issue relative to embedded SF-36 items and their associated scales is a similar problem in matching the VR-12 items to the SF-36 scales.

Without knowing how the items will actually be used in the context of SEER-MHOS linked data, the property we most desire for item scoring is to produce an estimate with the best possible external validity. This implies that when the rescored item is averaged across any particular MHOS subsample, possibly a subsample that substantially differs in characteristics and health status from the overall MHOS mean, the rescored item should produce the same mean score as its associated full SF-36 scale to the best possible extent.

The overall task of developing the scoring algorithm was divided into two main sections: (1) Matching scores of individual "embedded" SF-36 items to their corresponding scales, and (2) Matching VR-12 items and scales to the corresponding SF-36 scales (see Table 1). Our reasons for pursuing the first of these tasks were twofold. First, we wanted to determine the feasibility of our approach (that would also be used in a modified form for the second task) using a sample in which we did not have to be concerned about "sample evolution," as previously described. Second, SF-36 surveys not infrequently have one or more items with missing responses. A typical approach to calculating scales when items are missing from that scale is to average the remaining items scored mostly on a 0–100 scale [17], as long as at least half of the items have been scored. By rescoring individual items of each scale so that they best represent the scale calculated from all items, the results of the first task could permit a more accurate calculation of scale scores when items were missing and would even allow for calculation of the scale if only one item from a scale was present. This would allow us to use survey responses from a high percentage of respondents for matching the VR-12 scales to those of the SF-36.

Task 1: matching embedded SF-36 items to their associated SF-36 scales

For matching embedded SF-36 items to their associated SF-36 scales, we used the baseline surveys from cohorts 1–8 to define MHOS subsamples and the follow-up surveys from cohorts 1–6 in cases where responses for both the item and the corresponding SF-36 scale were complete for all subsample members.

The MHOS contains a variety of potential subsamples that vary by health status. If the subsamples are defined on the basis of the scale and the item, the results will be distorted, but we can find "independently defined" subsamples with other correlated variables. For the MHOS, these include age, gender, and responses to MHOS questions dealing with chronic conditions, physical symptoms (such as arthritis pain, shortness of breath, chest pain), and mental symptoms. Using these variables to describe a distance, we used a k means clustering algorithm to solve for $M = 200$ "clusters" (subsamples) plus one cluster that had missing data on one or more of these variables [18]. The cluster of responses was used to deal with errors in the independent variables.

Within each of the 201 cluster-defined subsamples, the difference between the item score and the scale score was used as a measure of how well the item and scale agree. If one averages the squared differences, weighted by subsample size and takes the square root of that quantity, the result is a measure of criterion validity, which we have termed cluster-weighted "extensibility." An extensibility value can be used in a way similar to a standard error of the mean [19]. The reason that this statistic is central to our approach (in both this and the second task of matching VR-12 items to SF-36 scales) is that we have no way to predict future use of the MHOS data and what subsamples might be chosen. We need a robust scoring methodology that can be applied to many circumstances with credible accuracy and a way to characterize the limits of such accuracy. The premise of this methodology is that we do not seek to estimate a scale value of a single respondent to the MHOS from one item but rather, a method that will estimate, from any subsample that we might encounter, the mean scale values from the mean values of a single item.

To evaluate and compare extensibility among rescoring methods, we conducted the following approach. Within each cluster and across all members of the cluster, we computed the average value of the scale based on the simple scoring method applied to all items from the scale, and the average value of the "estimated" scale based on a chosen scoring method applied to a single survey item from that scale. For a

Table 1 Crosswalk of individual items in common between and unique to the MOS SF-36 and the VR-12

Scales	SF36 items/VR12 items	Variable name (SF36)	# of response Choices	Variable name (VR12)	# of response choices
General health	General health	GH1	5	GH1	5
	Sick easier than other people	GH2	5		
	As healthy as anybody I know	GH3	5		
	Expect health to get worse	GH4	5		
	Excellent health	GH5	5		
Physical functioning	Vigorous activities	PF01	3		
	Moderate activities	PF02	3	PF02	3
	Lifting or carrying groceries	PF03	3		
	Climbing several flights of stairs	PF04	3	PF04	3
	Climbing one flight of stairs	PF05	3		
	Bending, kneeling, or stooping	PF06	3		
	Walking > 1 mile	PF07	3		
	Walking several blocks	PF08	3		
	Walking 1 block	PF09	3		
	Bathing or dressing	PF10	3		
Role physical	Physical health limiting time spent on activities	RP1	2		
	Physical health limiting amount accomplished	RP2	2	RP2	5
	Physical health limiting the kind of activities	RP3	2	RP3	5
	Physical health causing difficulty performing activities	RP4	2		
Role emotional	Emotional problems limiting time spent on activities	RE1	2		
	Emotional problems limiting amount accomplished	RE2	2	RE2	5
	Emotional problems limiting carefulness	RE3	2	RE3	5
Social functioning	Extent PH or EP interfered with social activities	SF1	5		
	Amount of time PH or EP interfered with social activities	SF2	5	SF2	5
Bodily pain	Bodily pain	BP1	6		
	Pain interfering with work	BP2	5	BP2	5
Vitality	Full of pep	VT1	6		
	A lot of energy	VT2	6	VT2	6
	Worn out	VT3	6		
	Tired	VT4	6		
Mental health	Nervous	MH1	6		
	Down in the dumps	MH2	6		
	Calm and peaceful	MH3	6	MH3	6
	Downhearted and blue	MH4	6	MH4	6
	Happy	MH5	6		

given cluster ‘*m*,’ we obtain the average of scale ($\overline{X}_m$) across cluster members scored in the simple manner [20].

Summary of notations

$\overline{X}_m$—average of SF-36 version 1 scale scored in simple manner within cluster *m*

$\overline{x}_m$—average of new rescored value of single item within cluster *m*

x_n—new rescored value of item for an individual *n* within a cluster

We used the index to refer to an item within the associated scale with the goal of judging how well a rescored value of such an item best matches the value of the scale. Applying the new scoring method to a single item, we then obtain an average estimate ($\overline{x}_m$) across the same cluster members. The squared difference between the two $\left(\overline{X}_m - \overline{x}_m\right)^2$ describes how far apart they are. The mean squared difference across all clusters describes how well the estimated scale values, based on the new scoring method applied to a single scale item, succeeded at matching the values based on the simple method derived from all scale items, and the square root of that quantity is what we have termed “extensibility.” A low value is desirable; that is, if the scale scores based on the new scoring method and the simple methods always agree, then that new method is extensible. The root-mean-squared difference describes the sum of statistical and non-statistical errors. Non-statistical errors come from a biased estimate, whereas statistical errors are due to random sampling variation. For rescoring methods that produce individual estimates, the statistical errors are described by the standard error, as calculated using standard formulas for the variance of a mean.

Our goal was to rescore the individual items so that they optimally represent the scale within the cluster. Cluster scoring methods assume that each response of each item within cluster m for individual n will take on some value x_n. For example, in the simple scoring scheme of an item with five response choices, x_n can have values of 0, 25, 50, 75, and 100. In the process of matching items to scales, the values of these response choices will be rescored. Within each cluster m, our choice of how to rescore each response for each item will determine the $\bar{x}_m$, which is the mean of the scored item within that cluster. If cluster m has N_m respondents and x_n is the rescored value of an item for an individual n within the cluster:

$$\bar{x}_m = \frac{1}{N_m} \sum_{n=1}^{N_m} x_n.$$

This will be compared to $\overline{X}_m$ that is the mean score, based on the simple scoring method, of the SF-36 scale that encompasses the item used to calculate $\bar{x}_m$, which is based on the new scoring method. The goal of this scoring method is to find optimal values for each response of each item for each cluster m so that the subsample size-weighted mean of $\left(\overline{X}_m - \bar{x}_m\right)^2$ for each item across all clusters is minimized. If N is the total number of respondents across all of the clusters (201 clusters in this case), N_m is the number of respondents with cluster m, and we weight $\left(\overline{X}_m - \bar{x}_m\right)^2$ relative to the mean cluster size N/201, then for each item, we seek to minimize the quantity.

$$S = \frac{1}{201} \sum_{m=1}^{201} \frac{N_m}{N/201} \left(\overline{X}_m - \bar{x}_m\right)^2 = \sum_{m=1}^{201} \frac{N_m}{N} \left(\overline{X}_m - \bar{x}_m\right)^2.$$

The square root of S is what we term cluster-weighted extensibility. In order to minimize these values, we used a non-linear least squares algorithm ("nl" in STATA [21]), which minimizes the sum of squared differences of a non-linear expression using a pseudo-regression technique that is fitted based on a modified Newton–Raphson algorithm [22]. The item values were parameterized in a way that forced the solution to be monotonic; for example, for the items in the physical function scale, the score for "limited a lot" has to be less than the score for "limited a little" in order to avoid unreasonable solutions corresponding to local minima of the objective function (the extensibility statistic). We also explored multiple starting values to assure ourselves we had the best local minimum extensibility. When extensibility values are reported the expected squared error due to random sampling is subtracted from "*S*." The cluster scoring is available upon request from the senior author.

Task 2: matching VR-12 items to the corresponding SF-36 scales

For the second task, matching of VR-12 items to the SF-36 scales required a somewhat different approach from that used to match extracted SF-36 items to their corresponding scales.

To define clusters for this particular analysis, our intent was to use variables that would likely have a similar meaning in cohorts 7 and 8 relative to cohorts 1–6, and on that basis, we chose age, gender, and baseline SF-36 scores.

We used the rescored SF-36 items from the first task to calculate SF-36 scales for the baseline and follow-up surveys for cohorts 1–6 and the baseline survey for cohorts 7–8.

SF-36 scales and corresponding VR-12 items were considered one at a time. Using Classification and Regression Trees (CART), we modeled follow-up scores in cohorts 1–6 for each scale based on the baseline SF-36, age, and gender. From this we obtained 100 clusters where the predicted SF-36 scale score was similar. We averaged the follow-up SF-36 score for each scale within each cluster using data that spanned cohorts 1–6.

We applied the cluster definitions obtained in cohorts 1–6 to individuals in cohorts 7 and 8 who had both baseline and follow-up surveys. In these cohorts, SF-36 was used at baseline and the VR-12 at follow-up. The mean expected SF-36 follow-up score for each scale within each cluster for cohorts 7–8 are those that were calculated based on cohorts 1–6 follow-up SF-36 scores.

From the above analyses, we have 100 clusters of respondents to the follow-up surveys in cohorts 7 and 8 in which we have an expected score for each SF-36 scale. We also know the observed VR-12 items responses to the follow-up survey in cohorts 7–8 for each individual within each cluster. The overall strategy was to rescore the response choices for each VR-12 item such that for each of the 100 clusters, the VR-12 item score across respondents within a given cluster best matches the expected corresponding SF-36 scale score.

In order to find the best match between VR-12 items and SF-36 scales, we again sought to minimize the cluster size-weighted mean square difference between the item and scale, as described in Task 1. To accomplish this, we again used the previously described non-linear least squares routine "nl" in Stata.

Evaluation of the new scoring algorithm

Our methodology was based on three kinds of calibrations/predictions or "steps." Step 1 was part of Task 1, and steps 2 and 3 are part of Task 2.

Step 1: SF-36 items were scored/weighted to match SF-36 scales. This matching was done cross-sectionally using cohort 1–8 baseline data and cohort 1–6 follow-up data

To evaluate this calculation, we compared the extensibility obtained using the equal interval method of scoring with that obtained after rescoring the SF-36 items to minimize extensibility across clusters.

Step 2: using CART, baseline SF-36, age, and gender in cohorts 1–6 was used to predict follow-up SF-36 scale scores. The prediction algorithms were then applied to develop predicted follow-up SF-36 scale scores in cohort 7–8

We evaluated the accuracy of our algorithm in predicting follow-up scales scores in a different set of cohorts from which the algorithm was developed. In our main analyses, we used data from cohorts 1–6 to predict follow-up scores in cohorts 7–8. To validate this methodology, we used a similar approach in which baseline and follow-up data in cohorts 1–3 were used to predict follow-up scale scores for cohorts 4–6. Specifically, CART analyses in cohorts 1–3 were used to model follow-up scale scores based on baseline SF-36 scores, age, and gender, from which 100 clusters were created. The same cluster specifications from cohorts 1–3 were then applied to the baseline SF-36 data from cohorts 4–6 in order to create predicted SF-36 follow-up scale scores for cohorts 4–6 for each cluster and each scale. Since the actual SF-36 follow-up scale scores for cohorts 4–6 are known, they can be compared to predicted scores.

Step 3: follow-up VR-12 items are scored/weighted to match the predicted follow-up SF-36 scale scores (from step #2) in cohort 7–8

We examined how the new scoring algorithm compared to the "simple" scoring method in matching VR-12 items to the corresponding SF-36 scales using extensibility as the criterion validity. Most of the items in the VR-12 are derived from the SF-36 and some could be described as VR-12 items. In scoring SF-36 scales in the simple manner, only those cases in which there were responses to all items within a scale were used. Four of the VR-12 scales comprise single items, and we also matched the 2-item VR-12 scales to those of the SF-36 by averaging the two items comprising those scales. Since the rescoring algorithm might be applied to data subsets quite different from the complete dataset for MHOS respondents, we used a new set of clusters to display comparisons between simply scaled and rescaled items, and we did not weigh the extensibility calculation by cluster size.

We also evaluated Step 3 to assess whether the extensibility results obtained are sensitive to the sample used to obtain predicted follow-up scores for cohorts 7 and 8 and the sample to which the rescoring algorithm is applied. For this analysis, we divided the sample comprising cohorts 1–8 into two parts, a larger part consisting of two-thirds of the data ("estimation sample"), and a smaller part with the remaining one-third ("test sample"). Using the methods described in Steps 2 and 3 (Task 2) of our main analyses, but applied to only the estimation sample, clusters were created in cohorts 1–6 and predicted follow-up SF-36 scales scores were calculated for cohorts 7–8 based on follow-up results in cohorts 1–6. In the estimation sample, VR-12 items for follow-up cohorts 7–8 were rescored to best match the predicted SF-36 scale scores. In the test sample, the same methodology was used to create new clusters and new predicted SF-36 scale scores for cohorts 7–8. We then assessed how well the VR-12 rescoring algorithm developed in the estimation sample worked in the test sample in matching the VR-12 item scores to the expected SF-36 scales scores. In particular, we examined the correspondence of extensibility between estimation and test samples with regard to the matching of VR-12 items to SF-36 scales.

Results

Profile of the MHOS cohorts

Table 2 is the profile of the analytic sample used in the analyses. There were 570,459 individuals in the SF-36 era (cohorts 1–6), 119,543 individuals in the transition era (cohorts 7–8), and 452,155 individuals in the VR-12 era (cohorts 9–12).

Rescoring of SF-36 items and extensibility of rescored SF-36 embedded items

Table 3 gives the SF-36 item scorings for the questions related to "moderate activities" and "climbing several flights of stairs" in the PF scale based on our analytic sample of elderly MHOS, English-language, mailed self-report surveys with all 10 PF items answered using: (1) the simple prorated scale values (0, 50, and 100), and (2) the cluster scoring method. In addition, Table A-1 of Supplemental Material 1 contains a table of the rescored items of the SF-36 that are

Table 2 Descriptive profile of Medicare Health Outcomes Survey (MHOS) cohorts

	SF-36 era (SF-36 at baseline and at follow-up)	Transition era (SF-36 at baseline and VR-12 at follow-up)	VR-12 era (VR-12 at baseline and at follow-up)
MHOS cohorts	1–6	7–8	9–12
Sample size	570,459	119,543	452,155
Age	74.3 ±6.0	75.4 ± 6.1	75.4 ± 6.3
Gender			
Male	41.7%	40.3%	41.0%
Race/ethnicity			
White	91.5%	90.4%	88.7%
Black	5.2%	6.2%	7.2%
Hispanic	0.9%	1.2%	1.5%
Other	2.4%	2.2%	2.6%
PCS scores	41.2 ± 11.7	40.5 ± 11.8	40.8 ± 11.7
MCS scores	52.9 ± 9.7	52.9 ± 9.7	53.2 ± 10.0

All included subjects were aged 65 and older at time of survey, English-speaking, self-administered survey, and completed the survey by mail

PCS physical component summary, *MCS* mental component summary

Table 3 Score equivalents of "moderate activity" and "climbing several flights of stairs" items (from the SF-36) for two scoring methods

Method	Yes, limited a lot	Yes, limited a little	No, not limited at all
Moderate activity (PF2)			
1. Simple scoring	0.00	50.00	100.00
2. Cluster	5.91	51.74	95.73
Climbing several flights of stairs (PF4)			
1. Simple scoring	0.00	50.00	100.00
2. Cluster	11.26	72.75	95.81

PF2 physical functioning item #2, *PF4* physical functioning item #4

substituted for its full-scale counterpart if scored in the most extensible manner, using the cluster method.

Rescoring the VR-12 items and extensibility of the rescored VR-12 items

Table 4 contains unweighted extensibility of both rescored and simply scored VR-12 items and scales based on 100 clusters of roughly equal sample sizes. (A brief item description associated with each item abbreviation is contained in Table A-2 of Supplemental Material 1 along with rescored item responses for each of the 12 items.) For example, the rescored VR-12 PF2 item (Moderate Activity) is slightly different from the SF-36 PF scale such that the extensibility or the root-mean-square difference is 2.1 points. For comparison, Table 4 also includes the unweighted extensibility calculated with the simple scoring of the VR-12, excluding the "role" items and scales. In all cases where comparisons are meaningful, the extensibility of the rescored items and

Table 4 Extensibility[a] (unweighted) of VR-12 items and scales

Item or scale	Extensibility	
	Simple scoring	Rescored
PF2	4.4	2.1
PF4	9.2	1.8
RP2	[d]	3.9
RP3	[d]	4.1
RE2	[d]	3.7
RE3	[d]	3.8
BP2[b]	8.3	1.5
GH1[b]	10.9	4.3
VT2[b]	3.4	1.5
SF2[b]	2.3	1.0
MH3	8.1	2.1
MH4	3.8	1.5
PF[c]	4.6	1.5
RP[c]	[d]	3.9
RE[c]	[d]	3.6
MH[c]	2.7	1.2

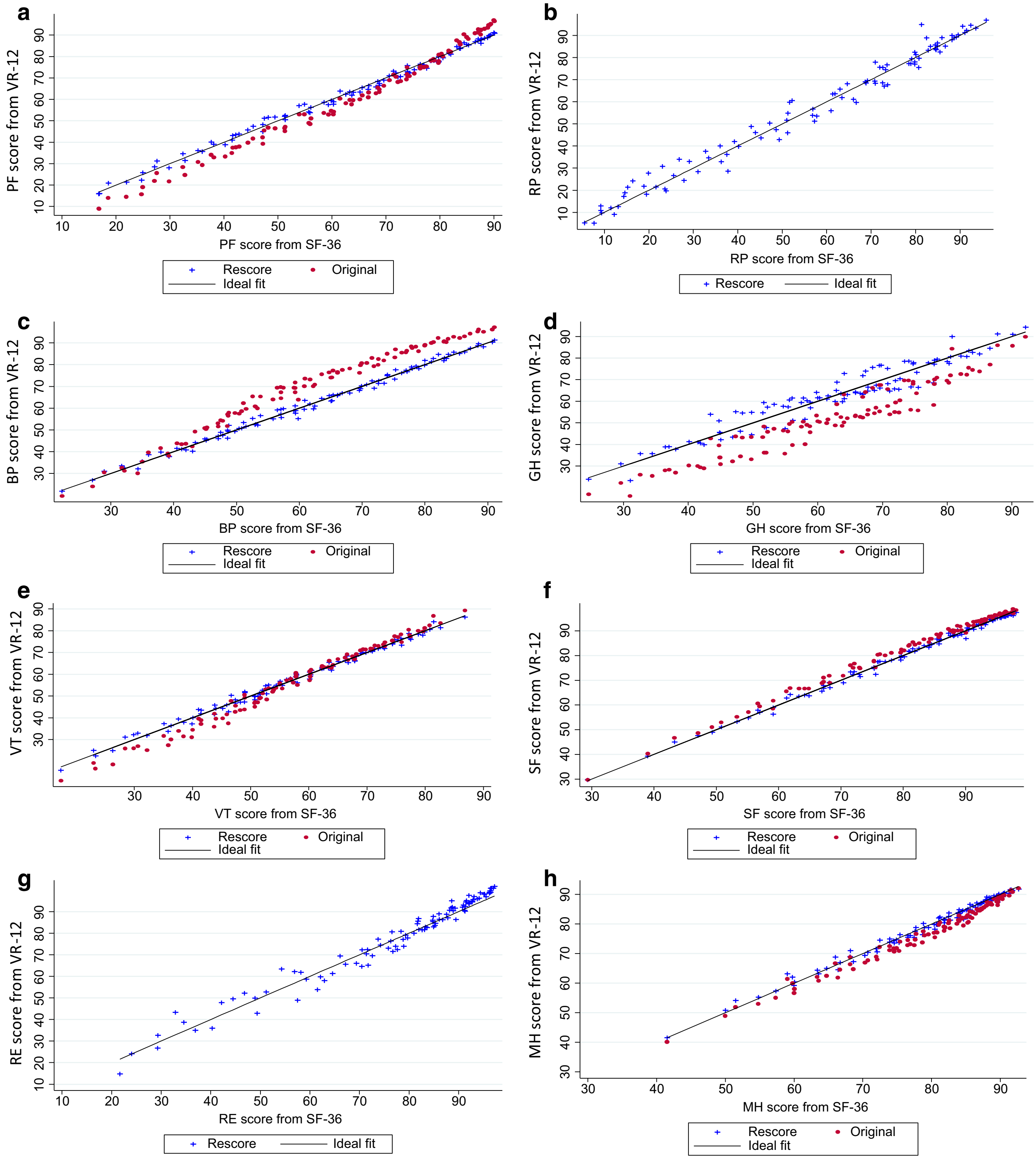
PF2 physical functioning item #2, *PF4* physical functioning item #4, *RP2* role physical item #2, *RP3* role physical item #3, *RE2* role emotional item #2, *RE3* role emotional item #3, *BP2* bodily pain item #2, *GH1* general health item #1, *VT2* vitality item #2, *SF2* social functioning item #2, *MH3* mental health item #3, *MH4* mental health item #4, *PF* physical functioning scale, *RP* role physical scale, *RE* role emotional scale, *MH* mental health scale

[a] Extensibility is a new validity measure calculated from the square root of the mean of the squared difference between the expected SF-36 scale score and the rescored VR-12 item

[b] One-item scale

[c] Two-item scale

[d] Not meaningful due to difference in response choices between the SF-36 and VR-12 surveys



Red dots: Scale scored using the original scoring of VR-12 items (0, 20, 40, 60, 80, 100) for the 100 clusters. Blue crosses: Scale score based on VR-12 item rescoring for the same clusters

Fig. 1 **a**–**h** For the 100 clusters applied to baseline cohorts 7 and 8, correlation between the estimated scores of the follow-up SF-36 scale and the scale scores for the VR-12. For the PF, BP, GH, VT, SF, and MH scales, the VR-12 scores are calculated by two different methods, simple and rescored. For the RP and RE scales, only the rescored VR-12 scales are displayed for the reasons described in the "Results" section text for Table 4. (Color figure online) scales was better (denoted by lower values) than with simple scoring. The "role" items were excluded because it is not meaningful to compare such items as simply scored due to the difference in responses choices between the SF-36 and VR-12 (yes/no vs. 5-point Likert scale). Supplemental Material 2 contains additional detailed information on how extensibility can be used and its applications for the researcher.

Figure 1a–h provides further illustration of the effect of rescoring on the association between SF-36 and VR-12 scales. For each of the 100 clusters derived from CART, the *x*-axis represents the "expected" follow-up SF-36 scale scores for cohorts 7 and 8 (on patients with complete survey data), based on the results from cohorts 1–6. The *y*-axis represents the observed follow-up VR-12 scale scores for cohorts 7 and 8 (on patients with complete and incomplete survey data), scored in both the simply scored and rescored manner. As previously noted, our methodology sought to rescore the VR-12 such that the observed follow-up VR-12 scores in cohorts 7 and 8 optimally matched the expected SF-36 scores across all clusters. If the match was ideal, all cluster points would fall along the line of identity. The data points representing the rescored VR-12 items fall relatively close to the line of identity or "ideal fit." For the PF, BP, GH, VT, SF, and MH scales, data points representing the simple scoring method for the VR-12 are also displayed. Compared to the points representing rescored values, the simply scored data points appear to deviate from the line of ideal fit to a greater degree. This is reflected in the differences in extensibility shown in Table 4.

Evaluation of the accuracy of predicted scale scores

As noted in Methods, we also assessed the accuracy of the algorithm used in predicting follow-up scale scores in cohorts 7–8 by using data in cohorts 1–3 to predict follow-up scale scores in cohorts 4–6 where actual scale scores were known. The results are presented in Figures A1a–A1h in supplemental material #3. The mean predicted SF-36 scores match quite well with the mean actual SF-36 scale scores across the clusters for six of the eight SF-36 scales, with somewhat less tight matching for the two role scales. The differences from the line of identity are consistent with the extensibilities observed from the matching of VR-12 items and scales to SF-36 scales. The less tight results for the role scales, while still following the line $y = x$, suggest that the yes/no responses are more difficult to accurately model, and this helps to explain in part why it is hard to match the responses of the VR-12 to these role scales.

Extensibility of estimation and test samples

In addition to the above results based on the full sample, we also calculated extensibility for the matching between VR-12 items and SF-36 scales based on a split estimation and test sample. Table 5 shows the correspondence of extensibility between estimation and test samples, which are within and around 0.6 of one another. When the extensibility is large, it may mean that the item content shifted a little in the VR-12. This would be true for the role items, which changed from yes–no in the SF-36 to "all of the time" to "none of the time" response choices in a 5-point Likert format in the VR-12. The general health item extensibility probably reflects the inability of the single item to represent the rest of the general health scale well; the SF-36 version has five items and the other four have a different response scale than the global item in the VR-12. Other items had acceptably low extensibility. These findings were generally similar to those from the original full MHOS sample (Table 4).

Items and scales across MHOS cohorts using the simple versus new scoring algorithm

Figure 2 illustrates the comparison of the simple compared to rescored items and scales across the 12 MHOS cohorts using PF as an example (other figures for other scales are available on request). The values in Fig. 2a, b were not adjusted for potential differences among cohorts in characteristics of the respondents and MAOs and other aspects of sample evolution. Figure 2a represents all those in the analytic sample who responded to the baseline surveys, and Fig. 2b represents those in the analytic sample who responded to the 2-year follow-up surveys. The

Table 5 Correspondence of extensibility values between estimation and test samples

VR-12 item[a]	Extensibility of estimation sample	Extensibility of test sample
PF2	2.04	1.73
PF4	1.45	1.06
RP2	3.29	2.73
RP3	3.56	3.25
RE2	2.61	3.03
RE3	2.77	2.85
BP2	0.94	1.42
GH1	3.90	3.58
VT2	1.12	1.35
SF2	0.00[b]	0.00[b]
MH3	1.74	1.84
MH4	0.94	1.12

[a] See abbreviations for Table 4

[b] Extensibility involves a small discount for the random sampling error within each cluster for both the VR-12 and SF-36 scales

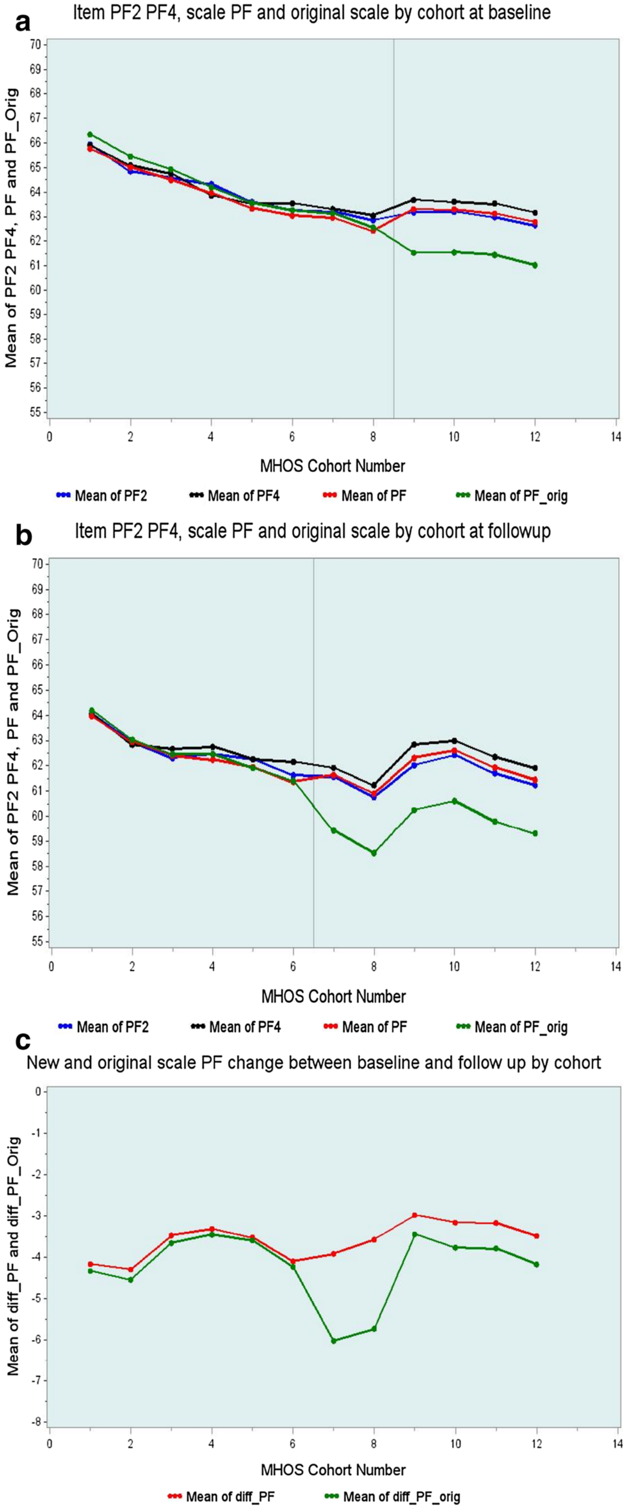


Fig. 2 The green line in **a**–**c** represents the original method of calculating the SF-36 and VR-12 scale scores for PF (physical functioning), while the red line represents the rescored method. In **a, b**, the blue and black lines represent the rescored versions of the PF2 (moderate activities) and PF4 (climbing several flights of stairs) questions, respectively. PF2: rescored "moderate activities" question of physical function scale. PF4: rescored "climbing general flights of stairs" question of physical function scale. Mean of PF: mean physical function scale calculated using rescored versions of each of the ten SF-36 and two VR-12 component questions. **a** The line for "Mean of PF" is based on 10 items in cohort 1–8 but only on two items in subsequent cohorts. **b** The line for "Mean of PF" is based on 10 items in cohort 1–6 but only on two items in subsequent cohorts. Mean of PF_orig: mean of physical function scale calculated using the original prorated response values on a scale from 0 to 100 instead of the rescored versions. Mean of diff_PF: mean of the difference of physical function scale using rescored versions of each of the ten SF-36 and two VR-12 component questions. Mean of diff_PF_orig: mean of the difference of physical function scale calculated using the original prorated response values on a scale from 0 to 100 instead of the rescored versions. (Color figure online) ▶

rescoring was applied to all items in both the SF-36 and VR-12 surveys. The lines representing rescored items (PF2 and PF4) and the rescored PF scale track closely to one another. Although they differ slightly due to differences in the method of scoring, the rescored version of the PF scale (red line) also tracks closely to the simple scoring of the scale (green line) during the SF-36 era. However, during the VR-12 era, the green line (simple scoring algorithm) substantially deviates from the other lines, while the red line (representing the rescored scoring algorithm) appears to track more smoothly between SF-36 and VR-12 era surveys.

Figure 2c represents the difference in scores between the baseline and follow-up surveys. During the era that the SF-36 was the baseline and follow-up survey (cohorts 1–6), the rescored and simple scoring lines track very closely to one another. For cohorts 7 and 8, where the SF-36 was the baseline and VR-12 the follow-up survey, there is substantial deviation between the two lines, reflecting the differences caused by rescoring of the VR-12. For cohorts 9–12, where the VR-12 was used for both baseline and follow-up, the red and green lines are again closer together.

Discussion

Our study used a novel methodology to create an algorithm for rescoring the scales from the VR-12 to numerically match those of the SF-36. Our cluster approach is aimed for group level interpretation. This method is applicable across a broad array of potential samples of individuals that differ in health status due to factors such as medical conditions, symptoms, and demographic characteristics. The extensibility values related to the matching of VR-12 items to SF-36 scales were generally credible, although somewhat less optimal for both physical and emotional role items and scales and the general health item. The scales of the VR-12 and SF-36 contain granular health status information that can be used along with the summary scores, physical PCS, and mental MCS, and the knowledge of information pertaining to scales may be of additional use in understanding how various factors affect the health status of patients [23].

The new bridging algorithm has important implications for missing data. One can use the embedded item rescoring based on best extensibility criteria. They provide reasonable estimates even if only one item of a 10-item scale is available. If more than one item is available, the results can be averaged across those present. If only a few items are missing the Modified Regression Estimation (MRE) accounts for the missing information by estimating the score based on the items that are available [23, 24]. Historically, two other methodologic approaches have been used to address this problem: regression and item response theory [25]. However, they have significant weakness. The regression answer is sensitive to the nature of the estimation sample. Since a particular item, physical measurement, or scale has response error, both the dependent and independent variable of the regression have error in them, so at best there is a difficult errors-in-variables problem and at worst there is an endogeneity problem if both variables come from the same samples and they are not assessed independently. The impact of error in the dependent variable is to bias the regression coefficient toward 0, so these answers suffer from regression to the mean bias. The item response theory answer solves the regression-to-the-mean problem, but there are other problems; in particular, what to do with values that are at the ceiling or the floor? In some cases this is not a severe problem, but in many cases we are trying to replace a whole scale such as the SF-36 vitality scale with a single item such as the Energy item. In these cases, the treatment of the floor and ceiling values is either very arbitrary, or (in some IRT implementations) dependent on the target data. The latter answer is not satisfactory because it creates a definition that varies from one target dataset to another, so datasets cannot be compared.

Several factors may limit the results of this study. First, while we were able to identify a methodology to minimize extensibility of items and scales of the VR-12, the magnitude of extensibility can be a limiting factor in the ability to detect differences in scale scores when MHOS cohorts that span the SF-36 and VR-12 eras are combined. Extensibility was noticeably higher for the general health item as well as the role physical and role emotional scales. For the "role" items, the higher extensibility was likely due to the difference in the number of response choices between the SF-36 and VR-12 surveys. Second, in combining health status scores across cohorts, it is important to take such sample evolution into account. In this report, we did not attempt to address the approaches that would be required to adjust for the variation in attributes, survey-related factors, changes in criteria over time defining disease diagnoses, and changes to treatment among cohorts and their members. Future studies can evaluate these factors. Third, our algorithm was developed from the MHOS in those 65 years of age or greater, it is possible that VR-12 surveys in a different setting with younger individuals could produce somewhat different results. Fourth, we did not revise the scoring of PCS and MCS because we did not want to compete with a previous solution for PCS and MCS commissioned by the National Committee for Quality Assurance (NCQA) on behalf of CMS.

In summary, we created a new bridging algorithm for rescoring VR-12 items and scales to match corresponding SF-36 scale scores that improved upon the "simple" method of scoring items and scales. In particular, through the use of "cluster" subsamples of the MHOS data and the application of extensibility, the algorithm that was developed should be applicable across a wide range of MHOS samples for future studies.

Data availability Several types of Medicare HOS data files are available for research purposes. Medicare HOS data files are available as public use files (PUFs), limited datasets (LDSs), and research identifiable files (RIFs). http://www.hosonline.org/en/datadissemination/research-data-files/

Author contributions AS participated in the conception and design. WR participated in the analysis and interpretation of data. SQ performed the statistical analysis. JR participated in the conception and design and interpretation of data. EK participated in drafting and revising it critically for important intellectual content. LK participated in the study design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Funding The study was supported by the Outcomes Research Branch/Healthcare Delivery Research Program Division of Cancer Control and Population Sciences, National Cancer Institute Contract # HHSN261201400530P. The views expressed represent those of the authors and not necessarily those of the National Cancer Institute, National Institutes of Health, and Boston University.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interest.

Ethical approval The Boston University Institutional Review Board (IRB) provided ethical review of this study to protect the rights and welfare of human subjects of research and to assure that human research is conducted according to applicable federal, state, and local laws and regulations, and the relevant policies of the Human Research Protection Program, Boston University.

References

1. Scott, R. E., & Saeed, A. (2008). Global eHealth—measuring outcomes: Why, what, and how a report commissioned by the World Health Organization's global observatory for eHealth. Retrieved September 10, 2017 from http://www.ehealth-conne

ction.org/files/conf-materials/Global%20eHealth%20-%20Measuring%20Outcomes_0.pdf.

2. Jones, N. III, Jones, S. L., Miller, N. A. (2004). The medicare health outcomes survey program: Overview, context, and near-term prospects. *Health and Qual Life Outcomes, 2*, 33.
3. Stewart, A. L., & Ware, J. (1992). *Measuring functioning and well-being: The medical outcomes study approach*. Durham: Duke University Press.
4. Usman Iqbal, S., Rogers, W., Selim, A., Qian, S. X., Lee, A., Xinhua, X., Rothendler, J., Miller, D., & Kazis, L. (2007). The Veterans Rand 12 Item Health Survey (Vr-12): What it is and how it is used. Technical report. Retrieved October 28, 2015 from http://www.hosonline.org/globalassets/hos-online/publications/veterans_rand_12_item_health_survey_vr-12_2007.pdf.
5. Kazis, L. E., Selim, A. J., Rogers, W., Qian, S. X., & Brazier, J. (2012). Monitoring outcomes for the medicare advantage program: Methods and application of the VR-12 for evaluation of plans. *The Journal of Ambulatory Care Management, 35*, 263–276.
6. Sprague, L. (2015). The star rating system and medicare advantage plans. *Issue Brief National Health Policy Forum, 854*, 1–10.
7. Warren, J. L., Klabunde, C. N., Schrag, D., Bach, P. B., & Riley, G. F. (2002). Overview of the SEER-medicare data: Content, research applications, and generalizability to the United States elderly population. *Medical Care, 40*(8 Suppl), 3–18.
8. Kent, E. E., Ambs, A., Mitchell, S. A., Clauser, S. B., Smith, A. W., & Hays, R. D. (2015). Health-related quality of life in older adult survivors of selected cancers: Data from the SEER-MMHOS linkage. *Cancer, 121*(5), 758–765.
9. Quach, C., Sanoff, H. K., Williams, G. R., Lyons, J. C., & Reeve, B. B. (2015). Impact of colorectal cancer diagnosis and treatment on health-related quality of life among older Americans: A population-based, case-control study. *Cancer, 121*, 943–950.
10. Stover, A. M., Mayer, D. K., Muss, H., Wheeler, S. B., Lyons, J. C., & Reeve, B. B. (2014). Quality of life changes during the pre- to postdiagnosis period and treatment-related recovery time in older women with breast cancer. *Cancer, 120*(12), 1881–1889.
11. Ware, J. E., Bayliss, M. S., Rogers, W. H., Kosinski, M., & Tarlov, A. R. (1996). Differences in 4 year health outcomes for elderly and poor chronically Ill patients treated in HMO and fee-for-service systems. Results from the Medical Outcomes Study. *JAMA, 276*, 1039–1047.
12. Kazis, L. E., Miller, D. R., Clark, J. A., Skinner, K. M., Lee, A., Ren, X. S., Spiro, A. 3rd, Rogers, W. H., & Ware, J. E. Jr. (2014). Improving the response choices on the Veterans SF-36 Health Survey role functioning scales: Results from the Veterans Health Study. *The Journal of Ambulatory Care Management, 27*(3), 263–280.
13. Ware, J. E., Kosinski, M., Bayliss, M. S., McHorney, C. A., Rogers, W. H., & Raczek, A. (1995). Comparison of methods for scoring and statistical analysis of SF-36 health profiles and summary measures: Summary of results from the medical outcomes study. *Medical Care, 33*(suppl 4), AS264–AS279.
14. Coste, J., Quinquis, L., Audureau, E., & Pouchot, J. (2013). Non response, incomplete and inconsistent responses to self-administered health-related quality of life measures in the general population: Patterns, determinants and impact on the validity of estimates—a population-based study in France using the MOS SF-36. *Health and Quality of Life Outcomes, 11*, 44.
15. Centers of Medicaid and Medicare Services. (2007). ‘Imputing the physical and mental summary scores (PCS and MCS) for the MOS SF-36 and the Veterans SF-36 Health Survey in the Presence of Missing Data. Retrieved October 28, 2015 from http://www.MHOSonline.org/surveys/MHOS/download/MHOS_Veterans_36_Imputation.pdf.
16. Ware, J. E. Jr., & Sherbourne, C. D. (1992). The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30*, 473–483.
17. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health Status Survey Manual and Interpretation Guide*. Boston, MA: The Health Institute, New England Medical Center. Retrieved from The Medical Outcomes Trust.
18. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Transactions on Pattern Analysis and Machine Intelligence, 24*, 881–892.
19. Crano, W. D., Brewer, M. B., & Lac, A. (2015). *Principles and methods of social research* (3rd ed.). New York, NY: Routledge.
20. Ware, J. E., Kosinski, M., & Keller, S. K. (1994). *SF-36® physical and mental health summary scales: A user's manual*. Boston, MA: The Health Institute.
21. nl—Nonlinear least-squares estimation. Retrieved November 3, 2016 from http://www.stata.com/manuals13/rnl.pdf.
22. Ypma, T. J. (1195). Historical development of the Newton-Raphson method. *SIAM Review, 37*(4), 531–551.
23. Rogers, W., Qian, S. X., & Kazis, L. E. (2004). Imputing the physical and mental summary scores (PCS and MCS) for the MOS SF-36 and the Veterans SF-36 Health Survey in the presence of Missing Data. Technical Report Prepared for the National Committee for Quality Assurance. Retrieved February 13, 2018 from http://www.hosonline.org/globalassets/hos-nline/publications/hos_veterans_36_imputation.pdf.
24. Spiro, A., Rogers, W., Qian, S. X., & Kazis, L. E. (2004). Imputing physical and mental summary scores (PCS and MCS) for the Veterans SF-12 Health Survey in the Context of Missing Data. Technical report prepared for the National Committee for Quality Assurance. Retrieved February 13, 2018 from http://www.hosonline.org/globalassets/hos-online/publications/hos_veterans_12_imputation.pdf.
25. Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker. ISBN 978-0-8247-5825-7.