RECORD, Volume 24, No. 3*

New York Annual Meeting
October 18–21, 1998

Session 148PD
Health Employer (Effectiveness) Data Information Set: Does it Work?

Track: Health/SOA Research
Key Words: Health Care Plans, Legislation and Regulation, Managed Care, Product Development

Moderator: THOMAS P. EDWALDS
Panelists: MICHAEL CHERNEW†
DENISE LOVE‡
LUIS MANUEL C. PAITA§
DENNIS P. SCANLON**
Recorder: THOMAS P. EDWALDS

Summary: This session includes an overview of the Health Plan Employer Effectiveness Data Information Set material presented at the 1998 Symposium on Managed Care. Panelists present the results of research undertaken to determine the effectiveness of HEDIS as a measure of health care delivery quality. The importance, or lack thereof, of these measures in member decision making is examined.

†Mr. Chernew, not a member of the sponsoring organizations, is Assistant Professor at the University of Michigan in Ann Arbor, MI.

‡Ms. Love, not a member of the sponsoring organizations, is Director of Health Data Analysis at the Utah Department of Health in Salt Lake City, UT.

§Mr. Paita, not a member of the sponsoring organizations, is Director of Research at the Office of Health Data Analysis/Utah Department. of Health in Salt Lake City, UT.

**Mr. Scanlon, not a member of the sponsoring organizations is Assistant Professor at the Department of Health Policy Administration at the Pennsylvania State University in University Park, PA.

Note: The charts referred to in the text can be found at the end of the manuscript.

Mr. Thomas P. Edwalds: I'm the senior research actuary for health and pensions at the SOA. In that role I have provided the staff support for two research projects that the SOA sponsored concerning HEDIS measures.

HEDIS stands for Health Plan Employer (Effectiveness) Data and Information Set, and it is essentially a prescription for a set of things that should be measured about managed care plans. It was put together by an industry and government group called the National Committee for Quality Assurance (NCQA) which started looking at some things that would make sense to measure and by considering what could be measured, and it has progressed over time. It's like a software package, in that it has a version number. The current version is 3.0.

Our studies were actually done with the most recent version, 2.5, because 3.0 is brand new, and there hasn't been any data collected from it yet, but we tried to focus on measures that are the same in 2.5 and 3.0. It's a big set. In 3.0, there are eight different domains of measures, but within each of those there's a fairly lengthy list of items to be measured. The specifications are about two inches thick. This is an area where actuaries can be very useful, because we're talking about things that need to be measured, validated, and made sense of. One of the things we're trying to do is figure out what needs to be measured.

The first pair of presenters is from the Utah Department of Health. Denise Love is the director of the Office of Health Data Analysis, and Luis Paita is the research director there. In their research they have looked at HEDIS measures. How reliable are they? What do they tell us? I'll let Denise take it from here.

Ms. Denise Love: I appreciate the SOA's support in what I consider an important research project. I will provide background about how the state of Utah has adopted and implemented HEDIS reporting and some of the barriers that we ran into as a state managing multiplan reporting. Luis Paita will discuss research findings from this study of the analytic utility of HEDIS.

Many performance measures in use these days are not rooted in science, and HEDIS is no exception. HEDIS was created to meet market demands as managed care penetration increased. Plans, purchasers, and states are looking for a way to define and measure value, not just price. HEDIS was developed to meet industry needs, not state, Medicaid, or scientific inquiry needs. We believe that this project, supported by the SOA, begins a much-needed effort to look at the science behind HEDIS.

States have to manage political and technical dynamics when implementing HEDIS. It's not enough to put a single-plan HEDIS set together. The state's accountability in

publishing a HEDIS data set is high, as it should be. We are managing the dynamics of proprietary tensions. We're managing the dynamics of the advocacy groups who want to know how Medicaid clients and other vulnerable citizens are faring in managed care programs. There's a lot of negativity about managed care, and there's a certain negative bias that many legislators have about managed care.

States like Utah already have the authority by law to collect, measure, and compare health data. We have been doing this with hospital data for many years. The addition into law of an independent validation of the office's results was driven by the fear that plans and others held about what we would do with any quality reporting results. This is just a representation of the political dynamics around HEDIS and public reporting.

Utah spent six months just planning managed care reporting, deciding what we were going to measure, what the unit of analysis would be, and how we were going to measure different aspects of plan performance. It took us two months to define a health plan. We never could get consensus, so we moved on. HEDIS was the result of this planning process. We knew that HEDIS was not perfect, but we didn't know how imperfect it was. Up front, we said we would use existing tools and HEDIS would be a transition tool to more perfect measures. States, including Utah, have limited resources. It makes sense to adapt existing measures instead of creating our own. HEDIS brought us a "package deal" that we could implement at a state level. In Utah, we integrated Medicaid with our commercial plans in a statewide HEDIS measurement project, which included enrollee satisfaction and audited HEDIS. In addition, we participated in piloting NCQA audit standards.

One of the largest problems in 1996 was that plans had varying degrees of experience with HEDIS reporting. Struggles with system upgrades and mergers, and their impact on the information systems and platforms, present a barrier. We have plans in Utah producing HEDIS through vendors, and some vendors have very few quality controls. Vendors may not be concerned with the implications of noncompliance to state reporting, which is a $5,000 fine per day. (We have one vendor that doesn't seem to be concerned about the plan's noncompliance in 1998.) This is no secret to many in this room, but I was surprised at the level of internal coding schemes and different software used in the plans. Also, unedited encounter data that were used to derive HEDIS measures, combined product lines, and other factors influenced data quality. So, we had a host of problems.

Plans fell short in producing HEDIS. Utah had 21 measures out of a potential of 71 that we could use to compare plans in the state of Utah for the first year and only 16 comparable measures for Medicaid. We put in a lot of work and energy to get to this point in the state of Utah. This was about a year and a half down the road. The

expectation was a product, and I, as the appointed state data agency director, had to make do with what we had or "spin straw into gold." We had to prove that the investment was worthwhile and that managed care plans in Utah could be accountable despite the challenges: tracking continuous enrollment, medical record abstraction problems, lack of provider information, and mixed product-line reporting. According to NCQA, these problems are consistent with national findings.

What should be the target for states that can't afford to produce the entire HEDIS set? That's what the SOA grant was intended to shed light on. We fell short of our goal to hold plans accountable for outcomes. We have to make up for the gaps. We are working with plans to improve their data reporting and data systems. Through this grant, we can set some analytic standards and continue moving toward our goal. Despite the problems in the state of Utah, we did derive HEDIS benchmarks that provide important context when speaking to legislators and others in the state.

How did Utah use HEDIS for quality measurement? Selected HEDIS measures, such as well-child care, drew a great deal of interest from the media and the plans. We found that we weren't as good in Utah as we thought, and that led to communitywide collaboration to improve well-child care HEDIS rates. If those data weren't made public and audited, I believe we would never have gotten to the community improvement part of this. For consumers, we derived a series of products from our limited HEDIS set, such as consumer report cards for Medicaid and commercial health plan survey reports. For policy makers and others, a HEDIS data set, an enrollee survey analytic file that Luis will talk about, and other formats are under evaluation.

At the time that we were struggling with what to do with these 21 measures, we saw the SOA's request for proposals, which seemed to be asking the same questions: We have HEDIS. Now what? Does this mean we can differentiate a good plan from a bad plan? What does this really mean? The SOA was looking at some questions to guide members in the HEDIS data-collection process: Should we do HEDIS annually? It costs a lot. The cost burden to plans in the state of Utah to compile the whole HEDIS data set is not insignificant. Should we rotate some of the measures or should we not? And, if we do, which ones should we rotate? Because we have problem measures, which ones should we work on and which ones should we skip altogether? Can other data serve as a proxy to HEDIS measures? Are there NCQA restrictions preventing a plan from reporting HEDIS measures? To what extent do enrollee satisfaction surveys reflect health plan performance?

Other questions which were of concern to Luis as research director included the following: How consistent are HEDIS measures across plan types, location, and enrollment composition? How consistent is that ranking across all the measures? What are the thresholds for rates? We were seeing some 2% differences and 7% differences. Also we were being asked by users about significance: Is that statistical significance or clinical significance, who sets those, and who decides? How do audited plans compare with unaudited plans? As HEDIS is cited by the media across the country (such as in *U.S. News & World Report*), analytic standards become more important.

What we were trying to get at with the SOA grant is the degree to which HEDIS process measures correlate with objective outcome measures or, in other words, predictive validity? How do subjective outcome measures correlate with performance? To what extent are subjective measures influenced by a health plan's performance, independent of enrollee characteristics or attributional validity (or what are the adjustment factors in enrollment when we compare plans)?

For this grant we proposed a study framework that Luis will go into in more detail. We are proposing this framework not only for this grant, but for other measure sets as well.

The data sources we used in our research over the last year include: NCQA's Quality Compass 2.5, the Administrative Solution Group's (ASG) HEDIS 2.5 data set (formerly Health Benefits America), and Mercer Management Group's HEDIS data. The ASG set included health plan descriptive information, satisfaction surveys, HEDIS measures, and precalculated data elements for 253 health plans. The Mercer data set that we used for this study had 27 HEDIS measures for 838 health plans, with 255 reporting at least 10 measures and 331 reporting at least one measure. Also, we used the state of Utah's patient level enrollee satisfaction survey.

The data that we used for this research project had limitations. First and foremost, it was unaudited HEDIS 2.5. There were missing measures and variances among data sources. NCQA began addressing the data quality issue through an audit in 1996. To account for the data problems, we tested all of the data for outlier values and validated selected measures across the three data sets. Table 1 summarizes the analytic data sets used in the study.

TABLE 1
HOW DO DIFFERENT HEDIS SOURCES
COMPARE IN TERMS OF ANALYTIC UTILITY?

Source	Number Of Plans With Valid Data	Completeness	Use
ASG	253	63% reporting Include enrollment data Include information for controlling factors	Analytic data
NCQA Quality Compass	211	Limited enrollment data Limited plan's structural information	Validation
Mercer	200	Only 27 HEDIS measures	Validation

The ASG data set was our primary research data set. It was a robust data set and more complete than the others. About 90% of the measures that we randomly cross-validated using the three data sources had the same value in both data sets. Before we delved into the study framework, we spent quite a bit of time cleaning up, cross-checking, and making sure that we understood the data. I'll let Luis talk about the statistical design and process.

Mr. Luis Manuel C. Paita: The study framework was effectively an analysis of the HEDIS measures categorized by different groups of related measures, which we call "measure sets." We categorized them into five boxes: (1) health plan profile, which includes measures of stability and of the structure of the health plan; (2) enrollment profile, which includes demographic characteristics and composition of the enrollee population of the health plans; (3) HEDIS process measures; (4) HEDIS objective outcome measures; and (5) HEDIS subjective outcome measures.

In this study, we made the decision to categorize clinical and administrative services as HEDIS process measures. The HEDIS process measures also included the effectiveness-of-care measures, including immunization rates and admission and readmission for major affective disorders. HEDIS objective outcome measures included utilization measures such as discharge rates, length of stay, and cost and utilization for outpatient services, ambulatory surgery, and emergency room utilization. Objective outcome measures also include some of the more clinical outcome measures, such as low and very low birth weight rates. The last box in this framework, the HEDIS subjective outcome measures, incorporates: retention rate, willingness to recommend, and satisfaction rate.

We addressed the questions that Denise mentioned by focusing on particular areas of this framework. We looked at internal consistency, predictive validity, and attributional validity. The degree to which measures within a set are correlated is

what we call internal consistency. Predictive validity looks at the correlation between HEDIS process measures and HEDIS outcome measures, both objective and subjective. Attributional validity refers to the strength of the residual effect of process measure sets on outcome measures adjusted for health plan profile and enrollment profile.

The first thing we looked at was internal consistency: How consistent is the relative ranking of health plans across HEDIS measures within measure sets? We did this by looking at the correlation matrix, containing nonparametric correlation between measures within a measure set, and then calculating the Cronbach's alpha reliability. This then became the basis for categorizing measure sets into high, moderate, or low internal consistency. To confirm the results from the correlation analysis and the level of Cronbach's alpha, we did a factor analysis. Most of the time, the results with all these three methods were consistent.

I'll illustrate what we did with the immunization rates, which included seven measures of immunization, overall immunization rate, and type-specific immunization rates. Table 2 is a correlation matrix showing very high pairwise correlations between the type-specific immunization and the overall immunization, which suggests high internal consistency. This is confirmed by a Cronbach's alpha of 0.95, which is very high. In the factor analysis, the type-specific immunization rates loaded together as one factor, again a confirmation of the high internal consistency of this measure set. Interestingly, the overall immunization rate did load with the type-specific immunization rates, meaning that health plans having a high immunization rate for a specific type did not necessarily have high immunization rates overall.

TABLE 2
IMMUNIZATION RATES

	DPT	OPV	Measles	Mumps	Rubella	H Influenza
All	0.88	0.85	0.78	0.77	0.79	0.86
DPT		0.92	0.88	0.85	0.85	0.85
OPV			0.88	0.86	0.88	0.82
Measles				0.98	0.97	0.81
Mumps					0.99	0.78
Rubella						0.81

All pairwise correlations are significant at $p < .001$.
Number of valid values range from 150 to 199.

Most of the measure sets were highly internally consistent, including the subjective measures of satisfaction and willingness to recommend; retention rate; all the immunization rates; and high-occurrence, high-cost diagnostic related groups

(DRGs). There was high internal consistency between rate, cost, and length of stay for all high-occurrence, high-cost DRGs. That means one could create an index of resource-use intensity of the health plan by combining rates, cost, and length of stay, without having to analyze the three separately.

The selected procedures, were also highly internally consistent, in terms of both overall rates and the rates by sex and by age. Health plans with high rates for angioplasty, for example, tend to have higher rates for cardiac cath, cholecystectomy, and laminectomy. There also was high internal consistency between length of stay measures. Inpatient, medical surgery cases, maternity, newborn, and the length of stay by age all were highly internally consistent.

Likewise for mental health utilization measures such as discharges per thousand, inpatient length of stay, chemical dependency utilization-related measures, major affective disorder (MAD) readmission rates, and mental health readmission rates, were all highly correlated not only within each of these sets but also across all of these measure sets.

High internal consistency was also observed among all age-specific measures: inpatient acute and nonacute discharges, ambulatory care visits, length-of-stay visits for inpatient acute medical surgery, and nonacute days. Finally, the all-ages category was strongly correlated with the age-specific measures. What this means is that a health plan that lacks the resources need not worry about age-specific measures. In analyzing a big database, one would do the data quality checks (e.g., checking the normality of the distribution) of the overall (all ages) measure instead of several measures. The same thing goes for age-specific measures for maternity discharge, c-section, and other newborns.

The measures by sex shows the same results as by age. Those health plans with a high rate for males also tend to have high rates for females. The both-sexes measure would suffice to analyze rates for selected procedures, mental health utilization, readmissions for MAD, chemical dependency, and readmission for chemical dependency.

Table 3 shows what it means when two measures would be highly correlated. In comparing the x-axis ambulatory surgery procedures per thousand with the y-axis ER visits per thousand, you'll see that health plans in the upper quartile for ambulatory surgery procedures also tend to have high average ER visits per thousand.

TABLE 3
EXAMPLE
ER VISITS /1000 VERSUS
AMBULATORY SURGERY /1000

Q1 (Low)	107.9
Q2	124.7
Q3	143.3
Q4 (High)	160.7

Ambulatory surgery procedures /1,000

The measure sets that we found to have moderate internal consistency, or an alpha level between 0.6 and 0.8, were preventive screening rates, including cholesterol, mammography, and cervical cancer screening; asthma admission and readmission; high-cost and high-occurrence DRGs; and rates, cost, and length of stay across DRGs. Total deliveries per thousand across age groups, vaginal deliveries across age groups, maternity care, and length of stay across age groups were also in the moderate consistency category.

The next thing we looked at was the consistency of HEDIS measures across plan types and enrollment composition to determine whether comparative analysis of performance of health plans must be adjusted for these factors. We looked at the relationship between health plan and enrollment profile with HEDIS process measures, objective outcome measures, and subjective outcome measures. After some preliminary analysis we reduced the analysis to five measures: (1) the type of health plan, (2) the regional location, (3) rate of primary care physician (PCP), (4) percentage change in total profit margin as a measure of stability, and (5) the age/sex composition of the enrollment population of the health plans.

Table 4 shows statistical significant relationships between health plan type and enrollment profile, and HEDIS measures. In general, we saw that there is some need to adjust for health plan type and enrollment composition if we are to compare health plans according to performance as measured by HEDIS.

TABLE 4
HP & ENROLLMENT PROFILE => PROCESS MEASURES

Measures	Plan Type	Plan's Region	Percent PCPs w/ Open Panels	Percent of Female Enrollees	Percent of Enrollees 65+	Percent of Enrollees 0-19
Overall Childhood Immunization	X	X	X	X		X
Composite Asthma Measure			X	X	X	
Average Preventative Screening Rate	X			X	X	
Heart-related DRGs: Discharges/1000				X		
I/P Discharges/1000			X			X
O/P Visit/1000		X		X	X	
Amb. Surg. Proc./ 1000			X		X	X
ER Visits/1000	X	X		X	X	
Mental Health Discharges/1000		X		X	X	X

Next, we looked at predictive validity, the relationship between process measures and objective outcome measures, between process measures and subjective outcomes, and between objective and subjective outcomes. Between process measure and objective outcomes, one of the most consistent results was the relationship with immunization, which we consider a process measure, and: pneumonia and pleurisy rates; the three utilization measures for bronchitis and asthma, namely discharge rate, cost, and length of stay for ages 0–10; inpatient nonacute discharge for ages 0–19; and inpatient nonacute length of stay for ages 0–19.

Another consistent finding was the high correlation between preventive screening measures and inpatient length of stay, outpatient visits per thousand, ER visits per thousand, and nonacute days per thousand. Briefly, utilization measures are affected significantly by preventive screening measures. Table 5 shows a positive significant relationship between inpatient nonacute length of stay and preventive screening rates.

TABLE 5
EXAMPLE
I/P NONACUTE ALOS VERSUS
PREVENTIVE SCREENING RATES

Q1	11.2
Q2	13.5
Q3	16.3
Q4	17.2

Preventive Screening Rates (Quartiles)

Probably more interesting than the previously discussed significant relationships are those that are not statistically related, where we expected a significant relationship to occur. For example, ambulatory follow-up after hospitalization for MAD should be related to readmission for MAD, but we found no statistical significance between these two measures. We expected a significant relationship between cholesterol screening and heart-related utilization measures, but we didn't find any. We expected a relationship between cervical cancer screening and hysterectomy, and we didn't see that. We expected a relationship between prenatal care, low birth rate, and very low birth weight, and we didn't see this statistical relationship. From this we conclude generally that there is a weak relationship between process measures and objective outcome measures, meaning low predictive validity.

Let's look at the relationship of process measures and objective outcomes measures with subjective outcomes. With satisfaction rate as the dependent variable, our multivariate analysis showed the following had significant effects: immunization rates, composite measure from the three preventive screening measures, discharge per thousand for ischemic attack, discharge per thousand for cerebral vascular procedure, discharge per thousand for laminectomy, outpatient discharge per thousand, ambulatory surgery per thousand, and maternity days per thousand.

With retention rate as the dependent variable, we found immunization rate, composite of preventive screening measures, length of stay for nonspecific cerebral vascular disease, outpatient per thousand, and inpatient length of stay for newborns and maternity days per thousand to have significant effects.

Table 6 shows what it means for process measures and objective measures to be related to subjective outcomes. The table shows a positive significant relationship between preventive screening rate and the subjective outcome measures.

TABLE 6
EXAMPLE
PROCESS <==>SUBJECTIVE OUTCOMES

	Percent Satisfied	**Percent Willing to Recommend**	**Retention Rate**
Q1	84%	77%	78%
Q2	84	80	80
Q3	89	86	84
Q4	92	90	94

Preventive Screening

After having seen significant relationship between the process measures, objective outcomes, and subjective outcomes, the next question that we raised was, would this relationship persist if we adjusted for health plan type and enrollment profile?

To answer this we performed a simple multiple regression of satisfaction rate on the process and objective outcomes measures that were correlated on a bivariate level with satisfaction rate. The results are as follows: With satisfaction rate as the dependent variable, the R-square was 0.68 and the significant effects were for immunization rate, outpatient per thousand, ambulatory surgery per thousand, and maternity days per thousand. Looking at retention rate as the dependent variable, the regression had an R-square of 0.31, and revealed only three factors with a significant coefficient: average of preventive screening rates, maternity days per thousand, and inpatient length of stay for newborns.

Those results tell us that the subjective outcome measures may be able to provide us with enough information to compare health plan performance because they are associated with some of the HEDIS objective measures and process measures. The relationship could be caused by factors that are not measured within the HEDIS measure set itself. It would require external, including clinical, data to be able to explain what those two relationships between clinical measures and subjective measures mean.

Enrollee-level data would allow an analysis of those relationships. We did that by gathering data from the Utah survey of enrollee satisfaction and using a similar framework. The primary dependent variables were overall satisfaction and five domains of satisfaction created using factor analysis. We wanted to see if there would be a residual effect if we adjusted for social location, which are socio-demographic measures, and health care needs, which are health status measures. The results showed this to be the case.

Chart 1 shows the relative odds of satisfaction. The striped ones are for the commercial population and the gray ones are for the Medicaid population. The x-axis shows the five domains of satisfaction and the overall satisfaction. The length of the bar indicates the relative odds of being satisfied (i.e., the satisfaction rate of those who perceive no access problem relative to those who perceive access problems). They're all greater than one, meaning that there is a significant difference between those who perceive problems with access and those who did not perceive problems.

Focusing on the results for quality of care, the commercial enrollees were more sensitive to having problems with access than were Medicaid enrollees. The former had much higher relative odds of dissatisfaction and were relatively more easily affected by having problems with access. We saw similar results when we compared enrollees who had utilized health care services and those who had not. Those who had utilized health care services were more likely to be dissatisfied. So, the more you use the system, the more likely you'll be to find problems.

To summarize, we found that health plans that perform well on one measure generally perform well also on other measures within the measure set, but, in general, a health plan's performance on one measure set is a weak predictor of its performance on another measure set. The broader the category you get into, the less relationship you'll see. At the health plan level, process measures are generally a weak predictor of outcome measures.

Another significant finding is that subjective measures of a health plan performance may be an indicator of a health plan's overall quality. A health plan's performance in process and objective measures influence subjective outcomes independent of the enrollment composition of health plan type. Moreover, at the enrollee level, satisfaction with the health plan is influenced by actual experience with the plan, independent of the enrollee's socio-demographic characteristics.

Ms. Love: I'll expound for just a few minutes on the recommendations and where we might want to go from here. From a public agency standpoint, we feel a systematic and standardized audit is very important, but just how important is not clear. We're hoping to create an ongoing dialogue with NCQA about the difference between an audited and an unaudited plan, because this became quite important to us in Utah. Some of you may work for the plans that were ranked in *U.S. News & World Report*, and the methodology used there did not match the methodology used in this study. *Newsweek* used yet another methodology. So, who wins? And how do those rankings compare? I think we need some talk on that, and an audit plays a role in how those plans stacked up.

Reduced measure sets and composite scores started this all. We had HEDIS measures and we had satisfaction measures. I went to Luis one day and said there has to be a way to do an index or a composite score so I can translate this myriad of data into something and communicate that to consumers, the plans, and legislators. We believe patient-level data is still needed. HEDIS does not cover all of the bases here, and one tool cannot fit all of our needs in this room, let alone the country. By gathering and continuing to require encounter-level data, HEDIS is just one piece of the puzzle, and the other data can allow us to analyze different units of analysis from physician to health plan to patient. It increases the measure sensitivity, and mixed measures are going to be very helpful as we start to drill down into the data.

We think that this study's framework should be continued, and we're talking with NCQA for HEDIS 3.0 and future measure sets. Again, benchmarks and what we're comparing become increasingly important. I think we need more guidance, if we're using this tool, to define what statistical significance and clinical significance are, and whether they should be applied across all measures. Is it one standard? I don't

think so. But those discussions must occur if we're to make this tool mean anything to the users.

Then we have the problem of imputing missing data. *U.S. News & World Report* apparently set the missing data to a mean if it was not applicable, to the minimum score if it was a not-report, and to the mean if it was unknown. Maybe that's right. Maybe that's wrong. I think we need to have some guidance about proper methodology. My research team kept coming to me throughout this study, as they were mining all this data and doing all the relationship testing, and asked, "What should be the ultimate measure that we're testing for? What is the ultimate single measure or measure set?" Those questions needs some more discussion in the health care community.

Mr. Edwalds: Our next pair of speakers, Dennis Scanlon from Penn State University and Mike Chernew from the University of Michigan, have done a study specifically looking at one very large corporation that reported something about HEDIS measures to all its employees during open enrollment. The company said, "Here are the plans you have to choose from and some information about the 'quality' of those plans." What they wanted to find out is whether or not anybody cared.

Mr. Dennis P. Scanlon: Mike and I embarked on this study as a continuation of some studies we've been doing in this area, looking at the relationship between HEDIS measures and health plan choice. Specifically, the research questions we were interested in asking were whether or not employee decisions regarding health plan choice are related to two things: (1) HEDIS-based, employer-reported health plan ratings that were constructed by the employer and disseminated to employees and retirees, and (2) individual HEDIS measures.

I first want to emphasize that HEDIS was developed in 1989 by a series of large employers who got together and said, "We want to know more about the value of the health plans that we're purchasing. We know a little bit about cost, but we want to know the level of quality we're getting for that cost." So, this came out of the purchasing sector. HEDIS is the most prominent data set. It's standardized and, starting in 1999, virtually all health plans that want to be accredited will have to report HEDIS data. This is something that's not going away any time soon.

The methods we used to answer these research questions were to take enrollment data from a single *Fortune* 100 employer and relate actual health plan choices to the HEDIS measures and the firm health plan ratings that were disseminated. I won't go into a lot of detail about the specific methodology, but we analyzed this from an economic framework, using the concept of utility maximization, and our specific

empirical estimates are derived by conditional logistic regression, which are the results that I'm going to be reporting today.

The data that we used in this study is a combination of HEDIS 2.0 and 2.5 data collected from all health plans. The reason we used the earlier versions and not 3.0 is that this was the underlying data used by the corporation to develop the ratings that were disseminated to their employees. However, for the ratings that were developed and for the measures that we specifically look at, the differences between 3.0 and the earlier versions were minor. One difference is that HEDIS 3.0 includes some additional outcome measures, for example, whether or not beta blockers are prescribed post-myocardial infarction. That's something that was not included in HEDIS 2.0 and 2.5. However, that's also a measure that was not included in the ratings that were disseminated by the firm. Those ratings included preventive-care measures or surgical-care measures, which, by and large, were pretty much the same between the versions that we looked at and the new version. There have been some major changes with the evolution of HEDIS since the SOA's Managed Care Symposium in the spring of 1998 and that has led to version 3.0 essentially being outdated as of 1999. So, this is a constantly evolving data set.

We also used in our data the five performance ratings developed by the firm, which were labeled as follows: participant satisfaction, preventive care, medical treatment, surgical care, and physician quality. Under each of these ratings a plan could receive one, two, or three diamonds on each domain; one diamond signified "needs improvement," two diamonds meant "average," and three diamonds signified "above average performance." It was these ratings that were reported to employees during open enrollment.

Employees were given two things. The first was a chart that said "comparing your medical options." This was customized for employees and retirees based on their geographic location. The five or six plans that employees could choose from listed the five ratings that the employees were given on these five different domains, as well as the out-of-pocket price that was relevant to the employees for each of those plans. So, they were given one spreadsheet with all the relevant data that the firm felt might influence their plan choice.

In addition to that, for each plan, they were given a more detailed description of each plan. Certain assumptions were used to develop the methodology for developing the ratings. For example, for surgical care, the description says: "Many medical problems can be treated without surgery, and health plans are rated on their ability to perform only surgical procedures that are medically necessary." What that essentially says is that plans that were evaluated higher did fewer surgeries. That's in the fine print. If you didn't read the fine print, as an employee

or retiree, you wouldn't necessarily understand that. Many people would take issue with the fact that the rating essentially is based on the volume of surgical procedures done without justification for something that we, in research, call case-mix severity or the illness level of the population. Nonetheless, this was the methodology that the firm employed, and, as researchers, we were trying to take a look at whether or not we could detect a correlation between the enrollment choices that people made and these ratings that were developed. However, the methodology and the underlying assumptions certainly are key.

We had three specific hypotheses. First, we assumed that employees might focus only on price and model type. Most of the literature to date on health plan choice said that these factors seem to matter in health plan choice. We know that there are many different model types across managed care organizations: staff models, group models, point-of-service models, PPO models, and the ABCs of HMOs, which many of you are probably familiar with but yet may not completely understand.

Our second hypothesis was that, aside from price and model type, other plan attributes might matter. For example, one might hypothesize that the information provided by the employer could have an impact on plan choice, and that was something we were interested in looking at.

Our third hypothesis was that informal plan information may dominate. When you read *U.S. News & World Report* and ask people what they look at when they decide to choose a health plan, often the first thing they mention is the experiences of friends, family, and colleagues. Could this be lumped under this category of informal plan assessments, which very well might affect plan choice? Unfortunately, we, as researchers, have no measurable variable for that type of information to include in our analysis, although we think such information does have an impact.

There are other reasons consumers may not use specific plan ratings given by the employer. Past research has found that individuals generally report that information would be very useful or essential to plan choice. For example, in one survey, consumers said information on physician quality, preventive care, and satisfaction would be very important. However, when you look at whether or not consumers used this information in a revealed preference scenario, you find there's no good evidence that they do. Why might this be the case? First, many consumers that don't have a choice of health plan so the notion of choice is not applicable in their situation. Second, consumers may not be aware of information. Even in our employed-environment case, where this was mailed to employees as part of their materials for open enrollment, it's possible that they ignored the information and simply filled out the plan selection sheet and mailed it back in. Consumers may be

overwhelmed by health care information, just as they are for any good or service. There's so much information available that they may not use it. And, again, consumers may care more about attributes that are not captured by the information, such as the informal plan information that they get from other sources.

Most of the literature to date has focused on a methodology using focus groups and survey methods, and certainly there are some advantages to doing that. One is you have direct measurement. You can sit people down in a room and give them all these vignettes and scenarios and ask them how they would make choices given this type of information. You have, therefore, greater control over the research design and few data problems. You have essentially no missing data because you can make sure people in the room answer the questions. The problem with this type of research methodology, however, is that people don't always do what they say. When you sit people down in a focus group and give them some hypothetical situations, they are not evaluating the real doctors and plans, and they're not spending their own real money. Therefore, some would argue that you have to be cautious in interpreting the results of focus group and survey methods.

We took a different approach. Several statistical issues are important for what we were trying to do. First of all, there are many HEDIS measures. In particular, 24 HEDIS measures constituted the ratings that we used in our study. The HEDIS 3.0 data set has about 75 measures. So, to look at what is important and what is not important in a plan choice model, you ultimately have to get personal and be selective. We employed factor analysis to come up with a set of measures that seemed to associate well together. We incorporated these HEDIS-based factors into our choice models to see if, in fact, there was any relationship between choice in these HEDIS-based factors.

There are also missing data problems, and this is notorious throughout the early stages of HEDIS. Many plans were told that they suddenly had to measure themselves. They had to come up with about 75 HEDIS measures, and that's a very difficult thing for plans to do, for two reasons: (1) they weren't used to doing it and (2) they didn't necessarily have the infrastructure and support to do it. So, during the early stages of HEDIS for which 2.0, 2.5, and even 3.0 would apply, there are missing data problems, either several measures that contained a lot of missing data across plans or several plans that had lots of missing data across measures. So, we had to figure out a way to adjust for that.

We decided that, if there were particular measures that had a lot of missing data across plans, we would drop those measures from our analysis. If there were particular plans that reported less than 50% of the data, we would get rid of those

plans from our analysis. And we would employ imputation to come up with data for the remaining missing data.

Then there were other factors that might influence plan choice. We tried to come up with some data using the interstudy data set to have control variables in our analysis.

Let's talk a little bit about the setting within which this choice occurred. It was employee choice at a *Fortune* 100 firm. Choices were made in October 1995 for enrollment in 1996. Choices were made in this firm as part of a flex-benefit system, where the firm defined the choice set that each relevant employee had. If you lived in Boston, the firm determined the five or six plans in Boston that you could choose among, and that choice was determined based on geography and business unit.

The firm also set the prices, and it's important to note that the prices we observed were not equal to true plan premiums. In fact, we wouldn't want true plan premiums. What we observed and included in our analysis is the actual out-of-pocket cost to the employee as part of a flex-benefit system. In many cases, the employer heavily subsidizes the total plan premium, but what's relevant to employees is the out-of-pocket premium, so that's what we included in our analysis.

We observed the choice set, that is, the plans one can choose from, the prices for each plan one can choose from, and the model type of that plan. We lumped model type into something we called integrated versus nonintegrated plans. Integrated plans were those that had more integration between the provider network and the insurance financing component of the health plan. These would be staff-model or group-model HMOs, such as the Kaiser-Permanente model that many of you may be familiar with. The nonintegrated plans have less integration between the provider network and the insurance function. These would be, for example, independent practice associations, where the plans contract with individuals who essentially contract with multiple plans and operate as autonomous physicians for the most part.

We also included the plan-specific HEDIS measures and the firm ratings. I want to emphasize that the way we set this up from a modeling perspective is that choice was modeled relative to the other plans that one had in their choice set. This is important because what matters, we would argue, in choice of plan is the price of the five or six plans. Employees are thinking, "How is the plan that I am enrolled in this year priced relative to the other options? Or, when I look at the five or six options, how do these so-called ratings or HEDIS or quality measures vary across plans that I choose from?" They are not about how the plan compares to a plan in

another market because that's really not part of the relevant choice set. Methodologically, that's how we set up our analysis.

The sample included 9,729 non-union employees choosing managed care family coverage. We chose family as opposed to single coverage because many of the HEDIS measures are more relevant to families. The immunization, prenatal care, and surgical care rates are in some cases relevant to children, women, or men between ages 45 and 64 who are more likely to be users of, for example, coronary artery bypass graft surgery.

The average age in our sample was 41 years. The average number of choices in the choice set was 4.7. Even though, nationally, there are many small employers who give their employees very little choice, in our sample there was a significant amount of choice. The employees in plans cut across 68 markets, with an average of 143 employees per market (geographic location and business unit). In our study, we controlled for the out-of-pocket price charged to the employee and the model type.

Let's get to the results. We found that price matters, and the odds ratios that we generated from our study indicate that employees are between 6% and 11% less likely to choose a plan priced $10 per month more than the mean price relative to an otherwise identical plan. Price seems to make a difference, even with only a $10 incremental difference in price per month, or a $120 differential per year. Model type seems to be less important in influencing plan choice.

The issue we want to stress is that this result may be correlated with other factors. It may not be the plan ratings, per se, versus the correlation that exists between those important unobserved plan variables and how they are correlated with the information that was given. Another way of saying this is it may not be that people looked directly at these plans and said, "This is how I'm going to make my plan choice." It may be that they made their plan choice based on how they would always make their plan choice in the absence of this specific information, and it happened that we were picking up some correlations because there are correlations between what employees value and what the firm reported. And the reason we can say that, incidentally, is because we got some statistically significant results that are, in some sense, nonsensical, which I'll talk about in a moment.

We found that choice is generally not positively correlated with the firm-reported ratings. Satisfaction, however, does have the hypothesized sign. We found that, with the satisfaction rating, employees were 17% more likely to choose a superior plan relative to an average plan, everything else constant. Plans that had a "needs improvement rating" were 57% less likely to be chosen compared to those with an

average rating. There's a particular rating where we do observe the hypothesized signs.

Let's discuss surgical care. We found that the firm rating system scored plans with high rates of surgical procedures poorly. Nevertheless, enrollees seemed to opt for plans with high rates of surgical care procedures. I mentioned the importance of potential informal information and how we're able to make that cautious conclusion in our study. Employees tended to enroll in plans that were rated worse in the plan-developed rating on surgical care.

We have to think about what "worse" means. Worse means plans that actually do more surgeries based on the methodology and assumptions that the firm came up with. Essentially, consumers are enrolling in plans that do more surgeries and hence are rated poorly. Some of the plan traits that might be important to consumers in making a choice are (1) easy access to specialists, and (2) easy access to procedures when you want them. That may be inversely correlated with the ratings that were developed by the firms. Finding this inverse relationship with the plan ratings says to us that (1) there may be some problems with the methodology used to construct these ratings, but (2) some of these underlying measures might be important.

Therefore, we brought in the individual HEDIS measures to see if we could test some of those hypotheses. We found that, in some cases, those did seem to check out. For example, if we look at c-section rates, we found that there was an increased probability of enrolling in a plan with higher rates of c-sections. That doesn't necessarily say that women employees value c-sections or want a c-section, but that the choice of plan may be correlated with underlying, unobserved measures, such as ease of access and referral to specialists and that sort of thing.

Ms. Stephanie B. Byrne: In your model type, are you including lock-in HMO options relative to point-of-service options, are all the plan types lock-in HMO, or are they all point-of-service?

Mr. Scanlon: When you say lock-in HMO, I'm assuming you mean whether one must lock-in to a provider panel.

Ms. Byrne: Right.

Mr. Scanlon: Or do they have an option?

Ms. Byrne: Right. No opt-out versus an opt-out product. Are they both included in this?

Mr. Scanlon: Both were included, and we tested whether or not that made a difference by including a point-of-service dummy in our models. We found that the conclusions largely didn't differ based on that feature. In addition, by having that integrated variable, in which the more integrated plans would provide less flexibility in terms of choice of physician than the less integrated plans, we're controlling for some aspect of what you're calling a lock-in characteristic.

Mr. Chernew: The issue is not only are you locked in or not, but also how broad is the panel that you're locked into? That's why we used integration.

Mr. Scanlon: Let's get back to the results and the importance of informal information. Regarding the so-called ratings that included measures of physician availability, we found that enrollees were less likely to enroll in plans with a high percentage of physicians accepting new patients. Again, the firm in the construction of the relevant rating assumed that physicians that had open panels that were accepting more patients, would be a better choice. Plans were rewarded for these things in the construction of the ratings. Plans that had fewer physicians accepting new patients were penalized. When we related enrollment to choice, we found that consumers essentially enrolled in plans that had fewer physicians accepting new patients, which is counterintuitive to the way the ratings were constructed.

To explain this apparent contradiction, the analogy that we often use is that you don't see long lines at bad restaurants, and you might not expect to see long lines at the offices of bad physicians. In other words, people may prefer physicians who are not accepting new patients or that have long waiting times simply because they are the more popular physicians or the physicians that they actually want to go to.

Again, this gets at the relationship between informal information and the methodology used to construct these ratings. Interesting enough, we compared our findings with another study in which the researchers moved the physician availability rating out of the satisfaction measure and into another measure in the year that we studied. The result was a reversal on the satisfaction rating sign that we found.

Generally, we find that employees seem to know a lot, for better or for worse, informally, and that informal information has some impact, we believe, on the health plans they choose. We believe they like low prices, and that has also been emphasized by other research studies. Also, easy access and popular physicians are important, and we think this comes out in the relationship between choice and the surgical ratings and also the underlying HEDIS measures, particularly the percentage of physicians accepting new patients.

Informal information seems to dominate, and that's important for understanding these rating constructions, because much of the statistical relationship that we're seeing between these ratings and choice very likely is driven by correlations between those ratings and the unobserved variables that we can't control for in our study. If we were just to rely on statistical significance alone, I could stand up here and say we found a lot of statistical significance, but the problem is that statistical significance in these findings are counterintuitive, which leads us to look further and use this underlying HEDIS data to examine what really is going on. We think we were able to tease some of this out a little bit.

This begs the question: What is the value of performance measurement? Is performance measurement in the HEDIS data set more valuable for employers' internal use or for incenting plan performance? Another way of phrasing this is, given our findings, should we just scrap this whole performance measurement HEDIS data set initiative in providing ratings to consumers?

I don't think we should do that for two main reasons. One is because the science of performance measurement and rating construction is still relatively new, and we're working out the kinks to find out what consumers want and how best to provide consumers with the information. Two, even if consumers don't use this information, there is potential value in doing measurement and holding plans accountable. For example, this information may be useful for firms trying to determine which set of plans they're going to contract with, for feeding back this information to plans, and holding them accountable relative to their competition.

HEDIS has undergone some major changes, even in the past couple months. Let me just highlight a few. Up until now, HEDIS measures have not officially or formally been incorporated into the NCQA accreditation process. Beginning with 1999, in a program called Accreditation 1999, the actual scores of plans on HEDIS measures will now count towards whether a plan is or is not accredited, which is more of an impetus for plans to report this data and to report it accurately. The second change is that a survey called the Consumer Assessment of Health Plan Survey (CAHPS) has now been formally incorporated into the HEDIS data set. This has changed the satisfaction survey that was part of 3.0. The satisfaction or consumer measures that are included in HEDIS will now utilize the CAHPS data set.

From the Floor: I wanted you to go back to the effect of price. Could you go over what the effect of a $10 increment in price would be?

Mr. Scanlon: We modeled this using an odds ratio, which is the probability that one would choose a plan with a given characteristic relative to an otherwise

identical plan with the same characteristics, except for a difference in the one that you're interested in examining When we look at the impact of price, assuming that all the quality ratings and everything else are the same between plans, we found that employees were between 6% and 11% less likely to choose the plans that had the higher price. I give that as between 6% and 11% because we estimated different models: models with the ratings alone, models with the HEDIS measures alone, and models with combinations of both. It depends on what model you're looking at, but that's fairly consistent with some prior work that we looked at. Employees were about 10% less likely to choose a plan given the relevant change in price.

Mr. Michael W. Fedyna: This is a question for Denise. With the NCQA issuing three-year accreditations, does that have any impact on the requirements that you would impose to collect data for state or other purposes?

Ms. Love: I don't know because I don't know how many of our plans will choose to be accredited. The state of Utah and other states are not requiring accreditation because of the cost and possible backlash from the plans. If plans participate and become accredited, it will make my job easier. We will know what to require as far as a measure set. But, in light of the fact that we will probably have two to three leading plans being accredited and the rest not accredited, we still have to define our reporting requirements through regulation.

Mr. Thomas X. Lonergan: I have a question regarding the difficulty that the health plans had in collecting the data. Did you encounter a fair number of health plans, especially network models, that had not historically done a good job of collecting information, especially on services provided by capitated providers, where they don't have obviously an incentive to provide the information?

Ms. Love: Yes. Global fees and carve-outs took their toll and resulted in no prenatal-care measures for Medicaid at all because our Medicaid program is capitated. Immunizations were a nightmare and continue to be a troublesome measure because in Utah, like Nevada and other states, people opt-out and go to the public health clinic and the data are not fed back to the chart. As each plan comes into a doctor's office to do their abstractions, they have piles of charts, and it's almost too crowded in the doctors' offices.

Mr. Chernew: One of the big issues with the HEDIS accreditation has been to try and make it easier for plans to do this. You see a lot of the measures like cholecystectomy rate and so on, and they've moved away from using only administrative data. But they have this chart abstracting process, what they call the hybrid method, to actually get the data, which is intrusive and somewhat expensive.

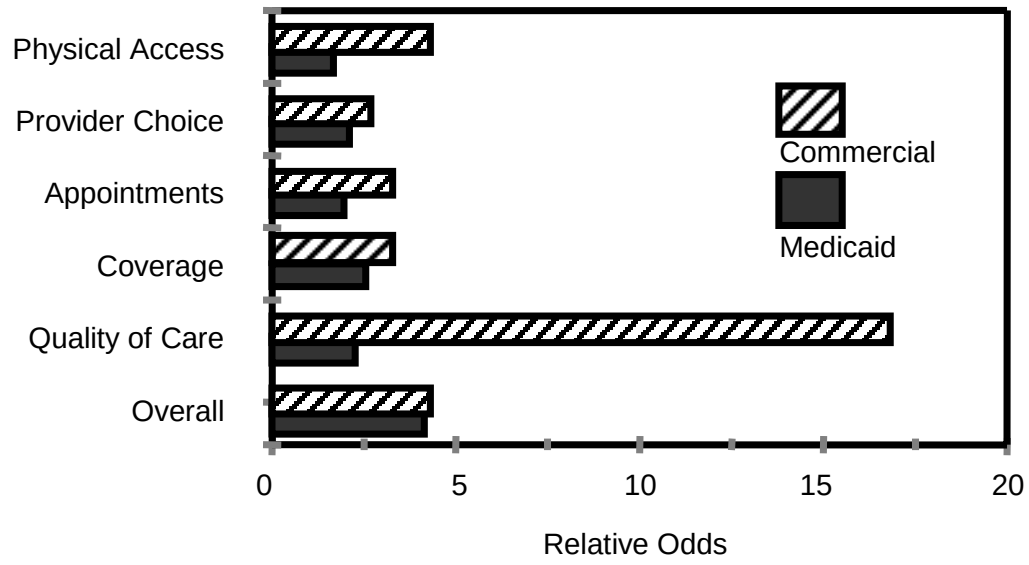
Mr. Scanlon: The other thing that's important to keep in mind, is that our studies are looking at HEDIS 3.0 and then 2.5 and 2.0. This is still a relatively new process, and those in the room who are part of health plans can attest to this. Anyone who works with this data knows that there's a lot of measurement error, and, as plans continue to report this data, they're going to get better at it. Also, given the Accreditation 1999 program, they're essentially required to report this in an audited form. Some things are happening that will make this more comparable. It'll be interesting to see if any of these results hold up once the data collection and reporting process improves.

Mr. Chernew: There was a SOA conference in Minneapolis in May 1998 that devoted an extraordinary amount of time to the process of plans—going through what you should do in January, in February, in March and in April, just like the old "chow-chow-chow" commercial. Tom was there, and he can put you in touch with people who can answer a lot of questions about the process of doing HEDIS, which is an industry in and of itself now.

Ms. Love: HEDIS will continue to evolve, but states, Medicaid agencies, and advocacy groups will continue to see a myriad of measures coming down. The new diabetic measure in HEDIS 3.0 is concerning me, because I believe we'll see smaller and smaller numbers as we try to get at those clinical measures in a plan population, and the data won't be useful.

CHART 1
ENROLLEE-LEVEL SATISFACTION

Odds of Being Satisfied with HMO
[Enrollees w/ no problems w/ access relative to those with at least one]



Note: Results are from logistic regression that included in addition, social location and health status variables.