DOI: 10.1111/j.1475-6773.2005.00470.x

Case-Mix Adjustment of the CAHPS® Hospital Survey

A. James O'Malley, Alan M. Zaslavsky, Marc N. Elliott, Lawrence Zaborski, and Paul D. Cleary

Objectives: To develop a model for case-mix adjustment of Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Hospital survey responses, and to assess the impact of adjustment on comparisons of hospital quality.
Data Sources: Survey of 19,720 patients discharged from 132 hospitals.
Methods: We analyzed CAHPS Hospital survey data to assess the extent to which patient characteristics predict patient ratings ("predictive power") and the heterogeneity of the characteristics across hospitals. We combined the measures to estimate the impact of each predictor ("impact factor") and selected high impact variables for adjusting ratings from the CAHPS Hospital survey.
Principle Findings: The most important case-mix variables are: hospital service (surgery, obstetric, medical), age, race (non-Hispanic black), education, general health status (GHS), speaking Spanish at home, having a circulatory disorder, and interactions of each of these variables with service. Adjustment for GHS and education affected scores in each of the three services, while age and being non-Hispanic black had important impacts for those receiving surgery or medical services. Circulatory disorder, Spanish language, and Hispanic affected scores for those treated on surgery, obstetrics, and medical services, respectively. Of the 20 medical conditions we tested, only circulatory problems had an important impact within any of the services. Results were consistent for the overall ratings of nurse, doctor, and hospital. Although the overall impact of case-mix adjustment is modest, the rankings of some hospitals may be substantially affected.
Conclusions: Case-mix adjustment has a small impact on hospital ratings, but can lead to important reductions in the bias in comparisons between hospitals.

Key Words: Case-mix, hospital care, CAHPS hospital survey, patient surveys

The Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Hospital project is an extension of the CAHPS project, in which the Agency for Healthcare Research and Quality (AHRQ) funded a consortium of investigators to develop patient surveys to assess consumer experiences of health

care (Homer et al. 1999; Hargraves, Hays, and Cleary 2003; Daniels et al. 2004; Landon et al. 2004). The CAHPS Hospital project developed surveys to assess the experiences of patients recently discharged from acute care hospitals.

Results of the CAHPS Hospital surveys will be used to compare quality among hospitals, to support decision making by patients, physicians, and payers, and to facilitate quality improvement in hospitals. When making such comparisons, there are at least two reasons why it might be desirable to adjust CAHPS Hospital scores (Zaslavsky et al. 2001). First, some processes of care are likely to vary with patient characteristics. For example, it might be more difficult to communicate clearly with less educated patients or patients who take more medication (Hargraves et al. 2001; Zaslavsky et al. 2001). Varying distributions of these characteristics across hospitals might affect the rate of problems with care. Second, patients' characteristics can influence how they respond to survey questions. For example, a younger patient might be more sensitive to waiting time and thus give lower scores than an older patient with fewer time constraints.

Without adjustment for case-mix, reports and ratings of hospital care may be misleading. Furthermore, hospitals would have an incentive to attract patients likely to give higher ratings and avoid those most likely to report problems. Case-mix adjustment uses statistical models to predict what each hospital's ratings would have been for a standard patient or population, thereby removing from comparisons the predictable effects of differences in patient characteristics that are consistent across hospitals.

Age and self-rated general health status (GHS) typically have the strongest and most consistent associations with patient-reported problems, with greater satisfaction among older patients and those with better self-perceived health (Cleary and McNeil 1988; Cleary et al. 1989; Ware and Berwick 1990; Ehnfors and Smedby 1993; Charles et al. 1994; Arnetz and Arnetz 1996; Rosenheck, Wilson, and Meterko 1997; Woodbury, Tracy, and McKnight 1998; Hoff et al. 1999; Hargraves et al. 2001; McNeill et al. 2001; Jenkinson, Coulter, and Bruster 2002; Thi et al. 2002; Wilson et al. 2002). Similar

---

Address correspondence to Paul D. Cleary, Ph.D., Professor, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899. A. James O'Malley, Ph.D., Assistant Professor of Statistics, Alan M. Zaslavsky, Ph.D., Professor of Statistics, and Lawrence Zaborski, M.S., are with the Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA. Marc N. Elliott, Ph.D., is with the RAND Corporation, 1776 Main Street, M5N, Santa Monica, CA.

predictors are important for evaluations of health plans (Zaslavsky 1998; Elliott et al. 2001; Zaslavsky et al. 2001).

There is some evidence that other characteristics, such as education, marital status, income, and sex are related to survey responses about health care (Ehnfors and Smedby 1993; Charles et al. 1994; Rosenheck, Wilson, and Meterko 1997; Hoff et al. 1999; Thi et al. 2002) but those results are not consistent (Cleary and McNeil 1988). Lengths of stay and readmission have also been associated with hospital ratings and reports of care; however, these are not appropriate case-mix adjustors because they could be affected by the quality of care. In this paper, we assess which patient characteristics should be used in a model for adjusting CAHPS Hospital scores when making hospital comparisons.

Data and Methods

Survey

The CAHPS Hospital pilot survey included 33 questions about patients' experiences with various aspects of care (e.g., "During this hospital stay, how often did you have to ask for pain medication?") and three questions that elicit overall ratings of the hospital, doctors, and nurses, as well as a question about whether the patient would recommend the hospital to others (Elliott et al. 2005). There also are 13 questions about patient characteristics.

Sampling

Patients were selected at each hospital using random sampling within service (medical, surgical, obstetric). Eligible patients were adult (aged over 18 years) medical, surgical, and obstetric patients who had an overnight stay and were discharged between December 2002 and January 2003. Patients were excluded from the study if they had a psychiatric diagnosis, were under age 18 at the time of their admission to hospital, were not discharged to home, or were missing data needed for identification and surveying. Sampling fractions were calculated to yield equal numbers of patients from each service, although this was not always possible (e.g., some hospitals did not provide obstetric services).

CAHPS Hospital survey questionnaires were mailed to all sampled patients. Telephone follow-up, or mailing of replacement questionnaires, began about 4 weeks after the mailing of the survey.

Data

After excluding patients who had an undetermined service or hospital affiliation, the sample comprised 19,720 patients discharged from 132 study hospitals (Goldstein et al. 2005). We removed from analysis a single hospital with only eight responses and a hospital that had no medical service responses. We confined our analysis to respondents who received medical (37 percent), surgical (40 percent), or obstetric services (23 percent) at one of the 130 remaining hospitals, leaving a final sample of 19,683 respondents. Twenty-eight of those hospitals had no respondents receiving obstetric services. The number of respondents per hospital ranged from 28 to 512. Only the most recent hospital stay was retained for patients that had multiple hospital stays.

Outcome Variables

To identify potential case-mix adjusters, we analyze the extent to which patient characteristics predict overall ratings of "nurse," "doctor," and "hospital." These outcomes were chosen because they are regarded as the patient's summary of the more topic-specific report items and because they are more subjective and therefore likely to be sensitive to reporting effects (Hargraves et al. 2001; Kim, Zaslavsky, and Cleary 2005). For each outcome the hospital is the unit of analysis. Analyses of the nurse and doctor items enable us to learn about the corresponding aspects of hospitals' performance, which are not necessarily captured by the hospital rating, and therefore may reveal important case-mix effects that would otherwise be missed.

Predictor Variables

The variables from the CAHPS Hospital pilot survey available as case-mix adjustors are: hospital service, self-reported general health status (GHS), self-reported mental health status (MHS), age, gender, education, whether Spanish is spoken at home, if a proxy helped complete the questionnaire, race, and the patient's Diagnostic-Related Group (DRG) code assigned by the hospital. The service variable has three categories indicating if the hospital stay was for surgery, obstetrics, or other medical services. Service was available both from patient self-report and from hospital records. We use DRG codes in these analyses because they were available for all patients and more accurate.

GHS and MHS had a 5-point response scale (excellent, very good, good, fair, poor). Age had eight categories, mostly 10-year intervals, from 18 to greater than 80 years of age. Education was a 6-category ordinal variable (eighth grade or less; some high school, but did not graduate; high school

graduate; some college or 2-year degree; 4-year college graduate; more than 4-year college degree).

Race/ethnicity was represented by separate indicator variables for white, black, Hispanic, Asian, and Native American or Hawaiian. Respondents could check multiple race categories, so we assigned them to a group using the following order of priority: Hispanic, black, Native American or Hawaiian, Asian, and white; thus, a respondent that checked both Hispanic and black was categorized as Hispanic. Spanish language indicates if Spanish is the language mainly spoken at home, endorsed by about half the self-reported Hispanics.

Proxy help and proxy answer indicate if the patient required help completing the questionnaire or had the questions answered for them, respectively. Finally, the DRGs assigned at admission to the hospital were grouped in 20 Major Diagnostic Categories (MDC), providing a profile of the patient's condition specific to the inpatient stay in question. We did not consider variables for case-mix adjustment that were characteristics of the hospital or determined by the hospital's actions (e.g., length of stay), as adjusting for such variables might obscure real differences in quality between hospitals.

Statistical Methods and Analysis

Our criterion for selection of case-mix adjustors is the "impact factor," which is the product of two measures: predictive power (the strength of the relationship between the candidate adjustor and the outcome variable at the individual level) and heterogeneity factor (the amount of variation among hospitals in the adjustor variable) (Zaslavsky 1998). Predictive power quantifies the improvement in model fit (R^2) attributable to a variable; unlike tests of statistical significance, it does not depend on sample size. The heterogeneity factor measures the extent to which the characteristic is unevenly distributed across hospitals and therefore potentially a source of bias in comparisons. A variable, such as gender, could be highly predictive of responses but have little impact on case-mix adjustment because its distribution is relatively homogeneous across hospitals. Conversely, a variable could have quite different distributions in different hospitals but be unrelated to the rating. By combining both predictive power and heterogeneity into a single measure, the impact factor is more informative than purely predictive measures such as R^2; it approximates the magnitude of the incremental adjustments due to adding a variable to the case-mix model.

To select a core set of predictor variables we screened potential adjustors using stepwise regression; this exploratory technique is appropriate because we seek to identify a nonredundant set of variables that predict ratings of hospitals, not to test hypothesis about predictors. To select a parsimonious model, the inclusion and exclusion *p*-value criteria were set at 0.005, and to check that no important variables were omitted we compared the model against alternatives generated by an all-subsets regression. A further validation of the model was performed by randomly partitioning the data set into halves, refitting the model on one half and predicting the ratings in the other half, and comparing the accuracy of the predictions to those when the model is fit to the full data set. Nine separate regression models were run, for the three overall ratings in each of the three services. Variables selected in any of the models formed a core set eligible for final selection.

We estimated the predictive power, heterogeneity factor, and impact factor across all services for each case-mix variable on each CAHPS Hospital score. Interactions between service and the other case-mix variables were also assessed. When interactions were included as case-mix predictors, the corresponding main effects were also included. We also performed separate analyses for each service.

We measured predictive power by the incremental amount of variance explained by the predictor (represented as the partial $r^2 \times 1{,}000$) in a linear regression analysis given the other variables already in the baseline model, including dummy variables for each hospital. We measure the heterogeneity of the predictor variable across hospitals as the ratio of between-hospital to within-hospital variance of the residuals when the variable is regressed on the same baseline variables. The product of the predictive power and heterogeneity factor is proportional to the impact factor, used to assess which variables are both important predictors of CAHPS Hospital ratings and are sufficiently variable across hospitals to warrant case-mix adjustment (Zaslavsky et al. 2001), as described above. We required a minimum impact factor of 1 for a variable to be included.

For the impact analysis, we treated ordinal variables (such as age, health status, and education) as linear effects. This assumes that the effect on ratings of a change between consecutive categories is uniform across the scale (e.g., the difference between ratings from those in poor versus fair health is the same as that between those in good and very good health). This approximation is convenient for calculating the impact of ordinal variables, but might not be the optimal specification if the uniformity assumption is incorrect. For each ordinal variable in the baseline model, we tested the uniformity assumption by comparing (with an *F*-test) the baseline model with one that recoded the linear

variable as a set of dummy variables. Unless the categorical specification significantly improves on the linear specification, the latter can be used with no detectable loss of accuracy. After identifying the case-mix predictors we then tested interactions of adjustors with service, to determine which coefficients differed significantly across services, to arrive at our final model.

We used the CAHPS macro (AHCPR 1999) to compute mean nurse, doctor, and hospital ratings for each hospital adjusted for the various sets of predictors in these final models. These are predicted mean ratings for each hospital if they all had the same case-mix. By examining the changes in the predicted values for each hospital across models, we can determine how much each model adjusts for the relevant inter-hospital differences in case-mix.

To evaluate the overall impact of case-mix adjustment on each CAHPS Hospital score we compared the unadjusted scores to scores adjusted for variables selected for two of the three ratings for each service. We used two measures of the importance of adjustments to any rating variable: the ratio of the standard deviation of adjustments to the unadjusted standard deviations of the hospital means, and Kendall's τ correlation between the adjusted and unadjusted hospital rankings of the scores. Larger standard deviation ratios reflect greater impact. Kendall's τ is directly related to the proportion of pairs of hospitals that switched ordering as a consequence of case-mix adjustment.

The standard case-mix adjustment model relies on the assumption that the adjustors do not interact with hospital. If this assumption does not hold, the choice of covariate values will affect comparisons between hospitals (e.g., the ranking of hospitals), and therefore different reports may be needed for different types of patients. We use *F*-tests to evaluate if there is significant heterogeneity in the case-mix coefficients across hospitals.

RESULTS

Variable Screening: Stepwise Regression Modeling

For analyses of the nurse, doctor, and hospital ratings there were 16,745, 16,744, and 16,840 observations respectively with complete data for the dependent and all independent variables; 368 additional observations with missing values only for proxy help or gender were added to analyses of "doctor" when these variables were eliminated from models. Many of the missing values arose because the final two facing pages (containing 10 items) were left blank by 1,053 respondents. Because missingness of this block was not associated with the other ratings or concentrated in particular hospitals, we

treat these data as missing completely at random and removed the corresponding cases from the analysis.

The distribution of ratings was concentrated at the high end of the scale, with 38 percent, 48 percent, and 36 percent of patients providing ratings of 10 for the nurse, doctor, and hospital items respectively; 60 percent of ratings on each item were 9 or higher. Consequently, the distribution of hospital mean ratings is concentrated toward the high end of the scale (65–99 percent of hospital means exceeded 8 across the nine rating-service combinations).

The main effect of service in the pooled model indicates that compared to surgical patients, medical patients gave lower overall doctor ratings but relatively similar nurse and hospital ratings, whereas obstetric patients gave more positive ratings for nurse and hospital but similar ratings for doctor. Service, GHS, MHS, age, education, being non-Hispanic black, and Spanish language are highly predictive for each of the nurse, doctor, and hospital ratings (Table 1). Male and proxy help also met the $p < .005$ threshold for the nurse and hospital ratings, but not for doctor. Hispanic, Asian, Native American, and proxy answer were not predictive of any rating.

In models fit separately by service (data not shown), medical patients gave a lower doctor rating but relatively similar nurse and hospital ratings to surgical patients, whereas obstetric patients gave more positive nurse and hospital ratings but similar doctor rating. Healthy (general and mental), older, less educated, non-Hispanic black, and Spanish speaking patients tended to give higher ratings. Males gave significantly more positive nurse and hospital ratings than females, but not doctor rating. Proxy help respondents gave lower ratings for nurse and hospital than patient respondents; this effect was attenuated for doctor rating.

Analysis of variance calculations for the ordinal GHS, MHS, age, and education variables indicate that with the exception of MHS for the hospital rating, the linear specification accounts for at least 88 percent of the variation explained by the categorical specification, so these variables may be adequately represented on a linear scale. The categorical specification is significantly ($p < .005$) better than the linear version only for age and education for some ratings, reflecting small departures from linearity.

Interactions with Service

The most pronounced interactions between the case-mix variables and service are between age and the doctor rating, non-Hispanic black and the hospital rating, and education and the nurse rating (Table 2). The regression coeffi-

Table 1: Stepwise Analysis (All Services)

	Nurse Rating		*Doctor Rating*		*Hospital Rating*	
Predictor Levels	*Parameter Estimate*	*Overall p-Value*	*Parameter Estimate*	*Overall p-Value*	*Parameter Estimate*	*Overall p-Value*
In						
Intercept	8.20	<.0001	8.74	<.0001	8.28	<.0001
Service						
Medical	0.04	<.0001	−0.42***	<.0001	−0.09	<.0001
OB	0.42***		0.03		0.41***	
Health status						
Excellent	0.24***	<.0001	0.28***	<.0001	0.32***	<.0001
Very good						
Good	−0.54***		−0.53***		−0.65***	
Fair	−0.31***		−0.19***		−0.34***	
Poor	−0.96***		−0.86***		−1.14***	
Mental health						
Excellent	0.17***	<.0001	0.24***	<.0001	0.19***	<.0001
Very good						
Good	−0.19**		−0.24***		−0.05	
Fair	−0.04		−0.09		0.01	
Poor	−0.48**		−0.74***		−0.35*	
Age						
18–24	−0.24**	<.0001	−0.30***	<.0001	−0.22*	<.0001
25–34						
35–44	0.10		0.14*		0.05	
45–54	0.24**		0.27***		0.25**	
55–64	0.53***		0.52***		0.56***	
65–74	0.70***		0.67***		0.76***	
75–79	0.79***		0.69***		0.85***	
80+	0.83***		0.76***		0.99***	
Education						
Eighth grade	0.35***	<.0001	0.30***	<.0001	0.42***	<.0001
Some high school	0.31***		0.26***		0.43***	
High school	0.22***		0.25***		0.29***	
Some college						
College	−0.15*		−0.03		−0.13	
College+	−0.20***		−0.23***		−0.22***	
Race/ethnicity						
Non-Hispanic black	0.41	<.0001	0.29	<.0001	0.43	<.0001
Spanish language	0.71	<.0001	0.61	<.0001	0.88	<.0001
Sex: male	0.24	<.0001			0.19	<.0001
Proxy help	−0.34	<.0001			−0.26	<.0001
Out						
Sex: male				.0348		
Race/ethnicity						
Hispanic		.1994		.0069		.0785
Asian		.2322		.0763		.0972

continued

Table 1: *Continued*

Predictor Levels	*Nurse Rating*		*Doctor Rating*		*Hospital Rating*	
	Parameter Estimate	*Overall p-Value*	*Parameter Estimate*	*Overall p-Value*	*Parameter Estimate*	*Overall p-Value*
Native American		.3018		.0708		.9721
Proxy help				.0124		
Proxy answer		.7628		.7459		.8087

*$p<.01$,
**$p<.001$,
***$p<.0001$. The baseline category of categorical predictors is the level with the blank cell.
OB, obstetrics.

cients indicate that the ratings for obstetric and medical patients increase more with age than for surgical patients. No other interaction effects are significant at $p<.005$. However, the interactions of service with age for the nurse rating, with MHS and non-Hispanic black for the doctor rating are significant at the 0.05 level, suggesting that additional interactions with service may exist.

Because of the interaction of some case-mix variables with service and because some of the case-mix variables do not apply to certain services (e.g., only females receive obstetric services and some of the DRG-based groups are only relevant to particular services), subsequent analyses are stratified by service. We also assume linear specifications of the ordinal case-mix variables (including age and education).

Diagnoses

Of the 20 MDC examined, only five applied to more than 5 percent of the sample: circulatory disorder, digestive disorder, muscle disorder, female reproductive disorder, and respiratory disorder. The prevalence of the other conditions was so low that their impact on case-mix adjustment would be minimal, even if they were predictive of the ratings. Therefore, we tested only the above five conditions (Table 3).

For surgery patients, having a circulatory disorder was an important predictor of higher ratings for nurse and hospital, while having a female reproductive disorder was a significant positive predictor of the rating for doctor. Due to very low prevalence, none of the medical conditions were predictive of the ratings for obstetric patients. For patients attending the hospital for general medical services, muscle disorder was negatively associated with all three

Table 2: Tests for Interaction of Case-Mix Variables with Service

		Nurse Rating		*Doctor Rating*		*Hospital Rating*	
Predictor	*Level of Service*	*Estimate*	*Overall p-Value*	*Estimate*	*Overall p-Value*	*Estimate*	*Overall p-Value*
Service main effect	Medical	0.259	.3989	−0.66*	.0015	0.08	.8922
	OB	−0.06		−0.65*		0.01	
Age	Medical	−0.06*	.0081	0.08***	**.0004**	0.00	.1409
	OB	−0.02		0.05		−0.09	
GHS	Medical	−0.04	.5152	−0.07	.1506	−0.08	.1301
	OB	0.01		−0.07		−0.01	
MHS	Medical	0.00	.8607	0.02	.0218	−0.01	.2475
	OB	0.03		0.14*		0.08	
Education	Medical	0.08*	**.0035**	−0.03	.4416	0.03	.1697
	OB	0.09*		0.01		0.06	
Male	Medical	−0.10	.1432	0.08	.4737	−0.06	.5185
	OB	−0.84		−0.07		−0.43	
Non-Hispanic black	Medical	−0.13	.2218	0.28	.0353	0.09	**.0033**
	OB	−0.24		0.05		−0.35*	
Spanish language	Medical	−0.12	.0510	0.14	.1221	0.05	.1604
	OB	0.30		0.34		0.30	
Proxy help	Medical	0.03	.8431	−0.10	.0762	−0.03	.9349
	OB	0.12		−0.46		−0.06	

The ordinal predictors were represented as linear scale variables when testing for the above interaction effects. Surgery was the baseline category for the main and the interaction effects. Boldface values indicate interactions that have $p < .005$, while *$p < 0.1$, **$p < .001$, and ***$p < .0001$ denote the p-values of individual contrasts.

OB, obstetrics; GHS, general health status; MHS, mental health status.

ratings, while circulatory and respiratory disorders had modest (positive) associations (p-values between .01 and .1).

The directions and significance levels of effects of GHS, MHS, age, education, non-Hispanic black, and gender (not shown) are similar across services, and largely in agreement with the results in Table 1. Hispanic was a strong predictor of ratings for general medical services whereas Spanish language was a strong predictor of ratings for obstetrics. Hispanic and Spanish language were never both in a model, since they largely explained the same ethnic variation.

Predictive Power, Heterogeneity Factors, and Impact Factors

Table 4 presents the predictive power, heterogeneity, and impact for the predictor variables that met the 0.005 threshold in the model for at least two of the specific service models. Hispanic was excluded as it did not meet this criterion.

Table 3: Stepwise Analysis (by Service) for DRG Conditions

		Nurse Rating		*Doctor Rating*		*Hospital Rating*	
Service	*Predictor*	*Parameter*	*p-Value*	*Parameter*	*p-Value*	*Parameter*	*p-Value*
Surgery	In						
	Circulatory disorder	0.41	<.0001			0.38	<.0001
	Female reproduction disorder			0.18	.0048		
	Out						
	Circulatory disorder				.0119		
	Digestive disorder		.1387		.6771		.3621
	Muscle disorder		.0813		.5711		.1683
	Female reproduction disorder		.7703				.0312
	Respiratory disorder		.2859		.1355		.8526
Medical	In						
	Muscle disorder	−0.45	.0023	−0.78	<.0001	−0.46	.0023
	Out						
	Circulatory disorder		.0259		.0842		.0599
	Respiratory disorder		.0939		.0201		.0109
	Digestive disorder		.3497		.8567		.1035
	Female reproduction disorder		.4182		.6625		.6604
OB	Out						
	All disorders are female reproduction						

DRG, Diagnostic-Related Group; OB, obstetrics.

The variables that have the greatest overall impact on one or more ratings are: age, non-Hispanic black, education, Spanish language, service, MHS, and GHS. Male and proxy help had relatively small impact. The results were consistent across ratings, although the impact factors for the doctor rating were typically larger than those for the other ratings.

The variables with the greatest predictive power do not necessarily have the greatest impact on the adjustment. For instance, GHS is highly predictive of each rating, but due to a homogeneous distribution across hospitals, has low impact. Conversely, the most heterogeneous variable, non-Hispanic black, has modest predictive power and is the second or third most important in terms of impact.

GHS and education were the only variables that had substantial impact for each service (Table 5). Although age had very high impact for both surgery and medical, it was not sufficiently predictive to even be considered as a

Table 4: Predictive Power, Heterogeneity, and Impact over All Services

				Nurse			*Doctor*			*Hospital*		
Predictor	*Mean*	*Coefficient of Variation*	*Predictor Heterogeneity Factor*	*Parameter Estimate*	*Predicted Power*	*Impact Factor*	*Parameter Estimate*	*Predicted Power*	*Impact Factor*	*Parameter Estimate*	*Predicted Power*	*Impact Factor*
Outcome heterogeneity factor					0.042			0.014			0.055	
Service												
Medical	0.37	0.27	0.054	0.04	0.09	0.11	−0.41	8.81	**33.39**	−0.08	0.35	0.35
OB	0.23	0.32	0.062	0.40	3.74	**5.53**	−0.04	0.05	0.22	0.42	4.34	**4.95**
GHS	2.53	0.02	0.003	−0.28	12.59	0.76	−0.27	13.14	**2.34**	−0.34	19.70	0.92
MHS*	2.17	0.04	0.008	−0.12	3.84	0.72	−0.17	8.17	**4.52**	−0.09	2.06	0.30
Age	4.54	0.07	0.048	0.16	15.55	**17.73**	0.15	14.57	**49.05**	0.19	21.44	**18.86**
Education	3.76	0.08	0.064	−0.14	7.26	**11.00**	−0.13	6.94	**31.07**	−0.16	10.67	**12.48**
Non-Hispanic black	0.10	1.63	0.375	0.35	2.15	**19.10**	0.22	0.92	**24.18**	0.38	2.65	**18.19**
Spanish language	0.05	1.36	0.131	0.66	4.15	**12.95**	0.56	3.32	**30.56**	0.83	6.90	**16.60**
Male	0.33	0.14	0.011	0.24	2.88	0.76	−0.07	0.29	0.22	0.20	1.93	0.39
Proxy help	0.10	0.48	0.029	-0.39	2.92	**1.99**	−0.18	0.66	**1.34**	−0.30	1.77	0.93

Boldface indicates predictors with impact factors that exceed the minimum threshold of 1.

*Due to collinearity between GHS and MHS, GHS was dropped from the model when the effect of MHS was evaluated and vice versa.

OB, obstetrics; GHS, general health status; MHS, mental health status.

Table 5: Predictive Power, Heterogeneity, and Impact by Service

				Nurse			*Doctor*			*Hospital*		
Predictor	*Mean*	*Coefficient of Variation*	*Predictor Heterogeneity Factor*	*Parameter Estimate*	*Predicted Power*	*Impact Factor*	*Parameter Estimate*	*Predicted Power*	*Impact Factor*	*Parameter Estimate*	*Predicted Power*	*Impact Factor*
Surgical												
Outcome heterogeneity factor					0.037			0.011			0.042	
GHS	2.55	0.04	0.010	−0.38	37.01	**9.89**	−0.40	56.26	**51.76**	−0.44	52.39	**12.43**
Age	5.08	0.07	0.055	0.17	17.44	**26.19**	0.10	8.12	**41.96**	0.17	18.85	**25.14**
Education	3.78	0.09	0.077	−0.19	13.11	**27.16**	−0.11	6.78	**48.41**	−0.19	14.75	**27.14**
Non-Hispanic black	0.08	1.76	0.366	0.36	1.94	**19.22**	0.01	0.00	0.10	0.33	1.77	**15.56**
Male	0.43	0.11	0.011	0.26	3.83	**1.10**	−0.12	1.09	**1.08**	0.20	2.48	0.63
Circulation disorder	0.20	0.79	0.209	0.41	4.91	**27.75**	0.12	0.60	**11.72**	0.40	5.24	**26.30**
OB												
Outcome heterogeneity factor					0.06			0.04			0.08	
GHS	1.74	0.03	0.008	−0.30	11.79	**1.75**	−0.24	8.16	**1.67**	−0.37	20.01	**2.18**
MHS*	1.71	0.04	0.006	−0.15	5.08	0.54	−0.27	18.04	**2.63**	−0.15	5.81	0.45
Education	4.10	0.08	0.087	−0.08	2.42	**3.82**	−0.09	3.65	**7.97**	−0.12	6.33	**7.35**
Spanish language	0.12	1.06	0.259	0.85	14.75	**69.34**	0.64	9.03	**58.63**	1.04	24.29	**83.78**
Medical												
Outcome heterogeneity factor					0.04			0.02			0.06	
GHS	3.04	0.03	0.009	−0.31	26.66	**5.81**	−0.32	25.62	**15.40**	−0.33	28.78	**4.62**
Age	5.60	0.07	0.049	0.13	11.71	**14.09**	0.19	24.69	**81.99**	0.19	25.84	**22.90**
Education	3.51	0.10	0.074	−0.10	4.51	**8.09**	−0.13	6.09	**30.11**	−0.15	8.97	**11.84**
Non-Hispanic black	0.12	1.45	0.386	0.24	1.05	**9.90**	0.31	1.74	**45.11**	0.44	3.49	**24.15**
Hispanic	0.06	1.42	0.165	0.15	0.30	**1.19**	0.38	1.73	**19.17**	0.36	1.59	**4.72**
Proxy help	0.15	0.41	0.035	−0.38	4.01	**3.44**	−0.25	1.64	**3.88**	−0.33	2.97	**1.88**
Muscle disorder	0.03	0.32	0.004	−0.44	1.56	0.14	−0.82	4.96	**1.19**	−0.47	1.70	0.11

Boldface indicates predictors with impact factors that exceed the minimum threshold of 1.

*Due to collinearity between GHS and MHS, GHS was dropped from the model when the effect of MHS was evaluated and vice versa.

OB, obstetrics; GHS, general health status; MHS, mental health status.

case-mix adjustor for obstetrics. Non-Hispanic black had significant impact in surgery and medical, but not in obstetrics, whereas Spanish language and MHS had major impacts only on obstetrics. Circulatory disorder (surgery only) was the only MDC to substantially impact case-mix adjustment; although muscle disorder was highly predictive the homogeneity of its distribution across hospitals meant it had modest impact. Male (only for surgery) and proxy help (only for medical) had modest impact.

Overall Impact

We quantify the overall impact of case-mix adjustment on hospital-level ratings by the ratio of the standard deviation of the adjustment to the standard deviation of the means and by Kendall's τ. The standard deviation ratios ranged from 0.17 to 0.28, indicating that the adjustments were modest but not negligible compared to the differences among hospitals. Furthermore, ratios of the maximum adjustments to the standard deviations of the means ranged from 0.47 to 1.09, suggesting that although the effect of the adjustment was small for most hospitals, it was important for some. Kendall's τ is between 0.82 and 0.88, meaning that the percentage of hospitals whose ordering would be changed by case-mix adjustment is between 9 percent and 5.9 percent for all services and ratings; surgery and obstetrics were most and least affected respectively.

Although the impact of the case-mix adjustment on between-hospital comparisons is of most interest, the amount of within-hospital variation explained by the case-mix model is a useful summary of model fit. The within-hospital R^2 was between 5.8 percent and 7.5 percent for the overall model and between 4.4 percent and 8.3 percent for the service-specific models; slightly more variation was explained for surgery than for obstetrics or medical.

Interactions with Hospital

The slopes for service and GHS in the overall adjustment model varied significantly ($p < .005$), as did the slope for GHS in the service-specific models. The slopes of several other case-mix variables varied significantly across hospitals in the overall model but not in the service-specific models.

DISCUSSION

Case-mix adjustment is a widely used method for making comparisons among health care providers fairer. Careful adjustment may assuage hospitals'

concerns that they may be disadvantaged in comparative ratings by an unfavorable patient population, contributing to acceptance of quality measures and making them more effective drivers of quality improvement.

In this study of patients discharged from hospitals in three states, service (surgery, obstetrics, medical) had a strong relationship to the ratings, and the proportion of patients in each service varies across hospitals, substantially affecting comparisons of hospitals. Service interacts with several other case-mix variables. Notably, age has a large impact on ratings for surgery and medical patients but little impact on obstetric patients, presumably because the ages of obstetric patients, e.g., those having child birth, are much more homogeneous.

To accommodate multiple interactions with service, we recommend fitting separate case-mix models for surgery, obstetric, and medical services. If a single report is to be made combining all services, the case-mix model should include interactions of each variable with service.

Besides service, age, non-Hispanic black, education, GHS, Spanish language, and circulation disorder (in surgery only) appear to be the other most important case-mix adjustors. Adjustment for GHS and education affected scores in each of the three services, while age and non-Hispanic black had important impacts for surgery and medical. Circulation disorder, Spanish language, and Hispanic affected scores for surgery, obstetrics, and medical respectively. The signs of the associations between the case-mix variables and the quality ratings cohere with previously reported results.

The limited impact of diagnostic categories is probably due to the small proportion of patients with each condition and thus the low variation between hospitals. For example, if the prevalence of a condition in a population is only 1 percent, then even if the proportion of patients with the condition varies substantially between hospitals the impact on case-mix adjustment will be minimal. The one exception is circulatory disorders, for which there is a relatively large number of cases. We do not know what characteristics of these patients or their experiences cause them to report more favorably. Because of the additional difficulty of collecting and coding diagnostic data, and because circulatory disorders only impact patients having surgery, we suggest more research in larger, more representative samples before recommending this variable or other diagnostic groups for case-mix adjustment.

The case-mix models explained a modest percentage of within-hospital variation, consistent with previous results which found that a similar set of variables only explained between 3 percent and 8 percent of the variation

in ratings about hospital care (Hargraves et al. 2001). However, the overall impact of the case-mix adjustments reported here, measured by the ratio of the standard deviation of the adjustment to the unadjusted ratings and Kendall's τ, exceeded the impact of a similar set of predictors on ratings of health plans (Zaslavsky et al. 2001).

Our case-mix models assume that the case-mix coefficients do not vary across hospitals. If they do, comparative inferences such as the ranking of hospitals could depend on the "standard" patient or population used to standardize the CAHPS Hospital scores; one hospital might perform better than another for some types of patients but worse for others. We found some evidence of hospital by case-mix interactions but when we fitted separate models to each service the only significant interaction was with age. However, the variation in the slope of age is relatively minor compared to the main effect of age. A previous study of case-mix adjustment for health plans suggested that variation in case-mix coefficients had little importance for adjustment of plan means but might indicate that comparisons of plans could be substantially different depending on the characteristics of the individual patient (Zaslavsky, Zaborski, and Cleary 2000). The heterogeneity of the age effect across hospitals should be evaluated again when larger datasets involving more hospitals are available.

The CAHPS Hospital pilot survey data represented only three states, so the effects of Spanish language or race may be different in other geographic areas. For instance, the relationship between Spanish language and reported experiences might be affected by the local concentration of Spanish speakers or by the specific Hispanic ethnicity (e.g., Mexican American, Cuban American) in the area. We tested whether the case-mix variables interacted with the region where a patient lives but did not find significant results. The consequences of such interactions would depend on whether the data were to be used primarily for local or national comparisons. Finally, we did not test if survey mode (phone versus mail) is needed for case-mix adjustment. Although there are differences by mode (Elliott et al. 2005), they are confounded with initial nonresponse, and patient experiences might be related to the mode they respond to, rather than whether or not they initially responded.

CAHPS Hospital scores will be used by patients and their providers to select hospitals, by hospitals to focus and monitor quality improvement efforts, and by policy makers to monitor and promote high-quality care. Because case-mix adjustment has the potential to prevent patient characteristics from confounding comparisons between hospitals, using adjustment models, such as the one specified here, is of crucial importance.

ACKNOWLEDGMENTS

The CAHPS II project is funded by the Agency for Health Care Quality and Research (AHQR) and the Centers for Medicare and Medicaid through cooperative agreements with Harvard Medical School, RAND, and the Research Triangle Institute. User support is provided through a contract with Westat. Additional information about the study can be obtained by calling the AHQR Clearinghouse at 800-358-9295. The authors thank project officers Chris Crofton, Chuck Darby, Beth Kosiak, and MaryBeth Farquahr for their active participation and helpful suggestions throughout the project and members of the CAHPS consortium for their role in the design and implementation of the data collection activities and helpful comments on an earlier draft of this manuscript.

REFERENCES

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

Arnetz, J. E., and B. B. Arnetz. 1996. "The Development and Application of a Patient Satisfaction Measurement System for Hospital-Wide Quality Improvement." *International Journal for Quality in Health Care* 8: 555–66.

Charles, C., M. Gauld, L. Chambers, B. O'Brien, R. Haynes, and R. Labelle. 1994. "How Was Your Hospital Stay? Patients' Reports about Their Care in Canadian Hospitals." *Canadian Medical Association Journal* 150: 1813–22.

Cleary, P. D., L. Leroy, G. Karapanos, and W. McMullen. 1989. "Patient Assessments of Hospital Care." *Quality Review Bulletin* 15: 172–9.

Cleary, P. D., and B. J. McNeil. 1988. "Patient Satisfaction as an Indicator of Quality of Care. Inquiry." *Inquiry* 25 (1): 25–36.

Daniels, A. S., J. A. Shaul, P. Greenberg, and P. D. Cleary. 2004. "The Experience of Care and Health Outcomes Survey (ECHO): A Consumer Survey to Collect Ratings of Behavioral Health Care Treatment, Outcomes and Plans." In *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment. Volume 3: Instruments for Adults,* edited by M. E. Maruish. Fairfax, VA: Lawrence Erlbaum Associates.

Ehnfors, M., and B. Smedby. 1993. "Patient Satisfaction Surveys Subsequent to Hospital Care: Problems of Sampling, Non-Response and Other Losses." *Quality Assurance in Health Care* 5: 19–32.

Elliott, M. N., C. Edwards, J. Angeles, and R. Hays. 2005. "Predictors of Item and Unit Nonresponse." *Health Services Research.* DOI: 10.1111/j.1475-6773.2005.00476.x. Available at www.blackwell-synergy.com.

Elliott, M. N., R. Swartz, J. Adams, K. L. Spritzer, and R. Hays. 2001. "Case-Mix Adjustment of the National CAHPS® Benchmarking Data 1.0: A Violation of Model Assumptions?" *Health Services Research* 36 (3): 555–74.

Goldstein, L., C. Crofton, S. Garfinkel, and C. Darby. 2005. "Why Another Patient Survey Matters." *Health Services Research* 40. DOI: 10.1111/j.1475-6773.2005.00477.x. Available at www.blackwell-synergy.com.

Hargraves, J. L., R. D. Hays, and P. D. Cleary. 2003. "Psychometric Properties of the Consumer Assessment of Health Plans (CAHPS®) 2.0 Adult Core Survey." *Health Services Research* 38 (6): 1509–27.

Hargraves, J. L., I. B. Wilson, A. Zaslavsky, C. James, J. D. Walker, and P. D. Cleary. 2001. "Adjusting for Patient Characteristics When Analyzing Reports from Patients about Hospital Care." *Medical Care* 39 (6): 635–41.

Hoff, R. A., R. A. Rosenheck, M. Meterko, and N. J. Wilson. 1999. "Mental Illness as a Predictor of Satisfaction with Inpatient Care at Veterans Affairs Hospitals." *Psychiatric Services* 50: 680–5.

Homer, C. J., F. J. J. Fowler, P. M. Gallagher, J. Shaul, M. Uyeda, A. Zaslavsky, V. Wilson, and P. D. Cleary. 1999. "The Consumer Assessment of Health Plans Study (CAHPS) Survey of Children's Health Care." *Joint Commission Journal on Quality Improvement* 25: 369–78.

Jenkinson, C., A. Coulter, and S. Bruster. 2002. "The Picker Patient Experience Questionnaire: Development and Validation Using Data from Inpatient Surveys in Five Countries." *International Journal of Quality in Healthcare* 14 (5): 353–8.

Kim, M., A. M. Zaslavsky, and P. D. Cleary. 2005. "Adjusting Pediatric CAHPS Scores to Ensure Fair Comparison of Health Plan Performances." *Medical Care* 43 (1): 44–52.

Landon, B. E., A. M. Zaslavsky, S. L. Bernard, M. J. Cioffi, and P. D. Cleary. 2004. "Comparison of Performance of Traditional Medicare vs Medicare Managed Care." *Journal of the American Medical Association* 291 (14): 1744–52.

McNeill, J. A., G. D. Sherwood, P. L. Stark, and B. Nieto. 2001. "Pain Management Ratings for Hospitalized Hispanic Patients." *Pain Management in Nursing* 2: 25–36.

Rosenheck, R., N. Wilson, and M. Meterko. 1997. "Influence of Patient and Hospital Factors on Consumer Satisfaction with Inpatient Mental Health Treatment." *Psychiatric Services* 12: 1553–61.

Thi, P. L., S. Briancon, F. Empereur, and F. Guillemin. 2002. "Factors Determining Inpatient Satisfaction with Care." *Social Science and Medicine* 54: 493–504.

Ware, J. E., and D. M. Berwick. 1990. "Patient Judgments of Hospital Quality." *Medical Care* 28 (9 Suppl): S39–44.

Wilson, I. B., L. Ding, R. D. Hays, M. F. Shapiro, S. A. Bozzette, and P. D. Cleary. 2002. "HIV Patients' Experiences with Inpatient and Outpatient Care: Results of a National Survey." *Medical Care* 40 (12): 1149–60.

Woodbury, D., D. Tracy, and E. McKnight. 1998. "Does Considering Severity of Illness Improve Interpretation of Patient Satisfaction Data?" *Journal for Healthcare Quality* 20: 33–40.

Zaslavsky, A. M. 1998. "Issues in Case-Mix Adjustment of Measures of the Quality of Health Plans." In *Proceedings of the Joint Statistical Meetings (Section on Government Statistics and Social Statistics), American Statistical Association*, pp. 56–65.

Zaslavsky, A. M., L. B. Zaborski, and P. D. Cleary. 2000. "Does the Effect of Respondent Characteristics on Consumer Assessments Vary across Plans?" *Medical Care Research and Review* 57: 379–94.

Zaslavsky, A. M., L. B. Zaborski, L. Ding, J. A. Shaul, M. J. Cioffi, and P. D. Cleary. 2001. "Adjusting Performance Measures to Ensure Equitable Plan Comparisons." *Health Care Financing Review* 22 (3): 109–26.

SUPPLEMENTARY MATERIAL

The following supplementary material for this article is available online:

APPENDIX S1. Procedure for Categorical Predictors.