UCLA

UCLA Previously Published Works

Title

The challenge of measuring intra-individual change in fatigue during cancer treatment

Permalink

https://escholarship.org/uc/item/8894s8vx

Journal

Quality of Life Research, 26(2)

ISSN

0962-9343

Authors

Moinpour, Carol M
Donaldson, Gary W
Davis, Kimberly M
et al.

Publication Date

2017-02-01

DOI

10.1007/s11136-016-1372-9

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The challenge of measuring intra-individual change in fatigue during cancer treatment

Carol M. Moinpour[1] · Gary W. Donaldson[2] · Kimberly M. Davis[3] · Arnold L. Potosky[3] · Roxanne E. Jensen[3] · Julie R. Gralow[4] · Anthony L. Back[4] · Jimmy J. Hwang[5] · Jihye Yoon[6] · Debra L. Bernard[4] · Deena R. Loeffler[7] · Nan E. Rothrock[11] · Ron D. Hays[8] · Bryce B. Reeve[9] · Ashley Wilder Smith[10] · Elizabeth A. Hahn[11] · David Cella[11]

Accepted: 18 July 2016
© Springer International Publishing Switzerland 2016

Abstract

Purpose To evaluate how well three different patient-reported outcomes (PROs) measure individual change.

Methods Two hundred and fourteen patients (from two sites) initiating first or new chemotherapy for any stage of breast or gastrointestinal cancer participated. The 13-item FACIT Fatigue scale, a 7-item PROMIS® Fatigue Short Form (PROMIS 7a), and the PROMIS® Fatigue computer adaptive test (CAT) were administered monthly online for 6 months. Reliability of measured change was defined, under a population mixed effects model, as the ratio of estimated systematic variance in rate of change to the estimated total variance of measured individual differences in rate of change. Precision of individual measured change, the standard error of measurement of change, was given by the square root of the rate-of-change sampling variance. Linear and quadratic models were examined up to 3 and up to 6 months.

Results A linear model for measured change showed the following by 6 and 3 months, respectively: PROMIS CAT (0.363 and 0.342); PROMIS SF (0.408 and 0.533); FACIT (0.459 and 0.473). Quadratic models offered no noteworthy improvement over linear models. Both reliability and precision results demonstrate the need to improve the measurement of intra-individual change.

Conclusions These results illustrate the challenge of reliably measuring individual change in fatigue with a level of confidence required for intervention. Optimizing clinically useful measurement of intra-individual differences over time continues to pose a challenge for PROs.

Keywords Measured change · Intra-individual change · Fatigue · Patient-reported outcomes · Cancer

✉ Carol M. Moinpour
cmoinpou@fredhutch.org

[1] Public Health Sciences Division, Fred Hutchinson Cancer Research Center [Emerita], Seattle, WA, USA

[2] Pain Research Center, Department of Anesthesiology, University of Utah, Salt Lake City, UT, USA

[3] Health Services Research, Georgetown University Medical Center, and Georgetown University Lombardi Comprehensive Cancer Center, Washington, DC, USA

[4] Seattle Cancer Care Alliance, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[5] Hematology/Oncology, Levine Cancer Institute, Carolinas HealthCare System, Charlotte, NC, USA

[6] Cancer Prevention Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[7] The Cystic Fibrosis Foundation, Bethesda, MD, USA

[8] Departments of Medicine and Health Services Research, University of California Los Angeles, Los Angeles, CA, USA

[9] Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[10] Outcomes Research Branch, National Cancer Institute, Bethesda, MD, USA

[11] Department of Medical Social Sciences and Center for Patient-Centered Outcomes, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Published online: 28 July 2016

Introduction

Fatigue is a common cancer-related symptom and adverse event reported by individuals treated for cancer [1, 2]. The importance of monitoring fatigue levels in cancer patients and survivors was recently emphasized in a new set of American Society of Clinical Oncology (ASCO) guidelines for screening, assessing, and managing fatigue [1]. The ASCO guidelines indicated that most cancer patients experience some degree of fatigue during treatment and about 30 % of patients have persistent fatigue. The value of routine, in-clinic assessment of multiple patient-reported outcomes (PROs), including screening for cancer-related symptoms such as fatigue, is of great interest [3–7]. Routine use of PROs in clinical settings potentially allows for (1) monitoring effects of treatment on individual patients as well as (2) assisting in disease management, both of which allow for the direct inclusion of patient experience into the care environment [7–9]. Fatigue screening approaches are recommended in the 2014 ASCO guidelines as routine practice, beginning at diagnosis and moving to more comprehensive assessment techniques for patients whose fatigue reflects moderate to severe levels; at this level, more clinical and laboratory evaluation as well as more comprehensive patient-reported measures is recommended.

Nunnally's text, Psychometric Theory, [10] is recognized as a key source for instrument development; a reliability of 0.95 was recommended if decisions are to be made regarding a single test score (page 246). Either a reliability level of 0.95 or $\geq$0.90 has continued to be recognized by instrument developers for use of PRO scores at the individual level [4, 11–13]. The precision of PRO measures used in oncology clinical settings has been debated for a number of years [11, 12, 14–17]. Although fixed length and CAT (variable length) measures developed using item response theory (IRT) have been described as more precise (per item) than commonly used PRO measures based on classical test theory criteria [8, 9, 18–20], few studies have been published supporting use of CAT measures (or any other measures) in terms of their reliable responsiveness to individual change. Lai et al. [19] have demonstrated such better precision of individual status with a CAT measure, but with respect to a single time point. The most commonly used measure of reliability has been coefficient alpha [21], but this measure uses item homogeneity to estimate systematic variation between individuals at static time points, not systematic differences in how individuals change [17, 22]. Important psychometric criteria for measured change in fatigue and other PROs, such as the precision of estimates of individual change, have been described [10, 12, 14–17, 22, 23]. Stringent reliability and precision criteria are just as important for measurement of individual change, but rate of change has seldom been considered as a measurement.

Hahn et al. [15] compared measurement error for PRO (e.g., SF-36 scales) and clinical (e.g., systolic and diastolic blood pressure) measures of patient status and concluded that these two types of measures were comparable with respect to measurement criteria and that within both types of measures, one finds high, good, and low reliability. Many clinical status measures show considerable measurement error (e.g., tumor size measurements and systolic hypotension); yet these measures are used consistently in medical care and research. Hahn et al. [15] suggested that given this widespread variation in measurement quality, use of both types of measures can complement each other (i.e., both are important for arriving at an accurate description of patient status). Methods addressing measured change at the individual patient level, such as those described in this report, need to be applied to longitudinal assessments using clinical measures. Measurement of trajectories of change is challenging regardless of the property being measured. Both clinical and PRO measures should be subjected to more research regarding tracking and interpreting change for individual patients.

In this exploratory, we sought to compare three measures of change in fatigue: IRT-based PROMIS Fatigue CAT [8, 9, 24, 25] and the 7-item PROMIS Fatigue 7a Short Form, hereafter referred to as the PROMIS 7a [19]; and the FACIT Fatigue [26, 27], a classical test theory-based, legacy measure of fatigue. This report compares the quality of change measurement for the three measures in a sample of patients undergoing chemotherapy treatment for cancer over a 6-month period.

Methods

Study design and sample

Two hundred fourteen patients with breast or gastrointestinal (GI) cancers were enrolled in this study through the Seattle Cancer Care Alliance (SCCA, Fred Hutchinson Cancer Research Center, Seattle, WA) and the Lombardi Comprehensive Cancer Center (LCCC, Georgetown University, Washington, DC). The study protocol was approved by Institutional Review Boards at both institutions for the following patient samples: women and men aged 21 years and older; up to three weeks pre- or post-initiation (first day of cycle 1) of the current chemotherapy regimen (intravenous or oral agents [IV or PO]) for any stage breast, gastric, colon, rectal, small bowel, esophageal, liver, bile duct, and gall bladder cancers; prior chemotherapy treatment allowed; US residence; ability to complete computer assessments in English (determined by research staff).

Study enrollment and data collection procedures

At both research sites, the Research Coordinator provided a laptop for enrolling patients; at LCCC, this laptop was made available upon request for future encounters with patients who did not have home Internet access. At SCCA, patients without home Internet access could complete the online fatigue assessments at the SCCA Patient and Family Resource Center. Regardless, all patients completed all assessments online through the PROMIS Assessment Center. Patients could be enrolled up to three weeks pre- or post-initiation (first day of cycle 1) of the current chemotherapy regimen. The Research Coordinator consented and registered each patient on the PROMIS® Assessment Center website [28] [http://www.nihpromis.org; http://www.assessmentcenter.net/] and then trained the patient to complete the first PROMIS® assessments online. Timing of the five follow-up monthly assessments was based on the date of study registration. Monthly assessments could be completed during the last week of the month (multiple log-ins possible during this week). Assessment dates were not always synchronized with treatment administration due to treatment delays, which could not be identified in time to revise the Assessment Center-scheduled assessments.

Participant incentives

Patients were offered a $50 (total) incentive for their participation in the study at two time points: after completion of at least two of the first three monthly assessments; similarly, upon completion of at least two of assessments 4, 5, and 6.

Adherence monitoring

The Research Coordinator at each of the two clinical sites provided email or telephone reminders at the beginning of the week in which the monthly online assessment was available and before the end of online availability if the patient had not completed that assessment during the first part of the open period. The Research Coordinator also called patients twice during the first 3 months to identify any difficulties accessing or responding via the website.

Patient-reported fatigue and sociodemographic measures

The PROMIS Fatigue CAT [8, 24, 25, 29] currently defaults to a minimum of 4 and a maximum of 12 items; this range typically ensures meeting the PROMIS standard error precision criterion [19]. This was the Assessment Center criterion when this study was conducted. We did not compare other stopping rules for two reasons: (1) we wanted to use the precision criteria currently in use by Assessment Center since this was a model assessment approach for measuring PROs in the multi-research site setting and (2) this project did not have sufficient funding to compare stopping rules or evaluate the use of different sets of items from the fatigue item bank. The CAT items were administered first, followed by the PROMIS 7a [30] [https://www.assessmentcenter.net/documents/InstrumentLibrary.pdf] and the FACIT Fatigue [26, 31]. Once a month for 6 months, patients were administered the additional covariate measures: PROMIS® Sleep Disturbance 8-item Short Form [32]; one PROMIS® Global Fatigue item [30, 33, 34]; a patient-reported performance status single item [26, 35, 36]; two questions about patients' physical activity [37, 38]; and the global rating of change in fatigue item (assessments 2–6). Our pilot test of the Assessment Center's administration of items for this Clinical Study confirmed that after answering the PROMIS Fatigue CAT, patients did not receive any duplicate items when the PROMIS 7a and the FACIT Fatigue Scale were presented.

This report also includes information for the following clinical status variables for patients from both research sites: cancer type, stage, and number of prior chemotherapy regimens. The Research Coordinators verified this information (medical records review or confirmation provided by clinical staff).

Statistical analyses for assessment of individual change

Precision and reliability of measured change

Measurement theory links an observed quantity to an unobserved or latent variable. For this report, we construe the observed measure as an individual's mean rate of change, or slope, on a PRO fatigue scale over six assessments. This pragmatic definition of rate of change generalizes the change score to multiple assessments and conveys to clinicians and patients an approximate degree of improvement or worsening. In the reliability context, the average rate of change for one person (as calculated from a chart or an Excel regression line) is the observed variable, and the true rate of change in fatigue is the unobserved attribute. Imprecision of measured change is given by the standard error of the estimated rate-of-change parameter, the expected deviation of the estimated attribute from the true attribute for one person, the typical amount by which the estimated rate of change is likely to be off from an individual's true rate of change.

If the error variance in a single assessment σ_e^2 is known, the standard error for an individual i's rate-of-change score $\beta_{1\mathrm{i}}$ can be directly calculated as the square root of the sampling variance $[V(\hat{\beta}_{1i}|\beta_{1i})]^{1/2} = [\sigma_e^2/\sum_k (T_k - \bar{T})^2]^{1/2}$ over the times of assessment T_k. If the variance $V(\beta_1)$ of the true rate of change in the population is further known, then the reliability of the measured rate of change can be directly calculated as $Rxx'(\hat{\beta}_1) = V(\beta_1)/[V(\beta_1) + V(\hat{\beta}_{1i}|\beta_{1i})]$, where the numerator reflects true variance in slope across members of the population and the denominator is the total measured slope variance (comprising true plus sampling variability).

In practice, particularly in common clinical monitoring scenarios, the average individual rate of change is easily estimated but the two variance quantities are unknown. To obtain efficient estimates for each of the three PRO measures, we assume a standard linear latent growth mixed effects model, with each individual patient i characterized by random effect intercept $\beta_{0\mathrm{i}}$ (starting point) and slope $\beta_{1\mathrm{i}}$ (rate of change) terms over the six ($t = 0$ to 5) time points of the study: $Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \varepsilon_{ij};\ \ \beta_{0i}, \beta_{1i} \sim MVN((\beta_0, \beta_1), \boldsymbol{\Psi})$, $\varepsilon_{ij} \sim (0, \sigma^2)$. In the population model, the two systematic random effects have means $\mathrm{B} = (\beta_0, \beta_1)'$ and unstructured bivariate normal covariance matrix $\boldsymbol{\Psi}$ (containing variances for the intercept and slope and their covariance), while the unsystematic error is assumed normal and serially uncorrelated with constant variance σ_e^2. To enable a more interpretable comparison of the precisions, we analyzed standardized scores for the three fatigue measures; this has no effect on reliability, relative precision, or other relevant psychometric properties. The key hypothesis is evaluated by comparing the relative precision and reliability for the three measures as described above and summarized in Table 1 [22, 39, 40].

For individual monitoring, precision is more important than reliability, but concepts of reliability are more familiar. The two indices provide complementary information. Precision describes the uncertainty in the measurement (estimation) of an individual's change, while reliability also reflects the heterogeneity of the attribute in the (infinite theoretical) population. Reliability is a population concept, capturing how variably members of the population are systematically changing relative to the total variability (including error) of rate-of-change measurements. Our estimate of reliability uses the classical definitional formula, which applies to any measured trait, state, or rate attribute β: $Rxx'(\hat{\beta}) = V(\beta)/[V(\beta) + V(\hat{\beta}_i|\beta_i)]$.

While we would prefer to emphasize the salience of precision in individual measurement, it is expedient for the purposes of this paper to present the more familiar reliability context as well. As noted above [10, 12, 14–17, 22, 23] and to our knowledge, PRO measures in use today have not examined the reliability or precision of a measure for assessing rate of change in individual patient status.

The average (linear) rate of change is indispensably relevant as a summary of individual improvement or worsening. It is therefore our primary psychometric target. For simplicity, we provisionally assume that the linear trend captures the important systematic change in the 6-month period. Although not our main focus, we also examine whether the fit of the measurement model can be improved by allowing polynomial time trends. Nonlinear trends are, however, much more difficult to interpret as summary attributes. As a pragmatic guide, we provide an estimate of the minimum number of assessments required for each of the measures to yield good measurement (conventionally and somewhat arbitrarily set at .90 for individual assessment; reliability depends on the intrinsic true variability, which varies across populations.) of individual change in fatigue [12].

Results

Table 2 summarizes baseline characteristics for 214 patients who enrolled in this study at the two locations with the following percentage of covariate information: SCCA in Seattle (62 %) and LCCC in Washington, DC (38 %). Patient covariate data along with form submission rate data were available for 213 of the 214 patients. The mean age was 52, and most patients were female (69 %). Table 3 lists baseline levels of major covariates collected for the study. Baseline fatigue levels for the single-item global item were mostly mild and moderate. These patients were physically active with 41 %, indicating that they walked more than an hour during a week; 66 % walked at a moderate to fast speed. Sixty-nine percent reported good performance status (no symptoms or some symptoms that did not require bed rest). Form submission rates for both sites (Table 4) were 100 % at baseline (assessment #1), dropping to 67 % by assessment #6. By the 6th assessment, form submission rates were higher for patients enrolled at the SCCA site (71 %) versus the LCCC site (59 %). Table 5 shows the number of patients receiving treatment at each of the six assessment times at the SCCA. The number of patients receiving treatment at the SCCA dropped after the third assessment; this was due primarily to shorter courses of chemotherapy. However, 22 SCCA patients began radiation treatment after their chemotherapy regimen during the remaining three assessment periods and were treated throughout the 6-month assessment period. Eighty-nine of the 214 patients (42 %) returned to answer

Table 1 Individual change attributes as distributed in the population and measured in a sample

Population data model	Population (prior) distributions	Scoring = Estimation of β_{1i}
Description of time trajectory for all hypothetical members of infinite population	Summary characteristics of the hypothetical infinite population (i.e., population parameters) that render observed data likeliest	Narrow scope of estimation: Measured time slope and uncertainty for person *i in the sample*
$Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \varepsilon_{ij}$	$\beta_{0i}, \beta_{1i} \sim MVN\left((\beta_0, \beta_1), \begin{bmatrix} V(\beta_0) & (sym) \\ C(\beta_0\beta_1) & V(\beta_1) \end{bmatrix}\right)$ $\varepsilon_{it} \sim \left(0, \sigma_\varepsilon^2\right)$	$\hat{\beta}_{1i} = \left[\sum_{j=1}^{T_i} \left(t_{ij} - \bar{t}_i\right)\left(Y_{ij} - \bar{\text{Y}}_i\right)\right] / \sum_{j=1}^{T_i} \left(t_{ij} - \bar{t}_i\right)^2$ (Measured time slope for person *i*)
		Conditional sampling variance (and standard error) of rate measurement $\hat{\beta}_{1i} \sim N\left(\beta_{1i}, V\left(\hat{\beta}_{1i}\vert\beta_{1i}\right)\right)$ (*unbiased*) $\sigma_{\hat{\beta}1i} = \left[V\left(\hat{\beta}_{1i}\vert\beta_{1i}\right)\right]^{1/2} = \left[\sigma_e^2 / \sum_{j=1}^{T_i} \left(t_{ij} - \bar{t}_i\right)^2\right]^{1/2}$
	Reliability of measured (estimated) rate of change $Rxx'\left(\hat{\beta}_1\right) = V(\beta_1) / \left[V(\beta_1) + V\left(\hat{\beta}_{1i}\vert\beta_{1i}\right)\right]$	

monthly assessments after a missed assessment, a fairly common experience. Only nine patients at SCCA and ten patients at LCCC went off study permanently; reasons for drop-out included death ($n = 8$), ill health ($n = 4$), no further verbal contact ($n = 3$), and other ($n = 4$).

Table 6 shows the reliability and precision estimates for measured change in fatigue at the individual patient level for each of the three fatigue measures ($n = 214$ patients): PROMIS Fatigue CAT, PROMIS 7a, and the FACIT Fatigue. The sample size for these analyses include one additional patient ($n = 214$) not included in Tables 2, 3, 4, and 5. The reliabilities mentioned in the title of Table 6 reflect the ratio of estimated variance of systematic individual differences in rate of change (i.e., how people in the population are estimated to differ in their individual rates-of-change) to total variance of measured individual differences in rate of change. These analyses are based on standardized (z-score) scales for either 6 or 3 occasions. Column 2 (population variance) is an estimate of true rate-of-change variability from mixed effects population models with correlated slopes and intercepts. Column 3 (sampling variance) is equivalent to the expected value of the squared discrepancy between unbiased maximum likelihood/ordinary least squares (ML/OLS) individual parameter estimates, assessed independently one person at a time, and the individual's true value. In the special case of linear trend, the sampling variance is the (pooled) measurement error variance divided by the sum of an individual's squared time deviations from the mean time point. Standard error of measurement results is shown in Column 4, which are square roots of respective sampling variances (e.g. $\sqrt{}.031143$ in Column 3 = the precision estimate of .176 in column 4). Uncertainties of individual measurements (estimates) of intercept and slope depend on the number and pattern of assessments for each person. Inspection of preliminary results suggested that patients displayed the most true change variability in the first 3 months, with greater stability thereafter. To allow the most favorable case for reliability to emerge, we therefore also ran models restricted to the first 3 months. Table 6 assumes availability of the full number of assessments (either 3 or 6). In practice, fewer assessments than this would yield worse precision. See Table 6 notes for additional statistical explanations for columns 2–5.

None of the three assessments met conventional benchmarks for excellent reliability (i.e., ≥0.90). Standard errors of measurement of change are also high, indicating large individual measurement uncertainty and poor precision of measured change at the individual patient level. As expected, each individual's rate of change was more precisely measured with six assessments than with three. Improved precision did not translate into enhanced scale reliability for change, since population rate of change variances also decreased with longer time spans; months 4–6 evidently introduced a clinical environment in which patients were somewhat more similar in how their fatigue changed. With the exception of CAT based on six time points, more complex polynomial time models did not notably improve the fit of the measurement model; thus, the linear fits of Table 6 are the best summaries of change measurement.

Discussion

The results of this study support previous concerns regarding use of PRO measures designed for group-level research to monitor change in individual patient status

Table 2 Patient characteristics ($n = 213^a$)

Age at diagnosis: mean (SD)	52.40 (10.80)
	n (%)
Gender	
Female	147 (69)
Male	66 (31)
Race	
White	162 (76.1)
Black or African American	17 (8.0)
Asian	16 (7.5)
American Indian/Alaska Native	2 (0.9)
Native Hawaiian/Pacific Islander	0 (0)
More than one race	9 (4.2)
Not provided (missing)	7 (3.3)
Ethnicity	
Hispanic or Latino	10 (4.7)
Not Hispanic or Latino	202 (94.8)
Not Provided (missing)	1 (0.5)
Education	
Less than high school	2 (0.9)
High school graduate or GED	13 (6.1)
Some college or technical/vocational school	58 (27.2)
College graduate	59 (27.7)
Some graduate school	15 (7.0)
Graduate degree	66 (31.0)
Marital Status	
Married or living with someone	148 (69.5)
Divorced	32 (15.0)
Separated	3 (1.4)
Widowed	3 (1.4)
Single (never married)	23 (10.8)
Not provided (missing)	4 (1.9)
Employment status	
Working full time	95 (44.6)
Working part time	18 (8.50
Full-time homemaker or family caregiver	15 (7.0)
Retired	39 (18.3)
Unemployed	13 (6.1)
Student	3 (1.4)
Other	12 (5.6)
More than one answered	17 (8.0)
Not provided (missing)	1 (0.5)
# People living near you/can count on for help	
0	7 (3.3)
1	14 (6.6)
2	33 (15.6)
3–5	65 (30.5)
6–9	29 (13.6)
10 or more	64 (30.0)
Not provided (missing)	1 (0.5)

Table 2 continued

	n (%)
Research site location	
Seattle Cancer Care Alliance	132 (62.0)
Lombardi Comprehensive Cancer Center	81 (38.0)
Cancer type	
Breast	89 (41.8)
Colon	79 (37.1)
Rectal	17 (8.0)
Colorectal	1 (0.5)
Small bowel	2 (0.9)
Gastric	7 (3.3)
Esophageal	7 (3.3)
Liver	5 (2.3)
Bile duct	4 (1.9)
Gall bladder	2 (0.9)
Cancer stage	
I	15 (7.0)
II	47 (22.1)
III	47 (22.1)
IV	103 (48.4)
Not provided (missing)	1 (0.5)
# Prior chemo regimens	
0	101 (47.4)
1	30 (14.1)
2 or more	23 (10.8)
Not provided (missing)	59 (27.7)

[a] Clinical data were not available for one patient

[12, 17, 41]. The issues raised in these analyses are not necessarily specific to the instruments tested; in fact, we are not aware of any fatigue instruments that have demonstrated better performance in measuring intra-individual change. Nor are these issues necessarily specific to PRO data. We agree with Hahn et al. [15] regarding the need to question the measurement of change in clinical and other outcome variables, with respect to how measured change is trusted as reliable when making decisions about individual patient status (e.g., responder versus non-responder).

An additional challenge for clinicians is that most guidelines for interpreting change (e.g., minimally important difference, or MID) in PRO measures are based on clinical trials involving large numbers of patients in different treatment groups; the cut point for change of interest is based on a comparison of averages obtained for each treatment group [42]. The challenge facing clinicians is how to apply findings based on averages for groups of patients to the patient facing the clinician [14, 17]. Hendrikx et al. [43] reported minimally important change

Table 3 Baseline covariate status ($n = 213^{a}$)

	n (%)
Global rating of fatigue	
None	30 (14.1)
Mild	85 (39.9)
Moderate	73 (34.3)
Severe	22 (10.3)
Very severe	2 (0.9)
Not provided (missing)	1 (0.5)
Physical activity	
Minutes/week walking	
Never takes walks	21 (9.9)
About 15 min	16 (7.5)
About 30 min	40 (18.8)
About 45 min	25 (11.7)
About 60 min	22 (10.3)
Longer than one hour	87 (40.8)
Not provided (missing)	2 (0.9)
Usual walking speed	
Never takes walks	13 (6.1)
Very slowly	8 (3.8)
Slowly	36 (16.9)
Moderately	111 (52.1)
Fast	29 (13.6)
Very fast	14 (6.6)
Not provided (missing)	2 (0.9)
Patient-rated performance status	
Normal activity without symptoms	56 (26.3)
Some symptoms, but not required bed rest during waking day	93 (43.7)
Require bed rest for less than 50 % of waking day	47 (22.1)
Require bed rest for more than 50 % of waking day	14 (6.6)
Unable to get out of bed	0 (0)
Not provided (missing)	3 (1.4)

[a] Covariate data were not available for one patient

cutoff values based on group-level data were not appropriate for monitoring change in individual patients due to misclassifications in such change; the authors also noted the failure to incorporate how patients value the change and the consequences of changing patient care based on these scores. This concern echoes what Donaldson noted in prior work [17].

As a practical example, consider the possible use of the PROMIS Fatigue 7a over the course of six assessments. This scale and time range featured the best precision of measurement in Table 6. (In this example, the low reliability reflects in part the low-population variability in true change over six assessments.) With a standard error of measurement of .160 (column 4, Linear PROMIS SF, 6 Times), the 95 % confidence intervals are roughly ±.320 about individual rate-of-change estimates, which are in units of expected change in standard scores per assessment time. A patient estimated to have a rate of change of .20 would display a cumulative increase of 1.00 standard deviation units over the course of the six assessments, a seemingly large effect. Yet the confidence limits on the rate-of-change measurement span .520 (.200 + .320) to −.120 (.200 − .320). Despite the large measured change estimate, given the wide confidence intervals around these estimates (and the fact that the interval includes zero), it would not be possible to confidently determine whether the patient was getting worse or better or staying the same. Individual assessments require excellent precision, with standard errors much smaller than the scores, the estimated attributes.

Despite the low reliabilities in this clinical application, adding additional times of assessment, more items/specific types of items, and different analysis methods may address the problems we observed. Others (Faes et al., Brandmaier

Table 4 Internet assessment submission rates

Assessment groups (*N*)	All	LCCC	SCCA
Assessment 1 completed	213	81	132
Assessment 1 missed	0	0	0
Assessment 1 off study	0	0	0
Assessment 2 completed	154	60	94
Assessment 2 missed	49	16	33
Assessment 2 off study	10	5	5
Assessment 3 completed	145	53	92
Assessment 3 missed	53	21	32
Assessment 3 off study	15	7	8
Assessment 4 completed	136	47	89
Assessment 4 missed	59	25	34
Assessment 4 off study	18	9	9
Assessment 5 completed	135	46	89
Assessment 5 missed	59	25	34
Assessment 5 off study	19	10	9
Assessment 6 completed	129	42	87
Assessment 6 missed	65	29	36
Assessment 6 off study	19	10	9
Assessment 1 submission rate	100	100	100
Assessment 2 submission rate	75.86	78.947	74.02
Assessment 3 submission rate	73.23	71.622	74.19
Assessment 4 submission rate	69.74	65.278	72.36
Assessment 5 submission rate	69.59	64.789	72.36
Assessment 6 submission rate	66.49	59.155	70.73

Submission Rate = Completed/Eligible to complete (completed + missed)

Submission rate data were available only for patients who completed the full set of measures; one patient did not complete the covariate data

LCCC Lombardi Comprehensive Cancer Center, *SCCA* Seattle Cancer Care Alliance

et al.) [44, 45] have provided efficient frameworks for investigating specific reliability and precision scenarios within common mixed model contexts. Below we suggest potential solutions, all of which require more research.

1. The low reliabilities reflect low rate-of-change variability in the population as well as imprecise individual measurement. Precisions (and therefore reliabilities) of change measurements may be improved by adding additional assessments. To attain excellent measured change reliabilities of .90 in our study, we use the Table 6 formula for sampling variance to estimate that the PROMIS CAT would need 15 total assessments over the same time range, the PROMIS 7a would need 14, and the FACIT would need 13, assuming the same underlying psychometrics (data not shown). Measurement can be further strengthened by sampling time points nearer the ends and the beginnings of longitudinal sequences (so that the sum of squared time deviations can be greater).
2. Adding additional items may or may not improve the reliability of change; this would depend on the items added. For example, true improvement in measurement of change may require the selection of items explicitly for their sensitivity to clinical change, as opposed to the typical psychometric development approach of maximizing individual differences at single time points. Item bank approaches such as those employed for PROMIS CAT and Short Form measures provide an excellent method for identifying such items; item banks can also be expanded to include more items that capture change.
3. Empirical Bayes (EB) scoring [46] would yield slightly better reliabilities and reduced sampling variabilities, but EB scores are conditionally biased, depend on applying weights from a population study, and allow each patient's score to be partly determined by the scores of other patients. This is somewhat at odds with the simple goal that each patient's change score should reflect only the observed data for that patient. Moving to slightly more complex models, reliabilities of residual gain scores (regression as opposed to difference) will be slightly better than scoring based on pure change [47]. Many other scoring variations are possible within a multivariate framework.

Applying any of these recommendations has meaningful implications for the overall feasibility of data capture and use in clinical care situations. The data from this study suggest that regarding patient/clinician decision making, the confidence limits of individual change assessment are so wide that reliable determination of changes in status over time may not always be feasible. Adding assessments can improve precision, but only until they become prohibitive due to patient burden and retest effects. We

Table 5 Number of SCCA patients receiving chemotherapy or chemotherapy + radiation at each assessment time point

	# PT on TX A1	# PT on TX A2	# PT on TX A3	# PT on TX A4	# PT on TX A5	# PT on TX A6	# PT off study
Chemo + RT	132	117	97	82	68	53	9
# PT lost at each A		15	20	15	14	15	

PT patient, *TX* treatment, *Chemo* chemotherapy [All types: Infusion and oral chemo; Chemo + RT (22 SCCA patients began RT post-chemo); Infusion + oral + RT], *RT* radiation therapy, *A* Assessment

Table 6 Estimated reliabilities of measured change[a] in three fatigue measures

$n = 214$	Population variance[c]	Sampling variance[d]	Standard error of measurement[d]	Reliability of measurement[e]
LINEAR CAT, 6 TIMES				
Intercept $T_0 = 0$	.414636	.285474	.534	.592
(Intercept $T_0 = 2.5$)	(.410391)	(.090826)	(.301)	(.819)
Linear slope	**.017738**	**.031143**	**.176**	**.363**
Within person		.544996 |*i*,*t*	.738238 |*i*,*t*	
Single assessment[f]				.455 |*t*
QUADRATIC CAT, 6 TIMES				
Intercept $T_0 = 0$	.494551	.387915	.623	.560
Linear Slope|T_0 = 0[b]	**.103431**	**.343221**	**.586**	**.232**
Quadratic	.004427	.012649	.112	.259
Within person		.472245 |*i*,*t*	.687 |*i*,*t*	
Single assessment				.528 |*t*
LINEAR CAT, 3 TIMES				
Intercept $T_0 = 0$	.651514	.33248	.577	.659
Linear slope	**.103597**	**.19949**	**.447**	**.342**
Within person		.38972 |*i*,*t*	.624 |*i*,*t*	
Single assessment				.610 |*t*
LINEAR PROMIS SF, 6 TIMES				
Intercept $T_0 = 0$	.533152	.235924	.486	.693
Linear slope	**.017712**	**.025737**	**.160**	**.408**
Within person		.450401 |*i*,*t*	.671 |*i*,*t*	
Single assessment				.550 |*t*
LINEAR PROMIS SF, 3 TIMES				
Intercept $T_0 = 0$	.830349	.219962	.469	.791
Linear slope	**.150426**	**.131977**	**.363**	**.533**
Within person		.263954 |*i*,*t*	.490 |*i*,*t*	
Single assessment				.736 |*t*
LINEAR FACIT, 6 TIMES				
Intercept $T_0 = 0$	.511570	.254444	.504	.668
Linear slope	**.023577**	**.027758**	**.167**	**.459**
Within person		.485757 |*i*,*t*	.697 |*i*,*t*	
Single assessment				.514 |*t*
LINEAR FACIT, 3 TIMES				
Intercept $T_0 = 0$	.693272	.252123	.502	.733
Linear slope	**.136008**	**.151274**	**.389**	**.473**
Within person		.302547 |*i*,*t*	.550 |*i*,*t*	
Single assessment				.697 |*t*

Table 6 continued

[a] Ratio of estimated variance of systematic individual differences in rate of change to total variance of measured individual differences in rate of change using model-based pooled estimates of within-person error. To make descriptive comparisons across scales easier, the analyses are based on standardized (z-score) scales over all time points available (either 6 or 3 occasions). This linear transformation has no effect on reliability calculations

[b] Instantaneous rate of change at $T = 0$ baseline

[c] ML (maximum likelihood) estimate of true variability from mixed effects population models with correlated slopes and intercepts

[d] Equivalent to expected value of squared discrepancy between unbiased ML/OLS (ordinary least squared) individual parameter estimate, assessed independently one person at a time, and individual's true value. Sampling variance given collectively by $diag(\Lambda'\Theta^{-1}\Lambda)^{-1}$, where Θ is the within-person sampling (measurement) error, an estimated population parameter, and Λ contains the constant, linear, and quadratic time contrasts. Standard errors are square roots of respective sampling variances. Uncertainties of individual measurements (estimates) of intercept and slope depend on number and pattern of assessments for each person. The table assumes availability of the full number of assessments (either 3 or 6). Standard errors increase, and reliabilities decrease, with fewer assessments. The square root of the within-person residual error is the model-based estimate of the scale's standard error of measurement at any given time point. For example, under the linear six assessment model the CAT has an estimated scale standard error of .742 in standardized units. This defines the typical error of measurement error expected in one assessment of one individual. The 95 % confidence intervals for the single assessment would be approximately ±(2 × .742) = ±1.484 standardized units

[e] Reliabilities defined classically as ratios of true population variance (Ψ) to measured variance. The measured variance is the sum of true variance and the sampling variance. The reliability is then $\rho_{bb'} = \psi_b/(\psi_b + diag_b(\lambda_b'\Theta\lambda_b)^{-1})$, where b is the intercept or slope (or quadratic coefficient)

Standard errors and reliabilities for Intercept depend on the time for which $T = 0$. Even though intercept is defined to occur at one (possibly hypothetical) time point, its estimation uses information from all time points. Taken above as $T = 0$ baseline, intercept is interpreted as initial status. As $T \rightarrow$ Mean (time), the intercept behaves more like the mean score across times, which can be highly reliable, though irrelevant for change. An example is provided for the linear CAT model with 6 assessment points. Setting $T = 2.5$, near the middle of the time range, yields highly reliable intercept measurement, but properties of the slope are unaffected

[f] With standardized measures, the model-based estimate of the reliability of a single assessment is one minus the residual or within-person error, equivalent to one minus the squared standard error of measurement of the scale. The term t refers to a single assessment at any single time

suggest designing studies to identify or develop precise change measures and the number of assessments needed to yield adequate reliabilities given the true variability in the population. For homogeneous populations, reliable measurement, whether of traits or of change, may simply not be possible. If all chemotherapy patients could show essentially zero fatigue that remained flat over time, this would be an ideal result, having zero reliability of measurement but excellent precision.

The sample size of 214 is not large for psychometric investigations, but easily exceeds conventional rules of thumb [10]. The calculated results involve ratios of variances and are essentially unbiased. Precision is more of a concern, though not a major one. This sample size yields a standard error of the measurement error variance ranging from 5 to 10 % in these scales, a margin too small to meaningfully alter the precision of estimation for the measured intercepts and slopes in Table 6. Reliability also involves uncertainty in estimated variances of the true slope and intercept attributes, but sampling uncertainty at this level can be subsumed under broader questions of how the reliability would change under populations having more or less attribute variation. Adding additional patients to such a study would not systematically affect our estimates of precision, but adding more numerous and widely spaced time points would improve precision. Psychometric study of change should include close consideration of the time design of the assessments.

Why did we fail to observe good to excellent reliability? We must first carefully distinguish the measured attributes in the reliability models. Conventional reliability studies investigate how consistently measurements can distinguish inter-individual differences in an unchanging trait or otherwise stable attribute. They assume no underlying change, or that everyone changes uniformly. We focused instead on the rate of change as the fundamental attribute to distinguish individuals. Conventional models that assume no change are clearly inconsistent with better measurement of how patients may be changing, and how reliably and precisely we can measure this change. As previously noted [12, 41], measurement of change can be much less reliable than static measurement.

Our study informs researchers that current PRO measures may lack the precision required to inform decisions about change in individual status, particularly with respect to medical decision making. We do not believe this finding is specific to the fatigue instruments used here to illustrate the problem. Any number of fatigue measures could have been used in this comparison and may well have come to the same conclusion. Indeed, we do not believe this issue is specific to PRO, as it may very well be the case in other clinical outcomes used to measure and monitor individual change. We suggest this be a subject of study more generally, as the field moves into increased tracking of individual change in research and clinical applications. Several decades ago, Cronbach and Furby [41] questioned whether we should be measuring change at all (particularly with respect to change or difference scores)! Current statistical methods now permit rigorous psychometric study of change, so our response to Cronbach and Furby's question is enthusiastically affirmative. Meeting rigorous psychometric criteria for change measurement, however, remains as challenging today as it was decades ago.

Acknowledgments Primary funding for the Clinical Study on Measuring Change in Fatigue was provided as part of a National Institutes of Health, Patient-Reported Outcomes Measurement Information System initiative for Grant # U01 AR057971 (Georgetown—4442-007-FHCRC): Arnold L. Potosky, PhD and Carol M. Moinpour, PhD, Co-Principal Investigators. The Patient-Reported Outcomes Measurement Information System® (PROMIS®) is an NIH Roadmap initiative to develop valid and reliable patient-reported outcome measures to be applicable across a wide range of chronic diseases and demographic characteristics. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. PROMIS II was funded by cooperative agreements with a Statistical Center (Northwestern University, PI: David Cella, PhD, 1U54AR057951), a Technology Center (Northwestern University, PI: Richard C. Gershon, PhD, 1U54AR057943), a Network Center (American Institutes for Research, PI: Susan (San) D. Keller, PhD, 1U54AR057926), and thirteen Primary Research Sites which may include more than one institution (State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, 1U01AR057948; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, 1U01AR057954; University of Washington, Seattle, PIs: Dagmar Amtmann, PhD and Karon Cook, PhD, 1U01AR052171; University of North Carolina, Chapel Hill, PI: Darren A. DeWalt, MD, MPH, 2U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, 1U01AR057956; Stanford University, PI: James F. Fries, MD, 2U01AR052158; Boston University, PIs: Stephen M. Haley, PhD and David Scott Tulsky, PhD (University of Michigan, Ann Arbor), 1U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD and Brennan Spiegel, MD, MSHS, 1U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, 2U01AR052155; Georgetown University, PIs: Arnold L. Potosky, PhD and Carol M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle),U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan DeWitt, MD, MSCE, 17 1U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, 1U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, 2U01AR052186). NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD, Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce B. Reeve, PhD, William Riley, PhD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein,MD, Ellen Werner, PhD and James Witter, MD, PhD. This manuscript was reviewed by PROMIS reviewers before submission for external peer review. The authors would like to thank the patients who participated in this study and who contributed data over the study period. We also appreciate the assistance of two individuals who helped set up the Clinical Study at FredHutch/SCCA: Denise L. Albano, MS and Christina G. Galer, MS. Chin H. Boo and Marjorie E. Alhadeff assisted with collection of SCCA clinical data. Kathleen Kealey and Nancy Williamson at the Fred Hutchinson Cancer Research Center provided key administrative assistance for the conduct of this study. We also appreciated assistance from Tania Lobo for statistical support at Georgetown University regarding form submission rates. In addition, we would like to thank Richard Gerson, PhD, Director of the PROMIS Technical Center that supports the PROMIS Assessment Center as well as Assessment Center staff Monica A. Prudencio and Linda Carrizosa.

Funding This work was supported by National Institutes of Health (Georgetown University, PIs: Arnold L. Potosky, PhD and Carol M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle), U01AR057971.

Compliance with ethical standards

Conflict of interest The authors have no conflicts of interest to disclose.

References

1. Bower, J. E., Bak, K., Berger, A., Breitbart, W., Escalante, C. P., Ganz, P. A., et al. (2014). Screening, assessment, and management of fatigue in adult survivors of cancer: An American Society of Clinical oncology clinical practice guideline adaptation. *Journal of Clinical Oncology, 32*(17), 1840–1850.
2. Berger, A. M., Lockhart, K., & Agrawal, S. (2009). Variability of patterns of fatigue and quality of life over time based on different breast cancer adjuvant chemotherapy regimens. *Oncology Nursing Forum, 36*(5), 563–570.
3. Davis, K., & Cella, D. (2002). Assessing quality of life in oncology clinical practice: A review of barriers and critical success factors. *Journal of Clinical Outcomes Management, 9*(6), 327–332.
4. Davis, K. M., Lai, J. S., Hahn, E. A., & Cella, D. (2008). Conducting routine fatigue assessments for use in clinical oncology practice: Patient and provider perspectives. *Supportive Care in Cancer, 16*(4), 379–386.
5. Alexander, S., Minton, O., & Stone, P. C. (2009). Evaluation of screening instruments for cancer-related fatigue syndrome in breast cancer survivors. *Journal of Clinical Oncology, 27*(8), 1197–1201.
6. Snyder, C. F., Aaronson, N. K., Choucair, A. K., Elliott, T. E., Greenhalgh, J., Halyard, M. Y., et al. (2012). Implementing patient-reported outcomes assessment in clinical practice: A review of the options and considerations. *Quality of Life Research, 21*(8), 1305–1314.
7. Wagner, L. I., Schink, J., Bass, M., Patel, S., Diaz, M. V., Rothrock, N., et al. (2015). Bringing PROMIS to practice: Brief and precise symptom screening in ambulatory cancer care. *Cancer, 121*(6), 927–934.

8. Reeve, B. B. (2006). Special issues for building computerized-adaptive tests for measuring patient-reported outcomes: The NIH's investment in new technology. *Medical Care, 44*(11 (Supplement 3)), S198–S204.
9. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11.
10. Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill Publishing Co.
11. Ware, J. E. J., Brook, R. H., Davies, A. R., & Lohr, K. N. (1981). Choosing measures of health status for individuals in general populations. *American Journal of Public Health, 71*(6), 620–625.
12. McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research, 4*(4), 293–307.
13. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl 1), S22–S31.
14. Donaldson, G. W., & Moinpour, C. M. (2002). Individual differences in quality-of-life treatment response. *Medical Care, 40*(6 Suppl), III39–III53.
15. Hahn, E. A., Cella, D., Chassany, O., Fairclough, D. L., Wong, G. Y., Hays, R. D., et al. (2007). Precision of health-related quality-of-life data compared with other clinical measures. *Mayo Clinic Proceedings, 82*(10), 1244–1254.
16. Fung, C. H., & Hays, R. D. (2008). Prospects and challenges in using patient-reported outcomes in clinical practice. *Quality of Life Research, 17*(10), 1297–1302.
17. Donaldson, G. (2008). Patient-reported outcomes and the mandate of measurement. *Quality of Life Research, 17*(10), 1303–1313.
18. Ware, J. E, Jr. (2003). Conceptualization and measurement of health-related quality of life: Comments on an evolving field. *Archives of Physical Medicine and Rehabilitation, 84*(4 Suppl 2), S43–S51.
19. Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation, 92*(10 Suppl), S20–S27.
20. Petersen, M. A., Aaronson, N. K., Arraras, J. I., Chie, W. C., Conroy, T., Costantini, A., et al. (2013). The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency. *Journal of Clinical Epidemiology, 66*(3), 330–339.
21. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
22. Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
23. Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K.-K. (2005). Evaluating the statistical significance of health-related quality-of-life change in individual patients. *Evaluation and the Health Professions, 28*(2), 160–171.
24. Garcia, S. F., Cella, D., Clauser, S. B., Flynn, K. E., Lai, J.-S., Reeve, B. B., et al. (2007). Standardizing patient-reported outcomes assessment in cancer clinical trials: A Patient-Reported Outcomes Measurement Information System initiative. *Journal of Clinical Oncology, 25*(32), 5106–5112.
25. Clauser, S. B., Ganz, P. A., Lipscomb, J., & Reeve, B. B. (2007). Patient-reported outcomes assessment in cancer trials: Evaluating and enhancing the payoff to decision making. *Journal of Clinical Oncology, 25*(32), 5049–5050.
26. Yellen, S. B., Cella, D. F., Webster, K., Blendowski, C., & Edward, K. (1997). Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *Journal of Pain and Symptom Management, 13*(2), 63–74.
27. Cella, D., Lai, J. S., Chang, C. H., Peterman, A., & Slavin, M. (2002). Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer, 94*(2), 528–538.
28. Gershon, R., Rothrock, N. E., Hanrahan, R. T., Jansky, L. J., Harniss, M., & Riley, W. (2010). The development of a clinical outcomes survey research application: Assessment Center. *Quality of Life Research, 19*(5), 677–685.
29. Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*(1 (Supplement)), 133–141.
30. Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64*(5), 507–516.
31. Cella, D., Hahn, E. A., & Dineen, K. (2002). Meaningful change in cancer-specific quality of life scores: Differences between improvement and worsening. *Quality of Life Research, 11*(3), 207–221.
32. Buysse, D. J., Yu, L., Moul, D. E., Germain, Anne, Stover, A., Dodds, N. E., et al. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep, 33*(6), 781–792.
33. Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Calla, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research, 18*(7), 873–880.
34. Weaver, K. E., Forsythe, L. P., Reeve, B. B., Alfano, C. M., Rodriguez, J. L., Sabatino, S. A., et al. (2012). Mental and physical health-related quality of life among U.S. cancer survivors: Population estimates from the 2010 National Health Interview Survey. *Cancer Epidemiology, Biomarkers and Prevention, 21*(11), 2108–2117.
35. Basen-Engquist, K., Bodurka-Bovers, D., Fitzgerald, M. A., Webster, K., Cella, D., Hu, S., et al. (2001). Reliability and validity of the Functional Assessment of Cancer Therapy—Ovarian. *Journal of Clinical Oncology, 19*(6), 1809–1817.
36. Heffernan, N., Cella, D., Webster, K., Odom, L., Martone, M., Passik, S., et al. (2002). Measuring health-related quality of life in patients with hepatobiliary cancers: The functional assessment of cancer therapy—Hepatobiliary questionnaire. *Journal of Clinical Oncology, 20*(9), 2229–2239.
37. Stewart, A. L., Hays, R. D., Wells, K. B., Rogers, W. H., Spritzer, K. L., & Greenfield, S. (1994). Long-term functioning and well-being outcomes associated with physical activity and exercise in patients with chronic conditions in the medical outcomes study. *Journal of Clinical Epidemiology, 47*(7), 719–730.
38. Moinpour, C. M., Lovato, L. C., Thompson, I. M, Jr., Ware, J. E, Jr., Ganz, P. A., Patrick, D. L., et al. (2000). Profile of men randomized to the prostate cancer prevention trial: Baseline health-related quality of life, urinary and sexual functioning, and health behaviors. *Journal of Clinical Oncology, 18*(9), 1942–1953.
39. Laird, N. M., & Ware, J. W. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963–974.

40. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
41. Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80.
42. Yost, K. J., & Eton, D. T. (2005). Combining distribution- and anchor-based approaches to determine minimally important differences: The FACIT experience. *Evaluation and the Health Professions, 28*(2), 172–191.
43. Hendrikx, J., Fransen, J., Kievit, W., & van Riel, P. L. (2015). Individual patient monitoring in daily clinical practice: A critical evaluation of minimal important change. *Quality of Life Research, 24*(3), 607–616.
44. Faes, C., Molenberghs, G., Aerts, M., Verbeke, G., & Kenward, M. G. (2009). The effective sample size and an alternative small-sample degrees-of-freedom method. *The American Statistician, 63*(4), 389–399.
45. Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C., & Lindenberger, U. (2015). LIFESPAN: A tool for the computer-aided design of longitudinal studies. *Frontiers in Psychology, 6*, 272. doi:10.3389/fpsyg.2015.00272.
46. Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent modeling: Multilevel, longitudinal, and structural equation models*. Boca-Raton, FL: Chapmen & Hall/CRC.
47. Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.