Composite Health Plan Quality Scales

Todd Caldis, Ph.D., J.D.

This study employs exploratory factor analysis and scale construction methods with commercial Health Plan Employers Data Information Set (HEDIS®) process of care and outcome measures from 1999 to uncover evidence for a unidimensional composite health maintenance organization (HMO) quality scale. Summated scales by categories of care are created and are then used in a factor analysis that has a single factor solution. The category of care scales were used to construct a summated composite scale which exhibits strong evidence of internal consistency (alpha= 0.90). External validity of the composite quality scale was checked by regressing the composite scale on Consumer Assessment of Healthcare Providers and Systems (CAHPS®) survey results for 1999.

INTRODUCTION

Although much attention has been given to development of quality measurement systems such as HEDIS® and more recently physician-centered quality measures being developed by the National Center for Quality Assurance (NCQA), CMS, and the American Medical Association, comparatively little research has focused on the problem of building composite measures of health care service quality from discrete and seemingly unrelated measures of clinical quality. This article tackles the problem of constructing a credible unidimensional composite HMO quality scale from cross-sectional HEDIS® scores reported by commercial HMOs in calendar year (CY) 1999. Evidence is uncovered that composite subscales by categories of care and a unidimensional composite HEDIS® quality measure are feasible.

BACKGROUND

The quality of medical care is likely to have many possible dimensions, but it is conceivable that the quality of medical care rendered by an HMO may also have an organizational dimension, reflecting a common approach across all types of medical care (Wholey et al., 2003). If a plausible summary measure of overall HMO output quality could be developed, it is likely to be useful to many actors: employers assembling sets of health plan choices to offer employees, consumers faced with health plan choices, HMO regulators needing a way to screen firms for regulatory attention, and researchers in need of quality variables to add a quality dimension to models of HMO behavior, e.g., in a model of HMO costs like the CMS risk- adjustment model.

Some research concerning composite health care quality measures has taken place ancillary to studies of consumer choice of health plans in which researchers have sought to determine whether consumers actually use quality information that is provided to them. The literature on consumer choice of health plans usually includes quality variables in binomial or

multinomial logit models of utility maximization (Chernew and Scanlon, 1998; Scanlon and Chernew, 1999; Feldman, Christianson, and Schultz, 2000; Chernew, Gowrisankaran, and Scanlon, 2001; Scanlon et al., 2002). Recent work by Chernew et al. (2003) that investigates employer decisions about which health plans to offer employees contains a good example of how researchers have used limited amounts of HEDIS® data to construct a quality variable by rolling several HEDIS® variables together and thereby obtaining a quality construct usable as a regressor to represent quality in a more general regression model. This literature does not address measurement properties of the quality measures that are implemented, and the measures implemented appear to be simply the best available proxy for the construct the researchers wished to include in their models of consumer behavior.

The most sophisticated HMO-level output quality scales so far reported have been constructed by researchers whose primary objective was to construct composite scales that might eventually be disseminated to consumers (Lied et al., 2002; Zaslavsky et al., 2002). Both of these studies relied on a single year of quality data taken from Medicare+Choice health plans. Lied and colleagues constructed a composite HEDIS® score for each health plan in a cross-section of Medicare+Choice plans by averaging percentile scores for 17 indicators chosen primarily from available HEDIS® effectiveness of care measures. Zaslavsky and colleagues used factor analysis to reduce 20 quality measures, consisting of 12 HEDIS® and 8 consumer satisfaction CAHPS® measures, into 4 interpretable constructs that explained 65 percent of the variance across all measures. The four constructs (office care, access/customer service, vaccinations, and clinical quality) were then used to

create averaged scales that exhibited good evidence of internal consistency. In each case Cronbach's (1951) alpha statistics computed to test the composite scales returned evidence of internal consistency, an alpha score of 0.88 for the scale constructed by Lied and colleagues, and alpha scores ranging from 0.86 to 0.96 for the scales constructed by Zaslavsky and colleagues.

These studies adopted different approaches to handle the problem of missing quality data, a significant obstacle to the construction of composite scales. Lied and others dropped 10 percent of the observations in their original sample when quality reporting was very limited or suspect. Only 64 of 160 remaining observations reported all 17 of the HEDIS® measures, and the mean number of reported measures in their final sample was 13.58 per observational unit. Each HMO's composite quality score was thus actually an average of its percentiles for the HEDIS® indicators that it happened to report.

To avoid losing HMO-level HEDIS® observations, Zaslavsky and others used procedures to impute missing HEDIS® scores, employing as predictor variables other reported HEDIS® scores for HMOs that did report the measure to be imputed and also special dummy variables keyed to patterns of missing variables in each included observation. This imputation method generated complete sets of HEDIS® measures that could then be used to compute factor analysis results. Although variances for observations with many imputed values were indeed larger than for observations with few missing HEDIS® scores, the effect of larger variances seemed to be muted when measures were standardized and averaged into composite measures.

In different ways both studies grappled with evidence that quality of care

is multidimensional. When Lied's group subjected their data to factor analysis, the researchers found evidence that the variance in the HEDIS® measures decomposed into three components rather than the single component scale that they had constructed, a result conceivably related to having disproportionately many measures from particular categories of medical care. For example, 6 of the 17 measures pertained specifically to diabetes care. Zaslavsky's group encountered a similar problem as a result of factor analyzing data from HEDIS® and CAHPS® together; the four factor structure that they uncovered is diffuse in its range, and it generally appears that HEDIS® and CAHPS® measures contribute to different factor loadings.

Thus, recent quality scaling work using HEDIS® data encountered difficulties when researchers used conventional scaling techniques to convert multiple measures into a smaller number of quality-related variables. Gaps in reported quality scores present a fundamental problem because algorithms needed to estimate factor analytic models require complete data with no missing values in any of the observations used. Imputation of missing values is unavoidable if the sample is to include the broadest possible set of HMOs.

Indiscriminate inclusion of all raw HEDIS® performance measures in a single factor analytic model may complicate identification of a unidimensional underlying organizational-level quality scale. As previously noted, the studies that have implemented factor analysis of HEDIS® data have included all available variables in the same factor analysis with disproportionate numbers of variables for certain disease categories such as diabetes.

An alternative approach is to implement a second-order factor model by first using HEDIS® variables to create subscales by categories of care (e.g., diabetes, women's care, childhood immunizations, etc.) and then using category-of-care subscales in an exploratory factor analysis intended to uncover underlying quality constructs at the health plan level (McArdle, 1980; Hershberger, 1994; Kline, 2005). Availability of CAHPS® survey results also makes it feasible to consider evidence for the external validity of a composite HEDIS-based HMO quality scale.

This article reports implementation of a two-stage approach to composite scaling of HEDIS quality data together with evidence for the internal consistency and external validity of the resulting unidimensional quality scale.

Data and Data Preparation

Data available for this study included 31 HEDIS® process of care and outcome measures designed and collected by NCQA from 380 HMOs concerning medical care provided to under age 65 commercial subscribers in CY 1999. The insured population covered by the reporting HMOs represented almost the entire under age 65 non-Medicaid, non-Medicare population that received its health care through HMOs in CY 1999. A list of the HEDIS® measures with descriptive statistics appears in Table 1. All of the measures are interval measures suitable for factor and correlation-based analysis and normality plots indicate that the measures appear to be normally distributed.

Reporting of HEDIS® measures by HMOs was often incomplete, though less incomplete than was the case in the study by Zaslavsky and others that also employed multiple imputation methods for missing data. Almost all HEDIS® measures in the data set had missing values, but the problem is most acute for the the battery of measures for asthma first introduced in

Care Measure	MEAN	S.D.	MIN	MAX	N
Child Immunizations					
Diptheria-Tetanus-Whooping Cough (DTP)	78.76	13.31	14.19	95.48	357
Measles-Mumps-Rubella	87.01	8.69	32.34	98.01	360
Oral Polio Vaccine (OPV)	82.61	12.10	21.23	97.57	357
Pneumonia Vaccine (HIB)	80.74	11.99	22.22	98.56	358
Hepatitis B Vaccine	75.13	15.07	6.71	96.25	356
Chickenpox Vaccine (VZV)	63.80	11.11	19.36	92.95	360
Adolescent Immunizations					
Measles-Mumps-Rubella	58.96	20.68	0.77	94.60	353
Hepatitis B Rate	34.32	19.47	0	81.51	350
Chickenpox Vaccine (VZV)	23.98	15.95	0	75.90	350
Women's Health					
Breast Cancer Screening Rate	73.37	7.15	40.04	88.66	368
Cervical Cancer Screening	71.72	9.24	31.24	91.19	368
Prenatal Care	84.59	13.42	2.31	99.99	354
Check-Up After Delivery	72.38	13.64	16.17	98.56	350
Coronary Care					
Control Blood Pressure	38.94	9.57	8.47	90.86	257
Beta Blocker After Heart Attack	85.03	11.00	44.55	99.99	233
Cholesterol Screening	68.97	11.55	25.74	93.50	279
Cholesterol Control	44.33	17.16	0	82.72	270
Diabetes Care					
Blood Sugar Testing (HbA1c)	75.02	11.66	13.2	97.13	367
Poor Control of Blood Sugar	55.22	16.61	0	99.99	355
Eye Exams	45.32	15.07	1.54	90.53	366
Cholesterol Screening	68.97	11.22	11.77	88.55	368
Cholesterol Control	36.74	11.55	0	76.12	360
Monitor Diabetic Nephropathy	36.08	14.41	6.05	96.58	365
Asthma Care					
Asthma Medication Management (Age 5-9)	57.53	13.75	0	93.28	263
Asthma Medication Management (Age 10-17)	54.78	13.97	0	96.80	285
Asthma Medication Management (Age 18-56)	59.40	14.74	0	96.36	337
Mental Health					
Follow-Up After Hospitalization—7 Days	46.86	16.61	0	90.97	301
Follow-Up After Hospitalization—30 Days	69.41	16.61	0	99.99	301
Practitioner Contact—Antidep Medication	58.74	10.23	5.5	88.00	304
Antidep Med Acute Phase	42.13	10.45	7.81	81.73	302
Antidep Med Continuation Phase	21.23	10.67	0	72.38	300

NOTES: HEDIS® is Health Employers Data Information Set. S.D. is standard deviation.

SOURCE: Caldis, T., Centers for Medicare & Medicaid Services, 2007.

the measurement year. Multiple imputation techniques using algorithms packaged in SAS® Proc MI and Proc MIANALYZE were employed to maintain the range of measures needed for the construction of a broad-based composite quality measure. The imputation algorithms assume that data are missing at random (MAR), which means that the value of a missing score is assumed independent of the likelihood that the score will be missing (Rubin, 1976, 1987, 1991; Allison, 2002; SAS Institute, 1999). If the MAR assumption is satisfied,

multiple imputation methods lead to estimates that are approximately unbiased and efficient, with standard errors that take into account variance added by imputation (Allison, 2002).

Implementation of multiple imputation methods occurs in two stages. First, the imputation algorithm is used to generate multiple versions of the data that is to be analyzed, with each version having its own randomly assigned values for missing scores. Second, another algorithm uses the multiple versions of the data to

compute a single set of relevant statistics with adjusted standard errors. Where, as here, correlations between measures are needed for factor analysis, a SAS-supplied variant of the algorithm is employed in order to correct for bias that would otherwise be present due to skewness of the distribution of sample correlation coefficients (SAS Institute, 1999).

In accord with the research strategy of generating a set of quality scales by categories of care and then using those scales to build a composite scale, multiple imputation methods were implemented separately for each battery of care measures: diabetes, heart, asthma, mental health, women's care, childhood immunization, and adolescent immunization. When an HMO had not reported values for a given battery of care measures, no imputation of values for that HMO for the battery of care occurred and that observation was ignored in constructing a scale for the battery of care. Once category of care scales had been constructed, multiple imputation was used again to fill in missing scales before proceeding with construction of the quality composite. Each implementation of multiple imputation methods employed the Markov Chain Monte Carlo method for imputing values for arbitrary non-monotone missing value patterns. Where missing data patterns were monotone, alternative imputation algorithms capable of exploiting the presence of monotonicity were implemented, but use of alternative algorithms for multiple imputation did not affect subsequent analysis.

Data available for external validation of HEDIS-based composite scales consisted of CAHPS® survey data for CY 1999 collected by NCQA from 361 out of the 380 HMOs that reported HEDIS® data. Specifically, for each of six CAHPS® composite measures for each HMO, the total percentage of respondents who gave their HMO a satisfactory or good rating was available. Depending on the CAHPS® composite this was the the percentage of respondents who said they were not having a problem or the percentage who said they were usually or always encountering the care or behavior covered by the CAHPS® composite.

Methods of Analysis

The methods of analysis consisted of exploratory factor analysis, summated scale construction and testing, and linear regression. Factor analysis was used to identify grounds for data consolidation through scale construction. Summated scales by categories of care and finally as a composite of the category of care scales are the principal output of the study and their internal consistency was validated using the Cronbach's (1951) alpha statistic. Linear regression was used to evaluate the external validity of the final composite quality of care scale by examining its relationship with CAHPS® survey results.

As in prior studies that have sought to reduce the dimensionality of HEDIS® health plan quality data, exploratory factor analysis was used to identify the component or components of variance common to multiple HEDIS® measures. Factor analysis is a statistical technique that originated in the field of psychometrics, but is used widely in social science research. The method seeks to explain the variation of observed random variables in terms of a smaller set of unobserved random variables, known as factors (Nunally and Bernstein, 1994). Algorithms that implement factor analysis model observed variables as the linear combination of unobserved variables and a random error term.

A matrix of correlations between observed variables is sufficient to compute a factor analysis model. As previously noted, the skewness of the distribution of

sample correlations was taken into account in assembling correlation matrices for factor analysis. Confidence intervals for each correlation in each matrix were also available, making it feasible in each case to compute models for low- and high-range correlations as well as for set of estimated correlations. In this way it was possible to verify that results did not vary substantially within the range of the confidence intervals for the correlation matrices.

The output of a factor model is a set of loadings or factor pattern which is the set of estimated coefficients that would be applied to unobserved variables to produce the observed variables. In each case the factor pattern was analyzed by standard procedures (Nunally and Bernstein, 1994). A HEDIS® measure was deemed to have a significant loading if it had a single loading on a single unobserved variable with an absolute value equal to or exceeding 40 percent and no significant loading on any other unobserved variable. Varimax (orthogonal) and promax (oblique) rotation methods were employed to enhance interpretability of factor patterns when the factor solution indicated that more than a single unobserved factor could be present. In multifactor situations, the meaning of each factor (each column in a factor pattern) was inferred from the pattern of meaning implied by the HEDIS® measures having significant loadings on that factor (Nunally and Bernstein 1994).

Summated scales were constructed based on the results of the exploratory factor analyses. As the term summated scale implies, such scales are constructed by summing the items that make the scale (Nunally and Bernstein, 1994; Spector, 1992). Measurement theory behind summated scale construction postulates that error terms associated with each variable included in a scale should cancel out due to their randomness, permitting the

true value of the construct to be revealed. The theory behind summated scales also implies that the variables making up a scale should be highly correlated with each other, and this intuition lies behind Cronbach's (1951) alpha, the standard test statistic for evaluating the internal consistency of summated scales (Spector, 1992; Hatcher, 1994). Generally, the alpha statistic will increase as the correlation between variables increases and the number of variables included in the scale increases. By convention, an alpha statistic equal to 0.7 is deemed minimally adequate evidence of internal consistency for a summated scale, though obviously the closer that the alpha statistic is to one, the stronger the evidence of internal consistency.

External validity of composite quality scales was investigated with linear regressions each of which used a composite quality scale as the dependent variable and a set of CAHPS® survey results as regressors. The set of regressors consisted of the six CAHPS® composites that involve aspects of health plan performance: (1) getting needed care, (2) getting care quickly, (3) physician communication, (4) staff courtesy, (5) customer service, and (6) claims processing. The presence of significant coefficients in the regression and the presence of explanatory power in the regression would be interpreted as evidence supporting the hypothesis that both HEDIS® and CAHPS® are measuring the same unobservable factor: health plan quality.

RESULTS

To document continuity with prior research results, an all measures factor analysis similar to the factor models in previous measurement studies of HEDIS® data included all of the 31 HEDIS® process of care measures, and as in those studies,

the all measures model pointed to a multi-dimensional factor solution whose pattern on promax rotation is shown in Table 2. The factor pattern was interpreted variable by variable in accord with the rule previously described, looking for variables that have a strong loading (absolute value of a loading > 40 percent) on a single unobserved factor. All immunization variables loaded strongly and exclusively on the second factor, as did the variable for controlling blood pressure. All diabetes and mental health variables along with some women's health and coronary care variables loaded strongly and exclusively on the first factor. Only asthma care variables loaded strongly and exclusively on the third factor.

The factor pattern for the all measures model shown in Table 2 is derived, as are all factor models in this article, from an analysis using standardized variables. The pattern demonstrated in another run with unstandardized variables was

Table 2
Factor Pattern for All Measures Factor Analysis

Care Measure	Factor 1 Four Categories	Factor 2 Immunizations	Factor 3 Asthma
		Percent	
Child Immunizations			
Diptheria-Tetanus-Whooping Cough (DTP)	-17	106	-2
Measles-Mumps-Rubella	-17	99	5
Oral Polio Vaccine (OPV)	-8	101	0
Pneumonia Vaccine (HIB)	1	96	-3
Hepatitis B Vaccine	-1	92	6
Chickenpox Vaccine (VZV)	15	42	1
Adolescent Immunizations			
Measles-Mumps-Rubella	26	72	-6
Hepatitis B Rate	30	50	2
Chickenpox Vaccine (VZV)	39	48	-4
Women's Health			
Breast Cancer Screening Rate	61	33	-15
Cervical Cancer Screening	48	54	-1
Prenatal Care	49	35	3
Check-Up After Delivery	50	43	6
Coronary Care			
Control Blood Pressure	25	41	31
Beta Blocker After heart Attack	55	31	5
Cholesterol Screening	61	30	6
Cholesterol Control	67	30	2
Diabetes Care			
Blood Sugar Testing (HbA1c)	57	33	15
Poor Control of Blood Sugar	75	8	19
Eye Exams	76	20	0
Cholesterol screening	72	18	3
Cholesterol Control	74	17	4
Monitor Diabetic Nephropathy	86	0	-3
Asthma Care			
Asthma Medication Management (Age 5-9)	16	1	85
Asthma Medication Management (Age 10-17)	12	-2	91
Asthma Medication Management (Age 18-56)	17	3	84
Mental Health			
Follow-Up After Hospitalization—7 Days	89	-1	0
Follow-Up After Hospitalization—30 Days	90	-1	3
Practitioner Contact—Antidep Medication	88	-20	10
Antidep Med Acute Phase	87	-19	14
Antidep Med Continuation Phase	64	-14	31

SOURCE: Caldis, T., Centers for Medicare & Medicaid Services, 2007.

substantively the same, which would be expected with variables not having widely discrepant variances. The all measures factor solution accounts for 90 percent of the variance contained in the 31 underlying measures. A scree plot raised the issue of adding a fourth factor to the solution, but the additional variance associated explained by a fourth factor was negligible (< 3 percent) and on examination its loadings were uninterpretable without even one variable having a strong and exclusive loading. On this basis a solution including more than three factors was rejected. Overall, the factor pattern provides evidence to support the existence of asthma care, immunization, and and a third factor that represents four different categories of care (diabetes, mental health, women's health, and coronary care). The three factors are labeled in Table 2 as Four Categories, Immunizations, and Asthma. The fact that so many categories of care were loading on a single factor was taken as a promising indication that the intended research strategy employing subscales would work as hoped.

In accord with the plan of research, factor analysis was performed by categories of care to examine whether measures whose subject matter is identical hang together in ways indicative of disease-specific or treatment-specific scales. Table 3 summarizes the results of eight different factor model estimates, displaying the loadings for a single-factor solution on the variables included in a given category of care model. Table 3 displays the proportion of total variance explained by each category of care factor solution and also the Cronbach's (1951) alpha statistics for the summated category of care scales that were created based on the factor analyses results obtained.

Sets of variables defined by categories of care turned out with minor and easily handled exceptions to lead to factor analyses that pointed to persuasive single factor solutions with strong loadings of comparable magnitude on most variables included in each set of estimating variables. The only set of variables that did not lead to a clear single factor solution was the battery of measures for mental health. The results of an analysis on all mental health variables pointed clearly to a two-factor solution, one defined by depression care and the other defined by followup after hospitalization. The implication of this result seemed to be to allocate these variables to separate factor models; thus, the loadings for these variables in Table 3 reflect the single factor solutions for the two different mental health factor models that were then implemented.

The proportion of variance reported by the SAS® algorithms that estimated the factor models frequently reported that the models as estimated explained more than 100 percent of the variance contained in the included items. In part this reflects the high correlation among measures and in part this reflects the possibility inherent in factor analysis models of extracting factors that explain either negative or positive shares of the variance contained within the correlation matrix used by the estimating algorithm, subject to the requirement that the variance explained by all factors sums to 100 percent (Hatcher, 1994). When this occurred, the proportion of variance explained was reported (Table 3) as simply 100 percent.

The fact that loadings within each given category of care tended to have broadly similar magnitudes may be interpreted as evidence that construction of summated scales is warranted. A summated scale was constructed for each of the categories of care (Table 3). With two exceptions

Table 3

Factor Loading Patterns for Category of Care Scales

Care Measure	Child Immunization	Adolescent Immunization	Women's Health	Coronary Care	Diabetes Care	Asthma Care	Mental Health I	Mental Health II
				Percent				
Child Immunizations								
Diptheria-Tetanus-Whooping Cough (DTP)	98							
Measles-Mumps-Rubella	92							
Oral Polio Vaccine (OPV)	99							
Pneumonia Vaccine (HIB)	96							
Hepatitis B Vaccine	91							
Chickenpox Vaccine (VZV)	49							
Adolescent Immunizations								
Measles-Mumps-Rubella		90						
Hepatitis B Rate		78						
Chickenpox Vaccine (VZV)		80						
Women's Health								
Breast Cancer Screening Rate			76					
Cervical Cancer Screening			89					
Prenatal Care			70					
Check-Up After Delivery			81					
Coronary Care								
Control Blood Pressure				37				
Beta Blocker After Heart Attack				45				
Cholesterol Screening				66				
Cholesterol Control				81				
Diabetes Care								
Blood Sugar Testing (HbA1c)					86			
Poor Control of Blood Sugar					80			
Eye Exams					63			
Cholesterol Screening					75			
Cholesterol Control					80			
Monitor diabetic Nephropathy					61			
Asthma Care								
Asthma Medication Management (Age 5-9)						91		
Asthma Medication Management (Age 10-17)						98		
Asthma Medication Management (Age 18-56)						92		
Mental Health								
Follow-Up After Hospitalization—7 Days							88	
Follow-Up After Hospitalization—30 Days							89	
Practitioner Contact—Antidep Medication								15
Antidep Med Acute Phase								94
Antidep Med Continuation Phase								72
Proportion Total Variance Explained	99	100	100	100	100	100	100	100
Alpha Statistic (Summated Scales)	0.95	0.87	0.86	0.70[1]	0.88	0.95	0.88	0.80[1]

[1]Summated scale excluded the variable with the smallest factor loading.

SOURCE: Caldis, T., Centers for Medicare & Medicaid Services, 2007.

satisfactory alpha coefficients were obtained using for each category of care all of the variables associated with the category of care. For coronary care and mental health II (depression care), the internal consistency of the summated scale could be substantially improved by excluding the variable with the smallest loading. In each case the magnitude of the outlier loading was substantially smaller than the loadings for the rest of the variables, and exclusion was consistent with the theory of summated scale construction. The alpha coefficients for coronary care and mental health II (Table 3) are for the summated scales that exclude the variable that had the lowest loading in the estimate for the corresponding factor model.

Six out of the eight category of care scales have internal consistency scores equal to 0.86 or better and two of the scales yielded high alpha coefficients substantially exceeding 0.90. Only one of the scales for coronary care, has a minimally acceptable alpha coefficient (Table 3).

Four alternative factor models were then estimated using the summated category of care scales created in the previous phase of the research and alternative summated composite scales were created for the set of variables in each model. In every case there was a clear single factor solution. The single factor loadings for each of the four different composite factor models appear in Table 4. In every model the single factor solution explains virtually all variance in the included items. Loadings on the scales for asthma care and mental health II were pronouncedly smaller in magnitude than those for the other scales. Unsurprisingly, removal of those items from composite models improved the internal consistency of the summated composite constructed from the remaining category of care scales.

Finally, a summated HEDIS® composite scale created from six category of care scales (Table 4) was used as the dependent variable in a regression that used CAHPS® composites as regressors (Table 5). The regression was significant and the adjusted R-squared statistic was 0.36. Coefficients on all variables were statistically significant or close to statistically significant. The coefficients for the physician communication and staff courtesy variables had negative signs. This suggests an interesting question bearing on the validity of CAHPS® results as quality measures: whether social circumstances of care that are perceived positively by patients are necessarily indicators of good medical care. However, the many positive coefficients in the regression and the statistical significance of all coefficients are sufficient to conclude that there is strong external evidence that the six category HEDIS® scale is a measure of the quality of medical care in the reporting health plans. In particular, the regression results provide reassurance that performance on the aggregate quality scale is not simply an artifact of which HMOs have good recordkeeping systems.

DISCUSSION

The results reported in this article are consistent with measurement theory. Reliable scales were constructed around

Table 4

Factor Loading Patterns—Single Factor Solutions for Composite Models

Care Measure	Factor Model 1 All Categories	Factor Model 2 All, but Mental Health II	Factor Model 3 All, but Asthma	Factor Model 4 All, but Asthma and Mental Health II
	Percent			
Child Immunization	86	85	87	86
Adolescent Immunization	78	78	78	78
Women's Health	85	83	83	83
Coronary Care	80	81	79	83
Diabetes Care	82	83	83	82
Mental Health I	66	61	64	62
Mental Health II	33	—	31	—
Asthma Care	21	24	—	—
Proportion of Variance Explained	100	100	100	100
Alpha Statistic	0.84	0.86	0.87	0.9

SOURCE: Caldis, T., Centers for Medicare & Medicaid Services, 2007.

Table 5

Validating Regression—HEDIS® Composite Regressed on CAHPS® Composites

Variable	Coefficient Estimate	Standard Error	t Value	Significance
Getting Needed Care	0.14	0.06	2.33	0.02
Getting Care Quickly	0.53	0.07	7.27	<.01
Physician Communication	-0.47	0.12	-3.79	<.01
Staff Courtesy	-0.24	0.13	-1.9	0.06
Customer Service	0.09	0.05	1.84	0.07
Claims Processing	0.09	0.03	2.62	<.01

NOTES: R-squared statistic is 0.36. Adjusted R-squared statistic is 0.35.

SOURCE: Caldis, T., Centers for Medicare & Medicaid Services, 2007.

categories of care and then analyzed together to construct a reliable measure of output quality. In this way, a broadly based composite measure emerged, reflecting six types of care. Validating regression results establish a strong connection between the six-category composite quality scale and the dimensions of consumer satisfaction constructed from CAHPS® survey data. The six-category scale exhibits plausible and persuasive evidence of internal consistency and external validity. Although it would be desirable to have a unidimensional scale that incorporated still more categories of care, the breadth of the scale that has been developed may plausibly be seen as a measure of quality at an organizational level.

Nevertheless, it is important to acknowledge significant limitations in the results obtained and a need for caution in how these results are used. First, this was an exploratory study on a single year of cross-sectional quality data for a single type of insured population. It is unknown whether the results will remain robust when the approach is used with additional years of quality data and for Medicare and Medicaid HMOs as well as for the commercial HMOs represented in this study. With new data a formal confirmatory factor analysis model should be evaluated, and because second-order factor models are difficult to identify econometrically, applied studies that incorporate quality scales within larger models could also provide important insight into whether the scales appear to add a plausible quality dimension (McArdle, 1980; Hershberger, 1994; Kline, 2005). For example, a composite HEDIS® scale (or category of care scales) could be included in a risk-adjustment model or other model of HMO economic behavior.

A second limitation on the substantive results obtained is the matter of missing data. No study of the measurement properties of HEDIS® measures has implemented a fully satisfactory solution to the missing data problem. The MAR assumption behind the multiple imputation methods implemented here is questionable: It is difficult to believe that in all cases the failure of an HMO to report a particular score had nothing to do with the level of the unreported score. Fortunately, the external validity regression provides tangible evidence that the composite HEDIS® scale is measuring something real despite the problematic nature of the MAR assumption. An actual program affecting payments to plans could not implement any kind of composite measure that relied on imputed data. It is to be hoped that the problem of missing data will disappear as HMOs become more and more accustomed to annual collection of HEDIS® data and purchasers of HMO services become more insistent that the data be thoroughly reported.

In light of the research limitations that have been discussed the immediate usefulness of the scales developed here is

limited. Certainly, the scales as they have been developed could be helpful as guides to HMOs on needed quality improvement, particularly the category of care scales. Beyond that possible application, the most immediate use of the scaling results is as a basis for further research to test and hopefully validate the approach. If the modeling approach implemented here in an exploratory way can be confirmed on a more substantial body of quality data, then the results would provide essential information needed for practical development of a pay-for-performance payment system for HMOs. Additional composite quality scaling research might also lead to refinements in the underlying components of aggregate scales, such as their relative weighting. More generally, if these methods can be broadly validated, they might be applied to the problem of physician performance measurement using the battery of physician-level process measures that NCQA is currently promulgating.

Although the results achieved here have limitations, they should be encouraging to policymakers and to institutions like CMS that are endeavoring to build the technical infrastructure needed to make pay-for-performance in health care a reality. The overarching message is that aggregate quality measures that would greatly facilitate the development of pay-for-performance payment systems may be practically feasible.

Much effort and expense have been expended developing the HEDIS® performance measurement system, but up until now evidence has been meager that multiple discrete HEDIS® measures can be assembled into a plausible organization-level composite measure of health care quality. The methods used in this study demonstrate that HEDIS® measures may carry information relevant both to the specific care contexts in which they arise and more generally to output quality at the organizational level. That the HEDIS®-based composite quality scale is highly correlated with composites in the CAHPS® consumer satisfaction survey provides important evidence of external validity, but also suggests the possibility of regularly using such methods as a check on the ongoing validity of both measurement systems. As policymakers become more determined to pay providers based on their performance, solutions to the problem of health care quality measurement are urgently needed. The results obtained here should encourage researchers and institutions that sponsor research to focus more attention on development and testing of aggregate health care quality scales.

REFERENCES

Allison, P.D.: *Missing Data.* Sage Publications USA. Thousand Oaks, CA. 2002.

Chernew, M. and Scanlon, D.P.:Health Plan Report Cards and Insurance Choice. *Inquiry* 35(1):9-22, Spring 1998.

Chernew, M., Gowrisankaran, G., and Scanlon, D.P.: *Learning and the Value of Information: The Case of Health Plan Report Cards.* National Bureau of Economic Research Working Paper No. 8589. November 2001.

Chernew, M., Gowrisankaran, G., McLaughlin, C., et al.: *Quality and the Employers' Choice of Health Plan.* National Bureau of Economic Research Working Paper No. 9847. July 2003.

Cronbach, L.J.: Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16(3):297-334, 1951.

Engberg, J., Wholey, D., Feldman, R., et al.: The Effect of Mergers on Firm's Costs: Evidence from the HMO Industry. *The Quarterly Review of Economics and Finance* 44(4):574-600, September 2004.

Feldman, R., Christianson, J., and Schultz, J.: Do Consumers Use Information to Choose a Health-Care System? *The Milbank Quarterly* 78(1):47-77, March 2000.

Hatcher, L.: *Step-by-Step Approach to Using SAS® Sytem for Factor Analysis and Structural Equation Modeling.* SAS Institute. Cary, NC. 1994.

Hershberger, S.L.: The Specification of Equivalent Models Before the Collection of Data. In Von Eye, A. and Clogg, C. (eds.): *Latent Variable Analysis*. SAGE Publications, Inc. Thousand Oaks, CA. 1994.

Kim, J.O. and Mueller, C.W.: *Factor Analysis: Statistical Methods and Practical Issues*. Sage Publications, Inc. Beverly Hills, CA. 1978.

Kline, R.B.: *Principles and Practice of Structural Equation Modeling, 2nd Edition*. The Guildford Press. New York, NY. 2005

Lied, T., Malsbary, R., Eisenberg, C., et al.: Combining HEDIS® Indicators: A New Approach to Measuring Plan Performance. *Health Care Financing Review* 23(4):117-129, Summer 2002.

McArdle, J.: Causal Modeling Applied to Psychonomic Systems Simulation. *Behavior Research and Instrumentation* 12(2):193-209, 1980.

Nunally, J.C. and Bernstein, I.H.: *Psychometric Theory, Third Edition*. McGraw-Hill, Inc. New York, NY. 1994.

Rubin, D.B.: Inference and Missing Data. *Biometrika* 63:581-592, 1976.

Rubin, D.B.:*Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons. New York, NY. 1987.

Rubin, D.B. and Schenker, N.: Multiple Imputations in Health-Care Databases: An Overview and Some Applications. *Statistics in Medicine* 10:585-598, 1991.

SAS Institute: *SAS Procedures Guide*. Version 8. 1999.

Scanlon, D.P. and Chernew, M.: HEDIS® Measures and Managed Care Enrollment. *Medical Care Research and Review* 56(Suppl. 2):56-84, 1999.

Scanlon, D.P., Chernew, M., McLaughlin, C. et al.: The Impact of Health Plan Report Cards on Managed Care Enrollment. *Journal of Health Economics* 21(1): 19-41, January 2002.

Spector, P.E.: *Summated Rating Construction*. SAGE Publications, Inc. Newbury Park, CA. 1992.

Wholey, D.R., Christianson, J.B., Finch, M., et al.: Evaluating Health Plan Quality 1: A Conceptual Model. *The American Journal of Managed Care* 9(2): SP53-SP64, June 2003.

Zaslavsky, A.M., Shaul, J., Zaborski, L., et al.: Combining Health Plan Performance Indicators into Simpler Composite Measures. *Health Care Financing Review* 23(4):101-115, Summer 2002.