Combining Health Plan Performance Indicators into Simpler Composite Measures

Alan M. Zaslavsky, Ph.D., James A. Shaul, M.H.A., Lawrence B. Zaborski, M.S., M.A., Matthew J. Cioffi, and Paul D. Cleary, Ph.D.

We investigated how the Consumer Assessment of Health Plan Study (CAHPS®) survey and the Health Plan Employer Data Information System (HEDIS®) measures from Medicare managed care (MMC) plans could be combined into fewer summary performance scores. Four scores summarize most of the variability in these measures, representing (1) care at the doctor's office, (2) customer service and access, (3) vaccinations, and (4) clinical quality measures. These summaries are substantively interpretable, internally consistent, and describe the majority of variation among units in the performance scores analyzed.

BACKGROUND

Numerous performance indicators currently are available from initiatives sponsored by CMS, such as the CAHPS® survey (Goldstein et al., 2001; Zaslavsky et al., 2001), managed care disenrollment surveys, Medicare HEDIS® (Schneider et al., 2001), and the Health Outcomes Study. Other measures, such as managed care disenrollment rates (Lied et al., 2001), can be derived from administrative data. Each available measure was developed with a somewhat different purpose in mind. Consumers often prefer separate reports of quality indicators (QIs). Clinicians prefer QIs that are specific and actionable. However, the proliferation of QIs makes it increasingly difficult to interpret what the numerous measures reveal about health plan quality, to prioritize them, and to create summary statements about health plan performance. In addition, many of these QIs are theoretically related and have high empirical correlations. Thus, fewer measures might provide a similar portrait of plan quality, and for certain purposes, it might be more useful to have a smaller number of scales that represent the majority of interplan variation. Even when measures are not closely related empirically, composite indices can simplify data reporting, although in this case weighting of the measures is more complicated substantively and statistically (Hurtado, Swift, and Corrigan, 2001).

A disadvantage of such composite QIs is that it might be harder to interpret a measure of a broad quality domain than a measure of a specific process of care, and the former are less likely to suggest foci for quality improvement.

A current CMS priority is to develop a smaller set of performance indicators that can be used to inform consumers about health care system performance and to provide staff at CMS with more concise information about variations in quality. In this article, we describe the first part of a project to help meet these needs. A major goal of the proposed analyses was to determine how information from multiple measures could be combined in a way that is meaningful and useful to consumers and

The authors are with the Harvard Medical School. The research in this article was supported under HCFA Contract Numbers 500-95-007 and 500-95-0057(TO#9). The views expressed in this article are those of the authors and do not necessarily reflect the views of the Harvard Medical School or the Centers for Medicare & Medicaid Services (CMS).

CMS officials. Other potential users of such composites include providers, health care administrators, State organizations, regulatory groups, and professional organizations.

METHODS AND RESULTS

Overview

If measures are strongly associated, there is relatively little loss of information when only a summary score is reported. Conversely, measures that are weakly associated each carry distinct information that is lost in the summary measure, and the summary measure itself depends greatly on how these distinct pieces of information are weighted together. Thus, our first task in developing summary scores was to determine which measures were most strongly associated.

In the first stage of the project, we focused on two sets of standardized measures that are collected from most of Medicare+Choice plans (M+C), CAHPS® survey and Medicare HEDIS®.

Starting in 1997, CMS mandated that health plans report on their quality using HEDIS® and CAHPS®. Medicare HEDIS® is the first tool to assess the quality of medical and mental health service delivery in Medicare health plans nationally. The MMC CAHPS® survey is the first standardized survey administered by a single vendor to enrollees in a national sample of managed care health plans.

Only a subset of plans provided data on all measures used in the summary indicators. For the remaining plans, we imputed the missing scores. The imputed values are estimates of what the plan's score would be, based on its scores on other measures. Once missing values are imputed, one can calculate composites for every plan.

We identified groups of highly correlated measures using factor analysis. To assess the internal consistency of the resulting composites we calculated Cronbach's coefficient alpha (Nunally, 1978). We evaluated several alternative weighting strategies and calculated the uncertainty due to imputation and sampling.

HEDIS® and CAHPS® Data

CAHPS® data are collected each year in a comparable manner from all M+C plans that meet certain criteria. Specifically, CMS draws a stratified sample of non-institutionalized Medicare beneficiaries who have been enrolled in an eligible plan for a minimum period (Goldstein et al., 2001). Eligible plans include all risk and cost health plans with Medicare contracts in effect on or before January 1 of the survey year, and in business for 2 years. For the analyses presented here, cases sampled from contracts that had ceased activity or been terminated prior to the survey and beneficiaries that left their plan or otherwise became ineligible before the survey was administered were deleted. Because the survey was conducted in a consistent manner by a single vendor, a mean score on every measurement was obtained from every eligible plan, although a few were too small to provide a sample of the required size.

The CAHPS® data that we used came from the third administration of the Medicare CAHPS® survey, conducted in fall 1999. These data include four overall ratings (of plan, doctor, care, and specialist), each reported on a 0-10 scale. The remaining CAHPS® data were based on 20 items that asked the respondent to report on specific treatment experiences and interactions with the health plan. These were summarized into five CAHPS® composite scores, (1) getting care you need—four questions,

(2) getting care quickly—four questions, (3) communication with providers—four questions, (4) courtesy and respect of doctor's office staff—two questions, and (5) paperwork, information and customer service—three questions, and three individual measures (pneumonia vaccination, received a flu shot last year, and advised to quit smoking). The CAHPS® measures were adjusted for case-mix differences among plans to remove the part of the scores related to beneficiary characteristics, such as health status, education, and age (Zaslavsky et al., 2001).

HEDIS® data are collected by plans. The definition of the denominator population varies from measure to measure, depending on clinical appropriateness and required duration of enrollment (Schneider et al., 2001). Some plans draw samples of eligible members for medical record abstraction for some measures. (More information on the definitions and eligibility standards for the HEDIS® measures can be found at the following Web site: http://hcfa.gov/medicare/opl047.htm)

HEDIS® guidelines distinguish between two different types of missing values for each measure: not applicable (NA) and no report (NR). Health plans report NA when they do not have a large enough population eligible for the measure to calculate a HEDIS® score (e.g., many measures require that rates be based on at least 30 members) or when CMS suppresses reporting of a rate that might allow a beneficiary to be identified because the number of eligible cases is very small. Health plans report NR when they choose not to calculate and report a rate or when the health plan's HEDIS® compliance auditor determines that a rate is materially biased (defined as a deviation of more than 5 percentage points from the true rate). In either case, the rate is not available, but because the NR values are determined by the choices and information capabilities of the plan, while the NA cases reflect data that are missing for reasons not under the control of the plan, they could give different information about the unobserved values. The 14 HEDIS® measures we analyzed are listed in Table 1 with the number and percentage of NA and NR missing values.

Units of Analysis

The usual unit of analysis for MMC quality measurement is the contract, representing the population served by a health plan either in a single State or in a few cases in a compact area crossing State lines. There were a few large contracts that were divided into multiple reporting units for CAHPS®, HEDIS®, or both. When the CAHPS® and HEDIS® reporting units within a contract could be matched up by examining the lists of counties in which their sample was located, we treated the matched CAHPS® and HEDIS® units as a single unit in the analysis. When the units could not be matched exactly because the definitions of areas within the contract differed too much for the two systems, we entered the CAHPS® and HEDIS® units as separate, unmatched units. Throughout this article, we use the term unit to refer to these units of analysis, the smallest identifiable for quality reporting. Although in most cases these units are a plan's enrollment in a State, in some cases there is more than one contract for a plan in the same State and/or more than one reporting unit within a contract. We did not have information on the extent to which these multiple contracts or areas represent administratively distinct operating units of a plan.

CAHPS® data were reported for 367 units. At least one HEDIS® 2000 (calendar year 1999) measure was reported for 299 units. Combining the data sets we have 393 units, of which 275 had both CAHPS®

Table 1

Number of Units[1] Missing Data, by HEDIS® Effectiveness of Quality of Care Measures[2]: 2000

Measure	Not Applicable		Not Reported		Total Missing	
	N	Percent	*N*	Percent	*N*	Percent
CHD						
Beta Blocker after AMI	107	35.4	12	4.0	119	39.4
LDL-C Screen after Acute Event	88	29.1	18	6.0	106	35.1
LDL<130 after Acute Event	88	29.1	24	7.9	112	37.1
Screen						
Breast Cancer Screen	20	6.6	8	2.6	28	9.3
Controlling High Blood Pressure (1st Year Measure)	1	0.3	39	12.9	40	13.2
Mental Health						
Followup after Hospitalization for Mental Illness						
7 days	143	47.4	19	6.3	162	53.6
30 days	143	47.4	20	6.6	163	54.0
Antidepressant Medication Management	130	43.0	30	9.9	160	53.0
Diabetes						
Lipid Profile	5	1.7	8	2.6	13	4.3
Lipid Control	5	1.7	19	6.3	24	7.9
Hemoglobin Testing	5	1.7	8	2.6	13	4.3
Eye Exam	5	1.7	10	3.3	15	5.0
Nephropathy Monitor	5	1.7	11	3.6	16	5.3
Poor Hemoglobin Control	5	1.7	26	8.6	31	10.3

[1] A unit is either a managed care contract, or a portion of a large contract defined for reporting purposes.

[2] *N*=302.

NOTES: HEDIS® is Health Plan Employer Data Information System. CHD is coronary heart disease. AMI is acute myocardial infarction. LDL is low-density lipids.

SOURCE: Zaslavsky et al., Harvard Medical School, Boston, Massachusetts, 2002.

data and some HEDIS® data, 24 had HEDIS® data only, 92 had CAHPS® data only, and 2 did not have CAHPS® or HEDIS® data. Only a few of these units (two with only CAHPS® data and eight with only HEDIS® data) represented cases where data were reported for the contract, but could not be matched for units within the contract. Our final analyses include only the 367 units with CAHPS® data.

Imputation of Missing Data

Some simple methods for dealing with missing HEDIS® data could yield biased results. For example, if we averaged the HEDIS® rates for which data were available, a unit could improve its score by failing to report the rate for a measure that is generally low compared with other rates. Similarly, if we used the mean score of all units with data in place of a score that is missing, we could be ignoring information contained in the other scores for that unit indicating that the unit is generally low or generally high and therefore, likely to be below or above average on the measure that is missing. These methods could also cause biases in estimates of correlations and factor analyses.

Modern statistical methods for dealing with missing data in surveys are applicable to this problem (Little and Rubin 1987; Rubin, 1987; Schafer, 1997). To illustrate the concepts underlying these methods, consider a situation in which only one variable, for example, beta-blocker utilization rate, is sometimes not reported. In that case, we might regress the beta-blocker rate on all the other variables in the data set, using the cases for which that rate was provided. We could then use the estimated

regression relationship to predict beta-blocker rates for the cases where the rate is unknown. Furthermore, we could quantify our uncertainty about that rate using the predictive error of the regression (i.e., the residual variation of the true values around the regression line).

In practice, the methods we use are more complex, for two reasons. First, different sets of variables are missing for different units. In many cases, more than one variable must be imputed for the same unit. Hence, imputation cannot be done using a single regression model. The procedures we used predict the missing values for each unit from the values of all CAHPS® and HEDIS® scores available for that unit.

Second, the imputed values are estimated with some error, and this should be taken into account in estimating how uncertain our estimates for each unit are. Thus, we used multiple imputation, creating multiple draws of each of the missing values, which can be combined to estimate the uncertainty due to missing data in the summary score.

Imputation is generally appropriate when the data are missing at random, but not missing completely at random, meaning that the missing values might be related to some observed predictors (Little, 1987). We were concerned that units missing data, especially if data are missing because of the plan's own actions or choices, might tend to have lower quality. For example, not reporting the rate of appropriate use of beta blockers might be an indicator of poor quality. While we did not know rates of beta-blocker use at units where it was not reported, we could test whether missing one score was a predictor of lower results on another score. For example, we divided units into two groups—those whose beta-blocker score was present and those whose score was missing—and then compared the mean breast cancer screening scores in those two groups. Generally, the units with more missing data tended to have lower scores on the observed variables, particularly when data were missing due to voluntary failure to report (NR) rather than because the eligible sample size was too small to meet minimum requirements (NA). Units with many NR values tended to score even lower than would be predicted by imputing them with knowledge of the other observed scores.

To better capture this pattern, after investigating a number of alternative specifications, we included three additional variables in the imputation model. One variable indicated whether any of the scores concerned with screening procedures were NR missing. Similar variables were defined for missing any diabetes care measure, or any mental health measure. After imputing values using these variables as well as the available measures, we compared imputed values to observed values for the units where a variable was reported, and found no systematic differences. Thus, this imputation approach adjusts for the predictable differences in scores associated with the pattern of missing data.

Factor Analyses

We conducted exploratory factor analyses of the HEDIS® rates and CAHPS® scores using the data sets completed by multiple imputation. Analyses were conducted with and without the four CAHPS® overall ratings. Results were similar with and without including the four CAHPS® overall ratings, so the ratings were included in subsequent analyses.

Oblique rotated (Promax) solutions were computed for 8 HEDIS® and 12 CAHPS® quality of care measures. HEDIS® measures for 7- and 30-day followup after hospitaliza-

tion for mental illness, antidepressant medication management, and followup treatment after an acute myocardial infarction (beta blocker prescription, LDL screen, LDL < 130 mg/Dl) were excluded after preliminary analyses because of the high percentage of missing data (Table 1). Oblique rotation was used because previous research suggested that it is unlikely that even measures in distinct groups would be uncorrelated.

In preliminary analyses (data not shown), the four factor solution was the most interpretable, and accounted for 65.6 percent of the variation in the measures. The eigenvalue for the fifth factor was less than 1.0, and it explained 2.9 percent of the total variance. With the exception of getting needed care and advice about quitting smoking, each measure loaded well on one of the four factors and the groupings were consistent with our previous research on the dimensions underlying unit-level quality measures in Medicare (Zaslavsky et al., 2000a; Zaslavsky and Cleary, 2001). The measure about quitting smoking is comprised of a single item and has not been correlated with other measures in previous investigations. In addition, other research we have done suggests that whether smoking cessation advice is offered depends a great deal on the geographical area in which the members are located, with very little influence of the plan in which they are enrolled.

The composite for getting needed care loaded on both the office care and access and customer service constructs. Since this measure was comprised of four survey items that may measure aspects of treatment and the plan, we re-analyzed the data using the individual items in the getting needed care composite.

The re-analysis also suggested that a four-factor solution, which explained 64 percent of the variation in the measures, was the most interpretable (Table 2). The first factor or construct that was identified, office care, includes seven CAHPS® measures of patients' experience with their treatment providers: ratings of personal doctors, specialists, and treatment overall, as well as reports about getting treatment quickly, getting necessary treatment, how well providers communicate, and courtesy, respect and helpfulness of the office staff. The second construct, clinical quality, is comprised of the eight HEDIS® measures: breast cancer screening, blood pressure control, and six measures related to diabetes treatment (hemoglobin testing, poor hemoglobin control, eye exam, lipid profile, and lipid control).

The third construct, access and customer service, concerns health plan administration, services, and access policies. The overall health plan rating, the composite about experience with plan administrative services, such as filling out paperwork and getting help from customer service, and three items from the getting needed treatment composite (getting a personal doctor, getting a referral to a specialist, treatment delays while wait for plan approval) loaded well on this construct.

The fourth construct, vaccinations, includes reports about getting vaccinations for the flu and pneumonia. As in the preliminary analyses, the measure of advice to quit smoking did not load well on any of the factors.

We computed Cronbach's coefficient alpha and corrected item-to-total correlations using unit-level data to evaluate the internal consistency reliability of the four composites (Table 3). The office care composite consists of 14 items, the access and customer service composite is composed of 7 items, vaccinations includes 2 items, and the clinical quality measure is composed of 8 measures. Reliability estimates of 0.70 or greater are generally

Table 2

Factor Pattern Matrix (Loadings[1]) for the CAHPS® and HEDIS® Measures with Individual Items Entered for the Getting Needed Care Composite

Measure	Office Care	Clinical Quality	Access and Customer Service	Vaccination
CAHPS®				
Personal Doctor Rating	[1]0.713	0.055	0.250	-0.429
Specialist Rating	[1]0.673	0.035	0.076	-0.133
Overall Care Rating	[1]0.914	0.046	0.132	-0.242
Overall Plan Rating	0.068	-0.002	[1]0.783	0.007
Getting Needed Care Items				
Get a Personal Doctor or Nurse	-.006	0.018	[1]0.705	-0.052
Get Referral to a Specialist	0.246	-0.003	[1]0.511	0.176
Get the Care You or a Doctor Believed Necessary	[1]0.539	-0.043	[1]0.189	0.252
Delays in Health Care Waiting for Approval from Health Plan	0.165	-0.024	[1]0.613	0.239
Getting Care Quickly	[1]0.855	-0.042	-0.162	0.335
How Well Providers Communicate	[1]0.956	-0.063	-0.068	-0.076
Office Staff	[1]0.864	-0.037	-0.155	0.208
Health Plan Administration	-0.029	0.037	[1]0.617	0.158
Pneumonia Vaccination	-0.095	0.069	0.102	[1]0.728
Flu Shot	-0.047	-0.002	0.086	[1]0.777
Smoking Advice	-0.148	0.197	0.199	-0.047
HEDIS®				
Mammography	0.062	[1]0.530	0.127	0.238
Blood Pressure Control	-0.030	[1]0.566	0.060	-0.153
Hemoglobin Testing	0.140	[1]0.791	-0.203	0.164
Poor Hemoglobin Control	0.135	[1]0.468	-0.102	0.163
Eye Exam	-0.050	[1]0.461	0.136	0.198
Lipid Profile	0.029	[1]0.896	-0.004	-0.181
Lipid Control	-0.072	[1]0.830	-0.041	-0.027
Nephropathy Monitor	-0.190	[1]0.486	0.134	0.098

[1] Loadings greater than 0.35.

NOTES: CAHPS® is Consumer Assessments of Health Plans Study. HEDIS® is Health Plan Employer Data Information System.

SOURCE: Zaslavsky et al., Harvard Medical School, Boston, Massachusetts, 2002.

considered acceptable (Nunally, 1978). Reliability estimates for the measures ranged from 0.86 (clinical quality and vaccinations) to 0.96 (office care).

In general, corrected item-to-total correlations were high (Table 3). Only two items had correlations with the total composite score less than +0.50; blood pressure control (clinical quality measure, +0.45) and waiting 15 or more minutes past the appointment time (office care, +0.45). The item about office wait times is the only item on the survey for which the response choice "Never"' is the most positive and the response choice "Always" is the most negative. Previous research has also shown that this item does not have a high correlation with the other measures. Excluding these items would have only a minor effect on the level of reliability for the corresponding composite measure so we have included them in our analyses.

Calculation and Weighting of Composites

To combine measures into summary indicators, we had to determine how to weight the different components. One approach is to weight items according to expert judgments about the importance of different measures. Since we did not have such information, we considered and compared several possible empirical approach-

Table 3

Internal Consistency Reliability Estimates and Item-to-Item Correlations for the Summary Measures

Summary Score	Measure	Measure Component Item	Item-to-Total Correlation
Office Care	Personal doctor rating	NA	0.66
	Specialist rating	NA	0.63
	Overall care rating	NA	0.87
	Getting care quickly	Got help/advice needed	0.85
		Saw provider as soon as wanted	0.71
		Got care as soon as wanted	0.75
		Office wait to see provider	0.45
	How well providers communicate	Listened carefully to patient	0.88
		Explained clearly to patient	0.82
		Showed respect for patient	0.84
		Spent enough time with patient	0.85
	Office staff	Office staff courteous	0.80
		Office staff helpful	0.86
	Getting needed care	Get the care you or a doctor believed necessary	0.68
Access and Customer Service	Overall plan rating	NA	0.81
	Health plan administration	Understanding written materials	0.56
		Get help from customer service	0.73
		Filling out paperwork	0.68
	Getting needed care	Get a personal doctor or nurse	0.58
		Get referral to a specialist	0.59
		Delays in health care waiting for approval from health plan	0.72
Vaccinations	Pneumonia vaccination	NA	0.76
	Flu shot	NA	0.76
Clinical Quality	Mammography screen	NA	0.62
	Blood pressure screen	NA	0.45
	Hemoglobin testing	NA	0.72
	Poor Hemoglobin control	NA	0.52
	Eye exam	NA	0.59
	Lipid profile	NA	0.68
	Lipid control	NA	0.71
	Nephropathy monitor	NA	0.54

NOTES: NA is not applicable because measure is a single item. Office care coefficient alpha equals 0.96. Access and Customer Service coefficient alpha equals 0.88. Vaccinations coefficient alpha equals 0.86. Clinical Quality coefficient alpha equals 0.86.

SOURCE: Zaslavsky et al., Harvard Medical School, Boston, Massachusetts, 2002.

es. CAHPS® items have different response scales (0-10, 1-4, 0-1, 1-3). We considered weights that are normed by the scale of the item, dividing the observed mean by some measure of the spread of the responses. Specifically, we divided each item score by the standard deviation (SD) of the unit means for that item, and summed. This scaling makes the SD of the unit means the same for every item before they are combined, giving them all equal weight. Two alternative scaling approaches, norming each item by the SD of the individual responses for that item, or downweighting items with low reliability, gave very similar results (correlation >0.90 for each score). For the clinical score, we considered similar weightings, with the additional alternative of a simple sum of the HEDIS® item scores (all on a comparable percentage scale). Again, the scores calculated with the various weightings were very similar (correlation >0.99).

Because the quantitative differences among the methods are small, we used the same method (norming by the SD of means across units) for both the CAHPS® and clinical summary scores. To make the

results more interpretable, we transformed the clinical summary score to match the original percentage scale, so a unit with a zero-percent rate on every measure would have a score of zero and one with a 100-percent rate on every measure would have a score of 100.

Uncertainty Due to Imputation and Sampling

The summary scores are based, to varying degrees, on imputed data, so there are two components of score variability, variability due to sampling error, and variability due to missing data. To estimate the component due to missing data, we calculated the SD of the multiple imputed values for each unit.

The imputation variance generally depends on how many scores are being imputed for each unit (because the scores that are observed rather than imputed do not vary from one imputation to the next) and how well we can predict the missing values from the observed information. In general, imputation variance should be larger for units for which more missing data are imputed.

The sampling variance for each HEDIS® measure depends on the sample size and the rate for the measure. The sample sizes vary from unit to unit and by measure, since each measure can have a different eligible population. We did not have information on the size of the denominator for each measure in each unit, but we were able to calculate the variance for each measure from the confidence intervals that were supplied. The various HEDIS® scores were calculated from independent samples, and therefore we could calculate the sampling variance of the summary measure using the usual formula for the variance of a sum of independent random variables,

$$V_{tot} = \Sigma W_i^2 V_i$$

where V_{tot} is sampling variance of the total, W_i is the weight for score i, and V_i is the sampling variance for score i. We combined sampling and imputation variance under the assumption that the errors are independent. In fact, they are probably positively correlated, because the sample estimates are predictors of the imputations, but it would be very complex to estimate the covariance term.

When we examined the square root of the combined sampling and imputation variances, clinical summary scores for the units with complete HEDIS® reports typically have the smallest standard errors (SEs) and those with no observed scores have the largest SEs.[1] All the units with complete data have SEs less than 1.88, and 80 percent of them have SEs between 0.74 (10th percentile of distribution of SEs) and 1.05 (90th percentile). On the other hand, for units with no HEDIS® data, SEs (arising solely from imputation variance) range from 5.09 to 10.72 (median = 7.77). However, of the 30 units with only a single missing HEDIS® score, 22 have SEs less than 2, as did 2 additional plans with more than one missing HEDIS® score. Thus, imputation enables us to generate a usable clinical performance summary for a substantial number of units with some missing data. Imputation generally has only a minor effect on the means or distribution of scores.

The SEs of the individual HEDIS® measures are more variable (Table 4) than those of the summary measure for units with complete data. For each individual measure, at least one-half the units have SEs falling into a narrow range between approximately 2 and 2.5 percent, probably because they used the sample size that is recommended as a HEDIS® minimum.

[1] Figures showing the distributions of SEs of the clinical summary score, for all units and by number of reported HEDIS® scores, are available from the authors.

Table 4

Distribution of Standard Errors for HEDIS® Measures Used in Summary Scores

Measure	*N* Valid	Minimum	Percentile 25th	Percentile 50th	Percentile 75th	Maximum
			Median			
Mammography	274	0.23	1.77	2.35	2.58	7.74
Blood Pressure Control	262	0.49	2.31	2.43	2.51	8.83
Hemoglobin Testing	289	0.28	1.78	2.06	2.35	7.30
Poor Hemoglobin Control	271	0	2.19	2.39	2.51	6.16
Eye Exam	287	0.28	2.28	2.43	2.53	8.38
Lipid Profile	289	0.35	2.02	2.24	2.47	7.48
Lipid Control	278	0	2.41	2.49	2.53	7.75
Nephropathy Monitoring	286	0.31	2.20	2.41	2.52	7.17

NOTE: HEDIS® is Health Plan Employer Data Information System.

SOURCE: Zaslavsky et al., Harvard Medical School, Boston, Massachusetts, 2002.

Some plans, however, collected larger or smaller samples due to their choice of methodology and the size of the populations eligible for each measure, and correspondingly have smaller or larger SEs for their measures. Thus, for each measure the smallest SE is less than 0.5 percent and the largest is more than than 6 percent. Combining several measures into a summary score smooths out the variation in the SEs, because the same units do not have the smallest or largest SEs on all items.

To be useful for quality reporting, the scores should be sufficiently precise to distinguish above- and below-average units. A useful standard for adequacy of precision is whether units that are moderately better or worse than average, relative to the overall distribution of plan quality, are likely to be reported as significantly better or worse using the conventional tests (2-sided 5 percent-level *t*-test) typical of CAHPS® reporting. Zaslavsky (2001) argues that reasonably good power is obtained for these tests when the variance of the means across units is approximately six times the estimation variance. (This corresponds to an interunit reliability of 0.83.) The variance of the unit scores is 83.5, so a reasonable standard would be that the variance of the estimate for each unit should be no more than approximately 7 (corresponding to a SE of 2.65). All of the units with complete HEDIS® data (220 units), or a single missing item (30 units), satisfy this criterion, as do some (13 out of 18) of those with two to five missing items, but none of those with more missing items than that. Thus, by this criterion 43 units with imputed data have adequate precision. With the 220 with complete data, we have 262 units, leaving 105 that do not meet the criterion.

Altogether, 224 out of 367 units are significantly ($p< 0.05$) different from the mean, with 133 significantly above and 91 significantly below. Because the SE of the score differs so much across plans, some units are significantly different from average with a smaller absolute difference than other units that were not significantly different from average.[2] Thus, hypothesis tests are not sufficient for determining which units are especially good or bad. Interestingly, most of the units with from 1 to 6 measures reported (from 2 to 7 missing) were significantly different from the average unit (20 out of 24). Although the SEs were comparatively large for these units, their scores were so extreme (usually low) that they could be distinguished from the average.

[2] Figures showing the SE of the clinical summary score plotted against the score itself are available from the authors.

Although the summary scores for the units with all HEDIS® data imputed are moderately dispersed, only 5 of these 93 units falls outside the ±2 SE limits, including one that had the lowest scores on 2 of the 3 CAHPS®-based summaries (and a low score on the remaining summary), and therefore had very low imputed HEDIS® scores as well. These units are unlikely to have a statistically significant difference from the mean for two reasons. First, the SEs are so large that only an extreme score would differ significantly. Thus, some units in this imputed-HEDIS® group have scores that are extreme (above or below average), but the imputation SEs are so large that they are not significantly different from the mean. Furthermore, the imputed scores are probably less dispersed than the true performance of this group of units. This is because the CAHPS® scores are only moderately predictive of which units are better or worse than average, so regression to the mean yields underdispersed mean imputations. Consequently, the clinical summary score has relatively little power to detect differences among these units.

There are no missing data on the unit CAHPS® scores, so only sampling variance need be considered. All of the CAHPS® measures for a given unit are taken from the same sample (with some selection due to skip patterns in the questionnaire), and therefore are independent. Therefore, estimation of the sampling variance of a unit mean must consider all the items at once. Although this calculation is complex, it is already implemented in the CAHPS® analysis program (Agency for Health Care Policy and Research, 1999); the only special modification required is to multiply the responses by the corresponding item weights before analysis.

The sampling variances of the CAHPS®-based composites are about as variable as those of the clinical composite when HEDIS® data are complete.[3] This reflects the uniform sizes of the CAHPS® samples. By the criterion previously mentioned (ratio of between-unit variance to sampling variance > 6), 310 of the 367 units have adequate precision for the office/doctor summary score, 351 for the customer/plan score, and 357 for the vaccination score. Thus, hypothesis tests can be almost as consistently used here to distinguish above- and below-average-performing units as with the clinical measures for units with complete data. Consequently, for each of the scores, a large number of units are significantly above or below average: out of 367 units, 192 for the office/doctor score (114 significantly high, 78 significantly low), 191 for the customer/plan score (109 significantly high, 82 significantly low), and 191 for the vaccination score (103 significantly high, 88 significantly low). The fraction of units that are significantly high or low is greater than is typical on the individual rating items or the standard CAHPS® reporting composites, because these summaries (except for the vaccination score) include more items and therefore are more reliable.

SUMMARY AND CONCLUSIONS

The 20 measures we analyzed can be used to create 4 summary measures that are internally consistent and have reasonable substantive interpretations. One summary, referred to as office care consists primarily of measures that reflect the behavior of clinicians and office staff. The summary called access and customer service, on the other hand, reflects primarily plan functions. The clinical quality summary consists of the HEDIS® indicators and vaccinations appears to be distinct from the previously mentioned factors.

[3] A figure showing the distribution of the sampling SEs of CAHPS®-based summary scores, omitting the five units with the largest sampling variances, is available from the authors.

It is not clear why the indicators of clinical quality should form a separate dimension of unit variability from the other factors. Finding factors that corresponded to the source of data was not a goal of the analyses. Previous analyses have shown that the HEDIS® and CAHPS® indicators are related in plausible ways (Schneider et al., 2001), but the analysis presented here shows that the HEDIS® measures are more strongly related to each other than they are to subsets of the CAHPS® measures. This might occur because there are some plan or health care system characteristics that affect multiple facets of clinical quality. This common variation in HEDIS® measures might also reflect differences in measurement and/or information systems. Whereas CAHPS® is administered uniformly by a single vendor for the entire country, each plan implements the measurement process independently for the HEDIS® measures. Thus, HEDIS® and CAHPS® measures might cluster separately because they are subject to different types of measurement error. Vaccination rates formed a dimension of quality distinct from the HEDIS® score, although like the HEDIS® measures it represents conformity to a specific clinical guideline. Vaccination rates are measured by survey, and therefore this measure does not rely on the varying capabilities of the plan information systems. Another possible explanation for the distinctness of the vaccination rate measures is that some vaccinations are often provided as a community-based service, rather than under the auspices of the plan or primary provider, and therefore their distribution might differ from that of other clinically-recommended services.

We modeled the relationships among the variables and the patterns of missing data so that we could calculate summary measures even when some of the constituent variables were missing. Thus, the maximum number of variables could be used to assess different health plans with comparable summary measures. By estimating the missing performance scores, comparative summaries could be made across units that had differing patterns of missing data. This is an essential capability when some units are either ineligible (due to small sample size) or unable to generate data for all of the HEDIS® measures. This method depends more on statistical models than does one that simply averages available data, but the statistical methods involved are objective and do not engender the known biases and opportunities for gaming that would occur with apparently simpler ad hoc methods of handling missing data. The alternative would be to limit sharply the number of HEDIS® measures used and not provide a summary of HEDIS® measures for units that did not have complete data on the remaining measures.

Our four quality measures could be further reduced, but to do so would involve combining data on distinct aspects of quality. The comparative results would therefore be more sensitive to the weights given to the different summaries among the current four, for example, the relative importance given to plan-driven (customer service and access to services) and doctor-driven (care received at the doctor's office) aspects of health plan quality. We see no basis for determining such weights, and therefore, do not recommend further reduction of the number of summaries beyond what we have presented.

Generally, summaries that combine more items should be more reliable than single items or composites of just a few items, as long as the weights are chosen well. The reliability of the CAHPS®-based measures was very good for each score, consistent with findings that ratings and composite measures previously used were quite reliable (Zaslavsky et al., 2000b). For the clinical score, there is a large range in

the precision of the estimates. This occurs in part because sample sizes were small at some units, particularly for some measures (such as use of beta blockers after an acute myocardial infarction that apply to narrowly defined patient subpopulations. However, the clinical score was sufficiently precise for every unit that reported all of the HEDIS® scores: several imprecise measures can be combined to obtain one score that is more precise. Even for some units whose sample sizes fell below the threshold for reporting of some individual measures, the limited data could have been useful (possibly in combination with imputed data) as part of a composite indicator. This is an important advantage of such indicators.

A more important cause of imprecision in the clinical scores was non-reporting of some measures by some units. Some of this non-response would not have occurred if units with small sample sizes had submitted their data, as previously suggested; pooling across years could also help to improve precision for such units. Given the limitations of available data, imputation makes it possible to compute scores even in the presence of non-response, but the SEs reflect the uncertainties of this imputation and are larger when more measures had to be imputed.

There is nothing wrong with reporting an imprecise measure, as long as the reader is always aware of the imprecision. However, if information about precision is not presented or is unlikely to be understood by the intended audience of a report, such imprecise measures could be misleading. Standard hypothesis tests are valid because imputation variance is incorporated into the test. However, non-significant results might be overinterpreted as indicating that the unit's performance is known to be about average, when in fact they are due primarily to lack of information. Thus, it might be desirable to suppress the clinical summary report for some units.

We see little reason for concern about public acceptability of imputation of one or a few of the HEDIS® scores, accounting for a small part of the summary measure, especially given the strong associations among the HEDIS® measures. For units with all of the HEDIS® measures unreported, the imputed scores represent a prediction from unit's CAHPS® data of what the HEDIS® measures might have been if they had been measured. Thus, while there is little additional information in these scores, they represent the best value that can be used for comparison to other units for which HEDIS® scores are available. Because the imputation methodology includes the uncertain of this prediction in the estimated SE, evaluation of these comparisons can be considered in the same terms of acceptable criteria for estimation error previously discussed. It also might be desirable to note for the reader those scores that depend largely on imputed data, but again the relevance of this information depends on the audience. Plans should be encouraged to provide all HEDIS® measures for which they are eligible, for example, by noting in public reports which plans voluntarily withheld some measures, so that they do not have an incentive to selectively report only the measures on which their performance is good.

As already indicated, CAHPS® measures were adjusted for case-mix differences among units. A study of HEDIS® data from commercial plans suggest that case mix also affects ratings on HEDIS, measures, and therefore might affect summary scores (Zaslavsky et al., 2000c). It would be worthwhile to investigate whether this is true for M+C plans. If this is so, case-mix adjustment of the clinical summary score should also be considered.

The technical results presented in this article do not answer the policy questions of whether and how such summary measures could be used. One potential advantage is their ease of use. If having 4, instead of 20, measures increases the likelihood that consumers and policymakers will scrutinize and respond to quality information, then summary measures definitely will be useful and important. This is likely to be increasingly important as a comprehensive set of indicators is developed for a national health care quality report (Hurtado, Swift, and Corrigan, 2001).

If summary indicators gain more prominence, it also might be argued that the burden of data collection could be reduced by relying on a smaller set of measures. On the other hand, the information in the individual measures could be important to the plans to support targeted quality improvement efforts. Furthermore, use of a smaller set of measures could give plans an incentive to game the system by focusing narrowly on measured services. Thus, there are significant policy reasons for maintaining a broad measure set even if composite indicators are emphasized in reports to consumers.

The usefulness and comprehensibility of the four proposed summary scores should be tested with different potential users. Such evaluations should test the perceived advantages and disadvantages of combining different measures. Any report using such summary measures would need to include a description of the methodological limitations of such summary measures. If such measures are to be widely used it will be important to develop simple ways of accurately conveying these limitations and to assess the impact of stated limitations on the perceived credibility of these summaries. Furthermore, the construction of the scores should be revisited as the Medicare quality measurement systems advance, especially as the completeness of HEDIS® measures is improved and new items are added to the CAHPS® instrument.

ACKNOWLEDGMENTS

We would like to thank Terry Lied, Amy Heller, Liz Goldstein, and collaborators at CMS. We would also like to thank Barents LLC, Westat, Inc., DRC, Inc., and the Picker Institute for their efforts in the CAHPS®-MMC implementation that generated the data on which this article is based.

REFERENCES

Agency for Health Care Policy and Research: *CAHPS®, 2.0 Survey and Reporting Kit*. Rockville, MD. 1999.

Goldstein, E., Cleary, P.D., Langwell, K.M., et al.: Medicare Managed Care CAHPS®: A Tool for Performance Improvement *Health Care Financing Review* 22(3):101-107, Spring 2001.

Hurtado, M.P., Swift, E.K., and Corrigan, J.M. (eds.): *Envisioning the National Health Care Quality Report*. National Academy Press. Washington, DC. 2001.

Lied, T., Sheingold, S., Landon, B., et al.: Voluntary Disenrollment in Medicare Managed Care Plans: A Study of Contributing Factors. Unpublished manuscript. 2001.

Little, R.J.A., and Rubin, D.B.: *Statistical Analysis with Missing Data*. John Wiley & Sons. New York. 1987.

Nunally, J.C.: *Psychometric Theory* (Second Edition). McGraw-Hill. New York. 1978.

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

Schafer, J.L.: *Analysis of Incomplete Multivariate Data*. Chapman & Hall Ltd. New York. 1997.

Schneider, E.C., Zaslavsky, A.M., Landon, B.E., et al.: National Quality Monitoring of Medicare Health Plans: The Relationship Between Enrollees' Reports and the Quality of Clinical Care. *Medical Care* 39(1):1313-1325, 2001.

Zaslavsky, A.M., Beaulieu, N.D., Landon, B.E., and Cleary, P.D.: Dimensions of Consumer-Assessed Quality of Medicare Managed Care Health Plans. *Medical Care* 38(2):162-174, 2000a.

Zaslavsky, A.M., Landon, B.E., Beaulieu, N.D., and Cleary, P.D.: How Consumer Assessments of Managed Care Vary Within and Among Markets. *Inquiry* 37(2):146-161, 2000b.

Zaslavsky, A.M., Hochheimer, J.N., Schneider, E.C., et al.: Impact of Sociodemographic Case Mix on the HEDIS® Measures of Health Plan Quality. *Medical Care* 38(10):981-992, 2000c.

Zaslavsky, A.M., Zaborski, L.B., Shaul, J.A., et al.: Adjusting Performance Measures to Ensure Equitable Plan Comparisons. *Health Care Financing Review* 22(3):109-126, Spring 2001.

Zaslavsky, A.M., and Cleary, P.D.: Dimension of Plan Performance for Sick and Healthy Members in the Consumer Assessments of Health Plans (CAHPS® 2.0.) *Medical Care* Forthcoming. 2002.

Zaslavsky, A.M.: Statistical Issues in Reporting Quality Data: Small Samples and Casemix Variation. *International Journal of Quality Health Care* 13(6):481-488, 2001.

Reprint Requests: Paul D.Cleary, Ph.D., Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899. E-mail: cleary@hcp.med.harvard.edu