Benchmarking Outpatient Rehabilitation Clinics Using Functional Status Outcomes

Pedro L. Gozalo, Linda J. Resnik, and Benjamin Silver

Objective. To utilize functional status (FS) outcomes to benchmark outpatient therapy clinics.

Data Sources. Outpatient therapy data from clinics using Focus on Therapeutic Outcomes (FOTO) assessments.

Study Design. Retrospective analysis of 538 clinics, involving 2,040 therapists and 90,392 patients admitted July 2006–June 2008. FS at discharge was modeled using hierarchical regression methods with patients nested within therapists within clinics. Separate models were estimated for all patients, for those with lumbar, and for those with shoulder impairments. All models risk-adjusted for intake FS, age, gender, onset, surgery count, functional comorbidity index, fear-avoidance level, and payer type. Inverse probability weighting adjusted for censoring.

Data Collection Methods. Functional status was captured using computer adaptive testing at intake and at discharge.

Principal Findings. Clinic and therapist effects explained 11.6 percent of variation in FS. Clinics ranked in the lowest quartile had significantly different outcomes than those in the highest quartile ($p < .01$). Clinics ranked similarly in lumbar and shoulder impairments (correlation = 0.54), but some clinics ranked in the highest quintile for one condition and in the lowest for the other.

Conclusions. Benchmarking models based on validated FS measures clearly separated high-quality from low-quality clinics, and they could be used to inform value-based-payment policies.

Key Words. Rehabilitation, physical therapy, quality measurement, benchmarking, profiling

Outpatient rehabilitation (therapy), including physical therapy, occupational therapy, and speech-language pathology, is a covered Medicare benefit used by approximately 4.7 million (13.5 percent) Medicare Part B Beneficiaries in 2010, at a cost of \$5.6 billion (Silver et al. 2013). During the 2006–2010 5-year period, Medicare outpatient therapy reimbursements grew at an average annual rate of 9.4 percent (Silver et al. 2013). This rapid growth was driven

primarily by increased service utilization per beneficiary (Ciolek and Hwang 2010). However, the value of this additional utilization remains unclear, given the significant variation in mean annual per-beneficiary expenditures.

Efforts to control outpatient therapy costs began with the Balanced Budget Act of 1997, which placed all therapy billing under the Medicare Physician Fee Schedule and set an annual cap on expenditures per beneficiary. However, the cap has come to be viewed by many as a coarse limitation that is not sensitive to the needs of the individual patient, and it has failed to curb excessive service use as an increasing number of patients exceed the cap each year (e.g., 15 percent in 2006 and 19 percent in 2010) (Silver et al. 2013). During the last decade, the Centers for Medicare & Medicaid Services (CMS) funded several projects to explore alternative reimbursement models. These included a Pay-for-Performance simulation model (Hart and Connolly 2006), the Short-Term Alternatives for Therapy Services (STATs) project (Ciolek and Hwang 2008), and the Developing Outpatient Therapy Payment Alternative project (Lyda-McDonald, Silver, and Gage 2012). One of the final STATs report recommendations was that “the outcomes resulting from provider interventions” be used in future payment efforts.

While these projects did not lead to direct policy changes, CMS’s recent Roadmap for Value-Based Purchasing (VBP) focuses on payment for efficient resource use with high quality of care (Center for Medicare & Medicaid Services 2013). In these models, treatment outcomes are a critical component of quality of care. Because the goal of outpatient therapy is to improve function, measurements of functional status (FS) are critical to determining the effectiveness of therapy treatment and can be used to create clinician and clinic risk-adjusted performance measures. These performance measures can be used to monitor quality, identify quality improvement approaches, improve accountability, and ultimately reduce practice variation and enhance FS outcomes of care across therapy providers.

Until recently, FS measures have not been required to be collected routinely or be submitted with outpatient therapy claims. As of July 1, 2013, CMS requires providers of Part B covered therapy services to collect FS measures

Address correspondence to Pedro Gozalo, Ph.D., Center for Gerontology and Health Care Research, School of Public Health, Brown University, 121 South Main St, Providence, RI 02912; e-mail: pedro_gozalo@brown.edu. Pedro Gozalo, Ph.D., Linda J. Resnik, Ph.D., and Benjamin Silver, B.A., are with the Department of Health Services, Policy, and Practice, School of Public Health, Brown University, Providence, RI. Linda J. Resnik, Ph.D., is with the Providence Veterans Administration Medical Center, Health Services Research, Providence, RI, and the Center for Gerontology and Health Care Research, School of Public Health, Brown University, Providence, RI.

as a first step toward future payment reform and quality improvement (Centers for Medicare & Medicaid Services 2012). These measures are categorical and CMS will need to determine to what degree they can be used for implementation of VBP payment policies. To this effect, it would be useful to have available results based on more detailed, validated FS measures as a point of reference. An important concern is that CMS's categorical measure may be too imprecise for VBP purposes and that a more detailed, validated FS measure may need to be incorporated into standard reporting practices in the future.

One possible assessment of FS that has been tested and validated over many years is Focus on Therapeutic Outcomes (FOTO) (Hart et al. 2010a). The FS measure collected by FOTO from 2000 to 2003 was used in the CMS P4P demonstration project to calculate patient risk-adjusted outcomes of therapy and to determine whether therapists achieved greater than, less than, or expected patient outcomes (Resnik and Hart 2003, 2004; Resnik and Jensen 2003; Resnik, Feng, and Hart 2006; Resnik et al. 2008). In the past decade, the outcome measurement system within the FOTO database has evolved considerably (Hart et al. 2010a). Currently, FOTO uses nine condition-specific FS instruments as well as a general overall FS measure.

Our analysis takes advantage of the FS measures in the FOTO dataset to address two important questions for the successful implementation of VBP in outpatient therapy care. Is there evidence of enough variation in provider outcomes to be able to benchmark outpatient therapy clinics? If so, are the facility rankings similar across patient conditions so that use of a single performance measure is sufficient to summarize the underlying provider quality, or should performance measures be evaluated separately for different conditions? The goal of our study was to utilize a large database of standardized functional status outcomes of outpatient therapy to examine benchmarking provider performance with common types of patient groups.

METHODS

Data Source and Study Population

This study used outpatient rehabilitation data from 538 clinics, involving 2,040 therapists and 90,392 patients drawn from the FOTO database during the 2-year period, July 2006–June 2008. The number of patients per clinic during these 2 years ranged from 30 to 1,943.

Focus on Therapeutic Outcomes collects a standardized set of data including demographics, intake and discharge functional status scores, and administrative data from outpatient rehabilitation services, as well as data on characteristics of health care providers and organizations (Swinkels et al. 2007). To our knowledge, the FOTO database is the largest outpatient rehabilitation outcomes database available for researchers in the United States.

Data may be collected by paper and pencil or by patient self-reported computerized adaptive testing (CAT) procedures. For the purposes of this study, we used only data from patients who completed the surveys via CAT.

Functional Status Measures

Self-reported FS was measured using either a body impairment-specific CAT (lumbar, shoulder, knee, cervical, foot/ankle, hip, wrist, or elbow) or a generic FS CAT measure (Hart et al. 2010a). FS measures were self-reported by patients at intake and at discharge from their rehabilitation episode. At the onset of therapy, patients, with assistance from staff, identify which body part (or neurological impairment) was their primary reason for rehabilitation. Each CAT administration yields a FS estimate that is transformed to a 0 (low-functioning) to 100 (high-functioning) metric. It should be noted that the mathematical equivalence of the FS estimates across the body part impairments has not been tested yet.

The development of each CAT, data supporting the discriminant validity of the CAT estimated FS measures, and the operating characteristics of the CATs have been described for persons with lumbar impairments (Hart et al. 2006b), shoulder impairments (Hart et al. 2006a; Wang et al. 2010b), knee impairments (Hart et al. 2008b), foot/ankle impairments (Hart, Mioduski, and Stratford 2005), and hip impairment (Hart et al. 2008a). CATs used for elbow, wrist/hand, or cervical regions (Hart and Connolly 2006) contain items from the SF-12 (Ware, Kosinski, and Keller 1996), SF-36 (Ware et al. 1993), items pertinent to patients with upper extremity impairments, items representing lower functional abilities, and items pertinent to specific impairments (Hart and Wright 2002). Items in the foot/ankle, knee, and hip CAT originated in the Lower Extremity Functional Scale (Hart, Mioduski, and Stratford 2005). Person reliability estimated using IRT methods (equivalent to Cronbach's alpha) was 0.92 for the lumbar CAT (Hart et al. 2006b) and 0.97 for the shoulder CAT (Hart et al. 2006a). Internal consistency (Cronbach's alpha) was reported as 0.96 for the hip CAT (Hart, Mioduski, and Stratford 2005), 0.97 for the knee CAT (Binkley et al. 1999) (Hart, Mioduski, and Strat-

ford 2005), and 0.96 for the foot and ankle CAT (Hart, Mioduski, and Stratford 2005). Reliability of the elbow wrist and hand measure or cervical CAT has not been reported.

The responsiveness, sensitivity to change, construct validity, and clinical interpretation of the FS measures generated by the lumbar, shoulder, hip, knee, and foot and ankle CAT administrations have been reported as strong (Wang et al. 2009a,b, 2010a,b). For example, in patients with lumbar impairments whose function was measured by the lumbar CAT, 66 percent attained FS change scores equal to or greater than a minimally detectable change (MDC) at the 95 percent confidence interval, and 71 percent of patients had FS change equal to or greater than the minimally clinically important improvement after outpatient rehabilitation treatment (Hart et al. 2010c). In patients with shoulder impairments, assessed using the shoulder CAT, 79 percent had FS change scores greater than the MDC, and 76 percent had changes greater than minimal clinically important improvement (Hart et al. 2010b). Effect sizes (ES) reported after treatment for patients with elbow, wrist, and hand were 0.94–1; while ES for patients with cervical impairments was reported as 0.88 (Hart and Connolly 2006).

Data Analysis

Model Development. Our outcome variable was FS at discharge from the rehabilitation episode. Hierarchical three-level linear regression models, with patients nested within therapists nested within clinics, were developed overall (all conditions), and separately for lumbar and shoulder, the two most common conditions among the nine patient impairment groups.

All models included the following risk-adjustment covariates: FS at intake, age, gender, symptom onset, number of surgeries, functional comorbidity index (FCI) (Groll et al. 2005), level of fear-avoidance, and payer type. For the overall model, indicator variables were used to control for the type of condition. The FCI is the only known index designed specifically to control for comorbid conditions that are hypothesized to affect functional status rather than mortality. The FCI uses patient self-report of the following comorbid conditions: arthritis, osteoporosis, asthma, chronic obstructive pulmonary disease, angina, congestive heart failure, prior heart attack (MI), neurological disease (multiple sclerosis, Parkinson's disease), prior stroke or transient ischemic attack, peripheral vascular disease, diabetes, upper gastrointestinal, disease (including ulcers and gastroesophageal reflux), depressed mood anxiety, visual impairment, hearing impairment, obesity (calculated from body

mass index [BMI] >30), and low back pain. A score of 0 indicates absence of any comorbid illness, and a score of 18 indicating the highest number of comorbid conditions. In the FOTO system, patients identify from a list of medical problems containing 17 FCI conditions (absent BMI) any problem that applies to them. Patients provide their height and weight from which their BMI is calculated. Presence of obesity (the 18th FCI) is identified if BMI is 30 or above. Although the FCI was originally scored by adding the number of “yes” answers to indicate the history of specific conditions, in this analysis comorbidity was assessed using quartiles of the FCI index. The use of quartiles of functional comorbidities, rather than the actual number of comorbid conditions was made to be consistent with the method that FOTO utilized for risk adjustment of the NQF-approved measures (Focus on Therapeutic Outcomes 2015).

In our models, age was categorized into quartiles to allow for a non-linear relationship. Symptom onset represented the number of days from condition onset until the beginning of therapy intervention, classified as acute (less than 21 days; reference group), subacute (22–90 days), and chronic (over 90 days). Symptom onset was identified by the patient at the time of intake. Number of surgeries for the impairment being treated was categorized as none, 1, or more. Fear avoidance was categorized as low or high using the Fear-Avoidance Beliefs Questionnaire for physical activities (FABQ-PA) (Waddell et al. 1993; Williamson 2006). High fear was operationally defined as FABQ-PA scores of 15 points or more, and low fear was operationally defined as FABQ-PA scores of 0 to 14 (Werneke et al. 2008).

Models also controlled for potential bias due to missing FS at discharge (loss to follow-up censoring) by using inverse probability of censoring (IPC) weighting (Robins, Hernan, and Brumback 2000). Calculation of the IPC weights was performed using a two-step procedure. In step 1, we fit a logistic regression model where the dependent variable takes the value of 1 if discharge FS measures for the patient are complete, and 0 if they are missing, using all patient baseline variables as covariates. In step 2, we used the inverse of the predicted probabilities of this logistic model as weights in our hierarchical regression outcome models. Thus, patients who, based on their characteristics, are less likely to have complete FS data were given more weight in estimating the effect model than those who are likely to have complete data. This approach is analogous to using survey weights, where subjects who are more likely to be selected into the study are given less weight in the analysis.

Theoretically, fraction/rate outcome variables, such as our discharge FS rate, are modeled with a (hierarchical) generalized linear model that allows for boundary values (Papke and Wooldridge 2008). In our case, however, the discharge FS outcome variable exhibited a mostly symmetric distribution around its mean with no values near zero and almost no values reaching the upper ceiling of 100, and linear models yielded very similar results and were faster to implement. Our results, therefore, show the estimates from the linear hierarchical models. For patient i receiving therapy from therapist j in clinic k , the hierarchical linear model of the discharge FS, Y_{ijk} , took the form:

$$Y_{ijk} = \beta X_{ijk} + \varphi_{jk} + \mu_k + \varepsilon_{ijk}$$

The models controlled for the set of risk adjusters X_{ijk} described above and used IPC weights to adjust for potential bias due to informative censoring. The error terms φ_{jk} and μ_k represent therapist and clinic random intercepts, respectively, that estimate the unexplained outcome variation attributable to the therapist and clinic, after adjusting for patient risk factors X_{ijk} . The random intercepts μ_k formed the basis for ranking the clinics. Estimation of the individual clinic intercepts μ_k standard error estimates was based on the empirical Bayes prediction method (Rabe-Hesketh and Skrondal 2008). For the single-condition models, we excluded clinics with less than eight patients with that condition to obtain reliable confidence interval estimates for our clinic random intercepts.

The estimated clinic random effects are displayed graphically for the overall model to more easily evaluate whether there is enough variability in the estimates to allow their use for benchmarking clinic providers. Similarly, the estimated clinic quality rankings, based on each of the two individual impairment groups, are shown against each other to illustrate to what degree clinic rankings are correlated across impairment groups. For this individual impairment comparison, we restricted our analysis to the 306 (57 percent) clinics with a minimum of eight observed outcomes in each condition.

RESULTS

During the July 2006–June 2008 2-year period, the FOTO database identified 90,392 patients who received outpatient rehabilitation from 2,040 therapists at 538 clinics, and who completed the self-reported surveys via CAT. The number of patients per clinic during these 2 years ranged from 30 to 1,943,

with an average (SD) of 168 (228.8) and median of 88 patients. Over a third of these patients, $N = 33,279$ (36.8 percent), did not complete a discharge survey, leaving a total of $N = 57,113$ (57.2 percent) in our outcome analytic sample. Among those in our analytic sample, average age was 54; almost two thirds (60.7 percent) were female; the most common conditions were lumbar (25.3 percent), shoulder (19.2 percent), and knee (17.0 percent); almost a third (30.6 percent) had surgery prior to therapy intake; and about half (50.5 percent) started therapy more than 90 days after the onset of the condition (Table 1). Patient baseline characteristics of those without a discharge survey were mostly similar to those who completed the discharge survey. The most notable differences were that they were almost 3 years younger (average age 51.3); experienced more lumbar and less shoulder and knee conditions; 3 percent more had onset beyond 90 days (53.2 percent); and 4 percent fewer had surgery (26.4 percent). Table 1 also shows the characteristics of the two most common conditions, lumbar and shoulder, used in our condition-specific benchmarking analyses. Patients with lumbar impairments improved an average of 14 points during treatment to achieve a functional staging level 4, interpreted as having little difficulty performing usual work or household activities and hobbies (Wang et al. 2010b). Patients with shoulder impairments improved an average of 17.9 points to reach a functional staging level 4, which is typical of someone who can perform routine daily activities using the affected arm with no difficulty (Wang et al. 2010b).

A comparison of the intake and discharge FS outcomes by impairment condition (Table 2) also shows modest differences in the intake FS measures among those whose discharge FS is observed and those with censored discharge FS. The logistic model used to construct the IPC weights shows that the probability of the discharge FS being observed is higher among older, female patients, those who had surgery, with lower number of comorbidities, low levels of fear/avoidance, and shorter times between onset of the condition and the start of therapy.

After adjusting for patient-level case mix, the clinic effect explained 9.1 percent of the total variation, while the therapist explained an additional 2.5 percent (Table 3). Higher intake FS was associated with higher discharge FS (close to .5 unit per additional intake FS unit), while later onset, higher number of comorbidities, and having Medicaid or workers compensation insurance (relative to those with HMO insurance) are among the factors most negatively associated with discharge FS. The clinic random intercept estimates showed clear and significant outcome differences between clinics. The estimated discharge FS attributable to clinics ranged from 0.4 to 24 units (in the 0–100

Table 1: Baseline Characteristics of Outpatient Rehabilitation Patients

	Analytic Sample			Censored Sample*	Total	IPC-Weighted Analytic Sample Total
	Lumbar	Shoulder	Total			
N	14,464	10,943	57,113	33,279	90,392	57,113
Age (mean ± SD)	54.2 ± 16.9	55.5 ± 14.9	54.0 ± 16.2	51.3 ± 16.2	53.0 ± 16.3	53.1 ± 16.3
Male (%)	41.1	45.4	39.3	39.5	39.4	39.4
Impaired body part (%)						
Lumbar	100		25.5	27.9	26.4	26.5
Shoulder		100	19.3	17.3	18.5	18.6
Knee			17.0	15.3	16.4	16.5
Cervical			11.5	12.0	11.7	11.7
Foot/ankle			9.5	9.3	9.4	9.4
Hip			6.7	6.7	6.7	6.7
Wrist/hand			5.0	5.4	5.2	5.0
Elbow			3.1	3.2	3.2	3.1
Ribs			2.1	2.5	2.2	2.2
Craniofacial			0.3	0.4	0.3	0.3
Onset (days) (%)						
0–7	6.1	4.5	5.0	4.8	5.0	5.0
8–14	7.9	5.4	6.7	6.5	6.6	6.6
15–21	8.8	6.6	8.0	7.7	7.9	8.0
22–90	25.2	29.6	29.8	27.8	29.0	29.4
91–182	12.4	19.9	14.8	15.0	14.9	14.9
>183	39.6	34.0	35.7	38.2	36.6	36.1
Functional comorbidity index (%)						
Low	18.4	30.5	26.1	26.0	26.1	26.4
Medium-low	25.6	24.7	25.2	25.1	25.2	25.2
Medium-high	35.7	29.5	31.8	31.0	31.5	31.4
High	20.3	15.3	16.9	17.9	17.3	17.0

continued

Table 1. *Continued*

	Analytic Sample		Censored Sample*	Total	Total	IPC-Weighted Analytic Sample Total
	Lumbar	Shoulder				
Had surgery (%)	17.3	30.6	26.4	30.6	29.0	29.3
Low fear/avoidance (%)	59.9	66.2	64.0	64.1	64.0	63.7
Insurance type (%)						
HMO	10.4	10.4	13.4	10.5	11.6	10.7
PPO	34.3	37.8	38.7	37.2	37.7	37.9
Medicare	25.5	22.1	18.3	22.5	20.9	21.2
Workers compensation	10.4	10.6	9.4	10.5	10.1	10.1
Indemnity	3.1	2.9	3.3	3.0	3.1	3.1
Medicaid	2.1	1.6	4.3	1.9	2.8	2.4
Other	14.2	14.6	12.6	14.4	13.8	14.6

*The censored sample consists of those patients with no discharge functional status measures.

Table 2: Functional Status at Intake and at Discharge, by Impairment Condition

	<i>Functional Status* (Mean ± SD)</i>		
	<i>Analytic Sample</i>		<i>Censored Sample[†]</i>
	<i>Intake</i>	<i>Discharge</i>	
Impaired body part			
Lumbar	50.2 ± 12.8	64.2 ± 16.3	48.9 ± 13.7
Shoulder	51.0 ± 15.3	68.9 ± 14.9	52.7 ± 14.8
Knee	42.6 ± 14.2	61.1 ± 16.7	44.3 ± 15.1
Cervical	51.6 ± 13.9	65.5 ± 17.7	50.6 ± 14.4
Foot/ankle	47.2 ± 14.2	64.2 ± 16.9	47.4 ± 15.3
Hip	48.0 ± 14.9	62.4 ± 17.2	47.9 ± 16.2
Wrist/hand	48.7 ± 11.0	60.7 ± 12.1	49.2 ± 12.0
Elbow	50.8 ± 9.8	62.6 ± 12.1	50.7 ± 11.1
Ribs	50.5 ± 13.0	65.0 ± 17.7	49.0 ± 13.8
Craniofacial	62.1 ± 17.5	73.6 ± 16.8	58.6 ± 20.0
Total	48.8 ± 14.2	64.4 ± 16.4	48.9 ± 14.6

*Functional Status ranges from 0 to 100.
[†]The censored sample consists of those patients with no discharge functional status measures.

scale); and performance differences between clinics ranked in the lowest and highest performance quartiles were significant at the 99 percent level. Figure 1 shows the clinic random intercept estimates (plus their 95 percent confidence interval estimate) that estimate the clinic-attributable average improvement for patients with all types of conditions (using the average clinic to center the distribution of clinics at 0). The figure shows data for one in every 10 clinics for clarity. The differences in confidence intervals among clinics indicate that some clinics may be achieving more uniform risk-adjusted outcomes for all their patients than other clinics. Of the 532 clinics ranked, a total of 145 (27 percent) clinics had their entire 95 percent CI not overlapping zero, 71 (13.3 percent) entirely below, and 74 (14 percent) entirely above zero. Censoring of discharge FS was 36 percent, but correcting for censoring had small effects in the rankings of most clinics (although a few clinics had relatively larger changes). Ranking was not associated with the number of visits and their duration (days from first to last visit). Using quartiles of clinic rankings, patients in clinics in the lowest ranking quartile received 12.2 visits over 47 days on average, compared to 11 over 42 days in the second quartile, 11.4 over 46 days in the third quartile, and 11.3 over 41 days among the highest ranked clinics. Among those without discharge FS, they received 7.4 visits over 32 days on average, and they were very similar across quartiles of clinic rankings.

Table 3: Hierarchical Model of Discharge Functional Status, All Conditions

<i>Variable</i>	<i>Coefficient</i>	<i>(95% CI)</i>
Intake functional status (per unit, 0–100)	0.47	(0.46, 0.49) [‡]
Age		
18–41	Ref.	
42–52	–2.24	(–2.95, –1.83) [‡]
53–64	–2.92	(–3.30, –2.54) [‡]
65–89	–3.03	(–3.60, –2.46) [‡]
Male	1.59	(1.36, 1.83) [‡]
Impaired body part		
Lumbar	–0.11	(–0.61, 0.40)
Shoulder	4.24	(3.73, 4.76) [‡]
Knee	0.25	(–0.31, 0.80)
Cervical	Ref.	
Foot/ankle	0.53	(–0.03, 1.10)
Hip	–0.36	(–1.01, 0.29)
Wrist/hand	–2.73	(–3.34, –2.12) [‡]
Elbow	–2.64	(–3.33, –1.95) [‡]
Ribs	0.01	(–0.97, 0.99)
Craniofacial	2.72	(0.31, 5.14) [†]
Onset (days)		
0–7	Ref.	
8–14	–3.12	(–3.83, –2.42) [‡]
15–21	–4.86	(–5.50, –4.22) [‡]
22–90	–2.38	(–3.17, –1.59) [‡]
91–182	–6.20	(–6.85, –5.56) [‡]
>183	–8.08	(–8.70, –7.46) [‡]
Functional comorbidity index		
Low	Ref.	
Medium-low	–0.97	(–1.32, –0.63) [‡]
Medium-high	–2.49	(–2.85, –2.12) [‡]
High	–5.34	(–5.76, –4.92) [‡]
Had surgery	–0.16	(–0.14, 0.45)
Low level of fear/avoidance	0.80	(0.56, 1.04) [‡]
Insurance type		
HMO	Ref.	
PPO	0.47	(–0.04, 0.98)
Medicare	–1.15	(–1.72, –0.58)
Workers compensation	–4.13	(–4.79, –3.47) [‡]
Indemnity	0.27	(–0.68, 1.22)
Medicaid	–4.36	(–5.35, –3.37) [‡]
Other	–0.66	(–1.22, –0.10) [†]
Intercept	72.69	(71.67, 73.71) [‡]
Random effects intercepts	Variance	% of Total variance
Clinic	17.87	9.09
Therapist	5.00	2.54
Residual	173.91	88.37

[†] $p < .05$; [‡] $p < .001$.

Figure 1: Ranking of Clinics Using All Patient Conditions: Clinic Random Intercepts and Their 95 Percent Confidence Interval*

Note. * Displayed every 10th ranked clinic to avoid overcrowding of lines.

Figure 2 shows the comparison of rankings obtained separately for lumbar and shoulder impairments, the two most prevalent conditions. The vertical and horizontal grid lines provide the quintiles of the rankings for each impairment category. While clinics in one quintile ranking for one impairment tend to be in a similar quintile for the other impairment (correlation between clinic rankings was 0.54), it is clear that some clinics ranked in the top quintile for shoulder (horizontal axis) are ranked in the lowest quintile for lumbar impairment (vertical axis), illustrating a lack of consistency in rankings across individual impairment categories.

DISCUSSION

A critical requirement for successful implementation of P4P reimbursement incentives is the availability of a quality measure with sufficient *variability*: defined as the ability to discriminate between high-quality and low-quality

Figure 2: Ranking of Clinics for Lumbar versus Shoulder Impairments*

Note.* Vertical and horizontal lines depict the quintiles of the rankings.

facilities with sufficient variation in quality across providers. The results of this study indicate that a validated measure of FS can be used to estimate risk-adjusted rehabilitation outcomes at the clinic level that provide enough variability to differentiate between clinics of high and low quality. Using the data from FOTO, our ranking of clinics using nine risk-adjusted, impairment-specific FS outcomes measures allows comparisons of providers both within and across different types of patient impairments. Our FS-based model was able to classify 14 percent of clinics as better than average and 13.3 percent as worse than average, which compares very favorably with other comparative provider models like Medicare's web-based Hospital Compare, which, for 2008, suggested that of 4,311 hospitals, none were worse than average, and nine were better than average for acute myocardial infarction (AMI) mortality (Silber et al. 2010). As CMS and other payers move toward pay-for-performance reimbursement mechanisms, these results can help inform their efforts.

It is not clear whether the claims-based outcome reporting system recently implemented by CMS in July 2013 will offer a similarly valid

measure of FS. That system requires providers of Part B covered therapy services to classify the primary functional status deficit for which the patient is seeking therapy, and indicate the percentage of impairment using a 7-point severity scale (Centers for Medicare & Medicaid Services 2012). Providers may use any measurement tool (regardless of the measures employed or the evidence base behind it) or simply clinical judgment in assigning a score for therapy outcomes. Medicare outpatient therapy claims do not require the provider to identify which measurement tool was used to calculate the severity rating. Research utilizing the claims-based outcome reporting system is needed to examine its validity. Although a 2013 MedPAC report recommended that Congress direct the Secretary to “use the information collected using this tool to measure the impact of therapy services on functional status, and provide the basis for development of an episode-based or global payment system” (Medicare Payment Advisory Commission 2013), there are threats to the validity of data derived from this system (Resnik 2013).

In our study, we observed wide variation in provider rankings by specific impairment types. This variation in ranking may be due, in part, to provider specialization. For example, a clinic that works primarily with patients who have shoulder impairments may achieve far better outcomes with those patients than they do with patients with other types of impairments. Pay-for-performance systems based on clinic rankings may inadvertently encourage such specialization; if a provider can maximize its quality ranking by only accepting certain impairment types of patients, it may be less inclined to treat other types of patients. As such, while specialization may improve quality and outcomes for some patients, it may also limit access to care for other patients. That said, in our analysis of shoulder and lumbar conditions, there was no correlation between specialization (percent of all patients that have a single condition within the clinic) and clinic ranking in that single condition. However, we acknowledge that the FOTO CAT dataset does not contain data from all patients treated at the clinic, and we have no way of knowing if the proportion of patients with specific types of impairments accurately reflects the true distribution of patients within each clinic.

The implications of this study extend beyond payment reform itself. Provider ranking based on risk-adjusted functional status reporting can be useful as a quality metric independent of a specific reimbursement mechanism. Internal quality improvement initiatives can be spearheaded based upon information gleaned from benchmarked quality rankings for specific patient conditions. Public reporting of quality rankings, for example, would allow patients to “vote with their feet,” choosing a provider based on its functional

status effectiveness. In this way, a patient's informed choice can yield a performance incentive similar to a payment reform.

However, measures of FS are already available with strong evidence of validity. Examples mentioned in the CMS rule-making include the FOTO instrument used in our study, the Activity Measure for Post-Acute Care, the American Physical Therapy Association's "Outpatient Physical Therapy Improvement in Movement Assessment Log" measure, and the American Speech Language and Hearing Association's National Outcomes Measurement System (Centers for Medicare & Medicaid Services 2012). For this study, we used data from FOTO because their measure was designed specifically for the outpatient population, is capable of spanning the entire spectrum of outpatient therapy needs, and the database is large, robust, and available for researchers. Further, FOTO's FS measures have been approved by the National Quality Forum as a measure of provider quality (National Quality Forum 2015). Our analysis did not incorporate service utilization patterns or costs, a core component of a VBP model. It is not clear from our analysis whether those providers who have better rankings were also providing more cost-effective care. Further study examining the relationships between provider quality rankings derived from FS and costs of care are warranted.

One limitation of our study is that clinic rankings were derived exclusively from patient self-reported functional status data. Therefore, we are unable to compare rankings generated using self-reported FS data to rankings generated using therapist ratings of functional performance because functional performance data were not available in the FOTO database. Although self-report instruments like FOTOS have a strong body of evidence supporting their ability to accurately measure FS and represent the patient perspective, performance measurement is an important dimension of functional assessment that is typically valued by payers. It is possible that clinic rankings would differ if they were derived from performance-based functional status measures, or a combination of self-report and performance measures. That said, there are ample data to demonstrate that self-report measures are reliable, valid, and moderately correlated with physical performance (Sherman and Reuben 1998; Sayers et al. 2004; Coman and Richardson 2006; Poole, Cordova, and Brower 2006; Denking et al. 2009; Farag et al. 2012; Papathanasiou et al. 2014).

The generalizability of our findings may be limited to patients who complete FS report by CAT and to clinics that support this mode of administration. FOTO FS measures are administered either by paper and pencil or by CAT. Our analysis was only of CAT data, and it did not include data collected by

paper-and-pencil method. To the best of our knowledge from discussions with FOTO, the selection of paper-and-pencil or CAT administration is made by the clinic and is not related to characteristics of patients. Patients who utilize the CAT measures are believed to be similar to those that use the paper-and-pencil survey forms. The dissemination of the CAT across clinics, however, was likely not uniform, with more technology-savvy clinics adopting CAT faster than others, which may have influenced which clinics are represented in our analyses. At present, most clinics use CAT surveys.

A further limitation is the generalizability of our findings to all outpatient therapy practices in the United States or to the clinic rankings specific to the Medicare population. We do not know how similar patients or clinics in the FOTO data are to all patients or clinics in the United States because no national comparison data are available. Furthermore, our study included therapy users of all ages, so it is unlikely but possible that clinic rankings generated from Medicare patients only could differ from those obtained using the more general FOTO population. In addition, because collection of FOTO outcomes data is voluntary, there is no guarantee that patients participating within each clinic are representative of the entire clinic population. Finally, while IPC weighting minimizes the imbalance created by censoring between the analytic uncensored sample and the population sample, this methodology may not address imbalances in other risk factors not available to us.

The FOTO data used for this project were collected independently of the incentives of a payment system. Should such measures be included in a future payment mechanism, it is possible that patterns of FS reporting may change. Patient self-report measures may minimize bias introduced by incentives of a payment system as patients are not directly impacted by the reimbursement. The degree to which payment incentives could influence censoring and outcome measures is not known and may well depend on the type and size of the payment incentives.

Lastly, we compared provider benchmarks using ranks derived from models employing different nonequivalent measures, and this may have impacted the findings of the overall-conditions model. The FS status scores in our analysis are not mathematically equivalent. A score of 50 for one of the measures does not necessarily equal a score of 50 for a different measure. That said, our overall-impairment conditions model includes condition-specific indicators to allow for differential rates of improvement and differences in measurement central tendency. Furthermore, our comparison of providers was based upon ranking, not absolute measures. For the single condition rankings, lack of comparability across conditions is not an issue.

CONCLUSIONS

The results of this study indicate that the use of validated measures of FS, such as those in the FOTO database, provide a good basis for estimating risk-adjusted rehabilitation outcomes at the clinic level with adequate variability that clearly separate high-quality from low-quality facilities. As CMS and other payers in the United States move toward VBP and P4P reimbursement models, we believe that these efforts should utilize validated measures of FS outcomes to assess quality. The benchmarking methods presented in this paper provide a viable method of estimating clinic quality and ranking outcomes at the clinic level. Such an approach could lay the groundwork for future VBP reform efforts.

ACKNOWLEDGMENTS

This study is dedicated to the memory of Dennis L. Hart. His encouragement and ideas were instrumental in the origins of the study, and his enthusiasm and academic curiosity will be missed.

Joint Acknowledgment/Disclosure Statement: The data utilized in this research were provided by Focus on Therapeutic Outcomes Inc (FOTO). The authors did contact FOTO to clarify aspects of their data collection and measurements, but the research of this study was conducted without additional guidance or financial support from FOTO or any other entity.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Binkley, J. M., P. W. Stratford, S. A. Lott, and D. L. Riddle. 1999. "The Lower Extremity Functional Scale (LEFS): Scale Development, Measurement Properties, and Clinical Application. North American Orthopaedic Rehabilitation Research Network." *Physical Therapy* 79 (4): 371–83.
- The Centers for Medicare & Medicaid Services. 2013. "Roadmap for Implementing Value Driven Healthcare in the Traditional Medicare Fee-for-Service Program." [accessed December 5, 2013]. Available at http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/VBPRoadmap_OEA_1-16_508.pdf

- Centers for Medicare & Medicaid Services. 2012. "Implementing the Claims-Based Data Collection Requirement for Outpatient Therapy Services. Section 3005(g) of the Middle Class Tax Relief and Jobs Creation Act (MCTRJCA) of 2012 (CMS-100-04, Transmittal 2622). Federal Register 77, No. 222 (November 16, 2013), pp. 68958–68978" [accessed on May 5, 2014, 2012]. Available at <https://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/Downloads/R2622CP.pdf>
- Ciolek, D. E., and W. Hwang. 2008. *Outpatient Therapy Alternative Payment Study 2 (OTAPS 2) Task Order*. Baltimore, MD: Centers for Medicare & Medicaid Services (CMS).
- Ciolek, D., and W. Hwang. 2010. *Short Term Alternatives for Therapy Services (STATS) Task Order: Final Report on Short Term Alternatives*. Baltimore: MD Computer Sciences Corporation.
- Coman, L., and J. Richardson. 2006. "Relationship between Self-Report and Performance Measures of Function: A Systematic Review." *Canadian Journal of Aging* 25 (3): 253–70.
- Denkinger, M. D., W. Igl, L. Coll-Planas, J. Bleicher, T. Nikolaus, and M. Jamour. 2009. "Evaluation of the Short Form of the Late-Life Function and Disability Instrument in Geriatric Inpatients-Validity, Responsiveness, and Sensitivity to Change." *Journal of the American Geriatrics Society* 57 (2): 309–14.
- Farag, I., C. Sherrington, S. J. Kamper, M. Ferreira, A. M. Moseley, S. R. Lord, and I. D. Cameron. 2012. "Measures of Physical Functioning after Hip Fracture: Construct Validity and Responsiveness of Performance-Based and Self-Reported Measures." *Age and Ageing* 41 (5): 659–64.
- Focus on Therapeutic Outcomes. 2015. "NQF Measure Specifications" [accessed March 31, 2015]. Available at <http://www.fotoinc.com/science-of-foto/nqf-measure-specifications>
- Groll, D. L., T. To, C. Bombardier, and J. G. Wright. 2005. "The Development of a Comorbidity Index with Physical Function as the Outcome." *Journal of Clinical Epidemiology* 58 (6): 595–602.
- 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. Washington, DC: Health & Human Services/Centers for Medicare & Medicaid Services.
- Hart, D. L., and B. D. Wright. 2002. "Development of an Index of Physical Functional Health Status in Rehabilitation." *Archives of Physical Medicine and Rehabilitation* 83 (5): 655–65.
- Hart, D. L., J. E. Mioduski, and P. W. Stratford. 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): 629–38.
- Hart, D. L., K. F. Cook, J. E. Mioduski, C. R. Teal, and P. K. Crane. 2006a. "Simulated Computerized Adaptive Test for Patients with Shoulder Impairments Was Efficient and Produced Valid Measures of Function." *Journal of Clinical Epidemiology* 59 (3): 290–8.

- Hart, D. L., J. E. Mioduski, M. W. Werneke, and P. W. Stratford. 2006b. "Simulated Computerized Adaptive Test for Patients with Lumbar Spine Impairments Was Efficient and Produced Valid Measures of Function." *Journal of Clinical Epidemiology* 59 (9): 947–56.
- Hart, D. L., Y. C. Wang, P. W. Stratford, and J. E. Mioduski. 2008a. "A Computerized Adaptive Test for Patients with Hip Impairments Produced Valid and Responsive Measures of Function." *Archives of Physical Medicine and Rehabilitation* 89 (11): 2129–39.
- . 2008b. "Computerized Adaptive Test for Patients with Knee Impairments Produced Valid and Responsive Measures of Function." *Journal of Clinical Epidemiology* 61 (11): 1113–24.
- Hart, D. L., D. Deutscher, M. W. Werneke, J. Holder, and Y. C. Wang. 2010a. "Implementing Computerized Adaptive Tests in Routine Clinical Practice: Experience Implementing CATs." *Journal of Applied Measurement* 11 (3): 288–303.
- Hart, D. L., Y. C. Wang, K. F. Cook, and J. E. Mioduski. 2010b. "A Computerized Adaptive Test for Patients with Shoulder Impairments Produced Responsive Measures of Function." *Physical Therapy* 90 (6): 928–38.
- Hart, D. L., M. W. Werneke, Y. C. Wang, P. W. Stratford, and J. E. Mioduski. 2010c. "Computerized Adaptive Test for Patients with Lumbar Spine Impairments Produced Valid and Responsive Measures of Function." *Spine* 35 (24): 2157–64.
- Lyda-McDonald, B., B. Silver, and B. Gage. 2012. *Developing Outpatient Therapy Payment Alternatives (DOTPA): Project Year 4 Annual Report*. Research Triangle Park, NC: RTI International.
- Medicare Payment Advisory Commission. 2013. *Mandated Report: Improving Medicare's Payment System for Outpatient Therapy Services*. Washington, DC: Medicare Payment Advisory Commission.
- National Quality Forum. 2015. "Measuring Performance" [accessed on April 30, 2015]. Available at http://www.qualityforum.org/Measuring_Performance/Measuring_Performance.aspx
- Papathanasiou, G., S. Stasi, L. Oikonomou, I. Roussou, E. Papageorgiou, E. Chronopoulos, N. Korres, and N. Bellamy. 2014. "Clinimetric Properties of WOMAC Index in Greek Knee Osteoarthritis Patients: Comparisons with Both Self-Reported and Physical Performance Measures." *Rheumatology International* 35 (1): 115–23.
- Papke, L. E., and J. M. Wooldridge. 2008. "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates." *Journal of Econometrics* 145 (1–2): 121–33.
- Poole, J. L., K. J. Cordova, and L. M. Brower. 2006. "Reliability and Validity of a Self-Report of Hand Function in Persons with Rheumatoid Arthritis." *Journal of Hand Therapy* 19 (1): 12–6, quiz 17.
- Rabe-Hesketh, S., and A. Skrondal. 2008. *Multilevel and Longitudinal Modeling Using STATA*. College Station, TX: Stata Press.
- Resnik, L., and D. L. Hart. 2003. "Using Clinical Outcomes to Identify Expert Physical Therapists." *Physical Therapy* 83 (11): 990–1002.

- . 2004. "Influence of Advanced Orthopaedic Certification on Clinical Outcomes of Patients with Low Back Pain." *Journal of Manual and Manipulative Therapy* 12 (1): 32–41.
- Resnik, L., and G. M. Jensen. 2003. "Using Clinical Outcomes to Explore the Theory of Expert Practice in Physical Therapy." *Physical Therapy* 83 (12): 1090–106.
- Resnik, L., Z. Feng, and D. L. Hart. 2006. "State Regulation and the Delivery of Physical Therapy Services." *Health Services Research* 41 (4 Pt 1): 1296–316.
- Resnik, L., D. Liu, D. L. Hart, and V. Mor. 2008. "Benchmarking Physical Therapy Clinic Performance: Statistical Methods to Enhance Internal Validity When Using Observational Data." *Physical Therapy* 88 (9): 1078–87.
- Robins, J. M., M. A. Hernan, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5): 550–60.
- Sayers, S. P., A. M. Jette, S. M. Haley, T. C. Heeren, J. M. Guralnik, and R. A. Fielding. 2004. "Validation of the Late-Life Function and Disability Instrument." *Journal of the American Geriatrics Society* 52 (9): 1554–9.
- Sherman, S. E., and D. Reuben. 1998. "Measures of Functional Status in Community-Dwelling Elders." *Journal of General Internal Medicine* 13 (12): 817–23.
- Silber, J. H., P. R. Rosenbaum, T. J. Brachet, R. N. Ross, L. J. Bressler, O. Even-Shoshan, S. A. Lorch, and K. G. Volpp. 2010. "The Hospital Compare Mortality Model and the Volume-Outcome Relationship." *Health Services Research* 45 (5 Pt 1): 1148–67.
- Silver, B., B. Lyda-McDonald, H. Bachofer, and B. Gage. 2013. *Developing Outpatient Therapy Payment Alternatives (DOTPA): 2010 Utilization Report*. Research Triangle Park, NC: RTI International.
- Swinkels, I. C., C. H. van den Ende, D. de Bakker, P. J. Van der Wees, D. L. Hart, D. Deutscher, W. J. van den Bosch, and J. Dekker. 2007. "Clinical Databases in Physical Therapy." *Physiotherapy Theory and Practice* 23 (3): 153–67.
- Waddell, G., M. Newton, I. Henderson, D. Somerville, and C. J. Main. 1993. "A Fear-Avoidance Beliefs Questionnaire (FABQ) and the Role of Fear-Avoidance Beliefs in Chronic Low Back Pain and Disability." *Pain* 52 (2): 157–68.
- Wang, Y. C., D. L. Hart, P. W. Stratford, and J. E. Mioduski. 2009a. "Clinical Interpretation of a Lower-Extremity Functional Scale-Derived Computerized Adaptive Test." *Physical Therapy* 89 (9): 957–68.
- . 2009b. "Clinical Interpretation of Computerized Adaptive Test Outcome Measures in Patients with Foot/Ankle Impairments." *Journal of Orthopaedic and Sports Physical Therapy* 39 (10): 753–64.
- Wang, Y. C., D. L. Hart, K. F. Cook, and J. E. Mioduski. 2010a. "Translating Shoulder Computerized Adaptive Testing Generated Outcome Measures into Clinical Practice." *Journal of Hand Therapy* 23 (4): 372–82, quiz 83.
- Wang, Y. C., D. L. Hart, M. Werneke, P. W. Stratford, and J. E. Mioduski. 2010b. "Clinical Interpretation of Outcome Measures Generated from a Lumbar Computerized Adaptive Test." *Physical Therapy* 90 (9): 1323–35.

- Ware Jr, J., M. Kosinski, and S. D. Keller. 1996. "A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity." *Medical Care* 34 (3): 220–33.
- Ware, J. E. Jr, K. Snow, M. Kosinski, and B. Gandek. 1993. *SF-36 Health Survey: Manual and Interpretation Guide*. Boston: The Health Institute New England Medical Center.
- Werneke, M. W., D. L. Hart, L. Resnik, P. W. Stratford, and A. Reyes. 2008. "Centralization: Prevalence and Effect on Treatment Outcomes Using a Standardized Operational Definition and Measurement Method." *The Journal of Orthopedic and Sports Physical Therapy* 38 (3): 116–25.
- Williamson, E. 2006. "Fear Avoidance Beliefs Questionnaire (FABQ)." *Australian Journal of Physiotherapy* 52 (2): 149.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.