Consumer Evaluations of Health Plans and Health Care Providers

Case-Mix Adjustment of the National CAHPS® Benchmarking Data 1.0: A Violation of Model Assumptions?

Marc N. Elliott, Richard Swartz, John Adams, Karen L. Spritzer, and Ron D. Hays

Objective. To compare models for the case-mix adjustment of consumer reports and ratings of health care.

Data Sources. The study used the Consumer Assessment of Health Plans (CAHPS®) survey 1.0 National CAHPS Benchmarking Database data from 54 commercial and 31 Medicaid health plans from across the United States: 19,541 adults (age ≥ 18 years) in commercial plans and 8,813 adults in Medicaid plans responded regarding their own health care, and 9,871 Medicaid adults responded regarding the health care of their minor children.

Study Design. Four case-mix models (no adjustment; self-rated health and age; health, age, and education; and health, age, education, and plan interactions) were compared on 21 ratings and reports regarding health care for three populations (adults in commercial plans, adults in Medicaid plans, and children in Medicaid plans). The magnitude of case-mix adjustments, the effects of adjustments on plan rankings, and the homogeneity of these effects across plans were examined.

Data Extraction. All ratings and reports were linearly transformed to a possible range of 0 to 100 for comparability.

Principal Findings. Case-mix adjusters, especially self-rated health, have substantial effects, but these effects vary substantially from plan to plan, a violation of standard case-mix assumptions.

Conclusion. Case-mix adjustment of CAHPS data needs to be re-examined, perhaps by using demographically stratified reporting or by developing better measures of response bias.

Key Words. Consumer satisfaction, interactions, response bias

INTRODUCTION

Case-mix adjustment of consumer ratings can provide more valid plan comparisons than unadjusted ratings by controlling for factors related to systematic response biases to questions about health care. Adjusted data are therefore potentially more appropriate for comparing the quality of care delivered. If members of a particular demographic group are less inclined than others to assign poor ratings to bad care and members of this group are disproportionately enrolled in some plans, case-mix adjustment for this systematic bias is useful when comparing assessments of different plans. In some ways case-mix adjustment is similar in its intentions to the epidemiologic technique of employing a standardized population. It attempts to estimate the ratings plans would have received if each plan had been rated by the same representative sample of the general population served by the complete set of plans.

Because of the high stakes associated with how plans are rated by consumers, considerable political sensitivity is related to the adjustment and presentation of these data. For example, some plans may be dissatisfied with and dispute their scores. Those plans whose relative scores drop after adjustment might be most likely to protest. On the other hand, inadequate adjustment of consumer ratings for case-mix differences could reduce the acceptance and usefulness of the results for beneficiaries, health plans, and purchasers.

Selection of Case-Mix Adjusters

The goal of case-mix adjustment is to eliminate response bias, defined here as differences in reports and ratings of care that do not correspond to actual differences in quality of care. Differences in response patterns may reflect both real differences in the quality of care and systematic biases in reporting, and it is difficult to separate these two factors. It is important to control for exogenous variables that lead to systematic response bias but to avoid adjusting for variables that reflect or have a causal link to the quality of the

Address correspondence to Marc N. Elliott, Ph.D., Statistician, RAND, 1700 Main Street, M-28, Santa Monica, CA 90407-2138. Dr. Elliott; John Adams, Ph.D., Senior Statistician; and Ron D. Hays, Ph.D., Professor are from the RAND Health Program, Santa Monica, CA. Karen L. Spritzer, M.S. and Ron D. Hays, Ph.D., Professor are from the University of California at Los Angeles. Richard Swartz, B.S. is from Rice University, Houston, TX. This article, submitted to *Health Services Research* on April 24, 2000, was revised and accepted for publication on November 27, 2000.

health plan. This critical concern with the causal nature of the association between ratings and potential "right-hand-side" variables distinguishes case-mix adjustment from inappropriate application of analysis of covariance or multiple regression.

Imagine two hypothetical extremes from a randomized experiment in which people are assigned to health plans. In the first situation there is no response bias, but true quality of care differs on the basis of a variable such as age. In the second case there is no difference in the quality of care by age within plans, but there is response bias related to age. The latter case would be ideal for case-mix adjustment by age because the entire effect of such adjustment would be to eliminate response bias without biasing estimates of true quality of care. In the former situation case-mix adjustment would mask actual differences in quality of care associated with age. To the extent that the truth lies between these extreme situations, the results will be a combination of the above consequences.

One class of potential case-mix adjusters that should be excluded is those that are endogenous, those variables that may themselves reflect satisfaction or quality of care. Utilization may be such a variable. While the Consumer Assessment of Health Plans (CAHPS®) survey requires a minimum length of enrollment for survey completion, there is still considerable variation in utilization in the surveyed population. Those who have experienced long appointment or office waits may be less likely to use the plan for optional procedures or for procedures for which other coverage is available. In this case an observed positive association between utilization and health care ratings might reflect actual differences in care, not response bias. To adjust for utilization in this case would mask actual differences in care. Length of enrollment with provider might be a similar endogenous variable.

Finally, when a case-mix model is constructed, the magnitude of biases must be weighed against the goal of model parsimony. In a regression context the effect of a case-mix adjustment variable on plan ratings is directly related to the product of the coefficient of that adjuster in a person-level regression and the difference between the mean of the plan in question and the overall mean on that adjustment variable. This means that for a case-mix variable to have a practical effect, the characteristic in question must both be strongly associated with ratings and vary significantly among plans. For example, even if the gender of respondents were strongly associated with ratings (and this were thought to represent response bias rather than true differences in care), the fact that most plans differ very little in their gender mix (the ratio of female to male members) dictates that gender would have little effect as a case-mix adjustment variable. Variables unlikely to influence ratings might be excluded on the grounds of parsimony.

The *CAHPS 1.0 Implementation Handbook* (AHCPR 1997a) recommends adjusting for age and health status when comparing consumer assessments of health plans. Younger people and those in poorer health tend to report more problems and less positive evaluations of health care than do older people and those in better health. It is generally believed that these consistent associations primarily reflect response bias rather than better care for the older people and those in better health. There is also evidence that higher education may be associated with less positive evaluations of health care (Fiscella and Franks 1999; Ware, Davies-Avery, and Stewart 1982; Fox and Storms 1981). A priori, one might suspect that those with better education receive better health care, so it might be reasonable to interpret any negative association between education and health ratings as response bias.

Choice of Form for Case-Mix Model

Multivariate regression is the common method of case-mix adjustment (Aharony and Strasser 1993; Cleary and McNeil 1988; Hall, Feldstein, Fretwell, et al. 1990; Kane, Maciejewski, and Finch 1997; Weiss 1988). In this approach the observed value minus the predicted value for each person in the sample is calculated, and the deviation represents the adjusted difference from the overall mean rating. Unbiased estimates of plan differences in the regression approach are based on the assumption that the regression model is correctly specified, the model is the same for all the plans, and the covariates are measured with negligible error. It is possible to test for the second assumption (testing for interactions between plan and the covariates) and the third assumption (estimating the reliability of covariates).

To the extent that the second assumption does not hold, one could argue that there really is no meaningful overall plan difference to be estimated. Under these circumstances stratified reporting may be necessary. For example, imagine that plan A has much better ratings for the elderly than plan B, but plan B has slightly better ratings than A for the young. This would be evident as an interaction between plan and age. It is arguably more plausible that this interaction reflects differential quality of care by age between plans than an interaction in response bias. The former interpretation would require only that plans provide care of different quality according to the age of the patient. The latter interpretation would require that plan memberships differ systematically in response tendencies by age—that some plans systematically attract easier-to-please younger adults but harder-to-please older adults than other plans. Thus, if such a significant interaction occurred, it might be more meaningful to stratify by age and compare plan A's ratings with plan B's within age groups.

Any attempt to specify overall plan differences when significant interactions exist would be tantamount to constructing a weighted average of different effects. While this might ultimately be necessary, one would need to consider carefully what mixing proportions to use as this will affect the relative performance of the two plans.

The CAHPS 1.0 approach uses a "health plan fixed effect" model (k-1 plan level dummies for k plans, with one omitted) to estimate the effects of case-mix adjusters and specifies age (in seven ordinal categories) and self-rated rated health status (poor, fair, good, very good, or excellent) as one degree of freedom (linear) terms.

METHODS

CAHPS Instruments

The CAHPS core survey is currently the national standard for measuring patient experiences with ambulatory care. CAHPS has been adopted by Medicare (Schnaier, Sweeny, Williams, et al. 1999), state Medicaid programs (Brown, Nederend, Hays, et al. 1999), and the National Council on Quality Assurance (NCQA) as part of its accreditation process (NCQA 1998).

The CAHPS instruments were developed by a consortium of investigators from RAND, Harvard Medical School, Research Triangle Institute, and Westat funded by the Agency for Healthcare Research and Quality and the Health Care Financing Administration. A major goal of this project was to produce survey instruments that could reliably and validly measure care as reported by health plan enrollees (Crofton, Lubalin, and Darby 1999). Survey results are primarily intended to inform consumers who are choosing health plans. CAHPSs have been developed for use in commercial and Medicare settings, for assessing adult and child care, and for administration via mail or telephone.

Data Source

The data used in this study are from the National CAHPS Benchmarking Database (NCBD 1.0), collected and administered by the Quality Measurement Advisory Service. The NCBD 1.0 is the first nationwide aggregation of CAHPS 1.0 data. The participants in NCBD 1.0 consisted of Medicaid and commercial sponsors who volunteered to be included. Surveys were fielded in 1997 and 1998. The Medicaid database includes 29 HMOs and two primary care case management plans from seven states. The commercial sponsors database includes information from 27 HMOs, eight physician provider organizations, three point-of-service plans, one fee-for-service plan, and 15 other unspecified health plans from six states. The Medicaid and commercial databases include 8,813 and 19,541 adult respondents, respectively (reflecting response rates of 52 percent and 63 percent, respectively). The Medicaid database also includes 9,871 responses from adults regarding their children (reflecting a 42 percent response rate).[1] The average plan had 284 adults responding about themselves in the commercial sample and 362 adults responding about themselves in the Medicaid sample, both near the CAHPS recommendation of 300 responses per plan. The average plan in the Medicaid sample has 183 adults responding about their children. The characteristics of respondents are described in Table 1.

We considered case-mix adjustment separately for three data sets: adult commercial (ACOM), adult Medicaid (AMED), and child Medicaid (CMED). As can be seen in Table 1, about one-third of the Medicaid samples and about one-fifth of the commercial sample represent ethnic groups other than non-Hispanic whites. The Medicaid samples are approximately 90 percent female, whereas the commercial sample is 59 percent female. Median ages are 25 to 34 years for adults in the Medicaid samples and 35 to 44 years in the commercial samples; median members of the former group are high school graduates, and median members of the latter group have one to three years of college education. Health status averages good for adults in the Medicaid sample and very good for both children in the Medicaid sample and adults

Table 1: Characteristics of Sample

	Adult Commercial	*Adult Medicaid*	*Child Medicaid*
Mean age level (1–7*)	3.37 (1.12)	2.35 (1.07)	2.49 (1.11)
Mean health level (1–5*)	3.74 (.89)	3.08 (1.08)	4.09 (.92)
Mean education level (1–6*)	4.36 (1.16)	3.09 (.99)	3.21 (1.02)
% female (parent/guardian)	59	90	91
% African American	7	18	16
% Hispanic	6	7	12
% Asian American	4	3	3
% other nonwhite	3	6	7
% non-Hispanic white	80	66	62

*Higher scores correspond to older age, better health, and more education.

in the commercial sample. The fact that the standard deviations of the three ordinal case-mix variables are nearly equal to one another in all three data sets suggests that the "levels," or "units," of the three case-mix variables are approximately equivalent and that magnitudes of case-mix coefficients may be compared directly across these three variables.

Measures

The dependent variables for this study consisted of four single-item global ratings (personal doctor, specialists, health care, and health plan) and 17 single items that define five reporting categories (access to needed care, promptness of care, provider communication, staff helpfulness, and health plan customer service). The four global rating questions and the individual questions corresponding to the five composites are shown in Table 2. All items were linearly transformed to a possible range of zero to 100 (with 100 reflecting more positive experiences with care) for presentation and comparison. Although some negative skewness is observed in these dependent variables, results using the variables in this form do not differ appreciably from results when these variables are transformed to symmetry.[2]

The independent variables for adult cases included age (in seven ordinal categories: 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75+), education (in six ordinal categories: eighth grade or less, some high school but did not complete, high school graduate or GED, one to three years of college, four-year college graduate, and more than four-year college degree), and self-rated health status (in five ordinal categories: poor, fair, good, very good, and excellent). For child cases the independent variables were parent/guardian age, parent/guardian education, and child's health status as rated by the parent/guardian. Daley and Shwartz (1994) argue that the relationship between age and many outcomes is unlikely to be linear and that carefully chosen age strata may result in better models. This argument might also hold for self-rated health and level of education. Nevertheless, extensive attempts to empirically determine superior functional forms for the bivariate relationships between these independent variables and ratings yielded very little improvement over the more parsimonious linear forms of the independent variables (treating the ordinal variables as linear). The latter forms were therefore chosen in all cases.

Models

Multivariate analyses were conducted using ordinary least squares regression on four case-mix models for each of the three data sets on each of the 21

Table 2: CAHPS 1.0 Questions from the NCBD Used in This Study

Item	*Response Options*
Single-Item Global Ratings	
Health Plan	
We want to know your rating of all your experience with your health insurance plan. How would you rate your health plan?	0–10 scale
Health Care	
We want to know your rating of all your health care in the last six months from all doctors and other health professionals. How would you rate all your health care?	0–10 scale
Specialty Care	
We want to know your rating of the specialist you saw most often in the last six months. How would you rate the specialist?	0–10 scale
Personal Doctor	
We want to know your rating of your personal doctor or nurse. How would you rate your personal doctor or nurse?	0–10 scale
Multi-Item (Composite) Reports	
Access to Needed Care	
With the choices that your health plan gives you, was it easy to find a personal doctor or nurse for yourself?	Yes (1), no (0)
In the last six months, was it easy to get a referral when you needed one?	
In the last six months, how often did you receive the tests or treatment you thought were needed?	Never (1), sometimes (2), usually (3), always (4)
In the last six months, how often did your health plan deal with approvals or payments without taking a lot of your time and energy?	
Provider Communication	
In the last six months, how often did doctors or other health professionals listen carefully to you?	Never (1), sometimes (2), usually (3), always (4)
In the last six months, how often did doctors or other health professionals explain things in a way you could understand?	
In the last six months, how often did doctors or other health professionals show respect for what you had to say?	
In the last six months, how often did doctors or other health professionals spend enough time with you?	
Staff Helpfulness	
In the last six months, how often did office staff at a doctor's office or clinic treat you with courtesy and respect?	Never (1), sometimes (2), usually (3), always (4)
In the last six months, how often was office staff at a doctor's office or clinic as helpful as you thought they should be?	

Continued

Table 2: *Continued*

Item	*Response Options*
Promptness of Care	
In the last six months, how often did you get the medical help you needed when you phoned the doctor's office or clinic during the day on Monday to Friday?	Never (1), sometimes (2), usually (3), always (4)
In the last six months, when you tried to be seen for an illness or injury, how often did you see a doctor or other health professional as soon as you wanted?	
In the last six months, when you needed regular or routine care, how often did you get an appointment as soon as you wanted?	
In the last six months, how often did you wait in the doctor's office or clinic for more than 30 minutes past your appointment time to see the person you went to see?	
Health Plan Customer Service	
In the last six months, how often did you get all the information or other help you needed when you called the health insurance plan's customer service?	Never (1), sometimes (2), usually (3), always (4)
In the last six months, how often were people at the health insurance plan's customer service as helpful as you thought they should be?	
In the last six months, how often did you have more forms to fill out for your health insurance plan than you thought was reasonable?	

ratings and reports. The four models tested are summarized in Table 3. Model 0 is simply an unadjusted comparison of plan means. Model 1 is the standard case-mix adjustment recommended in the *CAHPS 1.0 User's Manual* (AHCPR 1997b), adjusting for age and self-reported health status. Model 2 adds education to model 1, and model 3 adds age by plan, health status by plan, and education by plan interactions to model 2.

Analysis

All models were fit using ordinary least squares multiple regression. The statistical significance of interactions between a set of k plans and a given case-mix adjuster variable was tested as a block of k-1 terms for each case-mix variable using a partial F test. Statistical significance was assessed at the 5 percent level of significance using two-sided tests. Missing values were rare among the independent variables of interest but were replaced using multiple imputation when they did exist.

Table 3: Four Case-Mix Models Evaluated

Model	*Description*	*Right-Hand-Side Variables*
0	Unadjusted	*k-1* plan dummies
1	CAHPS 1.0 standard	Plan dummies, age, health status
2	CAHPS 1.0 standard + education	Plan dummies, age, health status, education
3	Plan interactions	Plan dummies, age, health status, education, age x plan interactions, health x plan interactions, education x plan interactions

RESULTS

Comparing Models 0, 1, and 2

The variables age and health status functioned very similarly in models 1 and 2, so the focus for exposition will be on model 2 (which also contains education). The health term was significant and positive almost uniformly (62 of 63 cases, where the 63 cases represent 21 items in each of the three samples). As seen in Table 4, the magnitude of the effect of health was substantial, 4.1 to 4.9 points (for a median item) of a possible 100 points per level of the five-point health scale. The coefficients on age were uniformly positive and significant in the commercial data sample (21 of 21 for ACOM) and were positive and significant in 34 of 42 cases in the Medicaid samples (no cases were significantly negative). As shown in Table 4, the magnitude of these coefficients was moderately large, 1.8 to 1.9 points per approximately ten-year level of age for a median item in the adult samples and 1.1 points per level of age (of the parent or guardian) for a median item in the child sample. This suggests that 20 to 40 years of age are approximately equivalent to one level of self-rated health in terms of propensity for positive ratings.

The effects of education were somewhat less consistent. In the adult samples education had a significant negative effect in 32 of 42 cases (and

Table 4: Median Item Coefficients (Model 2; Points/Level)

Health			*Age*			*Education*		
ACOM	*AMED*	*CMED*	*ACOM*	*AMED*	*CMED*	*ACOM*	*AMED*	*CMED*
4.1	4.9	4.3	1.9	1.8	1.1	−1.3	−1.3	−.5*

Note: ACOM = adult commercial; AMED = adult Medicaid; CMED = child Medicaid.

*Some coefficients were positive; median absolute value .8.

no significantly positive effects). Within the child data set education had a significant effect in only 12 of 21 cases, and this effect was negative in only half of those cases. As shown in Table 4, the magnitude of the education effect was also more modest than that for age or health status, about 1.3 points per level (not year) of education for a median item in the adult samples and .8 points per level (in absolute value) for a median item in the child sample.

Table 5 displays the proportion of variance (R^2) in person-level ratings explained by models 0 to 3 in the three samples for the median items. The R^2 for model 0 reflects the magnitude of difference between plans without case-mix adjustment. Plan membership explains 2.1% to 3.4% of this variance for a median item. The increase in R^2 after adding the two case-mix adjusters of model 1 is substantial, approximately equal to the proportion of variance explained by plan membership alone for a median item in each sample.

On the other hand, the increase in R^2 from model 1 to model 2 (adding education) was modest, approximately 2 percent to 10 percent of the model 1 R^2 for a median item in each of the three samples.

It is possible for case-mix adjustment to increase or decrease the variability of plan ratings (estimated mean plan scores on items). In the present data, however, a consistent tendency for models 1 and 2 to decrease the variability of plan ratings relative to model 0 was observed. Standard deviations of plan rating were typically reduced by three percent to six percent with reductions as large as 10 percent to 15 percent for some items.

Table 6 examines the amount of change in plan ratings in pairwise comparisons of models 0, 1, and 2.[3] The standard deviation of the differences in plan ratings between models was compared to the standard deviation of the plan ratings under model 0 for each item. When comparing models 1 and 2 to model 0, the changes for median items were 11 percent to 16 percent of the magnitude of the original standard deviation in plan ratings and 18 percent to 34 percent of that magnitude for the most-affected items. If we interpret these changes as elimination of response bias, models 1 and 2 eliminate response bias that would distort a typical plan's national percentile

Table 5: Proportion of Variance Explained (Median Item; Percentage)

Model	*Adult Commercial*	*Adult Medicaid*	*Child Medicaid*
0	2.5	2.1	3.4
1	4.8	5.6	5.7
2	5.2	5.7	5.8
3	6.8	7.6	7.4

by four points on a typical item and nine points on some items. One plan in 20 would otherwise experience response bias equivalent to ten points for typical items and 25 points for some items. The differences between models 1 and 2 were much smaller, ranging from 4 percent to 11 percent of the original (model 0) standard deviations of plan ratings for median items and from 10 percent to 18 percent of the original standard deviations of plan ratings for the most-affected items. Because CAHPS reports typically summarize ratings with "star" notations, we next examine how these changes in ratings translate to changes in stars.

CAHPS summarizes ratings by assigning plans to three categories: statistically significantly above the mean of other plans (three stars), statistically significantly below the mean of other plans (one star), and statistically indistinguishable from the mean of other plans (two stars). We assigned these ratings to each of the 21 items within each of the three samples of plans for models 0, 1, and 2. Two differences between what was done here and what is normally done in CAHPS should be noted. First, in CAHPS the 17 report items are averaged and combined into five composites for reporting and assignment of stars. The current approach kept them separate for consistency with other analyses in this article and because case-mix adjustment occurs at the level of the individual item. Second, the number of plans in the comparison sets (31 for the commercial sample and 54 in the Medicaid sample) is larger than in most CAHPS reports. This may affect the proportion of ratings assigned two stars. In the unadjusted sample (model 0) 69.2% of CMED, 61.7% of AMED, and 56.7% of ACOM item-plan combinations received two stars. These three proportions are significantly different by a chi-square test of homogeneity ($p < .05$ for each pairwise comparison).

We compared star ratings for models 0, 1, and 2. In no instance did a plan change two stars on any item between any pair of these three models (one star in one model and three stars in another). Changes of one star were

Table 6: Standard Deviation of Plan Rating Changes (Percentage of Plan Rating Standard Deviations Under Model 0)

	Median Item			*Most-Affected Item*		
Models compared	*ACOM*	*AMED*	*CMED*	*ACOM*	*AMED*	*CMED*
0, 1	11	16	15	20	32	26
0, 2	14	15	16	18	34	29
1, 2	11	6	4	18	12	10

Note: ACOM = adult commercial; AMED = adult Medicaid; CMED = child Medicaid.

common for models 1 and 2 when compared to model 0, with more change observed in the Medicaid samples than in the commercial sample. In the Medicaid samples 8 percent to 15 percent of plans changed by one star for median items and 23 percent to 39 percent changed for the most-affected items. In the commercial sample 5 percent to 6 percent of plans changed one star for median items and 11 percent to 13 percent changed for the most-affected items. Changes between models 1 and 2 were much less common, with 2 percent to 9 percent of plans changing ratings for median items and 9 percent to 23 percent changing for the most-affected items.

Assessing the Importance of Interactions

Table 7 presents results regarding the significance and magnitude of interactions between case-mix adjusters and plans. This is equivalent to testing whether case-mix adjusters have the same effect across plans. Overall about one-fourth of case-mix-by-plan interactions were significant. Health-by-plan interactions were significant in 22 of 63 cases, age-by-plan interactions were significant in 11of 63 cases, and education-by-plan interactions were significant in 12 of 63 cases. Lest multiple comparisons be thought to be a concern, even 11 significant results of 63 is extremely unlikely by chance ($p = .0003$). Overall at least one interaction (health, age, or education) was significant in 29 of 63 cases.

One way to assess the magnitude of the interactions is to look at the standard deviation of the plan-specific estimates of the coefficients. For health estimates the standard deviations for median items were 2.5 to 2.8 points (2.4 to 2.6 points for the median of items with significant interactions). For age estimates the standard deviations of median items were 2.3 to 2.8 points (2.4 to 4.6 points for the median of items with significant interactions). For education estimates these standard deviations were 2.2 to 3.0 points (1.8 to 4.7 points when restricted to items with significant interactions). These magnitudes are quite large when compared to the original model 2 estimates of the coefficients in Table 4. In cases with significant interactions the standard deviation of the plan-specific estimates for health were 67 percent to 94 percent of the corresponding model 2 estimated health coefficients in the median case; for age and education the standard deviations of the plan-specific estimates were one to three times the size of the corresponding model 2 estimates in the median case (because of the lesser magnitudes of these coefficients in model 2). To illustrate the magnitude of these interactions, a typical plan in the adult Medicaid sample has at least one report or rating item on which the national sample percentile for those in excellent or very good health differs

Table 7: Statistical Significance and Magnitude of Interactions (Model 3; Points/Level)

	Health			*Age*			*Education*		
	ACOM	*AMED*	*CMED*	*ACOM*	*AMED*	*CMED*	*ACOM*	*AMED*	*CMED*
Items with statistically significant interactions ($p < .05$)	9/21	7/21	6/21	5/21	4/21	2/21	4/21	3/21	5/21
Standard deviation of plan-specific coefficients, median item	2.8	2.8	2.5	2.3	2.8	2.3	2.3	3.0	2.2
Standard deviation of plan-specific coefficients as a percentage of mean plan coefficient, median of items with statistically significant interactions ($p < .05$)	82	94	67	205	355	111	202	338	352*

*Absolute value.

from their national percentile for those in worse health by ten points. One in 20 plans has an item on which their national percentile for the healthier stratum differs from their national percentile for the less-healthy stratum by 30 points.

Change in Variance Explained

As shown in Table 5, the increase in R^2 from model 2 to model 3 is substantial. It is about two-thirds of the magnitude of the increase in R^2 from model 0 to model 1 for a median item in each of the three samples. While the change from model 2 to model 3 involves the addition of a substantial number of right-hand-side degrees of freedom (90 for the commercial sample and 159 for the Medicaid sample), the adjusted R^2 increased uniformly from model 2 to model 3, reflecting the large sample sizes involved.

DISCUSSION AND CONCLUSION

Models 1 and 2

Given the importance of these ratings to consumers and other stakeholders, it is important that the consequences of case-mix modeling decisions be more fully investigated and understood. Case-mix model 1 (the CAHPS 1.0 model) and model 2 (adding education) substantially improve the proportion of variance in ratings that is explained. The three case-mix adjusters have statistically and practically significant effects for most items, and the magnitude is especially substantial for health status. Models 1 and 2 both reduce the variability of plan ratings. The use of either model may therefore reduce the difference between plans perceived by consumers when ratings and reports are summarized as means or proportions in bar-graph displays (McGee, Kanouse, Sofaer, et al. 1999). Model 2 reduces the proportion of plans declared significantly different from the average, suggesting that this model would reduce the difference between plans perceived by consumers when ratings and reports are summarized with the star system. One interpretation might be that model 2 in particular (and therefore education) serves to eliminate some false positive one-star and three-star ratings that had been produced by response bias in model 0. Overall, however, the differences between models 1 and 2 are small.

One should consider the special nature of the health status variable because there exists the possibility of causal relationships between it and plan selection. Consider two models of this relationship. In both models poor health status is associated with low evaluations. In the first model health status differs among plans because of a priori differential assignment to plans (or choice of plans) on the basis of prior health status. This may involve simple geographic convenience, conscious choice, or other factors. The second model is one in which pre-existing health status differences do not exist, but plan choice in fact causes differences in health status to develop.

The first model is the classic case-mix adjustment scenario. In the latter case, if plan choice actually causes changes in health status, one would argue against case-mix adjustment by health status (at least as measured subsequent to current plan enrollment) because true information on quality of care would be eliminated (perhaps in addition to some response bias). In this case one might wish to adjust for health status prior to enrollment in the current plan (because the causality of the second model would not apply here), but one would not want to adjust for changes in health status since enrollment or, by extension, health status measured after enrollment in the current plan. Obviously it is possible for situations to be a combination of the two models. To the extent that the second scenario is true, case-mix adjustment by health status is inadvisable because it reflects true differences in quality of care rather than response bias.

Interactions

In almost half of the cases an item had significantly different coefficients within plans for at least one of the three case-mix variables. In particular the health status coefficient differed within plans in one-third of cases. These differences in coefficients are of a substantial magnitude and constitute not only a theoretical but also a practical violation of the assumptions of case-mix adjustment. Unless one believes that the association between health status and response bias truly differs by plans, the conclusion is that this variability reflects differences in real experiences with care by demographic subgroups across plans.

One might argue that the estimated coefficients from model 2 are estimates of response bias and that the estimated interactions (plan-level deviations from this grand mean) are estimates of plan-level differences in relative experiences and quality of care by demographic subgroup. Suppose that the model 2 coefficient for health status in the regression regarding the overall rating of the health care plan had a coefficient of 5.0 points per level and that this coefficient differed significantly by plan, with an estimated coefficient of 1.0 point per level for plan A and 9.0 points per level for plan B. One might interpret this to mean that while there is a general tendency for the healthy to rate their health plans more highly than do the ill, the difference between the healthy and the ill is profound in plan B and small in plan A. This might suggest that plan B emphasizes the care of the healthy over the care of the ill to a greater extent than does plan A (which may actually emphasize the care of the ill). Under this interpretation we would conclude that the variation in care by demographic status is large relative to response bias in many cases.

At least four reactions to this violation of the assumptions of case-mix adjustment are possible. The first is to simply assign ratings based on the fully interacted model. This seems like a poor solution because it makes the questionable assumption that response bias varies by plan and would effectively eliminate the effects of case-mix adjustment where interactions exist.

A second approach is to continue to case-mix adjust by the model 2 coefficients where significant plan-by-case-mix interactions exist (the average case-mix effects), interpreting these as a measure of response bias, but remembering that under these circumstances the estimates of response bias are likely to be unstable and influenced by the set of plans that are included. Here one should also be aware that the global plan ratings might not well represent some demographic subgroups.

A costly but more satisfying third approach is demographically stratified reporting of ratings of care by plan. For example, for items on which interactions between plan and health status existed in a substantial and significant amount, one could report the ratings and stars for two strata for each plan: those with self-rated health of very good or excellent and those with self-rated health of poor, fair, or good. Within these strata one could traditionally case-mix adjust for other demographic variables that do not interact with plan. This approach is very informative to the consumer but does require sample sizes that are at least double those currently employed to retain the previous level of power to distinguish one- and three-star from two-star plans. Because such an approach involves additional information it is important to consider how to present the information in a targeted manner that would not overwhelm consumers.

A final approach that is not currently available is to try to develop an index or scale of items that are "pure" measures of response bias rather than demographic surrogates for this bias. This might be analogous to "severity" in traditional clinical-outcome case-mix adjustment. If such a measure could be constructed, it would eliminate the inherent difficulty of adjusting on the basis of a variable that simultaneously measures response bias and is associated with real differences in care by plan. Such items would attempt to measure a construct such as "general satisfaction" or "tendency to rate favorably." Perhaps items asking for ratings of the respondent's general life satisfaction, satisfaction with the United States health care system in general, or rating of some fixed quantity that does not vary substantially from person to person on a ten-point scale could play a part in such an index. There is some evidence, for example, that depression is associated with lower reported satisfaction with health care (Hays et al. 1994; Linn and Greenfield 1982). After cognitive testing, such a scale could be validated as a measure of response bias. If it functioned well it would (1) have strong, positive association with ratings and (2) not interact substantially with plans. As an initial step, several items like those suggested above are currently being piloted by CAHPS investigators. Finally, efforts are under way to replicate the above findings using the larger data sets and additional variables available with CAHPS 2.0 data.

ACKNOWLEDGMENTS

Support for this research was received from the Agency for Healthcare Research and Quality and the Health Care Financing Administration under contract No. 5U18HS09204-05. The authors express appreciation to Dale Shaller of the Quality Measurement Advisory Service for his assistance in obtaining the data for this study. The comparative CAHPS data used in this analysis were provided by the NCBD, a collaborative initiative of the Quality Measurement Advisory Service, The Picker Institute, and Westat. The NCBD is funded by the U.S. Agency for Healthcare Research and Quality under contract no. 290-95-2005.

NOTES

1. The NCBD 1.0 data do not provide frame information that would allow formal nonresponse analyses. The response rates observed are not unusually low for populations such as these. Finally, only differential nonresponse that varied systematically by plan would alter the observed results.
2. Only the former are reported here.
3. Table 6 does not contain results for model 3 because it is unclear how one would appropriately assign ratings and stars to plans in the presence of significant interactions.

REFERENCES

Agency for Health Care Policy and Research (AHCPR). 1997a. *CAHPS Implementation Handbook*. Rockville, MD: AHCPR.

———. 1997b. *CAHPS 1.0 User's Manual.* Rockville, MD: AHCPR.

Aharony, L., and S. Strasser. 1993. "Patient Satisfaction: What We Know and What We Still Need to Explore." *Medical Care Review* 50 (1): 49–79.

Brown, J. A., S. E. Nederend, R. D. Hays, P. F. Short, and D. O. Farley. 1999. "Special Issues in Assessing Care of Medicaid Recipients." *Medical Care* 37 (3): 79–88.

Crofton, C., J. S. Luliban, and C. Darby. 1999. "Foreword." *Medical Care* 37: 1–9.

Cleary, P. D., and B. J. McNeil. 1988. "Patient Satisfaction as an Indicator of Quality of Care." *Inquiry* 25: 25–36.

Daley, J., and M. Shwartz. 1994. "Developing Risk-adjustment Methods." In *Risk Adjustment for Measuring Health Care Outcomes,* edited by L. I. Lezzoni, pp. 199–238. Chicago: Health Administration Press.

Fiscella, K., and P. Franks. 1999. "Influence of Patient Education on Profiles of Physician Practices." *Annals of Internal Medicine* 131 (10): 745–51.

Fox, J. G., and D. M. Storms. 1981. "A Different Approach to Sociodemographic Predictors of Satisfaction with Health Care." *Social Science Medicine* 15 (A): 557–64.

Hall, J. A., M. Feldstein, M. D. Fretwell, J. W. Rowe, and A. M. Epstein. 1990. "Older Patients' Health Status and Satisfaction with Medical Care in an HMO Population." *Medical Care* 28 (3): 261–69.

Hays, R. D., G. N. Marshall, E. Y. Wang, and C. D. Sherbourne. 1994. "Four-year Cross-lagged Associations Between Physical and Mental Health in the Medical Outcomes Study." *Journal of Consulting and Clinical Psychology* 64: 441–49.

Kane, R. L, M. Maciejewski, and M. Finch. 1997. "The Relationship of Patient Satisfaction with Care and Clinical Outcomes." *Medical Care* 35 (7): 714–30.

Linn, L. S., and S. Greenfield. 1982. "Patient Suffering and Patient Satisfaction Among the Chronically Ill." *Medical Care* 20: 425–31.

McGee, J., D. E. Kanouse, S. Sofaer, J. L. Hargraves, E. Hoy, and S. Kleimann. 1999. "Making Survey Results Easy to Report to Consumers: How Reporting Needs Guided Survey Design in CAHPS." *Medical Care* 37 (3): 32–40.

National Committee for Quality Assurance (NCQA). 1998. *Accreditation '99: Standards for the Accreditation of Managed Care Organizations,* pp. 11–17. Washington, DC: NCQA.

Schnaier, J. A., S. F. Sweeny, V. S. L. Williams, B. Kosiak, J. S. Lubalin, R. D. Hays, and L. D. Harris-Kojetin. 1999. "Special Issues Addressed in the CAHPS Survey of Medicare Managed Care Beneficiaries." *Medical Care* 37 (3): 69–78.

Ware, J. E., A. Davies-Avery, and A. L. Stewart. 1982. "The Measurement and Meaning of Patient Satisfaction." *Health and Medical Care Services Review* 1 (1): 2–28.

Weiss, G. L. 1988. "Patient Satisfaction with Primary Medical Care: Evaluation of Sociodemographic and Predispositional Factors." *Medical Care* 26 (4): 383–92.