Crosswalk Functionality in MIPS Clinical Quality Measures #217-222 and 478

Background

Patient-reported outcome measures (PROMs) developed using modern measure science methods, known as ***item response theory (IRT),*** are key components of MIPS ***Clinical Quality Measures (CQMs)*** #217-222 and 478 which were developed and are stewarded by FOTO Patient Outcomes (**FOTO**). IRT-based PROMs offer multiple measurement advantages which include reducing patient response burden while simultaneously maintaining high measurement (i.e., score) precision when administering the IRT-PROM using computerized adaptive testing (CAT) or (fixed/static) short form.

However, ***"legacy"*** PROMs, which were developed using classical test theory methods and thus lack modern measurement advantages, are still commonly used by a proportion of today's healthcare providers such as rehabilitation therapists and medical physicians. To increase MIPS participation, FOTO research scientists developed crosswalk functionality so that clinicians may have the option to administer legacy PROMs to their patients for use in routine clinical care while reporting one or more of these CQMs for MIPS. Cross-walked scoring (linking methods) [1-3] originated primarily in the field of Educational Psychology and in recent years crossed over into the field of healthcare PROMs. Several crosswalks have been established by score-linking between PROMIS (IRT-based) and other PROMs.[4]

Statistically equivalent score-linking between an IRT-based PROM and a suitable legacy PROM is produced using advanced psychometric methods. The approach results in a ***crosswalk table*** that facilitates the conversion of a legacy PROM score to the metric of the IRT-based PROM. Once a legacy score is cross-walked to its equivalent IRT-based PROM score, the corresponding risk adjustment model for the CQM can be applied and the risk-adjusted predicted change score calculated, followed by the calculation of patient-level residual scores (i.e., observed change minus predicted change) from which the "performance met" criterion is applied (residual $\geq$=0). Depending on the measurement characteristics of the legacy PROM, an adjustment factor may be applied to the crosswalk-based residuals to balance the rates of performance met to those of the IRT-based PROM.

To score a MIPS CQM, CMS determines "***measure achievement points***" by comparing performance on a measure to a measure benchmark. CMS provided the formula for calculating measure achievement points in their posting of the MIPS Historical 2022 Quality Benchmarks.

This report assesses the impact of incorporating cross-walked PROM scores in the 7 FOTO CQMs on MIPS measure achievement points. This was done using residuals based on cross-walked scores compared to using residuals based on the currently designated IRT-based PROM score for each quality measure.

Methods

<u>First</u>, to create a reliable and valid crosswalk table, the two PROMs of linking interest need to be conceptually similar, i.e., they should have been psychometrically designed to measure a same or common construct (e.g., physical function). Practically, this conceptual link or common-construct requirement was assessed by calculating the inter-score correlation of the PROMs, which is expected to be of a magnitude of 0.7 or more.

<u>Second</u>, once this conceptual link was confirmed, the observed score equating method, using the R-based module "equate," was employed, using a linking sample, to first identify and then apply the "best" mathematical function most accurately predicting the FOTO IRT-PROM scores from legacy PROM scores. Additional detailed information regarding the observed score equating method is available on request.

<u>Finally</u>, a comprehensive set of analyses was conducted, using the linking sample as well as a separate validation sample, to test multiple levels of evidence evaluating the reliability and validity of the cross-walked scores, addressing the following questions:

(A) Using the linking sample, do predicted scores have the expected psychometric characteristics, including their inter-measure correlations and distributional features, compared to the observed, original (unconverted) scores derived from the crosswalk linking sample?

(B) Using validation sample #2, do patient-level residuals and rates of performance met correlate as expected between the two sets of scores?

(C) Using validation sample #2, are score associations at the group level for residuals, performance met, and measure achievement points supportive of minimal impact on the CQMs calculated using the IRT-based PROMs compared to legacy PROMs?

References
1. Albano AD, Rodriguez MC. Statistical equating with measures of oral reading fluency. <u>J of Sch Psych 50 (2012) 43–59</u>.
2. Albano AD. R Package 'equate' (Version 2.0-3): Observed-Score Linking and Equating. October 21, 2014.
3. Albano AD. equate: An R Package for Observed-Score Linking and Equating. October 21, 2014.
4. PROsetta Stone Linking Patient-Reported Outcome Measures. https://www.prosettastone.org/. Accessed January 15, 2022.

Results and interpretation pertaining to each CQM are provided on the following pages.

Results and interpretation pertaining to CQMs #217-219 for the Lower Extremity Functional Scale (LEFS) to the FOTO Lower Extremity Physical Function (LEPF)

(A) Table 1 demonstrates the following characteristics of a successful linking: The original inter-measure score correlation is successfully recreated in both the linking and validation samples; the original score distributional characteristics are successfully recreated in both the linking and validation samples; the distribution of individual score differences is centered on 0, indicating a balanced (unbiased) difference distribution; its SD is small, indicating score differences tend to be of low magnitude.

Table 1: Score-level analyses from the linking and validation #1 samples

	Linking Sample (n=9000)	Validation Sample #1 (n=1000)
*Pearson Correlation**	0.99 / 0.99	0.99 / 0.99
*Distributional Characteristics***		
Mean	51.3 / 51.3	52.7 / 52.6
SD	17.4 / 17.4	17.4 / 17.5
Median	51.5 / 51.4	52.1 / 52.2
Skewness	-0.12 / -0.12	-0.04 / -0.03
Excess Kurtosis	-0.11 / -0.11	-0.31 / -0.29
*LEPF Individual Score Differences: predicted minus actual ****		
Mean (SD); median	0.0 (1.9); 0.0	-0.1 (1.9); 0.0
*Data are Observed LEFS to FOTO LEPF / Predicted FOTO LEPF to FOTO LEPF scores		
**Data are Observed/Predicted LEFS to LEPF T scores		
***Differences are on the T-score metric (mean=50; SD=10)		

(B) A 2nd validation sample included a sample of 54,818 patients that responded to both the LEFS and LEPF measures (mean age/SD=54.0(19.2); range 14-89, 62.8% female). The patient level Pearson correlation between residuals derived from either the observed or predicted LEPF scores were 0.99, with 95.5% agreement on the performance met criterion.

(C) Table 2 below demonstrates a set of estimates at the group level, by different thresholds of minimum number of patients per group. Results demonstrate extremely high correlations between scores derived from the LEPF or linked from the LEFS, for either residuals, performance met, or measure achievement points, strongly supporting the accuracy (reliability) and validity of the linked scores. Mean differences in achievement points between the two sets of scores were negligible and not statistically significant, confirming that positive and negative differences well balanced (no score bias).

Table 2: Group level analyses

Group characteristics	Correlations	Achievement points

N groups	Case min	Mean / median cases per group	Residuals	Mean performance met	Mean achievement points	Observed / expected mean difference: mean (P-value)
1,004	10	49.2 (27)	0.994	0.951	0.935	-0.04 (0.319)
667	20	67.1 (43)	0.995	0.961	0.943	0.00 (0.997)
471	30	85.0 (56)	0.995	0.965	0.946	-0.03 (0.499)

Overall, these results demonstrate minimal impact of the inclusion of cross-walked scores on measure achievement points for CQMs #217-219. This suggests that this change is not substantive and does not require a benchmark reset.

Results and interpretation pertaining to CQM #220 for Modified Oswestry Disability Index (ODI) to the FOTO Low Back

(A) Table 1 demonstrates the following characteristics of a successful linking: The original inter-measure score correlation is successfully recreated in both the linking and validation samples; the original score distributional characteristics are successfully recreated in both the linking and validation samples; the distribution of individual score differences is centered on 0, indicating a balanced (unbiased) difference distribution; its SD is small, indicating score differences tend to be of low magnitude.

Table 1: Score-level analyses from the linking and validation #1 samples

	Linking Sample (n=20000)	Validation Sample #1 (n=7007)
Pearson Correlation*	0.74 / 0.74	0.74 / 0.74
Distributional Characteristics**		
Mean	49.0 / 49.0	48.5 / 48.6
SD	13.2 / 13.2	13.0 / 13.4
Median	49.7 / 49.4	48.9 / 48.2
Skewness	0.04 / 0.07	-0.06 / 0.06
Excess Kurtosis	1.03 / 1.18	1.03 / 1.45
Lumbar Individual Score Differences: predicted minus actual ***		
Mean	0.0	0.1
SD	9.5	9.6
Median	0.0	0.2
*Data are Observed ODI to Lumbar / Predicted Lumbar to Lumbar scores		
**Data are Observed Lumbar / Predicted Lumbar scores		
***Differences are a 0-100 scaling		

(B) A 2nd validation sample included a sample of 28,261 patients that responded to both the FOTO Low Back and the Modified Oswestry PROM (mean age/SD=57.9(17.2); range 14-89, 60.3% female). The patient level Pearson correlation between residuals derived from either the observed or predicted FOTO Low Back scores were 0.724, with 76.3% agreement on the performance met criterion.

(C) Table 2 below demonstrates a set of estimates at the group level, by different thresholds of minimum number of patients per group. Results demonstrate high correlations (>0.7) between scores derived from the FOTO Low Back or linked from the Modified Oswestry, for either residuals, performance met, or measure achievement points, supporting the accuracy (reliability) and validity of the linked scores. Mean differences in achievement points between the two sets of scores were small (<0.3) and not statistically significant, confirming that positive and negative differences well balanced (no score bias).

Table 2: Group level analyses

Group characteristics			Correlations			Achievement points
N groups	Case min	Mean / median cases per group	Residuals	Mean performance met	Mean achievement points	Observed / expected mean difference: mean (P-value)
298	10	91.1 (50)	0.871	0.760	0.703	-0.15 (0.226)
244	20	108.2 (60)	0.892	0.809	0.720	-0.24 (0.071)
202	30	125.6 (75)	0.907	0.834	0.771	-0.24 (0.300)

Overall, these results demonstrate minimal impact of the inclusion of cross-walked scores on measure achievement points for CQM #220. This suggests that this change is not substantive and does not require a benchmark reset.

Results and interpretation pertaining to CQM #221 for the DASH to FOTO Shoulder

(A) Table 1 demonstrates the following characteristics of a successful linking: The original inter-measure score correlation is successfully recreated in both the linking and validation samples; the original score distributional characteristics are successfully recreated in both the linking and validation samples; the distribution of individual score differences is centered on 0, indicating a balanced (unbiased) difference distribution; its SD is small, indicating score differences tend to be of low magnitude.

Table 1: Score-level analyses from the linking and validation #1 samples

	Linking Sample (n=2000)
Pearson Correlation*	0.81 / 0.82
Distributional Characteristics**	
Mean	50.3 / 50.3
SD	15.5 / 15.4
Median	52.0 / 51.3
Skewness	-0.43 / -0.43
Excess Kurtosis	1.24 / 1.20
ShoulderFS Individual Score Differences: predicted minus actual ***	
Mean	0.0
SD	9.3
Median	-0.2
*Data are Observed DASH to ShoulderFS / Predicted ShoulderFS to ShoulderFS scores	
**Data are Observed ShoulderFS / Predicted ShoulderFS scores	
***Differences are a 0-100 scaling	

(B) A validation sample included a sample of 7,915 patients that responded to both the FOTO shoulder and DASH measures (mean age/SD=55.1(16.6); range 14-89, 54.0% female). The patient level Pearson correlation between residuals derived from either the observed or predicted FOTO shoulder scores were 0.79, with 82.0% agreement on the performance met criterion.

(C) Table 2 below demonstrates a set of estimates at the group level, by different thresholds of minimum number of patients per group. Results demonstrate high correlations (>0.77) between scores derived from the FOTO shoulder or linked from the DASH, for either residuals, performance met, or measure achievement points, supporting the accuracy (reliability) and validity of the linked scores. Mean differences in achievement points between the two sets of scores were negligible (<0.3) and not statistically significant, confirming that positive and negative differences well balanced (no score bias).

Table 2: Group level analyses

Group characteristics			Correlations			Achievement points
N groups	Case min	Mean / median cases per group	Residuals	Mean performance met	Mean achievement points	Observed / expected mean difference: mean (P-value)
151	10	43.8 (25)	0.778	0.646	0.668	-0.21 (0.307)
93	20	62.6 (37)	0.868	0.763	0.721	-0.29 (0.226)
68	30	76.9 (46)	0.913	0.806	0.745	-0.29 (0.478)

Overall, these results demonstrate minimal impact of the inclusion of cross-walked scores on measure achievement points for CQM #221. This suggests that this change is not substantive and does not require a benchmark reset.

Results and interpretation pertaining to CQM #222 for the DASH to FOTO Elbow/Wrist/Hand (EWH)

(A) Table 1 demonstrates the following characteristics of a successful linking: The original inter-measure score correlation is successfully recreated in both the linking and validation samples; the original score distributional characteristics are successfully recreated in both the linking and validation samples; the distribution of individual score differences is centered on 0, indicating a balanced (unbiased) difference distribution; its SD is small, indicating score differences tend to be of low magnitude.

Table 1: Score-level analyses from the linking and validation #1 samples

	Linking Sample (n=900)
*Pearson Correlation**	0.81 / 0.81
*Distributional Characteristics***	
Mean	49.9 / 49.8
SD	15.3 / 15.2
Median	49.0 / 50.7
Skewness	-0.24 / -0.29
Excess Kurtosis	0.14 / 0.00
*EWH Individual Score Differences: predicted minus actual ***	
Mean	0.0
SD	9.4
Median	-0.1
*Data are Observed DASH to EWH / Predicted EWH to EWH scores	
**Data are Observed EWH / Predicted EWH scores	
***Differences are a 0-100 scaling	

(B) A validation sample included a sample of 3,366 patients that responded to both the LEFS and LEPF measures (mean age/SD=52.2(17.7); range 14-89, 61% female). The patient level Pearson correlation between residuals derived from either the observed or predicted EWH scores were 0.83, with 83.1% agreement on the performance met criterion.

(C) Table 2 below demonstrates a set of estimates at the group level, by different thresholds of minimum number of patients per group. Results demonstrate high correlations (>0.72) between scores derived from the EWH or linked from the DASH, for either residuals, performance met, or measure achievement points, supporting the accuracy (reliability) and validity of the linked scores. Mean differences in achievement points between the two sets of scores were negligible (<0.3) and not statistically significant, confirming that positive and negative differences well balanced (no score bias).

Table 2: Group level analyses

Group characteristics			Correlations			Achievement points
N groups	Case min	Mean / median cases per group	Residuals	Mean performance met	Mean achievement points	Observed / expected mean difference: mean (P-value)
72	10	35.1 (23)	0.822	0.721	0.735	0.24 (0.354)
42	20	49.8 (36)	0.912	0.817	0.835	-0.20 (0.448)
26	30	66.0 (49)	0.953	0.859	0.844	-0.20 (0.421)

Overall, these results demonstrate minimal impact of the inclusion of cross-walked scores on measure achievement points for CQMs #222. This suggests that this change is not substantive and does not require a benchmark reset.

Results and interpretation pertaining to CQM #478 for the Neck Disability Index (NDI) to the FOTO Neck

(A) Table 1 demonstrates the following characteristics of a successful linking: The original inter-measure score correlation is successfully recreated in both the linking and validation samples; the original score distributional characteristics are successfully recreated in both the linking and validation samples; the distribution of individual score differences is centered on 0, indicating a balanced (unbiased) difference distribution; its SD is small, indicating score differences tend to be of low magnitude.

Table 1: Score-level analyses from the linking and validation #1 samples

	Linking Sample (n=9000)	Validation Sample #1 (n=1000)
Pearson Correlation*	0.68 / 0.69	0.68 / 0.68
Distributional Characteristics**		
Mean	52.1 / 52.1	52.4 / 52.1
SD	12.4 / 12.3	12.2 / 12.5
Median	52.0 / 51.6	52.0 / 51.6
Skewness	0.16 / 0.14	0.03 / 0.22
Excess Kurtosis	0.90 / 0.89	1.10 / 1.02
LEPF Individual Score Differences: predicted minus actual ***		
Mean (SD); median	0.0 (9.7); 0.0	-0.2 (9.9); 0.1
*Data are Observed NDI to FOTO Neck / Predicted FOTO Neck to FOTO Neck scores		
**Data are Observed FOTO Neck / Predicted Neck scores		
***Differences are a 0-100 scaling		

(B) A 2nd validation sample included a sample of 11,210 patients that responded to both the FOTO Neck and NDI measures (mean age/SD=55.3(16.1); range 14-89, 66% female). The patient level Pearson correlation between residuals derived from either the observed or predicted FOTO Neck scores were 0.73, with 77.7% agreement on the performance met criterion.

(C) Table 2 below demonstrates a set of estimates at the group level, by different thresholds of minimum number of patients per group. Results demonstrate moderate to high correlations between scores derived from the FOTO Neck or linked from the NDI, for either residuals, performance met, or measure achievement points, supporting the accuracy (reliability) and validity of the linked scores. Mean differences in achievement points between the two sets of scores were negligible (<0.2) and not statistically significant, confirming that positive and negative differences well balanced (no score bias).

Table 2: Group level analyses

Group characteristics	Correlations	Achievement points

N groups	Case min	Mean / median cases per group	Residuals	Mean performance met	Mean achievement points	Observed / expected mean difference: mean (P-value)
226	10	44.4 (29)	0.828	0.698	0.595	0.13 (0.549)
145	20	61.6 (44)	0.786	0.661	0.602	-0.10 (0.699)
108	30	74.1 (53)	0.833	0.711	0.705	-0.10 (0.224)

Overall, these results demonstrate minimal impact of the inclusion of cross-walked scores on measure achievement points for CQM #478. This suggests that this change is not substantive and does not require a benchmark reset.