RECORD OF SOCIETY OF ACTUARIES
1993 VOL. 19 NO. 4B

HEALTH RISK ADJUSTERS

Moderator: ALICE ROSENBLATT
Panelists: KENNETH S. AVNER
BRUCE D. BOWEN*
KATHLEEN JENNISON GOONAN†
Recorder: RONALD D. REHKAMP

- Risk assessment
- Risk adjustment
- Actuarial applications
- Application to New York small-group reform
- Rand 36 and other methods
- Medical management applications

MS. ALICE ROSENBLATT: We'll be talking about risk adjustment and outcomes measurement, and we have a panel of three speakers. Ken Avner will talk about the actuarial aspects of risk adjustment. Bruce Bowen will discuss the use of self-reported health status as a risk-adjustment method. Kate Jennison Goonan will discuss outcomes management and the role of severity adjustments in managed care. I'll give a brief overview of the paper prepared by an American Academy of Actuaries Work Group on risk adjustment. Ron Rehkamp will be our recorder.

I'd like to start by telling you about our three panel members. Ken Avner is currently the vice president of actuarial at Blue Cross/Blue Shield of Illinois. He's been there for seven years and was previously with TPF&C/Tillinghast. He has spent most of his actuarial career working on managed care.

Bruce Bowen is an executive consultant at Kaiser Foundation Health Plan in the corporate offices. Previously he was the assistant director of medical economics and statistics at Kaiser. Prior to that, he was vice president of research and planning at Blue Cross of California. He also has experience as a professor of research methods and statistical analysis at the University of Michigan and at Arizona State University.

Kate Jennison Goonan is currently medical director of health services evaluation at Blue Cross/Blue Shield in Massachusetts. She is an M.D. trained in internal medicine at Massachusetts General Hospital. She did a fellowship in health services research at Massachusetts General Hospital, and she is the former director of quality indicators at the Harvard Community Health Plan HMO.

Many of you have been hearing about risk adjusters over the past three days, and I'd like to publicly thank the members of the American Academy of Actuaries Work Group on Risk Adjustment, which I chaired. The members of that work group

* Mr. Bowen, not a member of the Society, is Executive Consultant of Kaiser Foundation Health Plan, Inc. in Oakland, California.

† Ms. Goonan, not a member of the Society, is Medical Director of Health Services Evaluation of Blue Cross/Blue Shield of Massachusetts in Cambridge, Massachusetts.

included seven actuaries, two health economists, and a physician. The seven actuaries were John Bertko, Norman Crocker, P. Anthony Hammond, David Helwig, Bruce Pyenson, Geoffrey Sandler, and myself. In addition, panelist Bruce Bowen worked with us as did Sue Palsbo, a health economist from the Group Health Association of America (GHAA), and Dr. Michael Moore from the Jackson Hole Group.

"Health Risk Assessment and Health Risk Adjustment: Crucial Elements in Effective Health Care Reform" is the title of the paper prepared by the American Academy of Actuaries Risk Adjustment Work Group. The first thing we did was define health risk assessment, and you should basically think of that as a model. Some people like to think of it as a black box that measures the deviation from an average; i.e., a deviation from an average expected cost for health care. Risk adjustment then uses that assessment model to (1) make monetary transfers between carriers, with the intent to reduce the effects of inadvertent or intentional risk selection so the carriers can compete on the basis of true efficiency and not on their ability to select risk; (2) compensate carriers fairly and equitably; (3) maintain consumer choice on that same basis; i.e., without the impact of risk selection; and (4) protect the financial soundness of the system. In addition, risk adjustment can be used to do provider profiling or to make provider payments.

Risk adjustment can be done through premium, or it can be done by adjusting premium. For example, many of us are accustomed to making risk adjustments through premium by doing community rating by class or experience rating. There could be a prospective risk adjustment in which each carrier comes up with the premium it would charge without adjustment, and then there is a payment made or a payment received through a risk-adjustment transfer mechanism. There can also be some kind of back-end retrospective adjustment or a combination, such as a prospective method with a settlement at the end that would be a retrospective method.

The Academy Work Group said that risk adjustment was definitely needed, and if rating proposals were closer to community rating than rates are, they would be needed even more. Right now it looks like the Clinton proposal includes pure community rating and thus carriers will be highly motivated to avoid high risk. In addition, a goal of maintaining consumer choice through premiums or contributions that are not influenced by risk selection requires risk adjustment. If you don't use the risk-adjustment method, then risk selection will influence what the consumers see as the price tag.

I'm not going to go over all of the various risk-assessment methods. There are many, many different methods. Many models have been proposed. Health care economists have been working on many different models; for example, diagnostic cost groups and ambulatory care groups. Bruce Bowen will talk about the Rand 36 health status questionnaire. The current Medicare system uses the adjusting average per capita cost (AAPCC). Many of us have experience with medical underwriting. Community rating by class is an example of using a demographic assessment, and the Robinson Luft method is a series of equations that performs the risk assessment.

The Academy Group also recommended that further research is definitely needed. In particular, cost-benefit analyses need to be done to compare the various risk

assessment tools that are currently available. We need to compare how they do in terms of accuracy and also compare the cost involved in performing the calculations. Many of these methods rely on enormous amounts of data and manipulation of that data. We don't believe that enough research has been done using these methods in combination with a reinsurance system that adjusts for the outlier claims. A Society of Actuaries group chaired by Bill Lane will perform this research.

Other research is needed to model the financial impact of using risk adjustment; i.e., what kind of impact is it going to have on the carriers? There can be some solvency concerns. Such concerns would argue for prospective adjustments. Most of the methods that have been discussed include both a prospective and a retrospective piece. For those of you who are thinking about setting rates in a health alliance environment, one of the things you are going to need to consider is the impact of the risk adjustment. If there is a big retrospective portion for the risk adjustments, that leaves a lot of uncertainty.

With all of the methods available, we did not think that we could select a particular method and designate that method as a recommendation by an Academy Task Force to use for health care reform. In particular, that concern applied to the methods that use prior history. But we did know that reform was moving along. If a solution was needed within the next 18 months, we said that a nonvoluntary reinsurance mechanism, such as a high-cost medical-condition system, should be considered. We were referring to a system similar to the New York system.

An example of how a risk-adjustment mechanism would work is included in the Academy paper. If you have questions, you can call one of the members of the task force or the Academy, and we'll be glad to talk about it in more detail. We are looking for volunteers to work on risk adjustment. In particular, if your company has data that could be used to do risk-adjustment studies, I would certainly welcome your call as would Bill Lane or any of the other task force members.

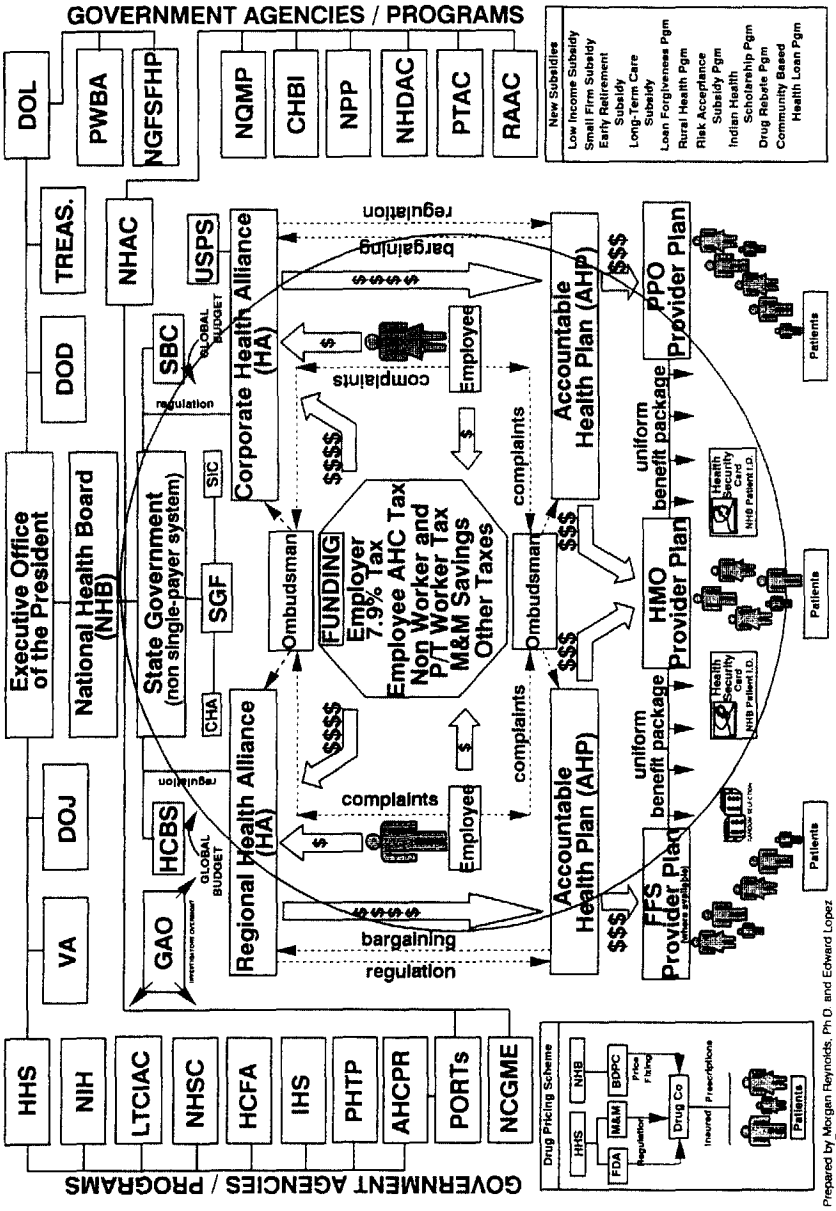
MR. KENNETH S. AVNER: This is an enormous topic that has an active research area. As time is limited, I will consider my task here to mention a number of the more important points from an actuarial perspective.

I want to start with a commercial for the Academy paper on risk adjusters. It should be the starting place for any actuary who wishes to follow or become involved in the continuing discussion of this topic.

My presentation will give a sampling of risk-adjustment methods available, with a focus on the ones in actual carrier use. Then I will outline approaches to evaluating the methods, giving some idea of the state of the art. Finally, like a good actuary, I will conclude with a couple of live examples with real numbers.

Chart 1 exhibits the entities envisioned in President Clinton's proposal. It was prepared by Dick Armey, Representative from Texas and Chairman of the House Republican Conference. It was published in a number of places, including the October 13, 1993 edition of *The Wall Street Journal*.

CHART 1
The Clinton Health Care Bureaucracy: "Simplicity" Defined



Before I get to the part in which we are interested, I want to share with you Armey's description of the National Health Board, located in a box in the middle near the top. He notes it is "a minor oversight board (according to Health and Human Services (HHS) Secretary Donna Shalala)" that regulates all aspects of the $900 billion health industry and oversees all government health care agencies and regulators." Minor?!

For us, we are interested in the responsibilities of an agency represented by a little box off on the right side, the Risk Adjustment Advisory Committee (RAAC). It "promulgates rules and regulations of the new national risk adjustment system, which adjusts premium payments to health plans to reflect the level of risk assumed for patients enrolled in comparison to the average population in the area."

Current practices: Let's consider some of the risk adjusters actually in wide current practice. The first example, AAPCC, is probably the most familiar to us and is probably the most researched. That is because it has been used for years in HMO Medicare risk contracts, and the Health Care Finance Administration (HCFA) has done demonstration projects and commissioned studies about it. Currently, it is in use by about 100 HMOs, covering about 1.3 million beneficiaries.

Let me outline how it is determined. You start with the U.S. per-capita costs for Medicare beneficiaries, which actually come in six flavors (separately for Part A and Part B for each of the aged, disabled, and end-stage renal disease (ESRD) classifications). It is adjusted to the county level based on five years of county-specific experience, if available. For each specific HMO, it is applied in 30 actuarial classes made up of age ranges and the sex, welfare status, and institutional status of the beneficiary.

What's the conventional evaluation of the AAPCC? After all, it has been in place for quite a while. First, it is generally agreed that it poorly predicts expenditures for individuals. That is why there was been so much research to improve it. I will show some statistics about that later. It explains about 1% of the variation of expenditures for individual beneficiaries.

Second, and it is hard to tell how much of this is real and how much of this is belly-aching by the HMO industry, it is believed to allow or encourage a significant selection bias in that there is nothing in the process that rewards you for caring for critically ill beneficiaries. This is related to the first point, but really is a distinct idea, which may become clearer when you see the retrospective components of adjustors I will talk about later. This concept of individual bias is a theme we see in risk adjustors over and over.

For those of us familiar with insurance, we know the answer should be that one cannot concentrate on the individuals, but rather we need to look at performance across large groups of people. With any individuals you may win or lose, but for the purpose of risk adjustment, it is only efficient to try to estimate the status of an entire group of people together. I admit there are arguments that each individual risk should be properly quantified. The real policy question is where to balance between efficiency and individual accuracy. Always remembering the limits of existing technology, we cannot perfectly predict a person's future utilization.

Before leaving AAPCC, I want to mention the view that, to some extent the limitations in the AAPCC have discouraged more participation in HMO Medicare risk contracting. On the other side, I have always thought that many HMOs have been cautious of entering a partnership relationship with the federal government. Maybe that deterred them from filing for risk contracts.

I would like to consider the second model only briefly. That model is the state reinsurance model, and it is discussed at length in the Academy paper for which I have already given a commercial.

This model arises from the NAIC Model Act on Small-Group Reform. For my purposes here, you can think of it as a standard reinsurance arrangement. There are issues of whether the reinsurance should be mandatory or voluntary and how contributions to the pool should be set, but I would like to surface two other concerns.

First, consider what I call "consistency problems with managed care." By this I mean the managed-care notion that it is the treatment modality itself that needs to be considered when estimating the cost of servicing an enrollee. We are not talking about two participants in the reinsurance pool in which the same treatment would be given by both, so the reinsurance is simply redistributing costs from random fluctuation or "selection bias." No, if the HMO gets the beneficiary, it will manage the resulting costs completely differently than the indemnity plan -- resulting in much lower costs for the HMO. If this is not appropriately recognized by the reinsurance mechanism, and it usually is not, the managed-care plan overpays for the risk adjustment, which then serves to subsidize the nonmanaged-care plans.

Similarly, most of these reinsurance plans, even when viewed between two similar plans, could be considered implicit subsidies of the inefficient by the efficient.

The second concern that I want to mention is the debate as to whether there is a need for an officially sanctioned reinsurance pool. The private-enterprise view would be that if reinsurance is desirable, the private sector is perfectly capable of providing it and will do a better job policing it than the public sector.

I should mention that Connecticut has probably enacted small-group reform most similar to what is described here.

My third example of current practices of risk adjustors concerns alternative (to fee-for-service) financing arrangements. Anybody who has worked in this area has dealt with adjustment methods such as stop-loss reinsurance or exceptional-condition payments. I would like to share with you one of my experiences, which is simple and understandable but shows how tricky these things can be.

In one part of our network we used to contract with a number of reasonably sized clinics for all physician services on a capitated basis. Basically, we computed a set of capitation rates for single and family participants and paid these same rates to all of the clinics. One year, one clinic appealed to us, complaining that this approach unfairly discriminated against it, because it services an area generally populated by families with a religious view that led to a large number of children per family.

So we did a study that compared our approach with using capitation per individual, age and sex adjusted. Sure enough, this clinic was correct. Its family size was significantly larger than anyone else's. But, we also found that as a whole, its population had younger adults. In fact, based on the individual capitations, it was actually slightly overpaid by our single and family approach. So in risk adjustment, especially prospectively, things are not always what you might guess.

I would like to move from a description of some current practices to a discussion of evaluation of risk adjustors. Without getting into the policy issues of how the adjustment process should actually work, let us focus on the classification schemes on which the adjustment can be based. What we need to do is classify individuals into categories that will then be used to determine which classification has the most need for health care resources. I have five criteria, the first four coming from the Academy paper, although I have changed the wording slightly to what I think are more standard and clearer descriptions.

The first criterion is accuracy or predictive power. What we would like are reasonable, homogeneous categories in terms of expenditures. But, as I mentioned before, there is a major question about what an appropriate expenditure is. We in the HMO business understand that carriers reimburse services that are treatments for certain illnesses; our question is how they know they are buying the correct services. That is an inherent problem when dealing with measurement of accuracy of an adjustor.

The second criterion is that it must be practical and understandable. I also include parenthetically that its administration should be efficient or low cost. Ideally, the data needed to classify individuals into adjustor categories have already been collected for other purposes or can be easily collected through the existing systems and processes. Ideally, the data would be easily available for audit and verification to protect against errors and fraud.

Consider the classification systems based on self-assessment and health status. Consider the diagnosis coding you get on your claims and how much effort continues to be put into trying to get it correct. How good are answers to English questions completed by individuals likely to be? Are you going to be comfortable adjusting premiums according to those answers? Because money will not flow to the participants based on their answers to these questions (as it does in claims payments), I think there is reason to be skeptical that the data will be sufficiently faultless. And, of course, very little of this data may be collected today.

My third criterion is timeliness and predictability. There is definitely a need for stability in the adjustors. Many dollars may be trading hands, and a carrier may not know how much it will end up with until after the period is over, after the services have already been rendered. Maybe this would be nothing new for insurance companies, but many HMOs would not be very comfortable with such an arrangement. That is not a good situation to be in if you are trying to provide the services within a budget. It would be contrary to good public policy to allow the adjustment to be a windfall or penalty after the fact, after the services were or should have been delivered.

The next one is no manipulation, by either the carrier, the provider, or the enrollee. Will the risk adjustors influence behavior? If a certain procedure is performed, will reimbursement change? Remember that when Medicare introduced the prospective payment system of diagnostic-related groups (DRGs), there were concerns that hospitals would encourage short-stay admissions and learn how to change coding to maximize reimbursement? Some of those concerns were well founded.

Finally, I come to reflect appropriate quality and care. I have talked about appropriate care already and how difficult that can be. But even more so is trying to reflect quality. There is quality from the standpoint of severity indexes and outcomes research. But what about the inherent quality of choice, even choices not taken? Isn't it worth something to know that if you had needed a referral you could have gone to an established "high-quality" provider for evaluation and treatment? We in the HMO industry generally consider this overrated, but it does cost more, and this is where I would include the issue of whether that extra cost should be allowed.

Now for the first set of numbers: the predictive power exhibit (Table 1). I use R^2 as a measure of the predictive power of the various adjustors' classification schemes. I do this because it is the one most commonly used in the literature and because most of us should be familiar with it. It is the same R^2 from multiple regression covered in the actuarial syllabus.

TABLE 1
Risk Adjustors
Evaluation
Predictive Power

Method	Estimate (R^2)
AAPCC	1%
CRG	up to 4%
DCG	4-16%
PACS	8.5%
ACGS	22%
Robinson-Luft	Groups of 1,000+
Maximum	15-20%

I do not really believe this is the correct metric to use in quantifying predictive power of risk adjustors for our purposes. R^2 puts a lot of emphasis on measuring individual point discrepancies, while as discussed before, we need to focus more on groups. There are other metrics used, but I am not really comfortable with them either. There is, for example, the predictive ratio, which is simply a ratio of what was predicted by the adjustor to be used, to what was actually used. There is also the product moment correlation, which is basically R^2 applied to a discreet variable instead of a continuous model. Anyway, more research is needed in this area.

As advertised, AAPCC shows up with 1%. Cost-related groups (CRGs), developed for Medicare were the first published method using clinical information and prior health services use. Developed by cluster analysis, it was found that prior expenditures account for more of the individual variable than any component of AAPCC. Included

also were assessments by physician panels of clusters where there is a high degree of physician discretion and the use of this information to build categories. This resulted in R^2 results of up to 4%.

Diagnostic cost groups (DCGs) can be thought of as CRGs taken another step. Also a result of Medicare pilot work, they have results of about 4%. In a study that incorporated continual updating, which is not practical for our purpose, they claimed to have reached R^2 of 16%.

The reason payment amounts for capitated systems (PACS) has an asterisk next to 8.5% is that the study I am quoting discarded the highest utilizing 1% of the population. I agree it is hard verifying data for that 1%, but there are many dollars up there, and it does not give a fair comparison if those data are summarily ignored.

Ambulatory care groups (ACGs) are quite well developed. They get a number as high as 22%, because they go beyond simply grouping by age, sex, and diagnostic codes. Unfortunately, they are only focused on ambulatory care, which is not where the big dollars are.

The Robinson-Luft approach was developed from the perspective of a large employer dealing with a multiple-choice situation. In those days, it was an issue of dealing fairly with an indemnity cover during the introduction of HMOs. Five years later, this idea looks good, being very similar to the risk-adjustor problem. But its research has been limited to large groups, so I could not find a study giving a comparison for this exhibit.

Finally, my bottom line is a reference to studies by the Rand Corporation and others, claiming that the highest one can hope to achieve prospectively is an R^2 in the 15-20% range. Back in my statistics class we could not claim to have an acceptable explanation of most of the variation with an R^2 that low. But the 20% is on an individual basis, and it is not clear that we really need to claim to have an explanation for most of the individual variation.

My first example is Arizona Medicaid. It is a demonstration project called the Arizona Health Care Cost Containment System (AHCCCS). Since 1982 it has been mandatory for all Arizona Medicaid beneficiaries other than Native Americans.

A very important reason why it appears to work so well is that roughly half the enrollees do not select their own carriers but instead have their carriers assigned randomly. So, in terms of getting homogeneous groups, half the people are assigned randomly. That really helps spread the risk.

What kind of risk adjustments are used? There are five that are readily identifiable: classifications, special payments, catastrophic reinsurance, specified-conditions reinsurance, and retrospective cost sharing.

The classifications are what you would expect in a Medicaid population: aid to families with dependent children (AFDC), social security income (SSI), Sixth Omnibus Budget Reconciliation Act (SOBRA), medically needy and medically indigent (MN/MI), and other children. The special payment provision is $4,000 for each SOBRA

delivery. The catastrophic reinsurance includes a threshold that varies by plan size, which seems to be good benefit design. The specified conditions reinsurance covers transplants and AIDS.

Retrospective cost sharing, as done in this plan is particularly interesting. It covers the needy or indigent population, which sometimes includes victims from accidents when they have spent their financial resources. In such a case, while still in the hospital, the person is assigned randomly to one of the plans. AHCCCS felt that such risk was too much for the carriers to accept without some sort of risk adjustment. It agreed to pay 50% of the cost for such individuals, and the responsibility of the carriers would be the remaining 50%, which could then be used in the catastrophic reinsurance stop loss. Because the plan guarantees six months of eligibility, the carriers believe that once they get the person out of the hospital, the financial exposure is reasonable.

What observations can we make about the Arizona experience? First, the risk adjustment and the entire process are very well developed. For example, to call its bidding process advanced probably understates it. It has complete, comprehensive data. Every encounter is recorded. And, the data is used. An actuary analyzes the carriers' data and estimates what bidding ranges it should expect. It doesn't share the ranges with the competing carriers, but it does share all the data with all bidders.

When the bids are received, they are reviewed individually. They request detail and justification. How many days are projected? How many visits? What are your projected unit costs? The accepted bids are in a tight band, usually less than 10% from top to bottom. That generally precludes serious selection driven by cost. But to me, the main reason the process works so well is because the 50% random assignment goes a long way toward homogeneous prospective populations for the competing plans.

The law affected business as of April 1993. It involves guaranteed issue, community rating, open enrollments, some limits on what the Blues could do (which is besides the purpose here), and preexisting-condition limitations.

What about risk adjusting? By regulation, this is done within seven geographical areas that cover the entire state, e.g., New York City, Buffalo, Utica. A demographic pooling fund is based on age, sex (for individual policies only), and whether a policy covers an individual or a family. So a family policy has the same weight, regardless of the sex of the members covered.

Finally, a specified medical conditions pool is funded on a per-capital basis. There are lump-sum payments for heart, liver, pulmonary, bone marrow, and pancreas transplants, and for neonates using more than 30 days of intensive care unit (ICU) care. Also, there are monthly pool payments for human immunodeficiency virus (HIV) disease and certain conditions requiring ventilator care. The lump-sum payments vary from $56,000 to $136,000, and the monthly payments range from $2,000 to $13,000.

The experience to date is sketchy. I have just gotten the numbers that are shown in Table 2. The risk sharing works on a quarterly cycle, and the second cycle has just

been completed. The first thing the state has discovered is that it needs to rescale the entire population, which is done by resetting the regional demographic factors. I've shown New York City and Buffalo. Originally they were both estimated at 1.03. But now they have been revised, so that the City is 1.08 and Buffalo is 1.04.

TABLE 2
Risk Adjustor Example
New York Experience to Date

Regional Demographic Factors		
City	Original	Revised
New York	1.03	1.08
Buffalo	1.03	1.04

Carrier Demographic Factors			
City	All Carriers		Large Carriers Percentage Point Range
	High	Low	
New York	1.14	0.79	17
Buffalo	1.47	0.82	14

How material is this? I was told that the dollars traded for the first quarter are about $10 million, which doesn't seem like very much. But based on the new factor, the estimate was that the second quarter would have a number at least triple and maybe quadruple that. Obviously, a transitional issue was not appreciated.

What is probably most interesting to us is the individual carrier demographic factors. Again, I've included only New York City and Buffalo. In New York, the factors range from 0.79 to 1.14. In Buffalo, the numbers go from 0.82 to 1.47. I don't know who that is, but that is one heck of a factor in Buffalo: 1.47! That is definitely a higher utilizing population, at least according to the risk-adjustment system.

Concerned that this was influenced by very small carriers' enrollments, I asked for a range determined by the top five carriers in each region. These factors were fairly well centered around the largest regional carrier. In New York City, this large-carrier range is 17 points; in Buffalo, it is 14 points. That still seems surprisingly large.

New York is in a similar situation. There's a catastrophic reinsurance with a threshold that actually varies by the plan size. Specific conditions' reinsurance for transplants and AIDS has a very interesting retrospective cost sharing. It covers the medically needy and the indigent population. Sometimes an accident occurs and somebody is medically needy. When in the hospital, that person is then assigned to one of the plans on a random basis. The carriers thought that was actually too much to take without some kind of risk adjustment, so they said they would pay 50% of the costs on that and the other 50% could be used against the stop-loss, if they want to run through that procedure. Actually, on the whole program there's a six-month guarantee of eligibility, so the carriers believe that once you get that person out of the hospital, you're usually in good shape.

The observations on Arizona are that it's well developed. It's already in the bidding process. I call that an understatement. These guys have complete, comprehensive data. They have every encounter. They use the data. A consulting actuary comes in and gives them ranges. "These are the numbers you should expect." They don't tell anybody the ranges, but they do give all the data to all the bidders and they tell them to do the analysis. When the bids come in, they sit down with all the bidders and they say, "Why did you come up with this number?" The bidders have to go through it in detail. "How many days do you think you're going to use? How many visits are you going to use? How much are you paying for that?" They go through it in gory detail.

There are tight capitation bands. The accepted are on a very tight band. It's not unusual to see the bids top the bottom 5%, maybe 10%. So that really wipes out that kind of selection in terms of cost. It has generally avoided homogeneous adverse selection and has been, instead, fairly homogeneous probably because of the 50% random assignment.

Now to my favorite, New York. I don't remember when the law was passed, I think in July or August 1992, but Regulations 145 and 146 came out in December 1992. It's individual small-group health insurance reform, and it actually includes Medicare supplement. I'm not going to talk about the Medicare supplement pool, which runs separately from this pool, although it's very similar to it. It was effective April 1993.

It involves guaranteed issue of community rating, open enrollments, some limits on what the Blues can do (but that's beside the purpose here), and preexisting condition limitations. What do they use for risk adjusting? By regulation, geographical areas. There are seven of them, including New York, Buffalo, Utica, and various places through the state. It covers the entire state. A demographic pooling fund is based on age, sex, if it's an individual policy, and then there's a single family. So family doesn't look at the sex, it just has family.

Then, finally, a specified medical conditions pool is funded on a per-capita basis. (I think the Academy paper talks about doing things this way.) There are lump-sum payments for transplants and neonates, and there are monthly payments for HIV positives and ventilator care (see Tables 3 and 4). Table 3 shows the lump-sum payments. You can see the size, about $80,000. Table 4 shows the monthly payment: $2,000-13,000.

The regional demographic factors in Table 2 are wrong. New York and Buffalo had both had 1.03. New York came in at 1.08 and Buffalo came in at 1.04. They were off by a significant amount in New York. I was told that the dollars being traded for the first quarter were about $10 million, which doesn't seem very much. But for the second quarter they figured that was going to at least triple and maybe quadruple. So that raises a whole transitional issue that people should think about.

TABLE 3

Medical Condition	Course of Medical Care	Pool Payment
Irreversible, progressive liver disease	Liver transplantation	$80,000
Irreversible, progressive heart disease	Heart transplantation	76,000
Irreversible, progressive pancreas disease	Pancreas transplantation	56,000
Irreversible, progressive lung disease	Pulmonary transplantation	136,000
Severe aplastic anemia	Bone marrow transplantation	120,000
Acute leukemia	Bone marrow transplantation	120,000
Chronic myelogenous leukemia (CML) in controlled (not blastic) phase	Bone marrow transplantation	120,000
Neuroblastoma, Stage III or Stage IV in complete remission	Bone marrow transplantation	120,000
Myelodysplastic syndrome	Bone marrow transplantation	120,000
Hodgkins disease	Bone marrow transplantation	120,000
NonHodgkins lymphoma	Bone marrow transplantation	120,000
Severe combined immune deficiencies (SCID)	Bone marrow transplantation	120,000
Wiskott-Aldrich Syndrome	Bone marrow transplantation	120,000
Other condition, approved by the superintendent in clinical situations where bone marrow transplantation has proven to be effective	Bone marrow transplantation	120,000
Neonate with birth weight of less than 1,500 grams	ICU care for more than 30 days	96,000

Source: New York State Insurance Regulation 146

TABLE 4

Medical Condition	Monthly Payment
HIV disease; the CD4 count is below 50 on 2 consecutive tests	$2,000
ALS leading to ventilator dependency for more than 30 days	13,000
Severe trauma leading to ventilator dependency for more than 30 days	13,000
Severe muscular dystrophy leading to ventilator dependency for more than 30 days	13,000

Source: New York State Insurance Regulation 146

Finally, for carrier demographic factors, in New York it went from 0.79 to 1.14, and in Buffalo it went from 0.82 to 1.47. I don't know who that is, but that was one heck of a factor in Buffalo, 1.47. That is a higher utilizing population, at least it is according to the risk-adjustment system that we're doing. I specifically asked for large carriers because it could be Podunk Life there at 1.42, which doesn't have any enrollment. I specifically asked for the large carriers, and they were very careful about saying that I could not share this data; this is not public. It's easy to pick out the winner and the loser under the system. So we negotiated a range that is fairly well centered around the number 1. In New York, the range from the top to the bottom on the large carriers is 17 points; in Buffalo it's 14 points.

MR. BRUCE D. BOWEN: I think it's useful to start by defining risk and selection bias. I have to remind my economist friends of these things, but I probably don't have to tell you. Basically, we're talking about the relative expected costs of a group of people. We're not talking about the underlying health risk of the population; that is epidemiological risk. We might like to do that, and I'm sure when Kay gets up here she's going to tell you how much she would like to do that, and the question is how and when we will ever be able to do that kind of thing.

My criteria are a little different than what you saw earlier, but not by very much. Many of the words are different. First is this accuracy. I'm an economist and I think of this in minimum variance kinds of terms. Most people think that the most important thing is that they be accurate, but I don't think it's the most important thing. Actually, I work for a very large HMO. We have very large numbers of people and many of you deal with that. We can tolerate a great deal of error as long as it's random.

What we're really worried about is bias. We want to make sure that it isn't always wrong in the same direction against the same group of people or the same group of carriers or whatever. This stuff about getting it exactly right, well, if you got it perfect then you wouldn't need to worry about how the error was distributed. But if it's not perfect, as it never seems to be, you have to worry about whether it's biased or not. And we're worried about bias as to efficiency of health plans among other things. If you're talking about a managed-care environment, you don't want to, in fact, devise a risk-adjustment scheme that rewards people who are inefficient. And that's one of the problems we see with some of the more traditional reinsurance kinds of pools that are based on high-cost cases.

There's a whole class of methodologies, and I like to divide them up into basically two kinds: those that deal strictly with demographics and those that deal with health status in some way, shape, or form. The demographic-based models include averages by risk classes. Risk classes are defined in some way and there are many different ways to do that. The AAPCC is one. New York has one. There are also some very complex models. Luft and Robinson has a six-equation multipart, conditional probability model that has legitimate equations in it and it's really fancy. But it uses the same kind of variables that are used in classification models and it doesn't buy you a whole lot more than these other kinds of models. You get some real benefit when you go to health status. There are many different ways to measure it. I'm sure you're all familiar with what I consider to be the worst possible way of measuring it, and that is last year's costs. That is a very good indicator, there is a lot

of accuracy, but it is very biased. If too much money was spent on someone last year, there's a good chance too much money will be spent on them next year. It isn't a very good method.

I'd like to talk about self-reported, health-status measures and, particularly, the Rand 36. It has some other names, by the way: a medical outcome study (MOS) SS36, Short Form 36, and a number of different things. It was originally developed by Rand as an outcome measure. It was developed for those people who couldn't come in for their final evaluation in the Rand health insurance experiment. So it was used to try to get at what kinds of outcome problems there were in the sample. There are many other measures that measure health status. Mathematical Policy Research developed some questions from the Medicare population in a recent report to the HCFA. It had, I think, three or four questions, a little instrument, and it asked people how well they felt and whether they had cancer or heart disease. It was sort of a combination of some Rand-36-type questions along with the old familiar underwriting questions what I call the, "have you ever? questionnaire." Have you ever had, or has a doctor ever told you that you had any of the following 342 conditions?

Rand 36 is basically 36 questions asking people about things in these areas. General health: Are you feeling healthy, better or worse than last year, a lot better, a little better? Physical functioning: Can you walk around the block, can you walk up the stairs, can you carry your groceries? It would be fairly easy to check up on these things if you wanted to audit this. Social functioning: Are you so sick you haven't been able to go to church or visit your grandmother? Role functioning: Do you need help in brushing your teeth? Mental health, energy/fatigue, and some pain questions are also asked. The answers are then aggregated to an overall score.

There are some problems with self-reported health status, and we should talk about those. Then we'll talk about some places where we've used some of these measures to try to estimate what's happening with risks. Regarding response-rate issues, one of the things that happens is you send out a questionnaire and ask people how they feel. You don't get them all back, so what do you do about that problem? Are those people healthier or sicker than the people who sent them in? Those people who don't return them are healthier.

We had a big controversy among those of us who are doing all this. Normally when you do survey research you don't have the luxury of knowing the value of your dependent variable. You're trying to estimate it with your questionnaire. Well, this questionnaire is just trying to measure the independent variables. We already knew how much the people cost. We were just trying to find out how healthy they were or how healthy they said they were. So we know exactly what the response bias is.

We knew they knew they were healthy. We as a health insurer sent this questionnaire out to people. We can speculate why people did or didn't send it back, but we really don't know. We didn't do any follow-up research with the nonrespondents to try to find out why. We prodded them a couple times, and if they didn't send it in, we went on to other things. But there are many good reasons that people can't respond. They're disabled, they're institutionalized, they don't speak the language, they're not mentally competent, and they're too young. So there are a lot of problems with self-reported health status.

I'm a little worried about self-reported health status as being gameable, at least in the long term. After people figure out what this is all about and how it works, it's possible to lie about your health status and alter the amount of money that your health plan might be given by the risk adjusters. I'm not worried about that in the short run, but in the longer term it could pose a problem.

Well, we've got some advantages. It's exactly the same measure for all health plans. This reduces the possibility that there is bias by health plans. Whenever you're asking for data to be submitted by plans, there are different reporting mechanisms, different data quality, different ways of dealing with this or dealing with that. It avoids those kinds of problems because you're collecting the data uniformly from everyone. It's relatively easy and inexpensive to administer. We did some large-scale studies with it, and got the costs down to less than $5 per person. If my memory serves me, it was about $2.40 per person to administer the questionnaire. So it's relatively cheap. I've been personally involved in a couple of studies using the Rand 36, or a subset of it, for risk assessment. That is the part of risk adjustment you do before you move the money around. That is where you calculate how much money you might move if you were to do it, only you don't actually do it. But we did a large-scale study at the Center for Health Research for part of the Northwest Region of Kaiser Permanente where we administered the entire survey, the entire 36 items. We also administered the underwriting questionnaire, a common one that we, in fact, used in some of our individual business, the "have you ever? questions." It turns out the Rand 36 predicted costs much better than the underwriting questionnaire did.

We currently have a project going with the Bay Area Business Group on Health, and I'm going to show you some results from that in a moment. Many researchers did this exotic demographic model and are involved in the process. Some people from Kaiser Permanente, including myself, are working on it, as are some consulting actuaries from Coopers and Lybrand. Nine Bay-area employers agreed to at least participate up to the point of looking at what the numbers are. They haven't agreed to move any money yet, so we're collecting some data. We've got some interesting results.

From the CHR Study we basically found considerable internal validity on those scales. We looked at whether they make sense. We ran some factor analyses and tried to find out if those scales really predicted costs in that area, and they, in fact, did. The estimates are relatively stable. We got prediction errors so let's talk about what prediction errors are and what I mean by them.

R^2 is just not the appropriate measure for what we're looking at. R^2 measures how well we predict each person. We need to find out how much more this group of people who chose this plan will cost than this group of people who chose that health plan. How well we do that is the measure of how good our models are, not how well we predicted what Bruce Bowen's costs are for next year, which is what our squared measure is. These people who are walking around saying "the AAPCC has an R^2 of 1%, and an ACG has an R^2 of 22% and, therefore, it's 22 times better" are crazy. That just doesn't make any sense. The question is, given a group of people, how well do you get the total amount right for that group of people versus for this group of people? That's what we have to look at. The prediction errors using Rand 36 in contrived groups that we constructed were predicting much better than 3%.

Worst cases were getting 3% error, within 3% of the total costs. So we're talking about some reasonable predictions using that.

From the Bay Area Business Group on Health study, we found that everybody expected this risk problem. These are nine of the largest employers in the San Francisco Bay area now. Remember, this is Chevron, Safeway, Bank of America. These are huge companies, so we're talking big groups here not small groups. But I think everybody expected these risk differences to be large among the health plans within these companies and they turned out to be small instead. I have been doing this for a while so I wasn't that surprised, but many corporate people were surprised that the risk differences weren't huge. When I saw the number that Ken just gave (1.47), I find it hard to believe, given what I've seen in the real world. That's not a believable number to me. I guess anything's possible.

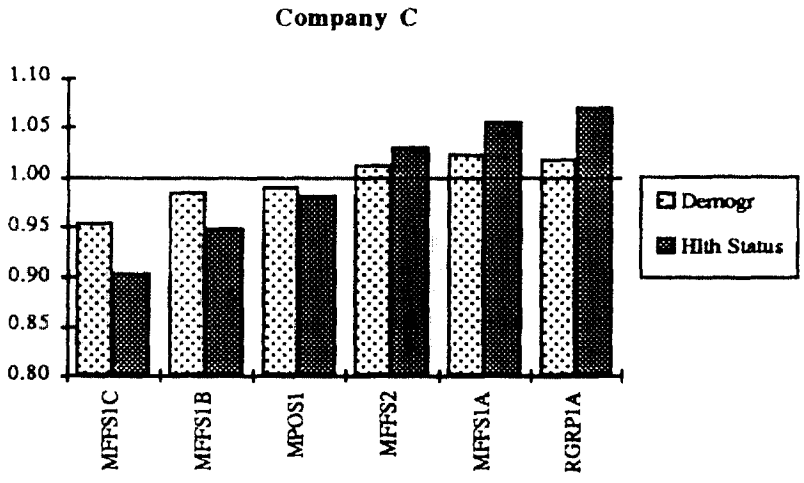
Chart 2 is the sample for a particular company. The left bar is the relative risk of this fee-for-service (FFS) plan. The relative risk or MFFS1C is 0.9 compared to the average risk of all the plans in that company. And that's being measured by a demographic model, age, sex model, average cost per cell. If you look at just these kinds of bars, you see there isn't very much difference between the ones with high risk and the ones with low risk. They're very compressed. When we add our health-status model by using a subset of the Rand 36, actually, only 13 questions out of the 36, the risk is spread out more. We think (and it's think at this point) because we haven't gotten the rest of the data that we need to answer this for sure. We believe that is because the health-status measure is able to better discriminate. We're getting too much averaging in those demographic health cells; that is, there's a lot of variance within any one cell and some more of that variance is explained when you can get self-reported health status in there. So there is more risk to explain than what the demographic models are capable of delivering.

Self-reported health status is not a perfect measure, far from it. If it is useful at all, it is probably useful in the short-term, until we can get some better things up and running. I'm looking forward to those population-based, epidemiological, underlying health risk kinds of models that Health Services Research people are working on. It is possible to game it in the long run, and it may be difficult to use for some subpopulations. We have no calibration for it on the uninsured, for example. For an insured person now, we've done a lot of these studies. We know what it means when they say they're healthy: they're in excellent health versus they're in very poor health. We know how much that's worth in dollars. We have data sets to calibrate it on. For uninsured, homeless people, whomever health care reform might encompass, we don't have calibration. It is unbiased to use and that's its big advantage.

MS. KATHLEEN JENNISON GOONAN: I have the daunting task to teach you everything you wanted to know about medical outcomes management in 20 minutes. Actually, I'd like to make every effort to "whet your appetite." How many of you currently have an outcomes measurement or management activity within your organization? Just two or three. I assume that most of you work in health care insurance. I can tell you that most managed-care organizations are scrambling to put outcomes measurement of some type in place, and they're under tremendous pressure to do so. I'm going to share with you a little bit about what that pressure is all about. But if there's one thing I can hope to convey to you it's that your

colleagues who are medical practitioners and health-services researchers are struggling with these issues and they need your help.

CHART 2
Health Risk Assessment
BBGH Results



Typically, I speak to clinician audiences on behalf of those who are very quantitatively oriented and attempt to convince clinicians that it's worth their while to understand population-based concerns. Clinicians are trained to focus on individuals. Epidemiological thinking and planning for populations does not come naturally. You think about populations, not individuals. Clinicians and actuaries have much to learn from each other. Your colleagues who are trying to actually implement medical management of outcomes could greatly benefit from your knowledge and perspective. I'm going to walk you through the "current events" in medical-outcomes research and how physician managers are attempting to use outcomes measurement in managed-care organizations to shape practice patterns of practitioners.

These are the three basic goals of outcomes management. We're trying to find valid, reliable, and meaningful measures of outcome. You will hear the word *outcome* used to refer to anything that reflects accurate measurement of performance. I think probably the better term would be *performance measurement* rather than *outcomes measurement*, because you'll typically find that people are measuring absolutely anything that they can find that is fair to measure, that can be measured accurately, and that has some bearing on information about what practitioners do and whether they're doing the right things in the right way. So *outcomes* has become a very generic term. In fact, it doesn't mean outcomes most of the time. Most of the time it means performance and whether or not health care providers are practicing consistently with identified standards. For example, there are national standards that

recommend that women over age 50 have annual screening mammograms. This is a standard we can measure health plans against. The percentage of women over age 50 in a health plan who have had a mammogram within the last two years is a measure of quality that can be measured. Defining valid and reliable measures of performance is what people are referring to when they talk about outcomes.

Using the information to improve performance is something all health plans are struggling with. What does it mean to have information about performance? How do we improve performance? I can ask a group of pediatricians what proportion of their patients is under age two have had their immunizations according to the American Academy of Pediatric guidelines. I will get one of two responses. One response I often get is that all will raise their hands. "All the children under age two in my practice have been immunized on time." Then I ask, How do you know? Clinicians will say, "I know the guidelines, I immunize them, that's what I do." But if I measure it and I find that 65%, or 50%, or 90%, it varies all over the map, have been immunized, they then will say they have never had that information before.

So having information about how well a plan, a group practice, or an individual is doing is not information that physicians have had before. In fact, the only physicians who know this kind of information about themselves are the people who, for instance, recently became board certified in surgery. In the past, they kept their own logs. When I was in residency training in the early 1980s, surgical residents kept logs. They recorded the number of surgeries they did, what happened to the patients, and any complications that occurred. Then when they sat for their boards, they discussed any complications they had and the patients' outcomes. The average physician doesn't know his or her rate of complication in any kind of quantitative way. So what we're doing, when we talk about measuring outcomes and using this information to improve, is asking clinicians to function in a way that's entirely new to them. Yet this is the fundamental purpose of outcomes measurement. Our first goal here is to have information that we can use for accountability. We can justify the cost of health care and explain why someone should buy insurance from us. This accountability is very important. We must do this. Clinicians understand this well now. I think most of the medical profession accepts that. They may not like it, but they accept it. But the notion then that they also try and use this information for self-improvement is even more foreign at this point.

Optimizing health status and function of population sounds great. However, average organizations do not know that our purpose is to improve health status. The purpose has been to treat disease, which is quite different. Here's an example: consider a community hospital that realizes one of the conditions it sees a lot of is children with head injuries. If it runs a public education campaign to get kids in helmets and reduce the head-injury rate, it is going to lose money. Now it may or may not choose to undertake that particular public service. A community hospital in Idaho has done just this. It did a very laudable job of reducing head-injury rates and complications from head injuries, which has significant review implications. Its ER use rate is down and it now has to contend with what to do to fill the gap. So there are many obstacles to adopting a community health-status focus.

To understand the field of risk adjustment for outcomes measurement, the major author to read is Lisa Iezzoni, MD. As she points out, the goal of risk adjustment is

to control for the confounding influence of patient severity in comparisons of outcome that might be related to severity." Outcomes include mortality, morbidity, function, the kinds of things Bruce was describing, cost, and satisfaction. These are all outcomes of care and are all things that are going to be confounded by severity.

We also face a serious problem, which is that there is no definition of severity. What does severity mean? Actuaries talk about severity as it relates to predicting resource consumption. This is the focus of your work. But severity also must be taken into account when we want to predict clinical outcomes.

Clinical outcomes generally fall into three categories. There are clinical outcomes such as unnecessary mortality, complications, and adverse events that are relatively rare events. Complications happen to a subset of people. People have a probability or a risk for adverse events, which varies depending on their underlying disease and severity. Then there are functional outcomes, such as the ability to work, or walk, or care for oneself. These are more experimental. We don't yet know how to use them for continuous improvement. There are some high-visibility pilot projects nationally, including ones being performed by the American Group Practice Association and the Managed Health Care Association (MHCA), an employer-based group. The MHCA project teams up health plans and employers to collaboratively figure out how to use functional outcomes to manage quality. Right now these measures are highly experimental and quite expensive. We don't know how we're going to use this information to improve care or to account for the quality of care.

I will describe a couple of examples that will highlight why this is so complex. Take, for instance, a patient who's 50 years old, who is otherwise healthy, but who has a lung nodule on a chest X-ray. This person is admitted to the hospital to diagnose whether that nodule is cancer. This is a perfectly healthy person; there is no co-morbidity. We go through a long diagnostic evaluation. Cancer is diagnosed. It is a resectable cancer. It could be successfully cured. The patient undergoes a surgical intervention and treatment. Keep in mind that this patient was symptom free when the nodule was found. This is a young, healthy person with a low risk of mortality and nationally significant health care resources. In fact, a study done in the mid-1980s showed that hospitalization for a person like this might cost $10,000-11,000 and maybe only have a 5% risk of mortality.

Compare this patient with another patient. Consider a 75-year-old with emphysema, a lifelong smoker with metastatic lung cancer. The only option is palliative treatment. In this situation, this is the last hospitalization. It may only cost $3,000. That person may die quickly, and we may do very little because it is futile. So one patient had a very high risk of mortality as opposed to another patient who had a very low risk of mortality, yet they had very different cost risks. One of the things you need to understand is that the models being used by people trying to do outcomes management and measurement on the clinical side need to predict much more than cost. We need to predict risk of mortality and risk of morbidity, and now purchasers want us to measure functional outcome. We need to factor in risk of some sort of functional status at the end of all treatment. Clinicians are trying to figure out how to apply outcomes orientation, risk of mortality, risk of true function, patient satisfaction, and so forth. Often they use models developed for your purposes to predict resource

use but use them to predict clinical outcome. As you know, this is very problematic. So there's a lot of confusion and frustration.

The other point that Bruce and Ken made but I want to reiterate is that all these models assume care being delivered now is appropriate. I hope you all are aware of the scientific literature available now that demonstrates significant variability in the rates of elective surgical procedures which vary by geographic area. There is little scientific explanation for this. There is also a tremendous urgency to answer questions about appropriate timing and indications for hysterectomy, prostatectomy, and other procedures. In fact, there is important developmental work going on now to involve patients in decision making that may change these rates. It entails providing patients with information about their risk of various outcome.

Let me just give you another colorful example that highlights how complex these things are. Rates of prostatectomy vary dramatically by geographic area with no clear-cut explanation, based on the indications for the procedure. There is a move to try to help patients participate when the decision is being made to have or not have surgery. This work is led by John Wennberg, M.D. and Al Mulley, M.D. One of the concerns within the profession is that patients are not fully informed of the probability of various outcomes. Some procedures would not be performed if the patients had full information about outcomes and participated in the decisions. To develop these decision-making tools, patient focus groups using market research techniques were held with men who were recommended for prostatectomy. Their input regarding their values, lifestyle priorities, and understanding of their disease was used to develop an interactive videotape to help patients decide whether to undergo a prostatectomy. In these patient focus groups, researchers found that sometimes as much as a third of the group thought that they were being recommended for prostatectomy because they had cancer. But the reason they were being recommended was because of benign conditions.

I'll give you another example. A successful prostatectomy usually results in retrograde ejaculation. The majority of the men in these focus groups did not know that retrograde ejaculation is a common outcome of a successful prostatectomy. To urologic surgeons, it is an unfortunate by-product of the surgery, not a complication. But to most men, this is a serious consequence and potentially a reason to decline surgery. The patient focus groups revealed fundamental miscommunication between patients and their physicians who have recommended prostatectomies. Patients did not have an accurate perception of their indications for surgery, and they did not understand what the consequences of the surgery were going to be. This example is more dramatic than others, perhaps, but I could tell you story after story of this level of dissonance between what patients need and want and their understanding of treatments and what the profession may or may not recommend. The challenge of considering patient preferences in treatment decisions augments the lack of specific guidelines for appropriate indications. When we consider models to predict outcome, we need to think of which outcome and what it means, base models on historic resource use and practice patterns, which vary dramatically and do not consider patient preferences.

I'll give one of the challenges we're up against. Length of stay for practically every DRG is dramatically shorter on the West Coast than on the East Coast. There is a

significant debate about whether this results from real differences in medical care or whether it reflects cost shifting from inpatient to outpatient settings in the West. Some would argue that California has gotten inpatient care out of the hospital and into the ambulatory setting. They further argue that the cost of hospitalization in California has gone up because much more is being done each day in the hospital. I have yet to see persuasive evidence to resolve this debate. Is managed care really further along in California? Are they actually managing resources better, or is it that they cost shift it from the inpatient to the outpatient settings? If we cannot yet answer these questions about utilization and managing care, how can we move to a more sophisticated level involving patient outcome preferences?

Now let's review some of the basic elements of outcomes measurement. Lisa Iezzoni identifies the five Ds (death, disease, disability, discomfort, and dissatisfaction) as the elements we must factor into our outcome models above and beyond dollars. Death has been used as an outcome fairly successfully, but its use is limited because it's not a highly sensitive measure of whether medical practice is good or not good. Second, she identifies disease, meaning the status of the patients when they come into the hospital or begin an episode of illness and what the status of the disease is when they leave. Third is disability, which refers to measures of whether people walk and take care of themselves. SF36 is an instrument to measure functional disability. The problem here is that SF36 is a generic tool. For example, consider people who have carpal tunnel syndrome and undergo surgery. SF36 has generic questions that ask about general physical functioning. It is too general to measure specific outcome of hand surgery for carpal tunnel syndrome. Most of the organizations using instruments like the SF36 for medical management are finding they have to augment it with diagnosis-specific measures. They have to add questions about the actual condition under study.

The final outcomes of interest are discomfort and dissatisfaction. You can see how complex outcome measurement and the need for risk adjustment can be. For example, consider the challenge of measuring outcomes of mental health care. For mental-health patients, good treatment may create dissatisfaction and unhappiness in the short run. Treatment changes the patient's views of themselves and the world around them. In this case, dissatisfaction is a good outcome.

Now let me just give you a couple of other examples of challenges we are tying to resolve. You probably know the difference between sensitivity and specificity. Sensitivity is the way we describe whether a diagnostic test really gives us information about a disease. What we're really trying to do here is use quality measurement as a diagnostic test for substandard medical care. We're trying to make judgments with these outcomes instruments about whether the medical care is good or bad. Presumably we're going to do something with the results of our measures. Either we're going to modify the clinicians' behavior, if we're a health plan, or we're going to take our employees out of the health plan and put them in another health plan, if we're a purchaser (an employer). Here we're trying to use risk-adjusted outcome tools to identify whether plans are truly providing bad care. We want to use risk-adjusted outcome measures as a diagnostic test for the quality of care being delivered in plans. As you can see, it is very complex, and much developmental work is needed.

Let me just give you a flavor for what some of the different interest groups in this field are doing to overcome these challenges. Various organizations are imposing performance measurements of all kinds on health plans and hospitals. The Joint Commission on the Accreditation of Health-care Organizations (JCAHO) has spent the last five years developing performance measurements that are going to be required of hospitals by 1996. They are well-designed measures. It has put a lot of thought and effort into making them valid and reliable. Risk adjustment has been dealt with in the design of the measures, to the extent possible. These will be required of hospitals, reported publicly, and will feed into a national database by 1996. A lot of effort has been made, and hospitals are currently gearing up to collect the information for these requirements. They're looking at clinical outcomes such as deaths and complications that shouldn't occur – the unusual negative events one would hope would not occur.

The National Committee on Quality Assurance (NCQA) and the health plan employer data information set (HEDIS) have the kinds of requirements that health plans are up against. The NCQA accredits HMOs. Every HMO and health plan in the country is under pressure from purchasers to meet the NCQA requirements. Many corporations that send requests for proposals to health plans expect NCQA or JCAHO accreditation.

HEDIS is the set of 60 measures that address financial, quality, access, satisfaction, and utilization performance. They measure rates such as the proportion of children in the health plan under age two who have been immunized, the cesarean section rate, the low-birth-weight baby rates, the mammography screening rate, and so on. A variety of measures like this are very simple measures. They are the lowest common denominator, those indicators that experts agree have some validity and reliability. They are designed not to require risk adjustment, meaning that they should be applicable to any population. However, for example, the low-birth-weight-baby rate will vary with the risk of the demographics of the population within a particular plan. Some populations have a higher frequency of low-birth-weight babies than others. The issues of risk adjustment have not been dealt with, and yet this is going to become some form of a national report card for health plans. In the next year, 22 health plans from around the country are going to be piloting these measures, trying to define how to perform these measures and how they can be used. The issues of risk adjustment will come up, and it will be interesting to see how they are dealt with.

Let me just mention something about the regional business plans. Bruce mentioned the one in the Bay Area. The premier example, which you may have heard of, is the Cleveland Area Choice Program. It has developed its own predictive models for mortality rates, and 35 hospitals have voluntarily started collecting not claims-based information, not secondary data, but primary data from the medical record. It has developed its own model for predicting mortality among general surgical and medical patients, and it uses the Apache System for ICU patients. It now releases documents and they are available in the public libraries and corner pharmacies. You can buy a small pamphlet that shows hospitals that perform above expected, at expected, or below expected regarding mortality statistics, satisfaction statistics, and efficiency scores. It's very simple to read. It just names the hospitals according to where they fall in that model, above or below expected. This project is being held up as the example. It is one of the best examples nationally in this field, and we can only expect it to expand.

MR. HARRY L. SUTTON, JR.: Ms. Goonan, we're trying to project differences in cost, and the actuaries are being asked to bless cost estimates, or will be in the future, and they have to estimate premium rates for their corporations. Do you see any way of relating outcomes studies in a short period of time, at least to whether different methods of practice will produce the same outcome at a lower cost? The most obvious part of the report card is if we have all normal deliveries instead of cesarean sections and you still have different prices for those in terms of longer stay in the hospital and higher fees for doctors, which many plans have gotten rid of. You can project what the effect might be actuarially. Some plans are advertising that cesarean sections are only 15% of deliveries instead of 25%, but those are very simple and kind of superficial ones. How do we measure whether something is a good outcome, or shouldn't we worry about whether the best methods of producing an outcome are going to be more expensive than some others? Will the government buy that in health care reform, if the best methods cost more money?

MS. GOONAN: That is a complex question. Let me try to address some of your concerns. One thought to consider is that we need to walk before we run. For instance, figuring out the relationship between prenatal care, low-birth-weight-baby rates, and cesarean section rates are all three measures in the HEDIS list. Imagine having all three of those measures on all plans in the country.

An important article came out last August that described the role of prenatal care in producing better pregnancy outcomes. It had been in review for five years, because the authors, who are the national experts on prenatal care, could not explain why, in some populations, prenatal care interventions appear to have a positive outcome, and in others there is no effect. So they deliberated for five years and then finally published it, stating that they could not explain why they see such a small effect. Meanwhile, employers will say they are impatient, that they need performance measurement now, and that they intend to make decisions with this information. So I guess the point I'd like to make in response to your very complex question is to look for small steps.

The other point I want to make is that as this information becomes available, it is sometimes called provider profiling, when it gets fed back to providers, you're going to see it skyrocket in all these health plans. Everybody is buying profiling systems. They're all going to be of poor quality, because data will be of poor quality for a few years. Then we're going to start seeing some real information flowing. You're going to see provider behavior change dramatically. As providers get information about how they compare with their peers (oftentimes, half the variance from the mean disappears just by telling them they're different from their peers), that's all going to affect what you do. So good luck with that one.

MS. DOROTHEA D. CARDAMONE: I'd just like to ask Bruce Bowen to explain Chart 2 again.

MR. BOWEN: Chart 2 is telling us the relative risk of each health plan within a company. Health plan MFFS1C has a relative risk as measured by demographic factors of 0.95. It has 5% healthier people than average for the other health plans in this particular company. This last group model HMO has about 1.07 or 0.7% higher risk than the average as measured by the health status indicator. The demographic

measures, the demographic models for measuring risk, compress the risk. That is, they show smaller deviations from the mean than does the measure that uses health status. We believe that's because the health-status measures are more accurate, and they're more accurately accounting for the variance. But we can't prove that yet. Give me another six months and maybe I can give you evidence.

MS. CARDAMONE: Wouldn't it help to put up the actual experience of each of those cohorts?

MR. BOWEN: That's exactly what we're collecting, but we were in a position where we didn't want to say how healthy people are who report they're healthy. We want to say how much they will use in the future, so this is all completely a prospective investigation. I was going to call it an experiment, but we measured the health status with the questionnaire. Now we're collecting the utilization data.

MS. CARDAMONE: Your comment on the variations being less than what you thought or what employers thought they would be, are you talking about HMO populations?

MR. BOWEN: Well, there are two fee-for-service plans, a point-of-service plan, another fee-for-service plan, and a group model HMO in this particular employer.

MS. CARDAMONE: Oh, because I have noticed in our own company plan that when you look at just the HMO populations, you're dealing with a much younger group. If you look at people in the HMO and their experience company by company, comparing it with the fee-for-service environment, you're going to have different populations to deal with. I could see groups coming together more in the HMO environment and having less variation than the fee-for-service group.

MR. BOWEN: But these are the people in an entire company, and these are the people who are in a fee-for-service plan. This is a fee-for-service plan, and this is another fee-for-service plan. So one fee-for-service plan is getting very low-risk people, and another fee-for-service plan in the same company is getting high-risk people. And an HMO is getting high-risk people, and an HMO is getting low-risk people. So this is within a single company. The people who choose both HMOs and fee-for-service plans (the light colored bars) are measuring that just adjusting for age and sex; the darker ones for health status.

MS. ROSENBLATT: I didn't go over the example of a risk-adjustment mechanism, but I want you to think about the following question. The example is included in the paper prepared by the American Academy of Actuaries. The intent of a risk-adjustment mechanism is to get premiums and contributions that remove the impact of risk selection so that you are comparing plans based on their administrative efficiency and their medical-management efficiency. Now in a competitive marketplace, if carriers can decide to do predatory pricing, what happens to the impact of the risk-adjustment mechanism?