Creating Unidimensional Global Measures of Physician Practice Quality Based on Health Insurance Claims Data

Grant R. Martsolf, Adam C. Carle, and Dennis P. Scanlon

Objective. To explore the extent to which commonly used claims-based process quality indicators can be used to create an internally valid global composite measure of physician practice quality.

Data Sources. Health insurance claims data (October 2007–May 2010) from 134 physician practices in Seattle, WA.

Study Design. We use confirmatory and exploratory factor analysis to develop theory- and empirically driven internally valid composite measures based on 19 quality indicators.

Data Collection Methods. Health insurance claims data from nine insurance companies and self-funded employers were collected and aggregated by third-party organization.

Principal Findings. Our results did not support a single global measure using the entire set of quality indicators. We did identify an acceptable multidimensional model (RMSEA = 0.059; CFI = 0.934; TLI = 0.910). The four dimensions in our data were diabetes, depression, preventive care, and generic drug prescribing.

Conclusions. Our study demonstrates that commonly used process indicators can be used to create a small set of useful composite measures. However, the lack of an internally valid single unidimensional global measure has important implications for policy approaches meant to improve quality by rewarding “high-quality physicians.”

Key Words. Quality measurement, physician practices, composite measurement

The quality of physician practice care in the United States is less than optimal (Institute of Medicine: Committee on Quality of Health Care in America 2001). Policy makers, practitioners, and researchers have proposed various interventions to improve quality, ranging from public reporting to pay-for-performance programs. To varying degrees, each intervention includes measurement to compare quality across providers. To increase the comprehension of reports, work has focused on creating composite (summary) scores

that combine information from multiple individual quality indicators into a single score (Lied et al. 2001; Zaslavsky et al. 2002; Jacobs, Goddard, and Smith 2005; Caldis 2007; Reeves et al. 2007; Scholle et al. 2008; Shwartz et al. 2008; Teixeira-Pinto and Normand 2008; Kaplan et al. 2009; National Quality Forum 2009; *Aligning Forces for Quality* 2010; Peterson et al. 2010; Romano, Hussey, and Ritley 2010). Many stakeholders have proposed the development of “global” composite measures, “super composites,” or “roll-ups” that measure the quality of care over multiple domains (Martsolf, Scanlon, and Christianson 2013). For example, a global composite might assess the quality of care delivered by physician practices across diverse areas such as patient experience, preventive care, and chronic disease care using a single unidimensional measure. Such a global composite measure holds conceptual appeal as it can identify the “best” physician practices for purposes of providing pay-for-performance incentives or allowing patients to choose practices in their market.

A number of public reporting organizations and payers have proposed the inclusion of a global measure into public reports and incentives-based payment systems. For example, we recently reviewed public reports of physician practice quality published by eight community-based multistakeholder organizations and found that six included a composite measure, two of which developed and presented a global composite score to combine information from across various clinical domains (Maine Health Management Coalition 2013; Massachusetts Health Quality Partners 2013; Minnesota Community Measurement 2013; Oregon Health Care Quality Corp 2013; Pacific Business Group on Health 2013; Puget Sound Health Alliance 2013; The Health Alliance of Cincinnati 2013; Wisconsin Collaborative for Health Care Quality 2013). Physicians will also be able to participate in Medicare’s Value-Based Modifier Payment program. One of the applicable measures is a single global composite score theoretically measuring a broad construct of clinical quality (Centers for Medicare and Medicaid Services 2013).

However, the creation of such global composite measures is not without risk. When multiple indicators measuring distinct aspects of quality are inappropriately combined into a single measure, the resulting composite measure

Address correspondence to Grant R. Martsolf, Ph.D., M.P.H., R.N., RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213; e-mail: martsolf@rand.org. Adam C. Carle, Ph.D., is with the Cincinnati Children’s Hospital Medical Center, University of Cincinnati, Cincinnati, OH. Dennis P. Scanlon, Ph.D., is with Health Policy and Administration, Pennsylvania State University, University Park, PA.

is not useful or even completely uninterpretable. For example, when indicators measuring unrelated constructs are included in a single score, the high score on some indicators could “hide” low scores on other indicators or vice versa. In this case, the composite measure does not provide a clear quality signal. Inclusion of invalid composite measures could actually hurt quality reporting by leading to physician practice misclassification (Friedberg and Damberg 2012). With so much attention on forming composite measures currently, it is critical to establish the validity of composites before using and interpreting them (Carle et al. 2011).

There have been a limited number of studies assessing the statistical validity of creating composite measures of physician practice quality. Many extant studies in this area have focused on forming composite quality measures of health plans (Lied et al. 2001; Zaslavsky et al. 2002; Caldis 2007), general hospital care (Teixeira-Pinto and Normand 2008), emergency department care (Liu, Jain, and Shi 2013), or cardiac surgery (O’Brien et al. 2007). A small number of studies have focused on physician practices but only tested composite measures for a single clinical domain (i.e., diabetes) (Lipner et al. 2007; Kaplan et al. 2009). To our knowledge, no published work has examined the validity or usefulness of forming a unidimensional global composite measure of physician practice quality that assesses care across discrete clinical domains using common Healthcare Effectiveness Data and Information Set (HEDIS) indicators and with data drawn from aggregated health insurance claims. In this paper, we sought to help address this gap by answering three research questions:

1. To what extent can HEDIS process indicators be used to measure a single construct for the purpose of creating an internally valid global composite measure of physician practice quality?
2. In the absence of a valid global composite measure, can HEDIS process indicators be used to create a set of valid composite measures that assess care across various clinical domains?

Conceptual Framework

Quality of care is a latent construct that must be assessed using indicators or measures. Two specific challenges arise when stakeholders attempt to use such indicators. First, quality of care (as a latent construct) is always measured with a certain degree of measurement error (i.e., reliability), meaning the same indicators will not always produce the same results under the same circumstances

within the same practice. If reliability is low, practices might be misclassified as high or low performers on any given quality report. Second, quality is also multidimensional, meaning that a large number of indicators must be used to assess it. Therefore, reports of physician practice quality can be very large and complicated, which renders them difficult for stakeholders (e.g., patients and payers) to use and interpret.

In order to make quality reports more reliable and interpretable, report sponsors often create composite measures to reduce the number of indicators included in a report. Many previous studies have demonstrated that composite measures for a single clinical condition are more reliable than each of the individual indicators included in the composite measure (Scholle et al. 2008; Holmboe et al. 2010; Weng et al. 2010). In the case of single clinical conditions, composite measures provide clear advantages over individual measures in terms of the reliability of measurement.

Reliability presupposes that composite measures are in fact valid measures of some underlying construct. If that is true, the measure should provide useful information about the overall “quality” of a practice across multiple indicators. This becomes more challenging when indicators across multiple clinical conditions are included. If the multiple indicators are not related to each other, the resulting composite measure may not provide useful information. For example, practice performance on indicators of diabetes and preventive care may be unrelated to each other. When indicators of these two dimensions of care are combined into a global composite measure, the resulting composite would provide very little information about the performance of the practice on either of the dimensions.

Therefore, the development and use of reliable and valid composite measures would have different implications for various stakeholders, particularly payers and consumers. Many provider organizations are especially concerned that unreliable quality measures would misclassify practices as above or below average performance compared to their peers. Payers organizing pay-for-performance programs would be particularly interested in ensuring that practices are reliably assessed in order to ensure buy-in from providers. Likewise, invalid global composite measures would create concerns for both payers and consumers. Although an invalid global composite measure would provide some signal about the overall quality of a practice, stakeholders would be unable to decipher what was truly driving practices’ scores on the composite. Payers would not be able to assess what they were actually incentivizing in the assessed practice. This concern may be even more important for consumers as they may be especially interested in understanding whether a high

overall score would provide a signal about any individual dimension of quality in which they might be interested.

Both reliability and validity are important for assessing composite measures. However, because reliability assessments must presuppose validity, this paper focuses on assessing the validity of a global measure of the quality of physician practices.

METHODS

Data

We used deidentified health insurance claims data aggregated across nine health insurance companies and self-funded employers in the Puget Sound region of Washington. The data cover roughly 60 percent of commercially insured individuals in that area. The Puget Sound Health Alliance (PSHA), a community-based health care collaborative organization, constructed the dataset to generate a community-level public report called the *Community Checkup* (Puget Sound Health Alliance 2010). Health insurance companies and self-funded employers directly submitted claims data to a third-party firm that aggregated and processed the data to create the physician practice report. The firm produced a dataset that included the physician, practice, and group to which each patient was attributed, as well as each patient's eligibility for and receipt of 19 indicators of physician practice quality. Patients were attributed to practices based on a strict decision hierarchy based on (1) the number of Evaluation and Management visits; (2) the sum of relative value units; and (3) the most recent visit to a physician practice.

Because each individual patient was generally only eligible for an average of two measures, we could not estimate a multilevel measurement model that used patient-level data. In our data, patient-level values not in the dataset are not "missing" in the usual sense; rather, the patient was not eligible for these measures. It would be inappropriate to impute values for a patient for a measure for which they were not eligible. For example, it would not be appropriate to impute a value for cervical cancer screening to a man. Therefore, we used the patient-level claims data to construct practice-level scores for each of the quality indicators. We calculated the practice-level scores as the number of patients that received the recommended care for that care indicator divided by the number of patients that were eligible for the indicator. Using multilevel models, we computed the between and within center variance for each quality measure (Goldstein, Browne, and Rasbash 2002). For each quality measure,

we used these values and a clinic’s quality measure specific sample size to evaluate the reliability of each clinic-level quality measure for each clinic (Snijders and Bosker 1999). Across quality measures and clinics, the reliabilities were high (25th percentile of reliabilities ≥ 0.98). The primary data used for this analysis were included visits from June 2009–May 2010.

The quality indicators are based primarily on the HEDIS specifications and include nine for chronic care guideline compliance, four for preventive care, and six for appropriate use of resources (Table 1). In certain limited circumstances, the PSHA deviated from HEDIS specifications. For example, the indicators related to coronary artery disease were based on specifications

Table 1: Descriptive Statistics for the 19 Individual Quality Indicators

<i>Quality Measure Description</i>	<i>Total Number of Care Opportunities</i>	<i>Average Care Opportunities Per Practice (StD)</i>	<i>Overall Performance across All Practices</i>
Chronic disease care			
Lipid lowering therapy for coronary artery disease	4,676	32.70 (32.52)	77.90%
Cholesterol management for heart disease	3,412	19.01 (18.70)	85.89%
Eye exam for diabetes	14,338	107.10 (94.96)	62.10%
Hba1c test for diabetes	14,338	107.10 (94.96)	89.39%
LDL test for diabetes	14,338	107.10 (94.96)	83.67%
Monitoring diabetic nephropathy	14,338	107.10 (94.96)	82.39%
Appropriate medication management for asthma	3,067	22.89 (22.40)	90.77%
Acute medication management for depression	3,922	29.27 (23.69)	65.08%
Continuation medication management for depression	3,922	29.27 (23.69)	49.86%
Preventive care			
Cervical cancer screening	66,598	497.00 (503.48)	74.30%
Breast cancer screening	70,196	523.85 (512.87)	72.97%
Chlamydia screening	11,046	84.97 (82.44)	38.28%
Colorectal cancer screening	33,548	250.36 (240.06)	43.58%
Resource use			
Generic NSAIDs	21,603	161.22 (152.67)	92.33%
Generic PPIs	24,330	181.57 (170.16)	68.55%
Generic SSRIs	23,045	171.98 (153.76)	79.56%
Generic statins	41,868	312.45 (299.24)	60.63%
Avoidance of imaging for back pain	8,334	62.2 (30.87)	86.60%
Avoidance of antibiotics for laryngitis	6,277	46.84 (50.00)	26.12%

developed by the American College of Cardiology and the American Heart Association. Likewise, the PSHA developed the indicators related to the use of generic drugs.

The primary focus of this study was on fully functioning primary care practices (i.e., not women's health-only practices or specific specialty practices such as cardiology). However, we did not have an indicator for practice type. Therefore, we focused on practices that had patients with the full range of applicable conditions. Therefore, we only included physician practices that had at least five eligible patients for each of the 19 individual quality indicators, similar to minimum thresholds used in a previous study (Scholle et al. 2008). The final analytical sample file included 134 practices.

Analytic Approach

Overview of Analytic Approach. We used measurement models (e.g., confirmatory factor analysis) to investigate the "dimensionality" of 19 specific physician practice quality indicators. In this case, *dimensionality* refers to the extent to which multiple indicators can be used to assess a single construct or multiple constructs. Specifically, the measurement model approach is used to assess the extent to which a single factor accounts for the observed covariance among indicators. Models that "fit well" do a good job of reproducing the observed covariance matrix (Bollen 1989; Hu and Bentler 1998; Brown 2006). We use a series of confirmatory and exploratory factor analyses to determine the "fit" of our models. Although a model may be "well-fitting," that does not necessarily mean that all of the *individual* indicators measure the underlying construct(s) well. Therefore, we also assessed "communality" for each measure in the model; communality describes the proportion of variation in each measure explained by the measurement model.

Analysis. To investigate the dimensionality of our included measures, we used confirmatory factor analyses (CFA) and exploratory factor analyses (EFA) (Bollen 1989; McDonald 1999; Carle and Weech-Maldonado 2012). In CFA, the hypothesized relationships between the variables are determined by the researcher. In EFA, the model is derived empirically from the observed covariance between indicators. Given that our first interest was in the validity of creating a single global composite from the 19 indicators, we first tested a single unidimensional CFA model. A well-fitting model would provide broad evidence for measuring quality using all 19 of the indicators in a single score.

However, we suspected that a single unidimensional model would not fit the data. Thus, we also developed a theoretically based model that specified three factors: chronic disease care, preventive care, and resource use. Table 1 shows our hypothesized grouping of the indicators.

In case a theory-based confirmatory did not fit well, we planned to use EFA to determine what types of composite measures could be validly produced from the 19 measures. (Bollen 1989) In EFA, one estimates a series of models with increasing numbers of factors and minimal constraints (e.g., allowing correlated factors, cross-loadings, etc.) to identify a model that fits the data well and maximizes parsimony (minimal number of factors). We estimated the EFA using promax rotation (Brown 2006). We chose an EFA model that fit the data well, was parsimonious, and created conceptually meaningful factors. We then used a CFA to test the final EFA model while allowing the indicators to load only on single factors. We identified indicators as “belonging” to a factor (i.e., construct) when they had a loading of >0.4 on a factor in the EFA.

For all models, we evaluated fit using validated guidelines (Hu and Bentler 1998, 1999). We used the root mean square error of approximation (RMSEA), Comparative Fit Index (CFI), and the Tucker–Lewis Index (TLI). We considered RMSEA values less than 0.08 acceptable and values less than 0.06 ideal (Hu and Bentler 1998, 1999). For the CFI and TLI, we considered values greater than 0.90 acceptable and values greater than 0.95 ideal (Hu and Bentler 1998, 1999). We considered a model as fitting well when the majority of indices indicated good fit.

After we established a model that had acceptable fit, we then assessed communality. We considered a value of greater than 0.15 as indicating that a model explained a minimally sufficient amount of variance in an indicator (McDonald 1999). All analyses used MPlus (Muthen and Muthen 2007). We also allowed the factors to correlate with each other in all models.

We also performed a number of validation checks for our final model. Acknowledging that our model development process involved exploratory and confirmatory steps in the same sample, we conducted a check in new data to assess the final model’s validity. First, we fit the final four-factor model using the same data from a completely different period (July 2008–June 2009). Second, we fit the final four-factor model using roughly 18 months of data (October 2007–June 2009), which allowed for larger sample sizes within each practice. In this analysis, practices were retained if they had at least 30 patients for each indicator.

RESULTS

Table 1 provides descriptive statistics for each of the indicators among the study practices. It lists each indicator, the number of care opportunities (i.e., patient-indicator pairs) for that indicator, the average number of care opportunities per practice, and the average performance on that indicator across practices.

We first tested a unidimensional model hypothesizing that the 19 indicators measured a single quality construct. We specified that all 19 of the individual indicators loaded on (i.e., measured) a single overall factor. Though statistically identified, the model did not converge despite increasing the number of iterations and attempting different start values. This strongly suggests that a single factor did not account for the observed covariance. We then specified the second model, a theoretically based model that specified three content-specific quality constructs: chronic disease care, preventive care, and resource use. We allowed indicators to load only on one factor. Although this model converged, it did not fit well (RMSEA = 0.69; CFI = 0.86; TLI = 0.84). Thus, we rejected both the unidimensional and three-factor models that were tested using CFA.

We then conducted an exploratory factor analysis (EFA). We fit several models with increasing numbers of factors. The fit indices and the pattern of the loadings suggested that a five-factor solution fit best (RMSEA = 0.038). The loading pattern is shown in Table 2. We used the EFA model's parameters as a guide to develop a five-factor CFA model that would allow indicators to load on only one factor. We specified indicators as measuring a factor by examining the size of an indicator's loadings across each of the five factors (the EFA allowed cross-loadings).

We labeled the resulting factors based on what it appeared they measured. The first factor measured Appropriate Use of Treatments (avoidance of antibiotics for laryngitis, avoidance of imaging for back pain indicators); the second Depression Care (acute medication management for depression, and continuation medication management for depression indicators); the third Generic Drug Use (generic nonsteroidal anti-inflammatory drugs (NSAIDs), generic proton pump inhibitors (PPIs), generic selective serotonin reuptake inhibitors (SSRIs), and generic statins indicators); the fourth Diabetes Care (the HbA1c test for diabetes, low-density lipoprotein (LDL) test for diabetes, and monitoring diabetic nephropathy indicators); and the final factor appeared to represent Preventive Care (cervical cancer screening, breast cancer screening, and colorectal cancer screening indicators).

Table 2: EFA Factor Loadings

	<i>Appropriate Use</i>	<i>Depression Care</i>	<i>Generic Drug Prescribing</i>	<i>Diabetes Care</i>	<i>Cancer Screening</i>
Lipid-lowering therapy for coronary artery disease	0.134	−0.062	0.002	0.236	0.114
Cholesterol management for heart disease	0.103	−0.043	−0.111	0.173	0.070
Eye exam for diabetes	0.011	0.18	0.307	0.164	0.373
HbA1c test for diabetes	−0.021	−0.001	−0.117	0.827	0.020
LDL test for diabetes	−0.091	0.027	−0.092	0.987	−0.076
Monitoring diabetic nephropathy	0.114	−0.046	0.196	0.607	−0.038
Appropriate medication management for asthma	−0.001	0.003	−0.085	0.175	0.202
Acute medication management for depression	−0.128	0.891	0.01	0.064	−0.054
Continuation medication management for depression	0.117	0.855	−0.036	−0.055	−0.031
Cervical cancer screening	−0.011	0.046	0.077	−0.017	0.837
Breast cancer screening	0.291	0.013	−0.205	−0.021	0.566
Chlamydia screening	−0.17	−0.049	0.271	0.144	0.016
Colorectal cancer screening	−0.157	−0.17	0.013	−0.059	0.837
Generic NSAIDs	0.148	−0.089	0.471	0.085	−0.229
Generic PPIs	0.199	−0.031	0.77	−0.13	0.056
Generic SSRIs	−0.064	0.002	0.806	−0.115	0.006
Generic statins	0.013	0.036	0.728	0.035	0.043
Avoidance of imaging for back pain	0.56	0.056	0.136	−0.038	0.068
Avoidance of antibiotics for laryngitis	0.475	−0.092	−0.009	−0.06	−0.204

Note. Bold indicates factor loadings greater than 0.4.

Final Model

As a final step in model development, we used CFA to fit the five-factor model resulting from the EFA. While this model fit well (RMSEA = 0.059; CFI = 0.934; TLI = 0.910), the parameters indicated one additional modification. The communalities showed that the Appropriate Use factor described almost all of the variance in avoidance of imaging for back pain but hardly any of the variance in avoidance of antibiotics. This suggested that the appropriate use factor was best measured by avoidance of imaging for back pain only and that avoidance of antibiotics was a relatively unreliable indicator of the factor. Because we were seeking to evaluate composites, which by definition include more than one indicator, we dropped the Appropriate Use factor. The resulting final four-factor model exhibited good fit (RMSEA = 0.067; CFI = 0.951;

Table 3: Mean Score and Communalities from the Final Four-Factor Model

Depression care	
Acute medication management for depression	0.67
Continuation medication management for depression	0.80
Cancer screening	
Breast cancer screening	0.73
Cervical cancer screening	0.37
Colorectal cancer screening	0.49
Diabetes care	
HbA1c test for diabetes	0.83
LDL test for diabetes	0.69
Monitoring diabetic nephropathy	0.26
Generic drug prescribing	
Generic NSAIDs	0.23
Generic PPIs	0.56
Generic SSRIs	0.63
Generic statins	0.52

TLI = 0.932). Furthermore, the factors satisfactorily explained the variance in the indicators based on the communalities (Table 3). In sum, our final model included four factors: Generic Drug Use, Diabetes Care, Depression Care, and Preventive Care.

Acknowledging that our model development process involved exploratory and confirmatory steps in the same sample, we conducted a check in new data to assess the final model’s validity. When we retested the model using an entirely different period, the model fit was satisfactory (RMSEA = 0.057; CFI = 0.944; TLI = 0.919). Also, when we retested the model on a large dataset that allowed for stricter practice inclusion criteria (81 practices with at least 30 observations per measure), the model fit was also satisfactory (RMSEA = 0.069; CFI = 0.945, TLI = 0.925).

DISCUSSION

We examined the creation of an internally valid global composite measure of physician practice quality from commonly used HEDIS indicators. Our results did not support the psychometric validity of a single unidimensional composite. Thus, we investigated the validity of creating multiple measures of physician practice quality the individual indicators. We did identify an acceptable multidimensional model, showing the presence of four more specific forms of physician practice quality rather than a single overarching construct of physician practice quality.

Although our results suggest that multiple HEDIS measures cannot be used to create a psychometrically valid unidimensional measure, researchers and practitioners (Nietert et al. 2007) have argued for the utility of a broad quality construct. Many organizations are pursuing the use of global measures of physician practice quality. Our results may call into question efforts to create and use single unidimensional measures of physician practice quality, as using such measures can lead to spurious conclusions about quality by hiding important aspects of quality and to increased physician misclassification by exacerbating the measurement error inherent in any given measure. Particularly, performance on an invalid global measure of physician practice quality may obscure practices' performance on more specific areas of clinical care.

Such concerns are more salient for different purposes for which composite measures might be used. For example, when composites are used to help patients choose a provider, especially patients with a number of health care needs, it may be especially important to ensure that the composite measures a single construct. Such an application would require a "reflective" composite that assumes that a single underlying construct such as "quality" describes the variation in the observed individual indicators, such is the approach used in this study (Coltman et al. 2008). The use of a unidimensional composite measure that does not measure a single construct might be more appropriate, however, if the measure is intended only to summarize observed performance across a small number of indicators that stakeholders consider to be representative of a good care. Such "formative" composite measures might be appropriate for instances such as pay-for-performance programs (Coltman et al. 2008). It is important to note, however, that previous research (Martsof, Scanlon, and Christianson 2013) suggests that the use of invalid composite measures may not have support across all stakeholder groups even for applications for which "formative" approaches might otherwise be appropriate. Purpose and context are important considerations when creating a single unidimensional composite measure of physician practice quality.

If unidimensional measures are to be used, some precautions should be taken to ensure multistakeholder buy-in. Entities that publicly report quality measures should include not only the unidimensional global measures but also other condition-specific composites as well as the constituent indicators. The reports might present the unidimensional measure and then allow report users to investigate other measures to better understand how practices earned their score on the unidimensional measure. Alternatively, payers might structure payments such that practices are paid based on the global measure as well as other condition-specific dimensions that are of specific importance to the

payer or the practices. However, rewarding practices for performance on the condition-specific composites would undermine the reasons that the global composite was created in the first place. This suggests that without further testing and development, such global measures may simply not be ready for extensive and high-reward usage.

Our results suggest that stakeholders can, however, use relatively smaller, clinically consistent indicator sets to create specific composite measures, such as the ones identified in our data. This finding was consistent across multiple checks, including varying sample size and using data from different years, suggesting that these findings are reliable, valid, and stable. Many of the composite measures of physician practice quality that are being formed in the field have generally been consistent with this suggestion, focusing on condition-specific areas, such as diabetes or heart disease. For example, practices in CMS's Accountable Care Organization will be assessed partly based on their performance on diabetes and coronary artery disease composite measures (Robert Wood Johnson Foundation 2010; Foley & Lardner 2011). Although our work provides important evidence regarding the validity of creating composite measures of physician practice quality, more work remains. For example, only one study (Reeves et al. 2007) has investigated different ways (e.g., weighted average vs. simple averages) to calculate scores in the physician practice once the indicator groupings have been determined. In addition, one can use other approaches to examine the internal validity of physician practice quality measures. These include item response theory (IRT) (Hambleton and Swaminathan 1985) and latent class analysis (LCA) (Collins and Lanza 2010). We did not investigate LCA because it addresses identifying nominally (as opposed to quantitatively) different groups of practices according to different patterns of values on the indicators. To our knowledge, creators of composites are interested in scores that reflect higher and lower levels of quality, not qualitative categories of "scores." We did not use IRT because it is used with categorical data. Our indicators were continuous. Nevertheless, future work should continue to investigate both the empirical validity of grouping indicators in single or multiple composites as well as the best method for creating a score once the indicators to include in the score have been validated.

Before concluding, we note some limitations of our work. First, we had a limited number of indicators available. However, it is unlikely that a unidimensional construct would "emerge" by including *more* indicators, when a unidimensional model with the more limited set of indicators already failed to fit the data well. Second, we only had data from practices within a relatively limited geographical area. We also have very little information about the

characteristics of each individual practices. So, we do not know the representativeness of our data to wider geographical areas. Future research in additional samples and in data with a greater number of and more diverse types of indicators can address both of these concerns.

Third, the indicators included in this dataset were identified by the stakeholders involved in our project before our involvement. Their indicators were not selected *a priori* for purposes of creating a composite measure. Future research might focus on how to systematically and prospectively choose indicators that conceptually ought to measure a broad single, unidimensional construct. This work could help further identify settings and contexts in which broad unidimensional composite measures might be appropriate. Fourth, we included in the analysis practices that had small sample sizes for each indicator. Although there is precedent in the literature and practice for using five and thirty patient thresholds for including practices in a quality measure, other studies have suggested that some indicators need as many as 100 eligible patients to be considered reliable (Hofer et al. 1999). Although composite measures are known to be more reliable than individual indicators (Scholle et al. 2008), our study did not test the reliability of the composite.

Fifth, our analysis was based on practice-level scores estimated from patient-level observations. To the extent that the individual measures are unreliable at the practice-level, our “ecological” approach to the psychometric models may be inappropriate. However, in our data, the vast majority of practice-level reliabilities ranged between 0.9 and 0.99 (Appendix SA2).

Finally, we have not addressed external validity. The question still remains whether scores on a set of composite scores predict future performance and quality. Internal validity does not guarantee external validity (Carle et al. 2011). The possibility exists that internally valid composites still will not sufficiently predict future physician practice quality. In addition, like other composite measures, including those outside medicine (e.g., the overall Scholastic Assessment Test [SAT] score), the possibility exists that a single summary score may lack conceptual clarity (unidimensionality), but still sufficiently predict future performance and quality. Future research that prospectively collects data on a set of practices can address this important next step in the creation and use of quality measures.

In conclusion, composite measurement is an issue of growing policy and practice interest. Our findings show that commonly used HEDIS indicators cannot be used to create valid unidimensional measure of physician practice quality, despite the conceptual appeal of creating single overall

measure. However, our study demonstrates that commonly used HEDIS process quality indicators *can* be used to create a small set of useful composite measures. Stakeholders must use caution when creating composite measures, carefully considering the context and purpose of creating such measures.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This project has been funded by the Robert Wood Johnson Foundation through the evaluation grant of the Aligning Force for Quality project.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Aligning Forces for Quality. 2010. "Summary of Alliance Use of Composite Measures" [accessed on August 1, 2010]. Available at <http://www.rwjf.org/files/research/71824.pdf>
- Bollen, K. 1989. *Structural Equations with Latent Variables*. New York: Wiley.
- Brown, T. 2006. *Confirmatory Factor Analysis for Applied Researchers*. New York: The Guilford Press.
- Caldis, T. 2007. "Composite Health Plan Quality Scales." *Health Care Financing Review* 28 (3): 95–107.
- Carle, A. C., and R. Weech-Maldonado. 2012. "Validly Interpreting Patients' Reports: Using Bifactor and Multidimensional Models to Determine Whether Surveys and Scales Measure One or More Constructs." *Medical Care* 50 (9 Suppl 2): S42–8.
- Carle, A., S. Blumberg, K. Moore, and K. Mbwana. 2011. "Advanced Psychometric Methods for Developing and Evaluating Cut-Point-Based Indicators." *Child Indicators Research* 4 (1): 101–26.
- Centers for Medicare and Medicaid Services. 2013. "Summary of 2015 Physician Value-based Payment Modifier Policies" [accessed on July 23, 2013]. Available at <http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/CY2015ValueModifierPolicies.pdf>
- Collins, L. M., and S. T. Lanza. 2010. *Latent Class and Latent Transition Analysis for the Social, Behavioral, and Health Sciences*. New York: Wiley.
- Coltman, T., T. M. Devinney, D. F. Midgley, and S. Venaiak. 2008. "Formative versus Reflective Measurement Models: Two Applications of Formative Measurement." *Journal of Business Research* 61 (12): 1250–62.
- Foley, and L. Lardner. 2011. "CMS Proposed Rule Describes ACO Quality Performance Standards, Reporting Obligations, and Determination of Shared Savings

- Methodology" [accessed on November 22, 2011]. Available at http://www.foley.com/publications/pub_detail.aspx?pubid=8110
- Friedberg, M. W., and C. L. Damberg. 2011. *Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting: A Guide for Community Quality Collaboratives*. Rockville, MD: Agency for Health Research and Quality.
- . 2012. "A Five-Point Checklist to Help Performance Reports Incentivize Improvement and Effectively Guide Patients." *Health Affairs (Millwood)* 31 (3): 612–8.
- Goldstein, H., W. Browne, and J. Rasbash. 2002. "Partitioning Variation in Generalized Linear Multilevel Models" [accessed on February 22, 2016]. Available at <http://seis.bris.ac.uk/~frwjb/materials/pvmm.pdf>
- Hambleton, R. K., and H. Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Academic Publishers.
- The Health Alliance of Cincinnati. 2013. "Your Health Matters" [accessed on July 23, 2013]. Available at <http://www.yourhealthmatters.org/>
- Hofer, T. P., R. A. Hayward, S. Greenfield, E. H. Wagner, S. H. Kaplan, and W. G. Manning. 1999. "The Unreliability of Individual Physician "Report Cards" for Assessing the Costs and Quality of Care of a Chronic Disease." *Journal of American Medical Association* 281 (22): 2098–105.
- Holmboe, E. S., W. Weng, G. K. Arnold, S. H. Kaplan, S. L. Normand, S. Greenfield, S. Hood, and R. S. Lipner. 2010. "The Comprehensive Care Project: Measuring Physician Performance in Ambulatory Practice." *Health Services Research* 45 (6 Pt 2): 1912–33.
- Hu, L., and P. M. Bentler. 1998. "Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification." *Psychological Methods* 3 (4): 424–53.
- Hu, L., and P. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1–55.
- Institute of Medicine: Committee on Quality of Health Care in America. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: Institute of Medicine: Committee on Quality of Health Care in America.
- Jacobs, R., M. Goddard, and P. C. Smith. 2005. "How Robust Are Hospital Ranks Based on Composite Performance Measures?" *Medical Care* 43 (12): 1177–84.
- Kaplan, S. H., J. L. Griffith, L. L. Price, L. G. Pawlson, and S. Greenfield. 2009. "Improving the Reliability of Physician Performance Assessment: Identifying the 'Physician Effect' on Quality and Creating Composite Measures." *Medical Care* 47 (4): 378–87.
- Lied, T. R., R. Malsbary, C. Eisenberg, and J. Ranck. 2001. "Combining HEDIS Indicators: A New Approach to Measuring Plan Performance." *Health Care Financing Review* 23 (4): 117–29.
- Lipner, R. S., W. Weng, G. K. Arnold, F. D. Duffy, L. A. Lynn, and E. S. Holmboe. 2007. "A Three-Part Model for Measuring Diabetes Care in Physician Practice." *Academic Medicine* 82 (10 Suppl): S48–52.

- Liu, K., S. Jain, and J. Shi. 2013. "Physician Performance Assessment Using a Composite Quality Index." *Statistics in Medicine* 32 (15): 2661–80.
- Maine Health Management Coalition. 2013. "Get Better Maine" [accessed on July 23, 2013]. Available at <http://www.getbettermaine.org/>
- Martsoff, G. R., D. P. Scanlon, and J. B. Christianson. 2013. "Multistakeholder Perspectives on Composite Measures of Ambulatory Care Quality: A Qualitative Descriptive Study." *Medical Care Research and Review* 70 (4): 434–48.
- Massachusetts Health Quality Partners. 2013. "Quality Insights: Healthcare Performance in Massachusetts" [accessed on July 23, 2013]. Available at <http://www.mhqp.org/>
- McDonald, R. P. 1999. *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum.
- Minnesota Community Measurement. 2013. "Reports and Data" [accessed on July 23, 2013]. Available at <http://mncm.org/reports-and-websites/reports-and-data/>
- Muthen, L. K., and B. O. Muthen. 2007. *MPlus User's Guide*. Los Angeles, CA: Muthen & Muthen.
- National Quality Forum. 2009. *Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety-Composite Measures: A Consensus Report*. Washington, DC: NQF.
- Nietert, P. J., A. M. Wessell, R. G. Jenkins, C. Feifer, L. S. Nemeth, and S. M. Ornstein. 2007. "Using a Summary Measure for Multiple Quality Indicators in Primary Care: The Summary Quality Index (SQUID)." *Implementation Science* 2: 11.
- O'Brien, S. M., D. M. Shahian, E. R. DeLong, S. L. Normand, F. H. Edwards, V. A. Ferraris, C. K. Haan, J. B. Rich, C. M. Shewan, R. S. Dokholyan, R. P. Anderson, and E. D. Peterson. 2007. "Quality Measurement in Adult Cardiac Surgery: Part 2—Statistical Considerations in Composite Measure Scoring and Provider Rating." *Annals of Thoracic Surgery* 83 (4 Suppl): S13–26.
- Oregon Health Care Quality Corp. 2013. "Partners for Quality Care" [accessed on July 23, 2013]. Available at <http://www.partnerforqualitycare.org/>
- Pacific Business Group on Health. 2013. "OPA Report Cards" [accessed on July 23, 2013]. Available at <http://reportcard.opa.ca.gov/rc2013/>
- Peterson, E. D., E. R. DeLong, F. A. Masoudi, S. M. O'Brien, P. N. Peterson, J. S. Rumsfeld, D. M. Shahian, and R. E. Shaw. 2010. "ACCF/AHA 2010 Position Statement on Composite Measures for Healthcare Performance Assessment: A Report of American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (Writing Committee to Develop a Position Statement on Composite Measures)." *Journal of American College of Cardiology* 55 (16): 1755–66.
- Puget Sound Health Alliance. 2010. "Community Checkup" [accessed on December 13, 2010]. Available at <http://www.wacommunitycheckup.org/>
- Puget Sound Health Alliance. 2013. "Community Checkup" [accessed on July 23, 2013]. Available at <http://www.wacommunitycheckup.org/>
- Reeves, D., S. M. Campbell, J. Adams, P. G. Shekelle, E. Kontopantelis, and M. O. Roland. 2007. "Combining Multiple Indicators of Clinical Quality: An Evaluation of Different Analytic Approaches." *Medical Care* 45 (6): 489–96.

- Robert Wood Johnson Foundation. 2010. "Summary of Alliance Use of Composite Measures" [accessed on November 22, 2010]. Available at <http://www.rwjf.org/qualityequality/product.jsp?id=71824>
- Romano, P. S., P. Hussey, and D. Ritley. 2010. *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives*. Washington, DC: AHRQ.
- Scholle, S. H., J. Roski, J. L. Adams, D. L. Dunn, E. A. Kerr, D. P. Dugan, and R. E. Jensen. 2008. "Benchmarking Physician Performance: Reliability of Individual and Composite Measures." *American Journal of Managed Care* 14 (12): 833–8.
- Shwartz, M., J. Ren, E. A. Pekoz, X. Wang, A. B. Cohen, and J. D. Restuccia. 2008. "Estimating a Composite Measure of Hospital Quality from the Hospital Compare Database: Differences When Using a Bayesian Hierarchical Latent Variable Model versus Denominator-Based Weights." *Medical Care* 46 (8): 778–85.
- Snijders, T. A. B., and R. J. Bosker. 1999. *Multilevel Analysis*. London: Sage Publications.
- Teixeira-Pinto, A., and S. L. Normand. 2008. "Statistical Methodology for Classifying Units on the Basis of Multiple-Related Measures." *Statistics in Medicine* 27 (9): 1329–50.
- Weng, W., B. J. Hess, L. A. Lynn, E. S. Holmboe, and R. S. Lipner. 2010. "Measuring Physicians' Performance in Clinical Practice: Reliability, Classification Accuracy, and Validity." *Evaluation and the Health Professions* 33 (3): 302–20.
- Wisconsin Collaborative for Health Care Quality. 2013. "Performance and Progress Reports" [accessed on July 23, 2013]. Available at <http://www.wchq.org/>
- Zaslavsky, A. M., J. A. Shaul, L. B. Zaboriski, M. J. Cioffi, and P. D. Cleary. 2002. "Combining Health Plan Performance Indicators into Simpler Composite Measures." *Health Care Financing Review* 23 (4): 101–16.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.

Appendix SA2: Reliability Estimates for Individual Measures.