

ISSUES IN CASE-MIX ADJUSTMENT OF MEASURES OF THE QUALITY OF HEALTH PLANS

Alan Zaslavsky, Harvard Medical School
Department of Health Care Policy, Harvard Medical School, Boston, MA 02115

Abstract

In a health care environment increasingly dominated by managed care, measurements of the quality of care provided by health plans are important tools for consumer and purchaser choice. Two important quality measurement tools are the CAHPS survey of health plan consumer satisfaction, and the HEDIS data set, which compiles rates at which essential preventive, screening and chronic care services are provided to health plan members. Because the health plans may select for patients with different characteristics, comparisons between plans should be adjusted to allow for the effects of case mix on plan ratings. A typical approach is a regression adjustment using a linear or loglinear model. Some issues of methodology and interpretation are considered, with illustrations from case-mix adjustment of CAHPS and HEDIS. We consider variable selection, the implications of nonparallelism of the case-mix regressions, the use of contextual variables for case mix, and presentation of the magnitude of the case-mix effects.

Keywords: consumer satisfaction; HEDIS; risk adjustment; CAHPS survey, health services; regression adjustment

1 An overview of quality measurement for health plans

With the burgeoning of managed care arrangements for health care, measurement and comparison of the quality of health plans has become increasingly important. A number of measures of quality have been developed and are becoming increasingly important in the managed care market.

These measures can be broadly grouped into satisfaction measures and clinical measures. Satisfaction measures, are based on members' general ratings of satisfaction with different aspects of care, and reports about experiences with care. The survey items typically deal with the patient's interactions with both the provider of care and the health plan itself. For example, members may be asked to rate their health plan, their doctor, or the kind of care they get overall on a 0–10 scale. They also may be asked about experiences with specific services (getting appointments quickly, dealing with problems about bills, getting needed prescriptions) in a specific reference period such as the preceding

six months: did they never, sometimes, usually, or always get the service in a satisfactory manner when it was needed? Or they may be asked about problems: was getting referrals a big problem, a small problem, or not a problem at all?

An important recent initiative in measurement of satisfaction with health plans is the Consumer Assessments of Health Plans (CAHPS) initiative, supported by the Agency for Health Care Policy and Research. This project has drawn together three teams of experts on satisfaction surveys, centered at Harvard Medical School, RAND, and the Research Triangle Institute, to develop a standard set of instruments for satisfaction surveys. The core CAHPS survey has been in the field since 1997. Version 2.0 of CAHPS, which incorporates changes made to meet the needs of the plan accreditation process sponsored by the National Center for Quality Assurance (NCQA), is due for release in late 1998. A number of variations of the survey have also been developed for special populations such as Medicare beneficiaries, Medicaid beneficiaries, patients with chronic conditions, children, and users of mental health services.

Clinical measures are concerned with whether particular populations of patients received specific medical services that are regarded as appropriate for those populations. Typically these services fall in the area of preventive care (immunizations, well-child visits), chronic care (medications for people who have had heart attacks, eye examinations for diabetics) and screening (mammograms, cervical cancer screening). The relevant populations must be well-defined, there must be a clear consensus on the appropriateness of the target treatment for almost everyone in the population, and it must be feasible to identify the population and ascertain whether the appropriate services were provided. Data collection for these measures relies on a combination of administrative records held by the plans and reviews of medical charts. Since neither of this information sources is primarily designed for quality assurance purposes, each measure requires considerable difficulty and expense for the plans.

One important set of quality measures for plans has been developed under the aegis of the National Center for Quality Assurance. HEDIS (Health Plan Employer Data and Information System) includes a set of measures of "Effectiveness of Care." These

are voluntarily reported by health plans to NCQA, which publishes results for participating plan in a compendium of plan data, the Quality Compass. Development of these measures has been a complex and challenging process due need to develop consensus across a broad range of stakeholders (medical experts, plans, health care providers, purchasers, and consumers), to specify the measures in a precise way, and to make sure that plan data systems are capable of producing the required information in a sufficiently consistent manner.

Both satisfaction measures and clinical quality measures are of great importance to the many players in the world of health care. Many large purchasers of health care (such as large employers, business groups, and the Medicare program) already require health plans to provide satisfaction and clinical quality measures; in many cases, the purchasers sponsor the satisfaction surveys and require the plans to cooperate. Consumers who are able to choose among health plans are being presented with an increasing array of comparative quality data, although the impact of these data on consumer choice has not been established. In future years (starting in 1999), these measures will be considered in accreditation decisions by NCQA. Another rationale for quality measurement is the belief that plans will use the data to guide their own quality improvement efforts, motivated both by the desire to improve their effectiveness and the competitive pressures created by the dissemination of ratings.

I should also note that quality measurement is widespread for units of the health care system other than health plans, such as hospitals, medical groups, and even individual physicians. Both satisfaction surveys and clinical measures (such as mortality and morbidity after surgery) are used in these quality assessment systems. Although there are some differences in the kinds of information that are available and the appropriate measures, many of the issues I consider here (both the purely statistical and policy issues) are relevant to these other levels as well.

2 Why do we care about case-mix?

A common concern about quality measures is that they may be affected by characteristics of the plan's members that are not under the control of the plan. Some kinds of members may generally tend to report poorer experiences and express lower satisfaction than others do, or to be less likely than others to obtain the appropriate services, even within the same plan. If there is substantial variation across plans in the distribution of these kinds of members, then some plans may receive poorer quality ratings

because of the kind of members they enroll (their "case-mix") rather than because of the quality of the care that they provide.

A wide variety of factors have been found to affect ratings of various kinds. Among these are

- sociodemographic characteristics (age, sex, race/ethnicity, income, education),
- self-reported health status (physical, mental),
- functional abilities and limitations,
- diagnoses and conditions,
- reporting circumstances (e.g. use of proxy informant), and
- geographical context, i.e. characteristics of the person's area of residence.

Each of these affects a different set of measures, and each presents different problems of measurement.

These case-mix effects can be important for several reasons.

1. Reports that are affected by case-mix effects defeat an important purpose of quality measurement, which is helping consumers and purchasers to distinguish the plans that can provide the best care and service to the individual consumer or a particular collection of patients (e.g. employees of a particular employer). The part of a plan's ratings that are due to their *current* case-mix are irrelevant to this assessment.
2. Plans are often keenly aware of any special difficulties that are posed by the member populations they serve; if anything, they will overestimate the negative effect of these particularities. The credibility of the ratings is impugned if these factors, whose impact might be assessed by the plans in a subjective manner, are not considered objectively by those who prepare ratings.
3. If plans know that enrollment patterns affect ratings that have a growing impact on their reputations and competitive viability, they have an incentive to avoid enrolling members who are likely to lower their ratings. This may make it reduce the range of options available to these potential members, especially when these incentives coincide with incentives based on differential cost and difficulty of treatment for some classes of patients (such as chronically ill patients or those who live in areas with few medical facilities).

These considerations have led to widespread use of case-mix adjustment. The basic idea of case-mix adjustment is to use (more or less complex) statistical models to predict what the plan's ratings would have been if they all had the same case-mix, i.e. plan ratings with a standard patient or population of patients at every plan. Cas-emix adjustment potentially can address the problems identified above, to the extent that it can

1. remove effects of individual characteristics that would affect ratings on quality measures at any plan;
2. address effects that plans may regard as spurious, to give more credibility to the adjusted estimates; and
3. remove incentives to plans to avoid enrollment of patients that might negatively affect their unadjusted measures ("hard to treat", "hard to satisfy").

Three conditions must be met for case-mix adjustment to be appropriate to these objectives: the case-mix variables must be related (holding plan constant, i.e. within-plan) to the measures, there must be variation between plans on the variables, and the variables must be appropriate for adjustment because they are not themselves determined by the plan's actions. The last, nonstatistical, criterion may be the most controversial but it is essential to the objective of separating effects of differential enrollment from those of the quality of care provided. The following are examples of variables that we excluded by these criteria:

- Length of relationship with the plan. Members who have been enrolled with a plan for a longer time tend to be more satisfied, but the length of the relationship is likely to be (at least in part) a consequence rather than a cause of satisfaction. It would not be useful to reduce a plan's satisfaction scores because it is successful in retaining members.
- Utilization of medical services. Heavy users are often less satisfied, because they have more opportunities to have problems that will alter the generally positive initial assessment that most patients have of their health care. Utilization, however, is in part determined by the plan's policies on access to services. Measures of need that are less directly affected by the plan are more appropriate adjuster variables.

- Plan characteristics, such as type of ownership or benefit package. These may be predictive of satisfaction or clinical quality, but they are, of course, part of the conditions created by the plan, not something that members bring to the plan when they enroll. Furthermore, they are constant across the plan so they cannot be associated with measures *within* plan.

Obviously, an element of judgement is necessarily involved in deciding whether some variables are appropriate adjusters.

Case-mix adjustment is in many ways similar to risk adjustment of costs. The motivations (particularly reduction of incentives to selective enrollment), conceptual bases, and statistical methodologies are very similar. The objective in case-mix adjustment, however, is more elusive because it is a counterfactual — quality as measured by an existing plan on a hypothetical population. There is no "gold standard" for adjustment of quality measures that is comparable to the standard of accuracy of predicted costs. Another, practically important, difference is that mean costs are often driven by the long right tail of the cost distribution, while the influence of any single member on quality measures is usually limited.

Modeling of case-mix effects also is similar in form to theoretical assessment of the individual (member-level) determinants of quality. In case-mix adjustment, however, the individual-level effects are "nuisance effects" to be removed from the analysis; the criterion for model selection is essentially predictive rather than explanatory. Furthermore, case-mix involves additional criteria, both quantitative (between-plan variation on the case-mix variables) and qualitative (those laid out above). Finally, for purposes of adjustment, it is irrelevant whether the observed associations are due to actual variations in quality of care or measurement error. For example, we have found (consistent with other research) that educational levels are positively associated with clinical quality measures and negatively with satisfaction scores. It seems likely that the latter effect is due to higher expectations held by more educated patients, rather than poorer care experienced by that group. The poorer clinical quality levels experienced by less-educated patients also could be due in part to some differences in record-keeping where they obtain their care, but it is likely that the health care system is failing to get some needed services to patients who are less demanding and may face linguistic, cultural and physical barriers to access to care.

3 Basic statistical strategies for case-mix adjustment

A common approach to case-mix adjustment uses a regression model to perform what is called in other contexts a covariance adjustment. We first fit a regression model with plan-specific coefficients, either (for numerical outcomes) a linear model,

$$y_{pj} = \mu_p + \beta'x_{pj} + \epsilon_{pj}, \quad (1)$$

or (for dichotomous outcomes) a logistic regression model,

$$\text{logit } P(y_{pj} = 1) = \mu_p + \beta'x_{pj}. \quad (2)$$

Here, p represents plan, j member, y_{pj} the corresponding measure value, β a coefficient vector that is fixed across plans, x_{pj} a vector of individual-level covariates, and ϵ_{pj} a random disturbance with expectation 0. (I note that in some applications, it is useful to embed this model in a hierarchical framework in which the μ_p are related by a common prior distribution, but this is less important when sample sizes by plan are large and nearly equal.)

Once we fit this model, we calculate the adjusted plan rating, when we use a linear model, as the prediction for an “average” person:

$$\mu_p + \beta'\bar{x}.$$

The definition of \bar{x} has no effect on the differences between plans, but we can define \bar{x} as $(1/I)\sum \bar{x}_i$, the mean of the plan means on the covariates, so that the mean of the adjusted plan quality means is equal to the mean of the adjusted plan means.

With a logistic model, we can similarly calculate an adjusted plan rating as the prediction for “average” person,

$$\text{logit}^{-1}(\mu_p + \beta'\bar{x}).$$

A possibly more interpretable alternative is the prediction for an “average” population:

$$(1/n)\sum \text{logit}^{-1}(\mu_p + \beta'x_{pj}),$$

where the sum is over the combined sample of size n . This treats the pooled sample as representative of a standard population of members. Due to the nonlinearity of the logit transformation, this differs slightly from the prediction for the average person.

4 Three applications

The methodological issues that are considered here arose in the context of three quality measurement applications, two implementations of the CAHPS member satisfaction survey and one involving the HEDIS measures. I now describe these three applications.

4.1 Application I: The CAHPS Medicare Managed Care survey

The CAHPS-MMC survey is a modified version of the CAHPS survey designed specifically for Medicare beneficiaries enrolled in managed care plans. The survey was administered to samples from all active Medicare managed care plans in early 1998. Results from this survey administration are due for release in late 1998 while the second year’s round of the survey is still in the field.

The preliminary analyses described here included 97,921 valid survey responses from a simple random sample of enrollees in 232 Medicare managed care plans. (The final data set differs, with some additional responses and some exclusions due to termination of some plans.) Most of the items in this survey use either a 0–10 or a 1–4 (never/sometimes/usually/always) rating scale. A number of sociodemographic and health status variables were also collected on the survey instrument. Area of residence could be approximated by zip code, although Medicare mailing addresses do not always correspond to residential location. Results from this survey will be reported back to health plans.

4.2 Application II: The CAHPS Washington State employee survey

This survey was sponsored by the Washington State Health Care Agency. This implementation, using the core CAHPS instrument, had a similar survey design to the Medicare survey, although an additional stratified sample was drawn from some plans in one county with an especially large number of employees. A total of 8319 valid responses were analyzed from 20 plans. The survey took place during the summer of 1997 and the results were reported to state employees as part of their benefits packet during their open enrollment period during the fall of that year.

4.3 Application III: HEDIS reporting set measures

The HEDIS clinical measures are described above. Because plans usually report these measures as aggregate rates, a special data-collection effort was organized by NCQA with 10 plans that volunteered to participate in the study after intensive recruitment efforts by NCQA staff. In our analyses, we considered the 7 measures for which the most plans submitted data: child and adolescent immunizations, checkup after delivery, prenatal care in the first trimester, mammograms, cervical cancer screening (Pap smear), and retinal exams for diabetics. These outcomes, like the other HEDIS clinical measures, are all dichotomous.

Over 91,000 patient-level records for quality measures in 1996 were collected altogether with 6 to 9 plans per measure. Most of this information was extracted from administrative databases at the plans. Very limited patient-level was available: age, sex, and zip code of residence.

5 Issue I: Selection of variables for the case-mix model.

The general criteria for inclusion of variables in the case-mix model are specified in Section 2: association with outcome within plan, variation between plans, and appropriateness as case-mix. Furthermore, it may be desirable to select a parsimonious model, which can be easily explained to users of the ratings. In the case of the CAHPS surveys, a large number of variables were identified that satisfied the criterion of appropriateness. The next criterion was of statistical significance; if the coefficient of the candidate variables was not significantly different from 0, it obviously could not be estimated sufficiently precisely to be useful. A substantial number of variables met both of these criteria. Hence, a simple screening method is desirable which combines the first two criteria.

We developed a procedure which approximates the potential impact of a candidate variable in the regression by combining two standard analytical outputs. The incremental predictive power of a candidate added variable x_{new} , obtainable from a regression package, is

$$\Delta R^2 \approx \frac{b^2 \text{Var } x_{\text{new}} | x_{\text{old}}, \text{Plan}}{\text{Var } y | x_{\text{old}}, \text{Plan}} \quad (3)$$

where y is the outcome variable and x_{old} is the vector of variables already in model.

Using this and two within-to-between variance ratios (estimated using standard ANOVA or variance components software) we estimate the impact of the candidate variable on adjusted plan means. We multiply the numerator of (3) by

$$\begin{aligned} & V(x_{\text{new}}, \text{between})/V(x_{\text{new}}, \text{within}) \\ &= \frac{\text{Var } \bar{x}_{\text{new}, \text{plan}}}{\text{Var } x_{\text{new}} | x_{\text{old}}, \text{Plan}}, \end{aligned}$$

the ratio of between to within variances for the residuals of x_{new} from the regression on x_{old} . Similarly, we multiply the denominator of (3) by

$$V(y, \text{between})/V(y, \text{within}) = \frac{\text{Var } \bar{y}_{\text{plan}}}{\text{Var } y | x_{\text{old}}, \text{Plan}},$$

the corresponding variance ratio for the outcome variable. We get approximately

$$\frac{b^2 \text{Var } \bar{x}_{\text{new}, \text{plan}}}{\text{Var } \bar{y}_{\text{plan}}},$$

the variance of the adjustment relative to the variance of plan means.

In the CAHPS-MMC analysis, we calculated the variance ratios for each of the candidate variables, and ΔR^2 for each variable in the model for each of the key outcomes. (We defined x_{old} in the base model as age and self-reported health status, which were clearly important predictors in this as in many previous studies.) The products were used to screen the variables (Table 1). (Of course, ΔR^2 depends on what other variables are in the model, so this procedure is subject to the potential instability of any stepwise procedure.)

Education (individual) is both a strong predictor within plan and has a large between-to-within variance ratio. The same is true for health. Black and Hispanic have large between-to-within ratios, but are weak within-plan predictors, so their predicted impact is small. Sex is a moderately strong predictor of individual satisfaction, but there was little between-plan variability on the distribution of sex. Hence, inclusion of sex in the model would have had little impact on ratings of plans.

Interestingly, the variable for educational level of the respondent's zip code area had much less predictive power than personal educational level in the individual level model. The variance ratio (between/within), however, was much larger for the zip area variable. Consequently, the impact of the zip variable was comparable to that of the individually measured variable for education. The same is true for Black or Hispanic race/ethnicity measured contextually. We return to this fact in Section 7.

6 Issue II: Homogeneity of case-mix coefficients

The assumption of models (1) and (2) is that case-mix coefficients β are the same at every plan. This assumption is empirically falsifiable. There may be demonstrable heterogeneity, i.e. the data may support a model with different slopes β_p at the various plans.

Heterogeneity of slopes has important consequences. Technically, it means that adjusted differences between the plans, and even the ranking of the plans, will be affected by the choice of the "average" covariate values \bar{x} . More important, the interpretation of the case-mix results is considerably

complicated. Even if there is an agreed-upon standard patient, ratings adjusted to that patient do not characterize the differences between plans for other patients. In fact, the entire notion of a single adjusted rating for a plan is called into question. For example, if the coefficient of age varies by plan, older patients may be more satisfied at Plan A than at Plan B, while younger patients are better satisfied at Plan B than at Plan A. Arguably (although not practically), distinct reports should be prepared for each patient! Philosophically, large variation in the case-mix effects calls into question the appropriateness of regarding these as beyond the control of the plan. If some plans are able to provide equal care to supposedly “difficult-to-treat” populations, it might be appropriate (although difficult) to attempt to measure and rate the differences in care between populations at every plan.

We used two approaches to investigating heterogeneity of case-mix coefficients. One was an exchangeable random effects model, and the second a model with interaction effects for prespecified groups of plans.

6.1 Random heterogeneity of coefficients in Washington State CAHPS

To test for heterogeneity of coefficients, we fit an augmented model to the Washington State data in which the coefficients of age and health status, as well as the intercept, were allowed to vary randomly. We estimated their variance-covariance matrix by REML.

For each of the main outcome variables, the variance component for the intercept was large, indicating that the plans differed substantially along the single dimension incorporated in our model. The variance components for the age and health status variables were very small, with SD much smaller than the corresponding mean coefficient values. We interpret this finding as indicating that the case-mix effects are fairly uniform across plans.

We then considered a patient at one SD (of the distribution of covariates to individuals) from the mean on each of the two case-mix variables. We estimated the impact of between-plan variation in the coefficients on this patient’s predictions, compared to variation in the intercept. The effect of variation in the coefficients had a predictive SD of 27–32% of the predictive SD of the mean effects μ_p . We concluded that for patients who were moderately different from the mean, reporting of the general plan rating is an adequate summary of plan rankings.

A second assessment of the impact of heterogeneity considered the effect of varying slopes on adjust-

ment of plan means. We considered a hypothetical plan at one SD of the *plan means* from the grand mean, and estimated the effect of variation in coefficients across plans on the adjustment for that plan. The effect was very small, predictive SD less than 0.5% of the SD of the mean effects μ_p . We concluded that the impact of heterogeneity of coefficients on the adjustments was negligible.

6.2 Systematic heterogeneity of coefficients in Medicare CAHPS

Because of the tremendous geographical extent of the CAHPS-MMC study, we were interested in assessing possible heterogeneity of case-mix coefficients across regions. For these purposes, we defined eight regions (standard Federal regions, merging three regions that had few plans).

Our main model had three case-mix variables — age, health status and education — two of which (age and education) were specified as categorical variables with several levels. We compared, using standard ANOVA techniques, three specifications of the age by region interaction: no regional interactions, regional interactions with a linear trend age effect, and regional interactions with each of the category parameters in the national age model. Similar specifications were tested for education, and the single interaction with trend was tested for health status. We found that the trend interactions were significant but the categorical interactions were not. Including these effects did not excessively complicate our model.

A consequence of specifying our model in this way is that heterogeneity of coefficients could complicate comparisons between plans in different regions, but not in the same regions. We decided not to case-mix differences between regions, so each regions unadjusted and adjusted means were the same. Besides the difficulties outlined above, case-mix adjustments between disjoint regions seemed too conjectural. The most important comparisons were between plans that competed for the same members, and these comparisons were rarely affected by inter-regional adjustments.

6.3 When heterogeneity is a big story: the HEDIS measures

Our analysis of interactions in the HEDIS data was limited by the small number of plans (6 to 9 per measure), which limited the extent to which we could quantify between-plan variation in case-mix coefficients. We tested plan by predictor interactions for the 27 predictor-outcome pairs that had significant associations in models with a single case-mix variable. Of these, 5 interactions were signifi-

cant at the .05 level, although it was hard to establish the magnitude of the interaction effects.

We also investigated whether the associations are significant when tested against the plan by interaction effects, treating the plans as a random sample of clusters. For this we used robust variance estimation procedures available in standard software for the analysis of survey data. Here again, the results were consistent with the standard case-mix model but the results were inconclusive.

Interaction effects are a particularly important part of the story for the HEDIS measures. Most of the case-mix variables in the HEDIS analyses represent characteristics of areas (mean income, education, race/ethnicity) that have been identified as related to inequities in access to health care. Significant effects in the case-mix models suggest that these inequities persist even among patients that have the same insurance coverage. Furthermore, the measures are relatively objective, and arguably less subject to bias than the self-reports of the CAHPS surveys. Interactions between case-mix variables and plan, if present, are an indication that some plans are more successful than others in equalizing care across areas while members of other plans experience greater disparities. Whether or not it will be possible to rate each plan on this interaction, it would be interesting in future research to assess the general magnitude of the interaction effects.

7 Issue III: Using geographically-based (contextual) variables

In some data sets (such as those generated for our HEDIS project), there are few sociodemographic variables (only age and sex) for individuals. This lack reflects the limitations of the administrative and clinical records on which HEDIS is based. In that case, geographically-based (contextual) variables may be the best available source of information on characteristics that can be used for case-mix adjustment. Contextual variables based on zip code statistics are often available (whenever there is an address). If confidentiality issues are not an obstacle, exact addresses can be geocoded to block groups, for which census data are also available. In our analyses, we used 6 such variables (chosen out of a longer list of variables after a principle components analysis), defined as the percentage of residents of the zip code who were (1) receiving public assistance income, (2) college-educated, (3) Black, (4) Hispanic, (5) Asian, or (6) in an urban area.

Even where individual-level variables are available, as in the CAHPS surveys, contextual variables may provide additional information. Knowing that

a person lives in an area with a high percentage of college educated people is different information from knowing that the person is college educated himself, and both may be useful predictors.

Conversely, we cannot deduce effects of individual level characteristics from “ecological” analyses (regressions on areas means) without additional strong assumptions. Hence, these effects must be interpreted carefully as relating to area rather than personal characteristics. Nonetheless, contextual variables are valid predictors that can be used in a case-mix model. While *predictive* power is essential, *interpretation* of the coefficients is not important to case-mix adjustment.

The usefulness of these variables was borne out in our analyses. In the HEDIS analyses, percentage college educated was positively associated with six of our seven measures and percentage receiving public assistance income was negatively associated with the same six measures. These effects are consistent with previous research on medically underserved populations.

In the CAHPS-MMC analysis, we found that contextual educational level (percent with college education in zip code area) was a significant predictor even after controlling for individual educational level as reported on the survey. As noted earlier, contextual education was a less powerful predictor of individual satisfaction than was individual education, but the between-to-within variance ratio is much larger for individual education, so the overall impact of the contextual variable was almost as large as that of the individual variable. This is a typical effect that can also be seen when comparing other pairs of related individual and contextual variables. The contextual variable loses some predictive power by averaging over individual differences, but because the service areas of plans tend to be geographically concentrated, much of the between-plan variation is captured in the contextual variables.

A potential (but at this point entirely conjectural) problem with use of contextual variables is that their effects may be especially subject to regional variation. For example, in one area Hispanics may live in a downtown area close to major teaching hospitals and in another they may live in rural areas where even clinics are rare. In order to determine the extent to which this happens, and whether it happens more with contextual than with individual variables, we need to analyze diverse samples and test models with interactions.

8 Issue IV: Summarizing the impact of case-mix adjustment

After case-mix adjustment has been carried out, we want to summarize the magnitude of its effects, both absolutely and relative to typical between-plan differences. In particular, we would like to know how much impact adjustment has on the ranking of the plans. We are also interested in the sensitivity of case-mix results to choice among alternative models. This information is important to users of the ratings because they bear on the extent to which they should be concerned about the adjustment in general, and details of the specification in particular.

Table 2 shows a typical display of the type we used in evaluating impact of case-mix adjustment in the CAHPS-MMC implementation. (Separate columns were displayed for each region separately, and for a series of other measures.) Models A0, A1, . . . B2 refer to a series of alternative models differing in the treatment of regional interactions and of the contextual education variable; of these, A2 (regional interactions and no contextual variable) was selected as the production model and the others were alternatives.

The first block of the table summarizes the magnitude of the adjustments, compared to the variability among plan means. The adjustment slightly reduces the variance of the means. Note that this result was not a mathematical necessity, and adjustment could have increased that variance as well. The magnitude of the adjustments averages about 14.4% of the magnitude of the differences among the plan means. The next block shows the largest adjustments up and down. Although the SD of the adjustments is not large, the largest adjustment is over 3/4 of the SD of the original means, certainly enough to be important to the affected plan.

The next block shows the Kendall correlation between the unadjusted and adjusted means. An interpretation of this statistic is that $(1 - .891)/2 = 5.5\%$ of the pairs of plans would switch their relative rankings due to adjustment. Another set of measures addresses changes in the “star ratings” of the plans. Plans were assigned one, two or three stars as they were significantly below the mean of plan means in their region, not significantly different, or significantly above. Of the 232 plans, 24 changed their star ratings due to adjustment. This measure, although it describes effects on a display which is of great interest to plans, nonetheless is largely a count of plans that were near the cutoff and were moved just across it by adjustment, so it is not very satis-

factory as a statistical measure. (A similar criticism applies to counting the number of plans whose clinical quality ratings move over or under a criterion for plan accreditation.)

The final block of Table 2 summarizes the impact on adjustments of the choice among models. The typical differences between models are small compared to the typical adjustments, indicating that model sensitivity is not excessive.

A graphical display (Figure) illustrates the effect of adjustment on HEDIS data for two measures, adolescent immunizations and cervical cancer screening (Pap smear). These displays show that most plans have modest adjustments but a few, especially one whose unadjusted score is worst on both measures, would receive very substantial adjustments. Examination of the data for this plan shows that it has membership that is very concentrated in areas with high public assistance recipients and concentration of Hispanic residents.

9 Conclusion

Case-mix adjustment involves a series of difficult and possibly controversial decisions. On the other hand, it is becoming part of the expected standard for presentation of quality ratings. The case-mix effects themselves are of intrinsic interest as indications of differential quality of care (whether received or perceived). It is important to remember, however, that the case-mix analysis is only partial. In particular, it does not address the correlates and predictors of quality at the plan level. For this, other analyses, also challenging, are required.

Reference note

At the time of preparation of this manuscript, a number of papers describing the projects mentioned here are in preparation or under review. Contact the author, zaslavsky@hcp.med.harvard.edu, for a current bibliography.

Acknowledgements

Support was received from the Agency for Health Care Policy and Research (grants HS09473-02 and U18-HS09205-03), and the Health Care Finance Agency (contract HCF-98-C-00057-004-0043). The projects described here have involved a multitude of talented researchers whose experience and insights I have drawn on and whose labors are represented in the results presented here. I particularly thank Paul Cleary and Arnold Epstein of Harvard University, principal investigators respectively of the core CAHPS (Harvard team) and QSPAN-HEDIS projects. I also thank our collaborators including Eric Schneider (Harvard School of Public Health), John Hochheimer and Joseph Thomp-

son (NCQA), Elizabeth McGlynn (Rand Corporation), Kathy Langwell (Barents), Sherm Edwards and John Rauch (Westat), David Veroff and Lee Hargraves (Picker Institute), Jack Fowler (University of Massachusetts), Mary Uyeda (Washington State Health Care Agency), and Elizabeth Goldstein (HCFA) and analysts at Harvard University including Jane Appleyard, Matt Cioffi, Lin Ding, Jim Shaul, Larry Zaborski, and Jie Zheng. The views expressed here are my own, and not necessarily those of the collaborators or the sponsoring agencies. Any mistakes and omissions are also my own responsibility.

Table 1: Screening variables for the case-mix model for “Overall rating of plan.”

Table 2: Display of impact of case-mix adjustment in CAHPS-MMC for “Overall rating of plan.”

SD of plan means: unadjusted	0.401
SD of plan means: adjusted (A2)	0.395
SD of adjustments	0.058
SD(adjustment)/ SD(unadjusted means)	0.144
Mean absolute adjustment	0.043
Largest adjustment up	0.312
Largest adjustment down	-0.255
Largest relative shift	0.567
Kendall correlation between adjusted and unadjusted means	0.891
“Star” ratings (unadjusted—adjusted)	
1-1	44
2-2	87
3-3	77
1-2	5
2-1	9
2-3	1
3-2	9
SD of difference of means between models	
A2-B2	0.029
A2-A0	0.010
A2-A1	0.012