**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** 0423

**Measure Title**: Functional Status for patients with hip impairments

**Date of Submission**: Click here to enter a date

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process |
| Efficiency | Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Focus on Therapeutic Outcomes Database

**1.3. What are the dates of the data used in testing**? Click here to enter date range

2000-2013

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: patient | other: patient |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

**Table 1.5: Measured Entities by level of analysis and Data Source**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis** | **Data source (years)** | Entities tested | | | |
| Patients | Clinicians | Clinics | States |
| a. Test-retest reliability | 2010 | 25 | NR+ | 7 | NR |
| b. Internal consistency | 2002-2004 | 444 | NR+ | 81 | 20 |
| c. Structural validity | 2002-2004 | 444 | NR+ | 81 | 20 |
| d. Construct validity (sample 1) | 2008-2010 | 5160 | NR+ | 519 | 26 |
| e. Sensitivity to change (sample 1) | 2008-2010 | 5160 | NR+ | 519 | 26 |
| f. Responsiveness (sample 1) | 2008-2010 | 5160 | NR+ | 519 | 26 |
| d. Construct validity (sample 2) | 2005-2007 | 8714 | NR+ | 257 | 31 |
| e, f. Sensitivity, responsiveness (sample 2) | 2005-2007 | 8714 | NR+ | 257 | 31 |
| g. Exclusions analysis | 2011-2013 | 55244 | 8,078 | 2182 | 48 |
| h. Risk adjustment modeling | 2000-2003 | 306,556 | 3447 | 552 | 40 |
| i. Performance patient level | 2005-2007 | 8714 | NR+ | 257 | 31 |
| j. Performance individual clinician level | 2011-2013 |  |  |  |  |
|  | 2011 | 3501 | 252 | 169 | 28 |
|  | 2012 | 5235 | 375 | 229 | 33 |
|  | 2013 | 11132 | 812 | 479 | 40 |
| k. Performance clinic/group practice level | 2011-2013 |  |  |  |  |
|  | 2011 | 3770 | 668 | 92 | 26 |
|  | 2012 | 5192 | 869 | 140 | 27 |
|  | 2013 | 10644 | 1679 | 287 | 36 |
| l. Missing data | 2011-2013 |  |  |  |  |
|  | 2011 | 23732 | 3989 | 1235 | 40 |
|  | 2012 | 30267 | 4813 | 1429 | 44 |
|  | 2013 | 56679 | 8022 | 2195 | 50 |
| + NR=not reported  \*Clinicians with 10+ patients with FS measures at intake & discharge.  \*\*Clinics with 10+ patient per clinician for small clinics (up to 4 clinicians) or 40+ patients for large clinics (5 or more clinicians), with FS measures at intake & discharge | | | | | |

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

a. For reliability testing we used data from 25 patients with hip impairments (all had chronic symptoms) from 7 clinics (characteristics shown in Table 1.61a below).

**Table 1.6a: Characteristics of Patients in Reliability Tests N=25**

|  |  |
| --- | --- |
| Female | 71% |
| Intake FS | 39 |
| Age (mean) | 76 yr |
| Surgery | 45% |
| Comorbities  None  1  2 or 3  4 or more | 33%  33%  23%  11% |
| Elevated fear | 67% |
| Payer  MedicareB  Patient  HMO  PPO  WC  Other | 62%  0%  12%  21%  5%  0% |

b. Internal consistency (person reliability) testing

c. Structural validity testing

Analyses b and c used data from 444 patients who were receiving outpatient rehabilitation or for a variety of hip impairments at 81 clinics in 20 states were analyzed seeing an orthopedic surgeon. (Hart & Mioduski 2005) Characteristics of these patients are shown in Table 1.6b.

**Table 1.6b: Characteristics of Patients in Structural Validity Analysis N=444**

|  |  |
| --- | --- |
| Age (mean ± SD, min, max in years) | 54 ± 17, 15, 88 |
| Age category |  |
| Age 14 to <45 (%) | 29 |
| Age 45 to 65 (%) | 45 |
| Age >65 (%) | 26 |
| Gender (% female) | 70 |
| Acuity of symptoms (%) |  |
| Acute | 13 |
| Subacute | 27 |
| Chronic | 60 |
| Surgical history (%) |  |
| None | 93 |
| One or more | 7 |
| Exercise History (%) |  |
| At least 3×/week | 34 |
| 1–2×/week | 24 |
| Seldom or never | 41 |
| Payer Source (%) |  |
| Indemnity | 11 |
| Medicaid | 2 |
| Medicare | 19 |
| Patient private pay | 0 |
| HMO | 43 |
| PPO | 15 |
| Workers’ compensation | 6 |
| Other | 4 |

d. construct validity

e. sensitivity to change

f. responsiveness testing

Analyses d used two samples (sample 1 and 2). The first sample was used in an internal FOTO study. Data was from 04/2008 to 04/2010 and included 5,160 patients who completed the hip short forms at intake, and 3,582 patients who completed the short form at discharge. All patients were treated for hip impairments. Characteristics of these patients are shown in the Table 1.6c below. The second sample utilized a sample of 8714 patients with hip impairments from 2005-2007 from 257 clinics in 31 states. ([Hart, Wang et al. 2008](#_ENREF_4)) Characteristics of these patients are shown in Table 1.6d.

**Table 1.6c Characteristics of patients in analyses d-f (sample 1)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Characteristic | | | Value | | |
|  | mean | SD | | min | max |
| Intake FS  Discharge FS  FSCH  Risk-adjusted FSCH Predicted  Risk-adjusted FSCH Residual | 46  60  15  15  0 | 16  18  16  4  14 | | 3  2  -56  2  -60 | 97  100  89  30  54 |
| Age (mean±SD, min, max in years) | 60 | 17 |  | 18 | 99 |
| Age category  Age 18 to <45 (%)  Age 45 to 65 (%)  Age >65 to 75 (%)  Age >75 (%) | | | 19  37  22  22 | | |
| Gender  % female | | | 64 | | |
| Acuity of Symptoms (%)  Acute (0 to 21 days)  Subacute (22-90 days)  Chronic (>90 day)s | | | 17  26  57 | | |
| Surgical History (%)  None  One or more  Missing | | | 65  35  0 | | |
| Payer Source (%)  Indemnity Insurance  Litigation  Medicaid  Medicare A  Medicare B  Patient private pry  Health Maintenance Organization  Preferred Provider Organization  Workers’ Comp  Auto Insurance  Other | | | 3  1  1  7  28  1  10  36  3  1  11 | | |
| Number of Functional Comorbidities (%) a  None  One  Two  Three or more | | | 31  22  17  30 | | |
| Fear of Avoidance  No fear  Intimidated | | | 54  46 | | |

**Table 1.6d Characteristics of patients in analyses d-f (sample 2) N=8714**

|  |  |
| --- | --- |
| Age, mean ± SD (range) (y) | 56 ±17 (18–102) |
| Age category |  |
| 18 to <45 (%) | 24 |
| 45 to 65 (%) | 39 |
| >65 (%) | 30 |
| Missing (%) | 7 |
| Sex (%) |  |
| women | 63 |
| Missing | 7 |
| Acuity of symptoms (%) |  |
| Acute (0–21d) | 15 |
| Subacute (22–90d) | 26 |
| Chronic ( >90d) | 52 |
| Missing | 7 |
| Surgical history (%) |  |
| None | 74 |
| One | 18 |
| Two | 4 |
| Three | 1 |
| Four or more | 2 |
| Missing | 1 |
| Exercise history (%) |  |
| At least 3 times a week | 21 |
| 1 to 2 times a week | 12 |
| Seldom or never | 19 |
| Missing | 48 |
| Payer source (%) |  |
| Indemnity | 2 |
| Litigation | \_1 |
| Medicaid | 2 |
| Medicare part B | 14 |
| Patient private pay | \_1 |
| HMO | 12 |
| PPO | 12 |
| Workers’ compensation | 3 |
| Other | 4 |
| Missing | 51 |
| Number of functional comorbidities (% |  |
| None | 21 |
| One | 22 |
| Three or more | 35 |
| Missing | 3 |

g. Exclusions analysis

Analysis used data from 2011-2013 combined from 55244 patients, 8078 clinicians, 2182 clinics and 48 states.

h. Risk adjustment modeling utilized data from 2000-2003, 306,556 patients, 3447 clinicians from 552 clinics in 40 states. ([Hart and Connolly 2006](#_ENREF_2))

i. Performance patient level

This analysis utilized the same sample described as sample 2 for analyses d-f and shown in (Table 1.6d).

j. Performance individual clinician level

Analysis j-was conducted using data drawn from the FOTO database of all patients 14 years and older with hip impairments in 2011-2013 in a FOTO internal study. Performance analyses utilized data from only those clinicians that had a minimum of 10 patients in the prior 12 months. In 2011 there were 3501 patients, 252 clinicians from 169 clinics and 28 states. In 2012 there were 5235 patients, 375 clinicians from 229 clinics in 33 states. In 2013 there were 11132 patients, 812 clinicians from 479 clinics in 40 states.

k. Performance clinic/group practice level

Analysis k-was conducted using data drawn from the FOTO database of all patients 14 years and older with Hip impairments in 2011-2013 in a FOTO internal study. Performance analyses utilized data from only those clinics that had a minimum of 40 patients in the prior 12 months for large clinics (5 clinicians or more) or a minimum of 10 patients per clinician for clinics with 4 or less clinicians. In 2011 there were 3770 patients, 668 clinicians from 92 clinics and 26 states. In 2012 there were 5192 patients, 869 clinicians from 140 clinics in 27 states. In 2013 there were 10644 patients, 1679 clinicians from 287 clinics in 36 states

l. Missing data analyses were conducted using data from all patients 14 years and older with hip impairments in 2011-2013. There were 23,732; 30,267 and 56,679 patients in years 2011, 2012 and 2013 respectively. Table 2b7.1 shows the characteristics of patients in each of these years. Data came from 1235 clinics from 40 states in 2011; 1429 clinics from 44 states in 2012, and 2195 clinics from 50 states in 2013.

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

Differences in the samples used for testing are explained in 1.5 and 1.6 above.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

a. Test-retest reliability: We assessed test-retest reliability of the functional status (FS) measures estimated using the paper and pencil survey process because we were interested in knowing if patients with hip impairments could respond to the FS items in a consistent manner. FS measures were assessed for their agreement using an Intraclass Correlational Coefficient (ICC(2,1) with 95% confidence intervals.

b. Internal consistency reliability: Internal consistency was evaluated using Cronbach’s alpha.

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

a. Test-retest reliability: The ICC(2,1) value with 95% confidence intervals was .96 (.90, .98) which is considered strong reliability, which supports using of the measure in clinical practice. Test-retest reliability of patients answering the Hip CAT has not been tested, but is planned for future study.

b. Internal consistency of the Hip CAT item bank was strong (Cronbach’s alpha=.96). ([Hart, Mioduski et al. 2005](#_ENREF_2))

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

c. Structural validity analyses: Data were analyzed to determine how well unidimensionality and local independence IRT assumptions were met. Presence of a dominant factor was assessed with exploratory factor analyses (EFA) of latent trait variables followed by confirmatory factor analyses (CFA).Eigenvalue analyses were conducted, and results were evaluated with scree plots. Model fit was evaluated using comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root-mean-square error of approximation (RMSEA). ([Hart, Mioduski et al. 2005](#_ENREF_2))

d. construct validity

e. sensitivity to change

f. responsiveness testing

Analyses d-f (sample 1) were conducted in a FOTO internal study which involved a secondary analysis of prospectively collected data from patients who completed the hip FS paper and pencil short form survey. Both intake and discharge data were used. Descriptive statistics were used to characterize the data. Functional status change was risk-adjusted by using a linear regression model where FS change was the dependent variable, intake FS was the covariate while controlling for the risk-adjustment variables of age, gender, symptom acuity, surgical history, number of functional comorbid conditions, payer and level of fear-avoidance beliefs of physical activities

Analyses d-e (sample 2) utilized a secondary analysis of data from patients who completed the FOTO Hip CAT. ([Hart, Wang et al. 2008](#_ENREF_4))We used known group construct validity methods to assess the ability of the CAT generated FS measures to discriminate groups of patients. The independent variables assessed included intake FS, age, symptom acuity, surgical history, condition complexity and prior exercise history.

d. Known group construct validity was assessed similarly in both testing samples. We used one-way ANCOVAs with functional status change as the dependent variable, intake FS as the covariate, with one ANCOVA for each risk-adjustment variable as the independent variable. Post hoc Sheffe analyses were run for significant main factors of the independent variable.

e. Sensitivity to change was assessed similarly in both testing samples. We used two distribution-based approaches. First, effect size statistics were estimated as follows: (discharge FS minus intake FS)/(intake FS standard deviation). Second, we assessed minimal detectable change (MDC), which is defined as change greater than measurement error. MDCs were calculated by calculating average measurement error (standard error or SE) at 10 levels of intake functional status, which represent conditional standard errors of measurement (CSEM). To calculate each CSEM, we estimated the average SE associated within each of the 10 scale ranges using the intake data and multiplied the average SE per scale range by 1.96 times the square root of 2. The proportion of patients with change scores greater than the MDC at the upper 95% confidence interval was reported.

f. Responsiveness was assessed similarly in both testing samples, using an anchor-based approach by calculating the proportion of patients whose FS change was greater than minimal clinically important improvements (MCII), which is change considered important to the patient.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

c. Structural validity testing:

Fit statistics were CFI =0.94, TLI =0.98 and, RMSEA = 0.21for the one-factor solution, demonstrating a unidimensional item pool with acceptable local independence. ([Hart, Mioduski et al. 2005](#_ENREF_2))

d. Known Group Construct validity

Results from the FOTO hip paper and pencil short form analysis (sample 1) and the Hip CAT analysis (sample 2) supported known group construct validity of the FS measures estimated. Briefly, in sample 1 patients who were younger, male, had acute symptoms, had no comorbid conditions, had low fear-avoidance of physical activities, did not have surgery related to the hip, or received benefits from patient private pay or health maintenance organizations reported better functional status change compared to patients who were older, female, had chronic symptoms, had four or more comorbid conditions, had elevated fear-avoidance of physical activities, had surgery, or received benefits from Medicaid, Medicare Part B, or Workers' Compensation. In sample 2, patients who were older, had more chronic symptoms, had more surgeries, had more comorbidities, and did not exercise prior to receiving rehabilitation, reported worse (i.e., lower) discharge FS compared to other patients in each independent variable after controlling for intake FS. ([Hart, Wang et al. 2008](#_ENREF_4))

e-f Sensitivity to change and Responsiveness

Results from analyses of samples 1 and 2 support the FS measures was sensitive to change and responsive. In sample 1 (paper and pencil measure), 65% of patients attained FS change scores equal to or greater than minimal detectable change (MDC) at the 95% confidence interval. 69% of patients attained FS change scores equal to or greater than minimal clinically important improvement (MCII).

In sample 2 (CAT measure) 61% of patients obtained discharge FS measures ≥ minimal detectable change (95% confidence interval) and 64% had change scores equal to or greater than the MCII. ([Hart, Wang et al. 2008](#_ENREF_4))

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Results supported known group structural validity, known groups construct validity, sensitivity to change and responsiveness of the FOTO (hip) PROM. .

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

g. Age exclusion: The FS measures were designed for adult patients aged 18 years or older. ([Hart and Connolly 2006](#_ENREF_1); Hart Wang 2008 ) FOTO has broadened its inclusion to include patients aged 14-18. We believe that this exclusion has face validity, given the maturity level of older teenagers. FOTO explored the impact of broadening its inclusion criteria by comparing values and significance of beta coefficients and predictive ability of risk adjustment models for data from patients 14 and above, and for patients 18 and above.

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**Table 2b3.2a Sensitivity analysis of exclusion criteria for age: Comparison of beta coefficients and R2 of risk adjustment models predicting FS at discharge (Data combined 2011-2013)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Sample includes patients 14+** | | **Sample includes patients 18+** | |
| **N=55244** | | **N=53104** | |
| **Beta, CI** | **P** | **Beta, CI** | **P** |
| **Constant** | 56.7 ,55.73 to 57.57 | P<0.001 | 55.3, 54.27 to 56.24 | P<0.001 |
| **Intake FS** | 0.5 ,0.53 to 0.55 | P<0.001 | 0.5, 0.54 to 0.55 | P<0.001 |
| **Age (continuous)** | (-)0.2 ,(-)0.17 to (-)0.15 | P<0.001 | (-)0.1, (-)0.16 to (-)0.14 | P<0.001 |
| **Acuity:** |  |  |  |  |
| 0-7 days (REF) |  |  |  |  |
| 8-14 days | (-)1.3 ,(-)2.07 to (-)0.48 | 0.002 | (-)1.1, (-)1.94 to (-)0.30 | 0.007 |
| 15-21 days | (-)2.2 ,(-)2.89 to (-)1.43 | P<0.001 | (-)2.1, (-)2.83 to (-)1.32 | P<0.001 |
| 22-90 days | (-)4.0 ,(-)4.65 to (-)3.34 | P<0.001 | (-)3.7, (-)4.40 to (-)3.05 | P<0.001 |
| 91 days to 6 months | (-)5.7 ,(-)6.36 to (-)5.00 | P<0.001 | (-)5.3, (-)6.03 to (-)4.63 | P<0.001 |
| Over 6 months | (-)6.9 ,(-)7.53 to (-)6.25 | P<0.001 | (-)6.5, (-)7.21 to (-)5.89 | P<0.001 |
| **Gender:** |  |  |  |  |
| Male (REF) |  |  |  |  |
| Female | (-)1.6 ,(-)1.80 to (-)1.32 | P<0.001 | (-)1.5, (-)1.73 to (-)1.24 | P<0.001 |
| **Payer:** |  |  |  |  |
| Indemnity Insurance | 0.7 ,(-)0.25 to 1.73 | 0.142 | 0.6, (-)0.44 to 1.56 | 0.274 |
| Litigation | (-)2.1 ,(-)6.12 to 1.89 | 0.300 | (-)2.3, (-)6.35 to 1.72 | 0.261 |
| Medicaid | (-)4.0 ,(-)4.87 to (-)3.23 | P<0.001 | (-)4.1, (-)4.93 to (-)3.20 | P<0.001 |
| Medicare A | (-)1.6 ,(-)2.28 to (-)1.00 | P<0.001 | (-)1.8, (-)2.39 to (-)1.11 | P<0.001 |
| Medicare B | (-)1.5 ,(-)1.81 to (-)1.16 | P<0.001 | (-)1.6, (-)1.95 to (-)1.29 | P<0.001 |
| Patient | (-)0.3 ,(-)1.58 to 1.04 | 0.686 | (-)0.4, (-)1.76 to 0.95 | 0.557 |
| HMO | (-)0.6 ,(-)1.02 to (-)0.22 | 0.002 | (-)0.7, (-)1.10 to (-)0.28 | 0.001 |
| Preferred Provider (REF) |  |  |  |  |
| Workers Comp | (-)5.1 ,(-)5.95 to (-)4.28 | P<0.001 | (-)4.9, (-)5.75 to (-)4.08 | P<0.001 |
| No Fault | (-)2.8 ,(-)7.08 to 1.56 | 0.210 | (-)2.6, (-)6.93 to 1.67 | 0.23 |
| Other | (-)1.0 ,(-)1.48 to (-)0.59 | P<0.001 | (-)1.0, (-)1.49 to (-)0.58 | P<0.001 |
| Early Intervention | (-)7.4 ,(-)33.61 to 18.86 | 0.582 | (-)7.5, (-)33.65 to 18.59 | 0.572 |
| School | 3.3 ,(-)5.00 to 11.61 | 0.435 | (-)0.9, (-)10.79 to 8.96 | 0.856 |
| No Charge | (-)2.7 ,(-)5.70 to 0.37 | 0.085 | (-)2.3, (-)5.31 to 0.81 | 0.15 |
| Auto Insurance | (-)2.4 ,(-)4.51 to (-)0.22 | 0.030 | (-)2.2, (-)4.34 to (-)0.04 | 0.046 |
| Medicare C | (-)1.7 ,(-)3.78 to 0.44 | 0.121 | (-)1.8, (-)3.88 to 0.32 | 0.097 |
| **Surgical history:** |  |  |  |  |
| No surgery (REF) |  |  |  |  |
| 1 or more surgeries | (-)0.7 ,(-)0.96 to (-)0.47 | P<0.001 | (-)0.7, (-)0.94 to (-)0.44 | P<0.001 |
| **Fear Avoidance-Physical:** |  |  |  |  |
| Not Elevated (REF) |  |  |  |  |
| Elevated | (-)0.9 ,(-)1.13 to (-)0.67 | P<0.001 | (-)0.9, (-)1.16 to (-)0.70 | P<0.001 |
| **Number of comorbidities:** |  |  |  |  |
| None (REF) |  |  |  |  |
| One | (-)1.3 ,(-)1.73 to (-)0.81 | P<0.001 | (-)1.2, (-)1.67 to (-)0.70 | P<0.001 |
| Two | (-)2.2 ,(-)2.68 to (-)1.81 | P<0.001 | (-)2.2, (-)2.66 to (-)1.76 | P<0.001 |
| Three or more | (-)4.4 ,(-)4.76 to (-)4.02 | P<0.001 | (-)4.3, (-)4.72 to (-)3.95 | P<0.001 |
| **Model R2** | **0.370** | | **0.354** | |

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Point estimates for risk adjustment variables were essentially unchanged. Comparison of model R-squared show little difference in predictive ability between the 2 models. The R-square of the model containing patients 14-<18 is very slightly higher than that of the model with exclusions (0.370 vs. 0.354), indicating improved model fit.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** 8 **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.   
 NA

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

j. Statistical risk adjustment: The FS measure and model for risk-adjusting the FS measure has not changed since the measure received time-limited endorsement (2008). The risk-adjustment model remains a multiple linear regression model with FS change as the dependent variable, intake FS as the covariate, and including the following risk-adjustment variables: age, symptom acuity, surgery, gender, payer, number of comorbid conditions and level of fear-avoidance beliefs of physical activities.

FOTO has been risk-adjusting their functional status outcomes measures for many years*,* (see ([Jette and Jette 1996](#_ENREF_9); [Jette and Jette 1996](#_ENREF_8); [Resnik and Hart 2003](#_ENREF_12); [Hart and Connolly 2006](#_ENREF_1); [Resnik, Liu et al. 2008](#_ENREF_13); [Resnik, Liu et al. 2008](#_ENREF_14); [Hart, Werneke et al. 2011](#_ENREF_5); [Hart, Werneke et al. 2011](#_ENREF_6); [Resnik, Gozalo et al. 2011](#_ENREF_11)) Each set of analyses supported the basic structure of the model.

Since the FOTO (hip) PROM estimated using the paper and pencil short forms are mathematically equated to the FOTO (hip) PROM from the computerized adaptive testing (CAT) application, we have not studied the risk-adjustment process for patients answering the paper and pencil short forms specifically. However, comparing the results of usability internal study on the paper and pencil short forms performed specifically for this submission to what we have published using the CATs, ([Hart and Connolly 2006](#_ENREF_1)) the results are similar, i.e., the beta coefficients for the risk-adjustment variables are similar between the two modes of application, i.e., paper and pencil vs. CAT.

The initial model for our risk-adjustment processes was first published following the CMS funded Pay-for-Performance grant. ([Hart and Connolly 2006](#_ENREF_1)) Between January 2000 and August 2003, data from 306,556 patients were entered in the FOTO data set from 552 participating outpatient physical and occupational therapy clinics in 40 states. These patients were treated by 3,447 therapists (74% physical therapists, 12% occupational therapists, and 14% other types of healthcare workers). Patient self-reported intake and therapist-reported discharge data were entered for each of the patients. Of these patients, 196,336 (64%) had complete episodes, i.e., patient self-reported intake and discharge data plus therapist-reported discharge data entered into the database. The sample was also cleaned by deleting patients who had staff entered data on number of visits and duration of treatment episode that appeared illogical, for example, duration >400 days, and probably represent data entry errors. This left a sample of 189,088 (62%) with clean data for analyses (Table 1 in the CMS Report). The 75 most common diagnostic ICD-9-CM codes, which represent 64% of the patients, are primarily common orthopedic conditions (Table 2 in the CMS Report). 3,025 patients (1.6%) of the sample had selected neurological impairments (e.g., cerebrovascular accidents and traumatic brain injuries), these patients represent a subsample of 3,025 patients.

The model is reassessed periodically using current data. Beta coefficients are adjusted as needed given the new data. However, since the samples are so large, the 95% CIs for each beta coefficient remain stable and require only minor adjustments.

**2b4.4. What were the statistical results of the analyses used to select risk factors?**When the results of this regression modeling were first described, ([Hart and Connolly 2006](#_ENREF_1)) our primary interest was the predictive validity of the model. To assess the predictive validity of our regression model, we randomly separated 189,000 patients into two equal samples, developmental and testing, developed the beta coefficients of the risk-adjustment variables using the developmental sample, and tested the strength of the prediction model using the testing sample. The R2 values for the developmental and testing samples were .36 and .35, respectively, which supports the models adequately controlled the variance of the data in both samples. The 95% CIs of the beta coefficients of all risk-adjustment variables estimated using both the developmental and testing samples were similar (P>.05), which supported the predictive validity (i.e., cross-validation) of the model.

In the testing sample, the predicted discharge FS (for which we used the beta coefficients developed using the developmental sample) was very close to the actual discharge FS (i.e., the average predictive ratio was 1.045; median predictive ratio 1.025), which supported the predictive validity of the regression model, although the model slightly over predicted the average discharge FS for the second sample (Table 8 in the CMS report).

The results of the modeling of both intake and discharge FS are displayed in http://www.fotoinc.com/science-of-foto/NQF0423.html. The same website provides the instructions for use of the tables and beta coefficients.

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

As explained above, FOTO uses multiple linear regression models to risk adjust functional status outcomes measures estimated from paper and pencil and CAT applications Three-level hierarchical linear models where patient self-report measures are nested within therapists and therapist data are nested within clinics have been studied. ([Resnik, Feng et al. 2006](#_ENREF_10); [Resnik, Liu et al. 2008](#_ENREF_13); [Resnik, Liu et al. 2008](#_ENREF_14)) However, application of more advanced models is complicated for clinicians using the paper and pencil surveys and do not have proprietary businesses (e.g., FOTO) automating the data analyses and reporting. So, for the current NQF measures, we used the less complicated multiple linear regression models for the NQF public domain surveys.

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

The R –squared of the model in our testing sample for patients (N=106,568) with all types of impairments was .354 ([Hart and Connolly 2006](#_ENREF_1))Table 9). Predictive ratio of the model for patients with hip impairments was 1.045(1.029,1.062)) (Table 8).

We recently examined our risk adjustment models to determine the impact of including patients ages 14-18 in the sample. The R-squared of the current model in our sample of patients 14 and older with Hip impairments from 2011-2013 was 0.370.

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
NA

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
NA

**2b4.9. Results of Risk Stratification Analysis**:

NA

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

The risk-adjustment model has moderate ability to control for differences in patient case-mix as guaged by the R-squared value. ([Hilfiker, Bachmann et al. 2007](#_ENREF_7))

**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

FOTO is in the process of exploring additional case-mix variables to determine whether or not their addition to the risk-adjustment model will provide greater explanatory power. This process is expected to be completed in early 2015.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

i. Performance at Patient Level

Performance of the FOTO (hip) PROM at the patient level was assessed three ways:

1. Calculating the patient’s risk-adjusted residual score after modeling. First, functional status change was risk-adjusted by using a linear regression model where FS change was the dependent variable, intake FS was the covariate while controlling for the risk-adjustment variables of age, gender, symptom acuity, surgical history, number of functional comorbid conditions, payer and level of fear-avoidance beliefs of physical activities. Using these risk-adjustment models we calculated residual scores for each patient, which we interpret as the amount of FS change beyond the predicted value, given their FS score at intake, age, gender, acuity, surgical history, comorbidities, payer and fear avoidance beliefs. If the residual score is greater than zero the patient changed more than expected, and if less than zero the patient changed less than expected. These residual scores were also be used to estimate performance at the clinician or clinic level (analyses j,k).
2. Calculating the % of patients whose FS scores exceeded the MDC and/or MCII by discharge (described above under f. Responsiveness testing).
3. Using a functional staging approach ([Wang, Hart et al. 2009](#_ENREF_18)) to assist clinicians in interpreting the clinical meaning of the patients’ measured FS scores at intake and discharge.

Analyses j-k Performance at the individual clinician and clinic level

We calculated patient level residual scores (described above in 2b51.j1) and aggregated scores by individual clinician (analysis k) or clinic (analysis l). At the individual clinician level, performance was evaluated only for clinicians that had a minimum of 10 patients in the previous 12 months. At the clinic level, performance was evaluated only for large clinics (5 or more clinicians) that had a minimum of 40 patients, and small clinics (1-4 clinicians) that had a minimum of 10 patients per clinician, in the previous 12 months. To examine statistical differences between entities (individual clinicians or clinic) performance scores, we plotted each clinic’s average aggregated patient residual scores (with their 95% confidence intervals) to examine whether or not there were statistically significant differences between clinics, or between each clinic and the national average. Since the mean residual score is hypothetically centered at zero, each entity can be compared to that standard which is the predicted clinic aggregated outcome. When the 95% CI for a clinic crosses zero, the clinic’s performance for that year is determined to be no different (statistically) than the predicted national average. If 95% CIs are below or above zero, the clinic’s performance for that year is determined to be worse or better than the predicted national average, respectively.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

j. **Performance Results at the Patient Level**

1. In sample 1, the mean residual score was 0, sd 14, and range (-)60-54.
2. In sample 1 using the paper and pencil short form version of the measure the intake FS measures averaged 46 (SD 16), discharge FS scores were 60 (SD 18), and FS change scores were 15 (SD 16), which produces an effect size ((discharge minus intake)/(standard deviation of intake)) of 0.88, which is considered large. 65% of patients attained FS change scores equal to or greater than minimal detectable change (MDC) at the 95% confidence interval. 69% of patients attained FS change scores equal to or greater than minimal clinically important difference (MCID). In sample 2, using the FOTO Hip CAT, 61% of patients obtained discharge FS measures ≥ minimal detectable change (95% confidence interval) and 64% had change scores equal to or greater than the MCII. ([Hart, Wang et al. 2008](#_ENREF_4))
3. Patients were classified Into Functional Staging Levels (Table 2b5.2.a) based on their intake (Rows) and discharge (Columns) functional status scores. ([Wang, Hart et al. 2009](#_ENREF_18))

In this sample, (55%) improved to the next (i.e., higher) functional stage from intake to

discharge, and 220 (5%) dropped to a lower functional stage

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2b5.2.a Functional Staging System: Expected Performance at Each Functional Stage Level** | | | | | | |
|  | Stage I | Stage II | Stage III | Stage IV | Stage V | Stage VI |
| FS score range | 0–19 | 20–29 | 30–37 | 38–47 | 48–62 | >62 |
| Activity | Physiologic ambulator | Limited household ambulator | Independent household ambulator | Limited community ambulator | Independent community ambulator | Active community ambulator |
| Running on even ground  Walking for a mile Walking 2 blocks Walking between rooms | Unable Unable Unable Unable | Unable  Unable  Unable  A lot of dif | Unable  Unable  A lot of dif Mod dif | Unable  A lot of dif  Mod dif  A little dif | Mod dif  A little dif  A little dif  No dif | A little dif  No dif  No dif  No dif | |

j. k. Performance Results at Individual Clinician Level

Clinician aggregated residual scores with 95% CIs were classified into three groups: low performance (clinics with 95% CI of residual scores below 0), average performance (clinicians with 95% CI of residual scores crossing 0), and high performance (clinicians with 95% CI of residual scores above 0). The distribution of clinician performance category by year is shown in Table 2b52.b.

**Table 2b5.2b Distribution of Clinician Performance Categories by Year**

|  |  |  |
| --- | --- | --- |
| **Year** | **Performance level** | **N Clinicians (%)** |
| 2011 | Low performance | 26 (10.3) |
| Average performance | 201 (79.8) |
| High performance | 25 (9.9) |
| Total | 252 (100.0) |
| 2012 | Low performance | 39 (10.4) |
| Average performance | 291 (77.6) |
| High performance | 45 (12.0) |
| Total | 375 (100.0) |
| 2013 | Low performance | 61 (7.5) |
| Average performance | 693 (85.3) |
| High performance | 58 (7.1) |
| Total | 812 (100.0) |

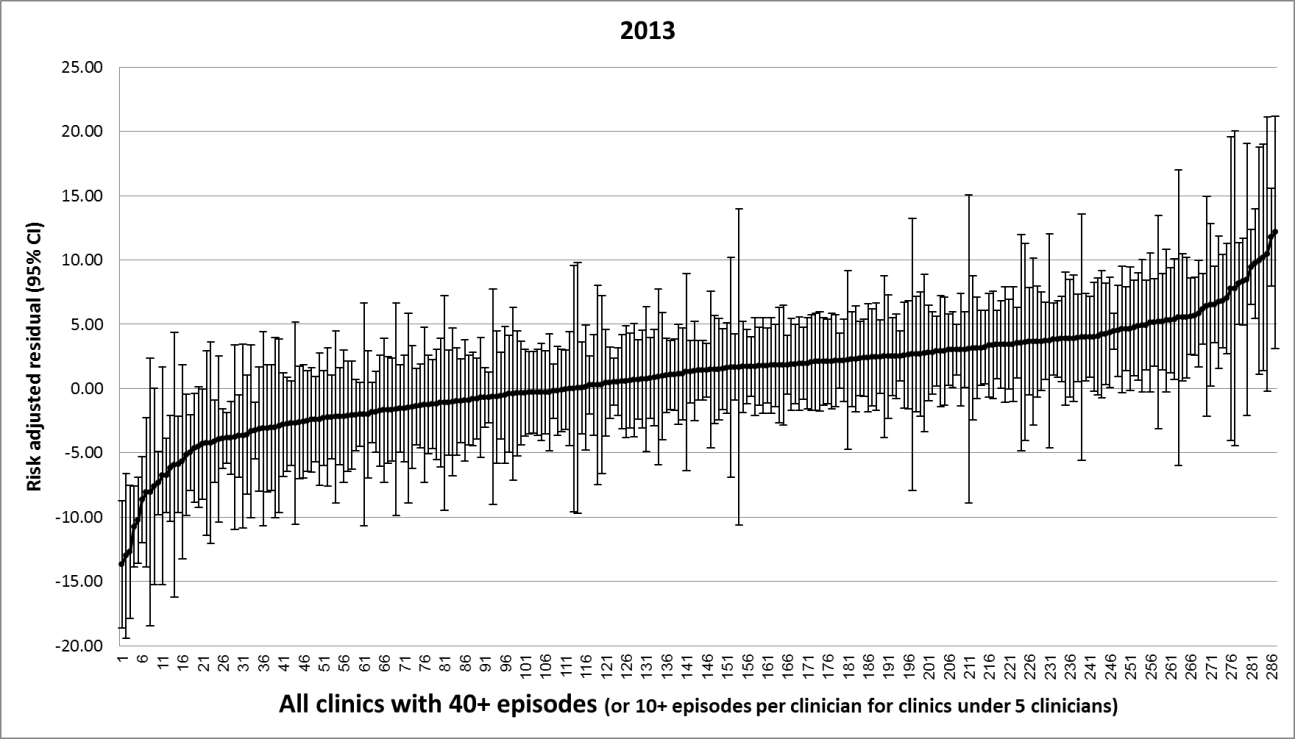
k. Performance Results at Clinic/Group Practice Level

Clinic aggregated residual scores with 95% CIs were classified into three groups: low performance (clinics with 95% CI of residual scores below 0), average performance (clinics with 95% CI of residual scores crossing 0), and high performance (clinics with 95% CI of residual scores above 0). The distribution of clinic performance category by year is shown in Table 2b52.b. The plot of clinics for 2013 is shown in Figure 2b52a.

**Table 2b5.2c Distribution of Clinic Performance Categories by Year**

|  |  |  |
| --- | --- | --- |
| **Year** | **Performance level** | **N Clinics (%)** |
| 2011 | Low performance | 5 (5.4) |
| Average performance | 65 (70.7) |
| High performance | 22 (23.9) |
| Total | 92 (100.0) |
| 2012 | Low performance | 18 (12.9) |
| Average performance | 99 (70.7) |
| High performance | 23 (16.4) |
| Total | 140 (100.0) |
| 2013 | Low performance | 21 (7.3) |
| Average performance | 216 (75.3) |
| High performance | 50 (17.4) |
| Total | 287 (100.0) |

**Figure 2b52a. Plot of aggregated residual scores with 95% CI bars in 2013**

****

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Using the data described here and in our publications ([Hart, Mioduski et al. 2005](#_ENREF_2); [Hart, Wang et al. 2008](#_ENREF_4); [Wang, Hart et al. 2009](#_ENREF_17)) we are able to demonstrate statistically significant and meaningful differences in FOTO (hip PROM) at the patient, individual clinician, and at the clinic level. These findings support use of FOTO’s Hip FS measure to derive FOTO’s risk adjusted, benchmarked effectiveness measure (PRO-PM) at the patient level, individual clinician level, and at the clinic level clinic.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

NA

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
 NA

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
NA

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
NA

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

With the free public domain measures, there are no methods in place to identify the extent or distribution of missing data or the bias associated with missing data. With the proprietary methods of data collection, there are methods available, free with both automated as well as manual methods of data collection for FOTO customers, including methods of performing non-participation audits for patients who do not collect data at intake, non-participation audits for patients who do not complete data collection at discharge, and electronic data entry quality assessment for inaccuracies and errors.

Historically, FOTO has assessed the potential impact of missing data by comparing the characteristics of patients with and without complete FS data at discharge. Of interest were specific patient characteristics known to be predictive of outcomes. The assumption was that if missing data were not missing at random, and a patient selection bias existed, patient with complete data might have characteristics of patients that on average achieve higher functional outcomes (e.g., younger, more acute conditions, more active exercise history). Small, but statistically significant differences between patients with and without complete data have been observed in several analyses.

Comparisons between patients treated for Hip impairments in 2011-2013 with and without complete FS data at discharge were made using t-tests or chi-square as appropriate (See table below).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Complete data (group 1)** | | | **Missing data (groups 2&3)** | | |  |
|  | **Year** |  | **Total N** | **% missing** |  | **Total N** | **% missing** | **P-value** |
| **Factors higher or more prevalent in patients with complete data** | | | | | | | | |
| Age (years): Mean(SD) | 2011 | **58.1 (18.2)** | 11970 | 0.0% | **54.2 (18.6)** | 5411 | 0.0% | <.001 |
| 2012 | **58.2 (18.2)** | 15600 | 0.0% | **54.0 (18.9)** | 7782 | 0.0% | <.001 |
| 2013 | **59.7 (18.0)** | 28711 | 0.0% | **54.7 (18.8** | 15192 | 0.0% | <.001 |
| Surgical history: 1 or more | 2011 | **39.1%** | 11850 | 1.0% | **30.4%** | 5338 | 1.3% | <.001 |
| 2012 | **38.0%** | 15522 | 0.5% | **28.6%** | 7749 | 0.4% | <.001 |
| 2013 | **38.0%** | 28671 | 0.1% | **26.7%** | 15162 | 0.2% | <.001 |
| Payer: Medicare B age 60 or more | 2011 | **27.8%** | 11929 | 0.3% | **19.8%** | 5408 | 0.1% | <.001 |
| 2012 | **27.7%** | 15564 | 0.2% | **20.0%** | 7778 | 0.1% | <.001 |
| 2013 | **37.6%** | 28711 | 0.0% | **26.0%** | 15192 | 0.0% | <.001 |
| **Factors higher or more prevalent in patients with missing data** | | | | | | | | |
| Intake FS: Mean(SD) | 2011 | **47.6 (15.0)** | 11970 | 0.0% | **48.7 (16.2)** | 5411 | 0.0% | <.001 |
| 2012 | **47.0 (14.8)** | 15600 | 0.0% | **48.7 (15.7)** | 7782 | 0.0% | <.001 |
| 2013 | **46.8 (14.4)** | 28711 | 0.0% | **48.6 (15.5)** | 15192 | 0.0% | <.001 |
| Payer: Medicaid | 2011 | **2.2%** | 11929 | 0.3% | **4.1%** | 5408 | 0.1% | <.001 |
| 2012 | **2.2%** | 15564 | 0.2% | **4.5%** | 7778 | 0.1% | <.001 |
| 2013 | **1.8%** | 28711 | 0.0% | **3.9%** | 15192 | 0.0% | <.001 |
| **Factors with similar values or prevalence between patients with complete or missing data** | | | | | | | | |
| Number of comorbidities: Mean(SD) | 2011 | **3.7 (2.9)** | 11952 | 0.2% | **3.7 (3.0)** | 5396 | 0.3% | 0.375 |
| 2012 | **3.5 (2.9)** | 15594 | 0.0% | **3.6 (3.1)** | 7780 | 0.0% | 0.050 |
| 2013 | **4.0 (3.1)** | 28711 | 0.0% | **3.9 (3.2)** | 15189 | 0.0% | <.001 |
| Acuity: Chronic - over 3 months | 2011 | **57.6%** | 11963 | 0.1% | **59.7%** | 5404 | 0.1% | 0.009 |
| 2012 | **56.3%** | 15597 | 0.0% | **59.5%** | 7776 | 0.1% | <.001 |
| 2013 | **58.0%** | 28684 | 0.1% | **58.3%** | 15169 | 0.2% | 0.616 |
| Gender (Female) | 2011 | **67.4%** | 11970 | 0.0% | **67.0%** | 5411 | 0.0% | 0.668 |
| 2012 | **66.9%** | 15600 | 0.0% | **68.4%** | 7782 | 0.0% | 0.028 |
| 2013 | **67.0%** | 28711 | 0.0% | **67.3%** | 15192 | 0.0% | 0.531 |
| Payer: Medicare B age under 60 | 2011 | **1.4%** | 11929 | 0.3% | **1.8%** | 5408 | 0.1% | 0.109 |
| 2012 | **1.7%** | 15564 | 0.2% | **2.3%** | 7778 | 0.1% | 0.001 |
| 2013 | **2.3%** | 28711 | 0.0% | **2.9%** | 15192 | 0.0% | <.001 |
| High Fear Avoidance at intake (Physical) | 2011 | **61.4%** | 11786 | 1.5% | **62.2%** | 5317 | 1.7% | 0.341 |
| 2012 | **55.3%** | 15461 | 0.9% | **55.0%** | 7739 | 0.6% | 0.672 |
| 2013 | **50.6%** | 28574 | 0.5% | **51.2%** | 15143 | 0.3% | 0.213 |
| Exercise history: 1 or more / week | 2011 | **68.3%** | 11612 | 3.0% | **68.1%** | 5274 | 2.5% | 0.850 |
| 2012 | **69.0%** | 15331 | 1.7% | **69.2%** | 7693 | 1.1% | 0.848 |
| 2013 | **66.8%** | 28628 | 0.3% | **67.4%** | 15150 | 0.3% | 0.204 |
| Difference not supporting potential for selection bias | | | | | | | | |
| Differences supporting potential for selection bias | | | | | | | | |
| Differences interpreted as not clinically important | | | | | | | | |

We found that patients with complete data had higher values or prevalence for characteristics that were predictive of both lower (age, surgical history) and higher (Medicare B payer age 60 or more) FS change therefore not supporting a systematic patient selection bias. Patients with missing data had higher values or prevalence for characteristics associated with lower FS change (higher intake FS, Medicaid payer) potentially supporting some patient selection bias. Patients with complete and missing discharge data were similar in terms of number or comorbidities, acuity, gender distribution, Medicare B payer under the age of 60, level of fear avoidance, and exercise history. Overall, these analyses were inconclusive and did not support a systematic patient selection bias.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

In an internal FOTO study, we examined the completeness of outcomes data of patients with Hip impairments admitted to therapy during 2011-2013. Data were extracted during July 2014 therefore all episodes of care had started at least 6 months prior to data extraction. For all years combined there were a total of 112,608 patients during this time period. 98.3% (110,678) had completed a FOTO (hip) PROM at admission, while 1.7% had a non-participation audit (NPA) indicating the reason that FS was not collected. We examined whether or not these patients had FS data collected; and if no data was collected we categorized the reason for missing discharge data. Reasons for missing discharge FS data included having an “open episode”, defined as patient not being discharged from therapy; having a NPA at DC, or unknown reason. For the 3 years combined the completion rate was 50.9%, while 49.1% had missing data. Open episodes accounted for 23.5% of missing DC FS. The results, stratified by year are shown in the table below.

**Table 2b7.2 a Summary of FS Data Collection at Admission and Discharge**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year at episode start | Patients at admission   N=112608 Women: 67.1% Age (Mn; SD; Range): 57.1; 18.5; 14-113 | | Patients after 6 months from admission N=110678 | | | |
| Has DC FS + N (%) % represents completion rate | Reason for Missing FS at DC N (%) | | |
| FS  N (%) | FS missing\* (NPA) N (%) | Open episode | NPA audit at DC | Unknown |
| 2011 | 23732 (99.5) | 114 (0.5) | 11970 (50.4) | 6351 (26.8) | 719 (3.0) | 4692 (19.8) |
| 2012 | 30267 (97.9) | 646 (2.1) | 15600 (51.5) | 6885 (22.7) | 3033 (10.0) | 4749 (15.7) |
| 2013 | 56679 (98.0) | 1170 (2.0) | 28711 (50.7) | 12776 (22.5) | 8280 (14.6) | 6912 (12.2) |
| **Total** | **110678 (98.3)** | **1930 (1.7)** | **56281 (50.9)** | **26012 (23.5)** | **12032 (10.9)** | **16353 (14.8)** |

+Only data from closed episodes with status/discharge FS are included in the aggregate dataset for calculation of risk adjustment coefficients.

\*FS missing at intake with NPA completed indicating reason for missing data

CR = Completion rate = Number of discharged patients who had FS at DC/ Number of patients who had FS at intake, includes episodes that are still “open”.

We also examined percent of complete outcomes data by clinic to determine the variability in completeness across clinics. There was substantial variation in the completion rate by clinic. Mean CR were 47.4%, 50.6% and 49.2% for years 2011, 2012 and 2013 respectively.

**Table 2b7.2 b Mean and SD of data completion by clinics (all clinics)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | N clinics | Patients with intake FS data  Mean N (SD) | Has DC FS | | Does not have DC FS  Mean N (sd) |
| Mean N (SD) | Mean CR(SD) |
| 2011 | 1235 | 19.2 (26.2) | 9.7 (15.4) | 47.4 (33.2) | 9.5 (13.5) |
| 2012 | 1429 | 21.2 (29.2) | 10.9 (17.5) | 50.6 (32.6) | 10.3 (14.9) |
| 2013 | 2195 | 25.8 (33.7) | 13.1 (19.2) | 49.2 (29.4) | 12.7 (17.7) |
| CR = Completion rate = Number of discharged patients who had FS at DC/ Number of patients who had FS at intake, includes episodes that are still “open”. | | | | | |

We repeated this analysis for clinics meeting the criteria selected for the performance analysis (who had at least 40 patients (for clinics with 5+ clinicians) or at least 10 patients per clinician for clinics with less than 5 clinicians) for the year tested with FS at intake & discharge treated for hip impairment. Completion rates for these clinics were higher at 67.4 %, 68.1 % and 65.0 % for the years 2011, 2012 and 2013 respectively. See Table below:

**Table 2b7.2 c Mean and SD of data completion by clinics with 40+ complete episodes per year (or 10+ complete episodes per clinician for clinics under 5 clinicians):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | N clinics | Patients with intake FS data (Dataset1) Mean N (SD) | Has DC FS (Group 1) | | Does not have DC FS Mean N (sd) |
| Mean N (SD) | Mean CR(SD) |
| 2011 | 92 | 74.8 (46.3) | 47.5 (27.5) | 67.4 (15.9) | 27.3 (24.0) |
| 2012 | 140 | 71.1 (53.5) | 45.6 (33.2) | 68.1 (15.8) | 25.6 (25.1) |
| 2013 | 287 | 75.1 (59.0) | 44.9 (33.4) | 65.0 (15.1) | 30.1 (30.4) |
| CR = Completion rate = Number of discharged patients who had FS at DC/ Number of patients who had FS at intake, includes episodes that are still “open”. (Group 1/ Dataset1) | | | | | |

We assessed whether missing data was a source of systematic bias by testing associations between clinic completion rates and clinic quality (as measured by clinic average residual scores after risk adjustment modeling) for clinics included in the performance analysis. Residual scores are the difference between predicted functional outcomes (given risk adjustment factors) and the actual outcomes. Existence of systematic bias was assumed to result in some associations between clinic completion rates and clinic quality. We examined Pearson Correlations between clinic CR and clinic average residual scores by year. No correlations were found between CR and residual scores.

**Table 2b7.2 d Relationship between Clinic Aggregated**

**Residual Score and Clinic Completion Rate (CR)**

|  |  |  |  |
| --- | --- | --- | --- |
| Pearson Correlations (r) | | | Clinic CR |
| 2011 | Residual | r | -.029 |
| Sig. (2-tailed) | .781 |
| N clinics | 92 |
| 2012 | Residual | r | .077 |
| Sig. (2-tailed) | .365 |
| N clinics | 140 |
| 2013 | Residual | r | -.096 |
| Sig. (2-tailed) | .103 |
| N clinics | 287 |

To examine whether there was an underlying pattern to the relationship between clinic completion rate and risk adjusted residual scores aggregated at the clinic level, we grouped clinics into 10 completion rate categories. Results shown below suggest that the relationship between CR and aggregated residual scores is not linear and has no strong pattern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 2b7.2e Average residuals at the clinic level by completion rate categories** | | | | |
| Year | CR category | Mean residuals | N clinics | Std. Deviation |
| **2011** | 30 to <40 | 6.3\* | 3 | 2.2 |
| 40 to <50 | 4.5 | 10 | 5.0 |
| 50 to <60 | 1.4 | 19 | 2.6 |
| 60 to <70 | -0.9 | 22 | 3.0 |
| 70 to <80 | 3.0 | 16 | 5.6 |
| 80 to <90 | 0.7 | 12 | 4.6 |
| 90 to Highest | 3.4 | 10 | 7.1 |
| **Total** | **1.7** | **92** | **4.8** |
| **2012** | 20 to <30 | 1.8\* | 2 | 5.6 |
| 30 to <40 | 3.7\* | 6 | 2.1 |
| 40 to <50 | 0.1\* | 7 | 3.3 |
| 50 to <60 | 0.5 | 23 | 3.7 |
| 60 to <70 | 0.2 | 42 | 3.9 |
| 70 to <80 | 0.9 | 31 | 3.8 |
| 80 to <90 | 1.5 | 14 | 6.4 |
| 90 to Highest | 2.8 | 15 | 8.3 |
| **Total** | **1.0** | **140** | **4.7** |
| **2013** | 20 to <30 | 3.6\* | 1 |  |
| 30 to <40 | 2.1 | 13 | 2.9 |
| 40 to <50 | 1.1 | 36 | 3.0 |
| 50 to <60 | 1.4 | 53 | 3.5 |
| 60 to <70 | 0.7 | 74 | 3.6 |
| 70 to <80 | 0.4 | 62 | 5.0 |
| 80 to <90 | 0.2 | 34 | 4.3 |
| 90 to Highest | 1.2 | 14 | 4.6 |
| **Total** | **0.9** | **287** | **3.9** |
| CR = Completion rate = Number of discharged patients who had FS at DC/ Number of patients who had FS at intake, includes episodes that are still “open”.  \* Average residuals for CR categories that had less than 10 clinics | | | | |

**.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Overall, the comparisons of characteristics of patients with and without complete outcomes data show no systematic pattern suggesting a selection bias in the collection of discharge FS data.

Overall completion rates (CR) by clinic were fairly stable between 2011-2013, but were lower than previously reported. We believe that this is because in our current analysis we categorized data as missing if patients had open episodes 6 months or more after admission, whereas in the earlier analyses patients with open episodes were not included in this calculation. While it is possible that some patients were still in active treatment 6 months or more after intake, we believe that this would be unusual. A more likely explanation is that these episodes were still open because in many cases the patient failed to return for a scheduled appointment and hence was not discharged and no outcome data or NPA was collected.

We also note that CRs were higher for clinics that had the minimum number of patients required to be included in the performance analysis. For all 3 years, these higher volume clinics had CRs more than 15% higher than all clinics combined. This suggests that greater familiarity with the FOTO data collection system is associated with higher completion rates. In addition, although the overall percent of clinic included in the performance analysis was low (10%), it increased with time, with 7% for 2011(92/1235), 10% for 2012 (140/1429), and 13% for 2013 (287/2195). This suggests an improved data collection process at the clinic level over time.

Our evaluation of the impact of missing data on provider level functional status found no linear relationships between clinic CR and clinic quality measured by aggregated residual scores. Residual scores of clinics by CR groupings containing less than 10 clinics should be interpreted cautiously, as small numbers may contribute to instability of estimates.

Recently, we have begun to explore the use of inverse probability weighting ([Robins and Finkelstein 2000](#_ENREF_15); [Seaman and White 2013](#_ENREF_16)) as a potential means to adjust for missing discharge data. We plan to examine whether or not the use of this method will change the coefficients in our risk adjustment models or influence the quality scores of providers.

FOTO is also taking several steps to improve completeness of outcomes data. Specific mechanisms that will be adopted in the next year include:

1. Generate the following performance measures to be included in FOTO’s benchmarking reports:
   1. Level of success at closing cases that are no longer active (> 6 months).
   2. Clinic’s patient survey completion rate at discharge.
   3. Require documenting reasons for missing data at intake and discharge using FOTO’s incomplete episode reports.
2. Encourage providers to link FOTO data with electronic health care records to validate non-participation audits.

**References:**

Hart, D. and J. Connolly (2006). Pay-for-Performance for Physical Therapy and Occupational Therapy: Medicare Part B Services Grant #18-P-93066/9-01, Health & Human Services/Centers for Medicare & Medicaid Services.

Hart, D. L., J. E. Mioduski, et al. (2005). "Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments." Journal of Clinical Epidemiology 58(6): 629-638.

Hart, D. L., J. E. Mioduski, et al. (2006). "Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function." Journal of Clinical Epidemiology 59(9): 947-956.

Hart, D. L., Y. C. Wang, et al. (2008). "A computerized adaptive test for patients with hip impairments produced valid and responsive measures of function." Archives of Physical Medicine and Rehabilitation 89(11): 2129-2139.

Hart, D. L., M. W. Werneke, et al. (2011). "Effect of fear-avoidance beliefs of physical activities on a model that predicts risk-adjusted functional status outcomes in patients treated for a lumbar spine dysfunction." The Journal of orthopaedic and sports physical therapy 41(5): 336-345.

Hart, D. L., M. W. Werneke, et al. (2011). "Using intake and change in multiple psychosocial measures to predict functional status outcomes in people with lumbar spine syndromes: a preliminary analysis." Physical Therapy 91(12): 1812-1825.

Hilfiker, R., L. M. Bachmann, et al. (2007). "Value of predictive instruments to determine persisting restriction of function in patients with subacute non-specific low back pain. Systematic review." European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society 16(11): 1755-1775.

Jette, D. U. and A. M. Jette (1996). "Physical therapy and health outcomes in patients with knee impairments." Phys Ther 76(11): 1178-1187.

Jette, D. U. and A. M. Jette (1996). "Physical therapy and health outcomes in patients with spinal impairments." Phys Ther 76(9): 930-941; discussion 942-935.

Resnik, L., Z. Feng, et al. (2006). "State regulation and the delivery of physical therapy services." Health Serv Res 41(4 Pt 1): 1296-1316.

Resnik, L., P. Gozalo, et al. (2011). "Weighted index explained more variance in physical function than an additively scored functional comorbidity scale." Journal of Clinical Epidemiology 64(3): 320-330.

Resnik, L. and D. L. Hart (2003). "Using clinical outcomes to identify expert physical therapists." Physical Therapy 83(11): 990-1002.

Resnik, L., D. Liu, et al. (2008). "Benchmarking physical therapy clinic performance: statistical methods to enhance internal validity when using observational data." Physical Therapy 88(9): 1078-1087.

Resnik, L., D. Liu, et al. (2008). "Predictors of physical therapy clinic performance in the treatment of patients with low back pain syndromes." Phys Ther 88(9): 989-1004.

Robins, J. M. and D. M. Finkelstein (2000). "Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests." Biometrics 56(3): 779-788.

Seaman, S. R. and I. R. White (2013). "Review of inverse probability weighting for dealing with missing data." Statistical methods in medical research 22(3): 278-295.

Wang, Y. C., D. L. Hart, et al. (2009). "Clinical interpretation of a lower-extremity functional scale-derived computerized adaptive test." Physical Therapy 89(9): 957-968.