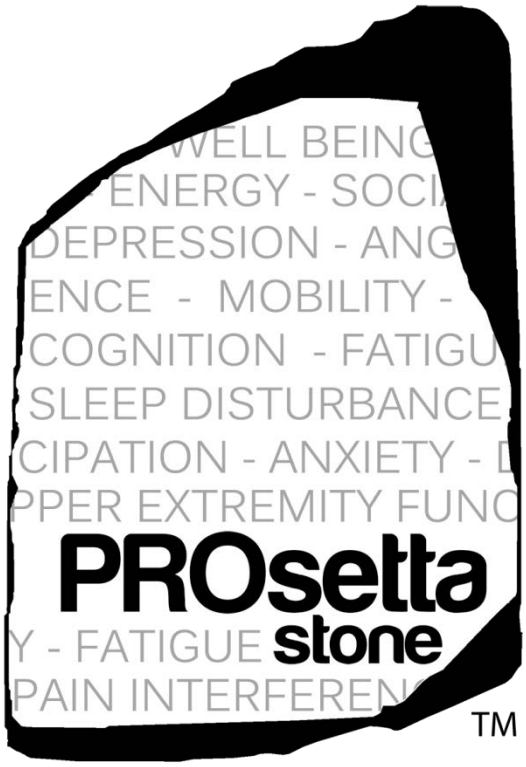

PROSETTA STONE® METHODOLOGY

A ROSETTA STONE FOR PATIENT REPORTED OUTCOMES

SEUNG W. CHOI, TRACY PODRABSKY, NATALIE MCKINNEY, BENJAMIN D. SCHALET, KARON F. COOK & DAVID CELLA

DEPARTMENT OF MEDICAL SOCIAL SCIENCES
FEINBERG SCHOOL OF MEDICINE
NORTHWESTERN UNIVERSITY

This research was supported by an NIH/National Cancer Institute grant PROSETTA STONE (1RC4CA157236-01, PI: David Cella). Authors acknowledge careful reviews, comments, and suggestions from Drs. Robert Brennan, Lawrence Hedges, Won-Chan Lee, and Nan Rothrock.

Table of Contents

1. Introduction

A common problem when using a variety of patient-reported outcome measures (PROs) for diverse populations and subgroups is establishing the comparability of scales or units on which the outcomes are reported. The lack of comparability in metrics (e.g., raw summed scores vs. scaled scores) among different PROs poses practical challenges in measuring and comparing effects across different studies. Linking refers to establishing a relationship between scores on two different measures that are not necessarily designed to have the same content or target population. When tests are built in such a way that they differ in content or difficulty, linking must be conducted in order to establish a relationship between the test scores. One technique, commonly referred to as equating, involves the process of converting the system of units of one measure to that of another. This process of deriving equivalent scores has been used successfully in educational assessment to compare test scores obtained from parallel or alternate forms that measure the same characteristic with equal precision. Extending the technique further, comparable scores are sometimes derived for measures of different but related characteristics. The process of establishing comparable scores generally has little effect on the magnitude of association between the measures. Comparability may not signify interchangeability unless the association between the measures approaches the reliability. Equating, the strongest form of linking, can be established only when two tests 1) measure the same content/construct, 2) target very similar populations, 3) are administered under similar conditions such that the constructs measured are not differentially affected, 4) share common measurement goals and 5) are equally reliable. When test forms are created to be similar in content and difficulty, equating adjusts for differences in difficulty. Test forms are considered to be essentially the same, so scores on the two forms can be used interchangeably after equating has adjusted for differences in difficulty. For tests with lesser degrees of similarity, only weaker forms of linking are meaningful, such as calibration, concordance, projection, or moderation.

2. The PRO Rosetta Stone Project

The primary aim of the PRO Rosetta Stone (PROsetta Stone®) project (1RC4CA157236-01, PI: David Cella) is to develop and apply methods to link the Patient-Reported Outcomes Measurement Information System (PROMIS) measures with other related “legacy” instruments to expand the range of PRO assessment options within a common, standardized metric. The project identifies and applies appropriate linking methods that allow scores on a range of PRO instruments to be expressed as standardized T-score metrics linked to the PROMIS. This preliminary report encompasses the first wave of 20 linking studies based on available PRO data from PROMIS (aka, PROMIS Wave I), Toolbox, and Neuro-QOL.

2.1. Patient-Reported Outcomes Measurement Information System (PROMIS)

In 2004, the NIH initiated the PROMIS[1] cooperative group under the NIH Roadmap[2] effort to re-engineer the clinical research enterprise. The aim of PROMIS is to revolutionize and standardize how PRO tools are selected and employed in clinical research. To accomplish this, a publicly-available system was developed to allow clinical researchers access to a common repository of items and state-of-the-science computer-based methods to administer the PROMIS measures. The PROMIS measures include item banks across a wide range of domains that comprise physical, mental, and social health for adults and children, with 12-124 items per bank. Initial concepts measured include emotional distress (anger, anxiety, and depression), physical function, fatigue, pain (quality, behavior, and interference), social function, sleep disturbance, and sleep-related impairment. The banks can be used to administer computerized adaptive tests (CAT) or fixed-length forms in these domains. We have also developed 4 to 20-item short forms for each domain, and a 10-item Global Health Scale that includes global ratings of five broad PROMIS domains and general health perceptions. As described in a full issue of *Medical Care* (Cella et al., 2007), the PROMIS items, banks, and short forms were developed using a standardized, rigorous methodology that began with constructing a consensus-based PROMIS domain framework.

All PROMIS banks have been calibrated according to Samejima's (1969) graded response model (based on large data collections including both general and clinical samples) and re-scaled (mean=50 and SD=10) using scale-setting subsamples matching the marginal distributions of gender, age, race, and education in the 2000 US census. The PROMIS Wave I calibration data included a small number of full-bank testing cases (approximately 1,000 per bank) from a general population taking one full bank and a larger number of block-administration cases (n= ~14,000) from both general and clinical populations taking a collection of blocks representing all banks with 7 items each. The full-bank testing samples were randomly assigned to one of 7 different forms. Each form was composed of one or more PROMIS domains (with an exception of Physical Function where the bank was split over two forms) and one or more legacy measures of the same or related domains.

The PROMIS Wave I data collection design included a number of widely accepted "legacy" measures. The legacy measures used for validation evidence included Buss-Perry Aggression Questionnaire (BPAQ), Center for Epidemiological Studies Depression Scale (CES-D), Mood and Anxiety Symptom Questionnaire (MASQ), Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F), Brief Pain Inventory (BPI), and SF-36. In addition to the pairs for validity (e.g., PROMIS Depression and CES-D), the PROMIS Wave I data allows for the potential for linking over a dozen pairs of measures/subscales. Furthermore, included within each of the PROMIS banks were items from many other existing measures. Depending on the nature and strength of relationship between the measures, various linking procedures can be used to allow for cross-walking of scores.

[1] www.nihpromis.org
[2] www.nihroadmap.nih.gov

2.2. The NIH Toolbox for Assessment of Neurological and Behavioral Function (Toolbox)

Developed in 2006 with the NIH Blueprint funding for Neuroscience Research, four domains of assessment central to neurological and behavioral function were created to measure cognition, sensation, motor functioning, and emotional health. The NIH Toolbox for Assessment of Neurological and Behavioral Function[3] provides investigators with a brief, yet comprehensive measurement tool for assessment of cognitive function, emotional health, sensory and motor function. It provides an innovative approach to measurement that is responsive to the needs of researchers in a variety of settings, with a particular emphasis on measuring outcomes in clinical trials and functional status in large cohort studies, e.g. epidemiological studies and longitudinal studies. Included as subdomains of emotional health were negative affect, psychological well-being, stress and self-efficacy, and social relationships. Three PROMIS emotional distress item banks (Anger, Anxiety, and Depression) were used as measures of negative affect. Additionally, existing "legacy" measures, e.g., Patient Health Questionnaire (PHQ-9) and Center for Epidemiological Studies Depression Scale (CES-D), were flagged as potential candidates for the Toolbox battery because of their history, visibility, and research legacy. Among these legacy measures, we focused on those that were available without proprietary restrictions for research applications. In most cases, these measures had been developed using classical test theory.

2.3. Quality of Life Outcomes in Neurological Disorders (Neuro-QOL)

The National Institute of Neurological Disorders and Stroke sponsored a multi-site project to develop a clinically relevant and psychometrically robust Quality of Life (QOL) assessment tool for adults and children with neurological disorders. The primary goal of this effort, known as Neuro-QOL[3], was to enable clinical researchers to compare the QOL impact of different interventions within and across various conditions. This resulted in 13 adult QOL item banks (Anxiety, Depression, Fatigue, Upper Extremity Function - Fine Motor, Lower Extremity Function - Mobility, Applied Cognition - General Concerns, Applied Cognition - Executive Function, Emotional and Behavioral Dyscontrol, Positive Affect and Well-Being, Sleep Disturbance, Ability to Participate in Social Roles and Activities, Satisfaction with Social Roles and Activities, and Stigma).

3. Legacy Instruments

The following instruments are widely accepted "legacy" measures that have been used as part of the initial validation work for PROMIS and Toolbox. Data were collected on a minimum of 500

[3] www.nihtoolbox.org

respondents (for stable item parameter estimation) along with at least one other conceptually similar scale or bank.

3.1. Mood and Anxiety Symptom Questionnaire (MASQ)

The Mood and Anxiety Symptom Questionnaire (MASQ) is a 77-item self-report questionnaire that assesses depressive, anxious, and mixed symptomatology. Three scales measure General Distress: depressive symptoms (12 items), anxious symptoms (11 items), and mixed symptoms (15 items). There are also anxiety-specific (Anxious Arousal, 17 items) and depression-specific scales (Anhedonic Depression, 22 items). Higher scores reflect greater levels of symptomatology. (Watson et al., 1995). For the current analysis, we used the Anxious Symptoms scale.

3.2. SF-36

The SF-36 is a multi-purpose, short-form health survey with 36 items. It yields an 8-scale profile of functional health and well-being scores as well as psychometrically-based physical and mental health summary scores and a preference-based health utility index. The SF-36 version 2 (Ware, Kosinski, & Dewey, 2000.) consists of items assessing physical functioning (PF; 10 items), social functioning (SF; 2 items), role limitation due to physical health (RP; 4 items), bodily pain (BP; 2 items), mental health (MH; 5 items), role limitations due to emotional health (RE; 3 items), vitality (VT; 4 items), general health perceptions (GH; 5 items), and reported health transition (1 item). The Physical Component Score (PCS) and Mental Component Score (MCS) range from 0-100 with higher scores indicating better health-related quality of life.

3.3. Center for Epidemiological Studies Depression Scale (CES-D)

The Center for Epidemiological Studies Depression Scale (CES-D) is a 20-item measure designed to assess depressive symptoms in the general population. Items are rated for the past week using a four-point scale for duration (from “rarely or none of the time” to “most or all of the time”). The CES-D has good psychometric properties and has been used in a variety of contexts, including community samples and clinical samples with both medical and psychiatric illnesses (Radloff, 1977).

3.4. Buss-Perry Aggression Questionnaire (BPAQ)

The Buss-Perry Aggression Questionnaire (BPAQ) is a 29-item self-report measure that includes four subscales: physical aggression (9 items), verbal aggression (5 items), anger (7 items), and hostility (8 items) (Buss & Perry, 1992). There is no time frame specified, and items are rated using a seven-point scale from “extremely uncharacteristic” to “extremely characteristic”.

3.5. Health Assessment Questionnaire (HAQ)

The Health Assessment Questionnaire (HAQ) was developed as a comprehensive measure of outcomes in patients with a wide variety of rheumatic diseases (Fries, Spitz, Kraines, & Holman, 1980). It should be considered a generic rather than a disease-specific instrument. The HAQ has been administered primarily in one of two versions, short HAQ-DI (Disability Index) or the Full HAQ. The HAQ-DI assesses the extent of a patient’s functional ability. It is composed of 20

items in 8 categories (Dressing and Grooming, Hygiene, Arising, Reach, Eating, Grip, Walking, Common Daily Activities).

3.6. Functional Assessment of Chronic Illness Therapy (FACIT)

The Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System is a collection of QOL questionnaires targeted to the management of chronic illness including cancer. The FACT-G (now in Version 4) is a 27-item compilation of general questions divided into four subscales: Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, and Functional Well-Being. It is considered appropriate for use with patients with any form of cancer, and has also been used and validated in other chronic illness conditions (e.g., HIV/AIDS, multiple sclerosis) and in the general population (using a slightly modified version). Validation of a core measure allowed for the evolution of multiple disease, treatment, condition, and non-cancer-specific subscales. FACIT subscales are constructed to complement the FACT-G, addressing relevant disease-, treatment-, or condition-related issues not already covered in the general questionnaire. Each is intended to be as specific as necessary to capture the clinically-relevant problems associated with a given condition or symptom, yet general enough to allow for comparison across diseases, and extension, as appropriate, to other chronic medical conditions. For the current analysis, we used the Fatigue scale. The Functional Assessment of Chronic Illness Therapy-Fatigue Scale (FACIT-Fatigue scale) is a 13-item questionnaire that assesses self-reported fatigue and its impact upon daily activities and function (Yellen, Cella, Webster, Blendowsky, & Kaplan, 1997). It was developed to meet a growing demand for the precise evaluation of fatigue associated with anemia in cancer patients. Subsequently, it has been employed in over 70 published studies including over 20,000 people, including cancer patients receiving chemotherapy (Berndt et al., 2005; Quirt et al., 2001), cancer patients not receiving chemotherapy (Quirt et al., 2001; Quirt et al., 2002), long term cancer survivors (Ng et al., 2005), childhood cancer survivors (Mulrooney et al., 2008), rheumatoid arthritis (Cella et al., 2005; Mease et al., 2008; Mittendorf et al., 2007), psoriatic arthritis (Chandran, Bhella, Schentag & Gladman, 2007), paroxysmal nocturnal hemoglobinuria (Brodsky et al., 2008), and Parkinson's disease (Hagell et al., 2006). It has also been validated in the general United States population (Brucker, Yost, Cashy, Webster & Cella., 2005; Cella, Lai, Chang, Peterman & Slavin, 2002). In all cases, the FACIT-Fatigue scale has been found to be reliable and valid.

3.7. BPI Severity and Interference

The Brief Pain Inventory (BPI) (Cleeland & Ryan, 1994) produces pain severity and pain interference scores ranging from 0 to 10 and higher scores indicate worse pain. There is a short and a long form. There are 15 questions on the Short Form BPI (9 questions, with the last question containing 7 parts). In PROMIS calibration testing, 11 of the 15 questions were administered (BPI items 1, 2, 7, and 8 were omitted). However, for some BPI items, PROMIS calibration testing used a one week recall period. This matches the recall period used by the BPI long form, but not the 24-hour recall period used in the BPI short form.

3.8. Generalized Anxiety Disorder Scale (GAD-7)

The Generalized Anxiety Disorder Scale (GAD-7) is a 7-item instrument developed with primary care patients and the goal of identifying probable cases of GAD (Spitzer, Kroenke, Williams, &

Löwe, 2006). Items are rated for the last two weeks, using a four-point scale for duration (from "not at all" to "nearly every day").

3.9. Kessler 6 Mental Health Scale (K6)

The Kessler 6 Mental Health Scale (K6) (Kessler et. al., 2003) is a measure of non-specific psychological distress. The K6 is a tool used for screening mental health issues in a general adult population. The scale was designed to be sensitive around the threshold for the clinically significant range of the distribution of non-specific distress in an effort to maximize the ability to discriminate cases of serious mental illness from the rest.

3.10. Patient Health Questionnaire (PHQ-9)

The Patient Health Questionnaire (PHQ-9) is a nine-item instrument designed for use in primary care settings (Kroenke, Spitzer & Williams., 2001). It is based directly on the diagnostic criteria for major depressive disorder in the Diagnostic and Statistical Manual, Fourth Edition (American Psychiatric Association, 2000). Items are rated for the last two weeks, using a four-point scale for duration (from "not at all" to "nearly every day"). The PHQ-9 has been adopted widely as a screening and diagnostic tool as well as a measure for monitoring treatment.

4. Linking Methods

PROMIS full-bank administration allows for single group linking. This linking method is used when two or more measures are administered to the same group of people. For example, two PROMIS banks (Anxiety and Depression) and three legacy measures (MASQ, CES-D, and SF-36/MH) were administered to a sample of 925 people. The order of measures was randomized so as to minimize potential order effects. The original purpose of the full-bank administration study was to establish initial validity evidence (e.g., validity coefficients), not to establish linking relationships. Some of the measures revealed severely skewed score distributions in the full-bank administration sample and the sample size was relatively small, which might be limiting factors when it comes to determining the linking method. Another potential issue is related to how the non-PROMIS measures are scored and reported. For example, all SF-36 subscales are scored using a proprietary scoring algorithm and reported as normed scores (0 to 100). Others are scored and reported using simple raw summed scores. All PROMIS measures are scored using the final re-centered item response theory (IRT) item parameters and transformed to the T-score metric (mean=50, SD=10).

PROMIS's T-score distributions are standardized such that a score of 50 represents the average (mean) for the US general population, and the standard deviation around that mean is 10 points. A high PROMIS score always represents more of the concept being measured. Thus, for example, a person who has a T-score of 60 is one standard deviation higher than the general population for the concept being measured. For symptoms and other negatively-worded concepts like pain, fatigue, and anxiety, a score of 60 is one standard deviation worse than

average; for functional scores and other positively-worded concepts like physical or social function, a score of 60 is one standard deviation better than average, etc.

In order to apply the linking methods consistently across different studies, linking/concordance relationships will be established based on the raw summed score metric of the measures. Furthermore, the direction of linking relationships to be established will be from legacy to PROMIS. That is, each raw summed score on a given legacy instrument will be mapped to a T-score of the corresponding PROMIS instrument/bank. Finally, the raw summed score for each legacy instrument was constructed such that higher scores represent higher levels of the construct being measured. When the measures were scaled in the opposite direction, we reversed the direction of the legacy measure in order for the correlation between the measures to be positive and to facilitate concurrent calibration. As a result, some or all item response scores for some legacy instruments will need to be reverse-coded.

4.1. IRT Linking

One of the objectives of the current linking analysis is to determine whether or not the non-PROMIS measures can be added to their respective PROMIS item bank without significantly altering the underlying trait being measured. The rationale is twofold: (1) the augmented PROMIS item banks might provide more robust coverage both in terms of content and difficulty; and (2) calibrating the non-PROMIS measures on the corresponding PROMIS item bank scale might facilitate subsequent linking analyses. At least, two IRT linking approaches are applicable under the current study design; (1) linking separate calibrations through the Stocking-Lord method and (2) fixed parameter calibration.

Linking separate calibrations might involve the following steps under the current setting.

- First, simultaneously calibrate the combined item set (e.g., PROMIS Depression bank and CES-D).
- Second, estimate linear transformation coefficients (additive and multiplicative constants) using the item parameters for the PROMIS bank items as anchor items.
- Third, transform the metric for the non-PROMIS items to the PROMIS metric.

The second approach, fixed parameter calibration, involves fixing the PROMIS item parameters at their final bank values and calibrating only non-PROMIS items so that the non-PROMIS item parameters may be placed on the same metric as the PROMIS items. The focus is on placing the parameters of non-PROMIS items on the PROMIS scale. Updating the PROMIS item parameters is not desired because the linking exercise is built on the stability of these calibrations. Note that IRT linking would be necessary when the ability level of the full-bank testing sample is different from that of the PROMIS scale-setting sample. If it is assumed that the two samples are from the same population, linking is not necessary and calibration of the items (either separately or simultaneously) will result in item parameter estimates that are on the same scale without any further scale linking. Even though the full-bank testing sample was a subset of the full PROMIS calibration sample, it is still possible that the two samples are somewhat disparate due to some non-random component of the selection process. Moreover,

there is some evidence that linking can improve the accuracy of parameter estimation even when linking is not necessary (e.g., two samples are from the same population having the same or similar ability levels). Thus, conducting IRT linking would be worthwhile.

Once the non-PROMIS items are calibrated on the corresponding PROMIS item bank scale, the augmented item bank can be used for standard computation of IRT scaled scores from any subset of the items, including computerized adaptive testing (CAT) and creating short forms. The non-PROMIS items will be treated the same as the existing PROMIS items. Again, the above options are feasible only when the dimensionality of the bank is not altered significantly (i.e., where a unidimensional IRT model is suitable for the aggregate set of items). Thus, prior to conducting IRT linking, it is important to assess dimensionality of the measures based on some selected combinations of PROMIS and non-PROMIS measures. Various dimensionality assessment tools can be used including a confirmatory factor analysis, disattenuated correlations, and essential unidimensionality.

4.2. Equipercentile Linking

The IRT Linking procedures described above are permissible only if the traits being measured are not significantly altered by aggregating items from multiple measures. One potential issue might be creating multidimensionality as a result of aggregating items measuring different traits. For two scales that measure distinct but highly related traits, predicting scores on one scale from those of the other has been used frequently. Concordance tables between PROMIS and non-PROMIS measures can be constructed using equipercentile equating (Lord, 1982; Kolen & Brennan, 2004) when there is insufficient empirical evidence that the instruments measure the same construct. An equipercentile method estimates a nonlinear linking relationship using percentile rank distributions of the two linking measures. The equipercentile linking method can be used in conjunction with a presmoothing method such as the loglinear model (Hanson, Zeng, & Colton, 1994). The frequency distributions are first smoothed using the loglinear model and then equipercentile linking is conducted based on the smoothed frequency distributions of the two measures. Smoothing can also be done at the backend on equipercentile equivalents and is called postsmoothing (Brennan, 2004; Kolen & Brennan, 2004). The cubic-spline smoothing algorithm (Reinsch, 1967) is used in the LEGS program (Brennan, 2004). Smoothing is intended to reduce sampling error involved in the linking process. A successful linking procedure will provide a conversion (crosswalk) table, in which, for example, raw summed scores on the PHQ-9 measure are transformed to the T-score equivalents of the PROMIS Depression measure.

Under the current context, equipercentile crosswalk tables can be generated using two different approaches. First is a direct linking approach where each raw summed score on non-PROMIS measure is mapped directly to a PROMIS T-score. That is, raw summed scores on the non-PROMIS instrument and IRT scaled scores on the PROMIS (reference) instrument are linked directly, although raw summed scores and IRT scaled score have distinct properties (e.g., discrete vs. continuous). This approach might be appropriate when the reference instrument is either an item bank or composed of a large number of items and so various subsets (static or dynamic) are likely to be used but not the full bank in its entirety (e.g., PROMIS Physical

Function bank with 124 items). Second is an indirect approach where raw summed scores on the non-PROMIS instrument are mapped to raw summed scores on the PROMIS instrument; and then the resulting raw summed score equivalents are mapped to corresponding scaled scores based on a raw-to-scale score conversion table. Because the raw summed score equivalents may take fractional values, such a conversion table will need to be interpolated using statistical procedures (e.g., cubic spline).

Finally, when samples are small or inadequate for a specific method, random sampling error becomes a major concern (Kolen & Brennan, 2004). That is, substantially different linking relationships might be obtained if linking is conducted repeatedly over different samples. The type of random sampling error can be measured by the standard error of equating (SEE), which can be operationalized as the standard deviation of equated scores for a given raw summed score over replications (Lord, 1982).

4.3. Linking Assumptions

In Section 5, we present the results of a large number of linking studies using secondary data sets. In each case, we have applied all three linking methods described in sections 4.1 and 4.2. Our purpose is to provide the maximum amount of useful information. However, the suitability of these methods depends upon the meeting of various linking assumptions. These assumptions require that the two instruments to be linked measure the same construct, show a high correlation, and are relatively invariant in subpopulation differences (Dorans, 2007). The degree to which these assumptions are met varies across linking studies. Given that different researchers may interpret these requirements differently, we have taken a liberal approach for inclusion of linkages in this book. Nevertheless, we recommend that researchers diagnostically review the classical psychometrics and CFA results in light of these assumptions prior to any application of the cross-walk charts or legacy parameters to their own data.

5. Linking Results

Table 5.1 lists the linking analyses included in this report, which have been conducted based on samples from two different studies: PROMIS and Toolbox (see Section 2 for more details). In all cases, PROMIS instruments were used as the reference (i.e., scores on non-PROMIS instruments are expressed on the PROMIS score metric); however, shorter versions of PROMIS were used in Toolbox.

Table 5.1. Linking by Study

Section	Study	PROMIS Instrument	Non-PROMIS Instrument to Link
5.1	PROMIS Wave1	Anxiety	Mood and Anxiety Symptom Questionnaire (MASQ)
5.2	PROMIS Wave1	Anxiety	SF-36 Mental Health (SF-36/MH)
5.3	PROMIS Wave1	Depression	Center for Epidemiological Studies Depression Scale (CES-D)
5.4	PROMIS Wave1	Depression	SF-36 Mental Health (SF-36/MH)
5.5	PROMIS Wave1	Anger	Buss Perry Aggression Questionnaire (BPAQ)
5.6	PROMIS Wave1	Physical Function	Health Assessment Questionnaire (HAQ-DI)
5.7	PROMIS Wave1	Physical Function	SF-36 Physical Functioning (SF-36/PF)
5.8	PROMIS Wave1	Fatigue	Functional Assessment of Chronic Illness Therapy – Fatigue Scale (FACIT-F)
5.9	PROMIS Wave1	Fatigue	SF-36 Vitality (SF-36/VT)
5.10	PROMIS Wave1	Pain Interference	Brief Pain Inventory Severity (BPI Severity)
5.11	PROMIS Wave1	Pain Interference	Brief Pain Inventory Interference (BPI Interference)
5.12	Toolbox	Anxiety	Generalized Anxiety Disorder Scale (GAD-7)
5.13	Toolbox	Anxiety	Kessler 6 Mental Health Scale (K6)
5.14	Toolbox	Anxiety	Mood and Anxiety Symptom Questionnaire (MASQ)
5.15	Toolbox	Depression	Center for Epidemiological Studies Depression Scale (CES-D)
5.16	Toolbox	Depression	Patient Health Questionnaire (PHQ-9)
5.17	Neuro-QOL	Anxiety	Neuro-QOL Anxiety
5.18	Neuro-QOL	Depression	Neuro-QOL Depression
5.19	Neuro-QOL	Physical Function	Neuro-QOL Mobility
5.20	Neuro-QOL	Physical Function	Neuro-QOL Upper Extremity