RECORD OF SOCIETY OF ACTUARIES
1995 VOL. 21 NO. 3A

RISK ADJUSTERS

Moderator: ALICE ROSENBLATT
Panelists: DANIEL J. DUNN*
HAROLD S. LUFT†
Recorder: STUART F. RUBINSTEIN

Results of risk adjusters' research will be discussed. How are risk adjusters incorporated into various state and federal proposals? What can be learned from the New York experience?

MS. ALICE ROSENBLATT: I work for the Boston office of Coopers & Lybrand, and I'll be the moderator of the session. I also chair the AAA Work Group on Risk Adjustment. For those of you who aren't aware of the AAA series of monographs, there have now been two done on the subject of risk adjustment, and we're hoping to do a third. I'm also very pleased that we have two health care researchers with us who have been doing much work on risk adjustment.

Hal Luft is professor of health economics and acting director of the Institute for Health Policy Studies at the University of California, San Francisco. He received his undergraduate and graduate training at Harvard University, majoring in economics, with a specialization in health economics. Prior to coming to the University of California, San Francisco in 1978, he was an assistant professor in the Health Services Research Program at Stanford University. He is a member of the Institute of Medicine of the National Academy of Sciences. His research has covered a wide range of areas, including applications of benefit cost analysis, studies of medical care utilization, the relationship between hospital volume and patient outcomes, regionalization of hospital services, adverse selection in multiple option health insurance settings, competition in the medical care market, and HMOs.

Dan Dunn is a senior research economist in the Department of Health, Policy, and Management at the Harvard School of Public Health. For the past year, he has served as a Senior Investigator on the SOA funded study to compare different methods of health risk adjustment. Previously, Dan spent eight years as technical director of the resource-based relative value schedule (RBRVS) project at Harvard, a study to develop an alternative physician payment system for the Medicare program. In addition to risk adjustment and physician payment, Dan's research interests include hospital reimbursement methods.

I will be discussing what some states are doing with respect to risk adjustment. Colorado includes groups of under 50 insureds in their risk-assessment method. There is no premium redistribution, meaning there is risk assessment, but no risk adjustment, that is,

---

*Mr. Dunn, not a member of the sponsoring organizations, is Research Economist of Harvard University in Cambridge, MA.

†Dr. Luft, not a member of the sponsoring organizations, is Acting Director, Institute for Health Policy Studies, School of Medicine at the University of California in San Francisco, CA.

there is not actually a transfer of funds between carriers. Florida has recently issued a request for proposal asking for assistance in reviewing various methods of risk adjustment. Kentucky implemented fairly aggressive legislation involving risk adjustment. All insured individuals and groups of less than 100 are risk adjusted. There is a statewide alliance; the risk adjustment involves a dual methodology. First, demographics are used, which include sex, retirement status, and COBRA. Age adjustments are allowed within premium rates, so age did not need to be included in the risk-adjustment mechanism. There is also a high-cost condition retrospective method. This method is closer to reinsurance than some of the other high-cost condition types of risk adjustment.

Minnesota is also dealing with the subject of risk adjustment. The risk adjustment work group in that state has developed a paper on the subject of risk adjustment that includes much information on the goals of a risk-adjustment mechanism, the issues connected with using the various risk adjustment methods, and so on In the State of Utah, there is a private alliance using a risk-adjustment method that depends on gender, family size, and high-cost claims.

There is much research being done on the subject of risk adjustment. There are a few Robert Wood Johnson Foundation funded research projects. One is being done with the Managed-Risk Medical Insurance Board (MRMIB) in California, which is the Health Alliance in California (HAIC). Another involves a group of large employers, called the Pacific Area Business Group on Health, who are looking at risk adjustment from the viewpoint of a large employer offering multiple choice plans.

There is also research funded by the Robert Wood Johnson Foundation for the Washington Basic Health Plan. Arlene Ash, who is a researcher with Boston University, is doing work funded by the Health Care Finance Administration (HCFA) on diagnostic cost groups (DCGs). This research involves the Medicare population and the use of DCG to refine the risk adjustment in the adjusted average per capita cost. Finally, Dan will be discussing the SOA project on the subject of risk adjustment being done by Coopers & Lybrand and Harvard University.

The following update on the New York risk-adjustment method is based on material from Bob Benedict, the Chief of the Accident and Health Rating Section of the New York State Insurance Department. On July 17, 1992, the governor of New York signed Chapter 501, which mandated community rating and open enrollment. The law also provided for the establishment of market stabilization mechanisms to prevent and dampen the disruption in individual and small group health insurance. On December 22, 1992, Regulations 145 and 146 were promulgated, which implemented community rating, open enrollment, and the risk-pooling mechanisms. On April 1, 1993, those community rating, open enrollment and risk pools took effect. In New York, there are seven geographic areas that form risk pools, and for each of those seven geographic areas, there are three pools that are part of the risk adjustment mechanism: demographic for Medicare supplement plans, demographic for non-Medicare supplement plans, and a specified medical condition list. The specified medical condition list is a list of very few conditions with a set payment for each condition. This pool applies to non-Medicare supplement plans only.

In terms of actual results in New York State, one of the key questions is whether the community rating law and the risk-adjustment mechanism caused a decline in the number

of insureds. The source for the following numbers is the New York State Insurance Department. Between September 30, 1992 and March 31, 1993, there was a decrease in the number of insureds in the small group, individual, and Medicare Supplement markets of 114,000, or a 5.1% decrease. However, from the period March 31, 1993 to January 1, 1994, there was a decrease of 3%, and from January 1, 1994 to July 1, 1994, there was a decrease of 2.7%. Thus, given the passage of the law on April 1, 1993, it does not appear, from these numbers, as if the law had a major effect on the number of insureds.

Another question is whether the move to community rating caused all the young, healthy people to leave the insurance market. Since the risk-pool mechanism is based on demographic factors of age and sex, there is now a collection mechanism to get all of that information and we can examine the average age/sex factor over time. These factors have been tracked by geographic region. In Albany, on April 1, 1993, the age/sex factor was 1.104. By July 1, 1994, the age/sex factor had decreased to 1.083. Similarly, in the New York City area, which is where the bulk of the population in the state is located, the age/sex factor has decreased from 1.077 to 1.072. There were some areas where there were increases. The factor for Buffalo increased from 1.038 to 1.067. That is a quick overview of state activity and of some of the research projects occurring.

DR. HAROLD S. LUFT: I'd like to focus on issues of biased selection, risk adjustment, and in particular, on two different areas in which this might be used—paying health plans and employers paying into pools of the kind that Alice just described.

Obviously, risk selection, including adverse selection, is going to occur unless health plan enrollees are allocated randomly, which doesn't happen. If it did, there wouldn't be a need for actuaries. The world isn't random. With different kinds of health plans, not just plain vanilla health insurance plans, but markedly different kinds of managed care plans, the opportunity for selection will become greater and greater. That's particularly the case if some health plans see it in their interest to try to select low-risk enrollees.

One of the questions that comes up is, can't we just risk-adjust at the level of the employer and the health plan, particularly if we're dealing with large employers that are used to being self-insured? They will say, "I want to set the rate; I know what my people cost; I want to negotiate with the health plan; and I want to adjust the role in various ways." This raises a series of technical questions. Can we go beyond age and gender? Dan Dunn will be talking about ambulatory cost groups, diagnostic cost groups, and so on. I'm going to argue that, for some purposes, age and gender may be the only factors we need to consider. For other purposes, whatever factors we throw in won't be nearly enough. And so it's going to appear that I'm talking out of both sides of my mouth. But then I'm a two-handed economist, and that's what I do. However, that is the nature of the problem that I think we all need to be dealing with.

Can we adjust at the level of the health plan and the individual employer? I'd like to present some data in order to consider this issue. The data on employees by employer and health plan within the Pacific Business Group on Health are about a year-and-a-half old, and the information is confidential, so you can't tell who the employers are. One employer offers 24 health plans. All but three of those health plans have under 2,000 enrollees. Another employer offers 25 health plans. All but four of those plans have over 2,000 enrollees. Only one out of the 15 employers has under 5,000 employees. All the others

have more than 5,000 enrollees. So we're talking about large employers, but when you get it down to the health plan level, the cells get too small to do a reasonable level of risk adjustment.

When looking at a group of about 5,000, some issues and problems arise in trying to do this sort of adjustment. The example I will use are the 5,000 enrollees from the Bank of America, who are enrolled in a fee-for-service, or what used to be a fee-for-service plan. At the subscriber-unit level, rather than the individual level, 17% have no expenditures in the preceding year which is not unusual; 12.3% have some inpatient expenditures, with an average cost around $15,000 per subscriber unit (Table 1). The distribution for this sort of population, not surprisingly, is highly skewed (Chart 1). That little blip on the right-hand side is because we truncated it to 50,000. So there are 49 families out of the 5,000, slightly under 1%, with $50,000 or more in expenditures. That's the group that I would argue we need to keep our eyes on and worry about.

TABLE 1
5,000 BANK OF AMERICA INSURANCE ENROLLEES (SUs)
(TRUNCATED) MEDICAL EXPENDITURES FOR ONE YEAR

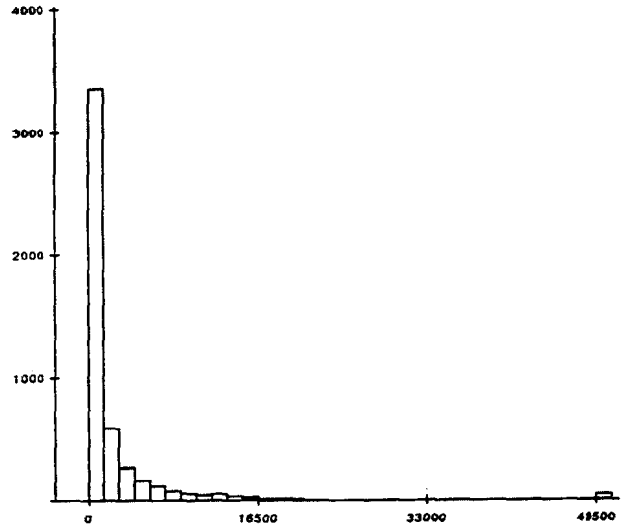
	Number of Enrollees (Percentage)	Average Expenditure	Median Expenditure	Coverage	Number of Enrollees (Percentage)
No Expenditures	832 (16.7%)	0	0	Employee Only	2,072 (41.4%)
Some Expenditures, No Inpatient Services	3,551 (71.0%)	1,557	779	Employee + Spouse	983 (19.7%)
Some Expenditures Inpatient Service	617 (12.3%)	15,149	10,902	Employee + Family	1,945 (38.9%)
49 enrollees (1%) with expenditures of $50,000 or more.					

In our study we used age and gender plus some other demographic data, like information that an employer might have, such as years with the company, job status, salary level, and so on; tier-single, two-party, family; and region within California because there are geographic differences both in cost and practice patterns; and we developed a moderately sophisticated model that takes into account the fact that 17% of these people have no cost, and a small number have very high costs. We used a four-equation model. I won't go into the econometrics of this, but it was a better model than the standard age and gender breakdown because it took into account the skewness of the distribution. We got what is not unusual with this kind of model, an R^2 of about 0.04. Economists look at R^2s.

Chart 2 illustrates what 0.04 is. There are 5,000 points, one for each of those subscriber units, carefully plotted—thankfully not by hand. There are a couple of stories here. The 49 families with expenditures of $50,000 or more are represented along the top. They tend to have somewhat higher predicted values, but not markedly so. What's more important is our minimum predicted value is about $650. Nobody is at zero, even though 17% of the population is truly at zero. And our maximum predicted value is about $10,000, even though we know that 1% of the group is at $50,000 or more. So, the models don't predict

terribly well, particularly at the tails. But, does that matter? After all, what we're all interested in is predicting at the group level, not at the individual level. If we can predict for groups, we're a good part of the way there.

CHART 1
DISTRIBUTION OF BANK OF AMERICA
SUBSCRIBER UNITS BY YEARLY EXPENDITURE



In Chart 3 we see what happens if we take these 5,000 enrollees and lump them into groups of 50. We get 100 groups of 50, which have been randomly assigned; think of them as relatively large "small" employers or groups of 50. The chart shows the average costs for each of the small employer groups. The predicted values spread somewhat, but not much. The observed range is between about 1,700 and a little over 6,000, whereas the predicted are in a fairly tight band. There is some positive correlation there. This is what age and gender, do for you. Most of what you're seeing with groups of 50 is random variability, or variability above and beyond age and gender, employment status, and so on, in groups of this sort.

Does that mean that we need to give up? I would argue no. If people are randomly assigned to groups, you don't need to worry about risk adjustment. You just enroll many groups, and they'll all wash out with random assignment, but they're not randomly assigned. Let's say we were to take, of those 5,000 enrollees, the 50 with the lowest predicted value, and put them into a group (Chart 4).

CHART 2
OBSERVED AND PREDICTED EXPENDITURES
FOR 5,000 SUBSCRIBER UNITS
IN THE BANK OF AMERICA FEE FOR SERVICE PLAN

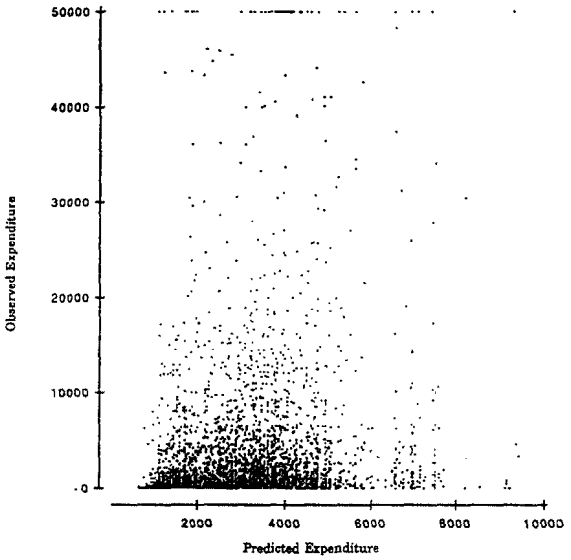


CHART 3
OBSERVED AND PREDICTED EXPENDITURES PER SUBSCRIBER UNIT
WHEN SUs ARE RANDOMLY ASSIGNED TO GROUPS OF 50

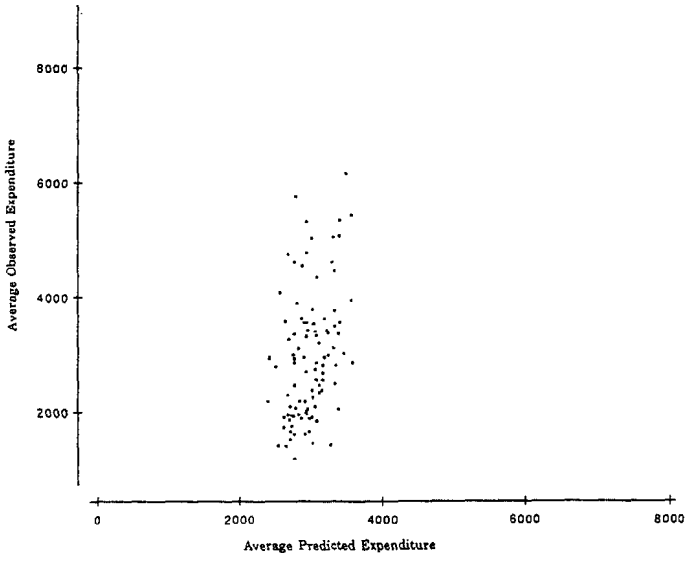
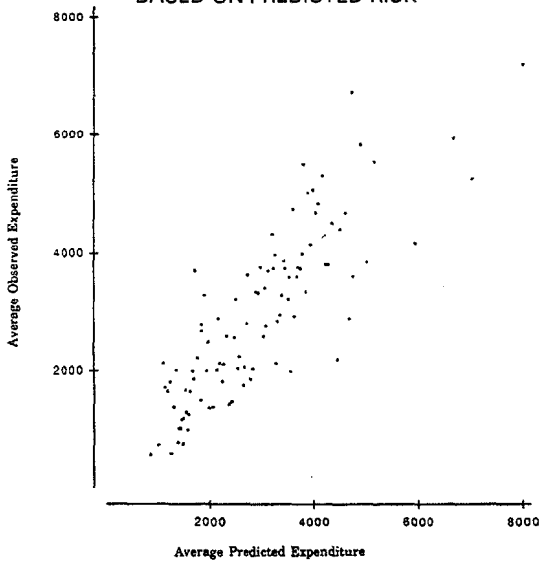


CHART 4
OBSERVED AND PREDICTED EXPENDITURES PER SUBSCRIBER UNIT
WHEN SUs ARE RANDOMLY ASSIGNED TO GROUPS OF 50
BASED ON PREDICTED RISK



Take the next 50, put them into the second group, all the way up to the top group of 50, the highest risk group. This is as skewed as we can get in terms of the predicted values. Then let's ask, how well do those predicted, at the group level, predict observed expenditures? The fit is much better. What we're seeing here are two things. The average predictive spread is now about as wide as the observed spread. Groups with low predicted values, in fact, on average, have low observed values. Young people tend to have low observed expenditures. People who are in the high predicted cost groups tend to have high costs. So even a relatively simple model works reasonably well. It took me five or six years to figure that out. Actuaries have been doing it for decades, and if only I'd read your literature, I would have known that age and gender work.

Let me argue that that applies when we're talking about employers, that is risk adjustments across different employer groups. If an employer group comes to me as a health purchasing cooperative, I can look at that employer, use measures of this sort, and feel reasonably comfortable in setting a premium for that employee group. Why? Because in general, people who work for employers are doing so and are hired by the employer because of their work characteristics which are unrelated to their health status. For whatever reasons they're hired by or working in that company, it is not because of their prior medical expenditures.

The situation is very different if we're looking at the choice of health plan within an employee group, within a health insurance purchasing cooperative (HIPC), or within one of these risk pools. Then people may very well choose the health plan that best fits their medical needs. And if we're talking about different kinds of health plans, if all the health plans were the same, or if they all had a $250 deductible, and the same benefit factors, and

if they all looked like vanilla ice cream, then there would be no selection. However, there is no reason to have multiple plans all selling vanilla ice cream. You're going to have vanilla, dutch chocolate, chocolate chip, strawberry, and other different flavors, and when you have different flavors, you're going to get differential selection. In that kind of situation, these sorts of risk models won't work. Dan will be talking about adding additional information, and I think additional information will help a great deal, but it won't help enough to deal with the kinds of subtle selection, particularly the kind of selection that might occur with those 49 families with expenditures of $50,000 or more. The very high-cost individuals are the ones we need to worry about.

Let me give you an example of subtle selection that led me into this area of research. I was on the faculty welfare committee for the University of California, which deals with health insurance. We started hearing that a number of employees who were being hired during the year would go to their doctor and contact their health plan. The doctor and the health plan would say, "You're not covered." The employee would say, "Well I filled out the form." When the employee went to the UCSF benefits office, benefits would say, "Yes, you filled out the form." The health plan would respond, "We never received the form. We never charged you for a premium, and you are not covered until we charge you for a premium." The paperwork was lost. Everybody who has dealt with bureaucracies knows that people lose paperwork. Most of the people who had their paperwork lost at UCSF seemed to be singled out by their type of lifestyle. It was very clear what was happening: the health plan was redlining within an employee group—a practice that is not legal. The choice then is, either we figure out how to be a cop with a big stick to make the health plans do what they ought to be doing, or we need to develop incentives to encourage the plans to take care of the people they ought to be taking care of—that is, pay them more for high-risk people, and pay them less for low-risk people. Hence, my research on risk adjustment.

As we get into finer and finer levels of risk adjustment, I believe we need to focus on changing the process and not just improving the technology, which is like setting more and more detailed rules for teenagers. It doesn't work. It becomes a game for them to figure out how they can beat you. The process change that I would argue for is the change to a zero-sum game. Every purchasing cooperative, be it Pacific Business Group on Health, California HIPC, the Kentucky plan, or whatever, must say to each of its health plan members, "We're going to put a certain amount of money on the table to be available for risk-adjusting contributions, and we will set higher rates for those plans that have higher-risk people. Of course, those plans with lower-risk enrollees would get less, because there's a fixed sum."

To risk adjust, the purchasing cooperatives could use age and gender, or other additional factors. But then what do we do with the people with HIV? What do we do with the small number of people with cystic fibrosis, or amyotrophic lateral sclerosis (ALS), or who are ventilator-dependent? Those are the cases that occur so infrequently, and yet are so expensive, that you can't capture them with a standard kind of risk-adjustment methodology. You need something more subtle. The more subtle thing might be the kind of high-risk pool that Alice mentioned New York is using, where for each person with AIDS, based upon their CD-4 count, you can get a monthly payment. If you're a managed care plan and are able to keep your costs down, you can actually make money on HIV people because you're being paid enough on a monthly basis.

You can set up the process whereby each health plan has an incentive to bring to the table information on high-risk people and say, "Next year, we'd like to add this category." The risk manager, who is just a referee, says, "OK, medical directors, have all your health plans come in with the information on how you would define these cases; you can pull charts and we can review and audit them. Also bring in your finance people and tell us how much it's going to cost, and then we'll figure out what the average payment will be. This is essentially reinsurance, but not cost-based reinsurance. We're not going to pay what you claim it costs you in an individual health plan; we're going to pay you the average cost for an efficient health plan to give you the incentive to do well. Then we will do one more thing because we're going to be clever enough to move this from the pure payment system to a quality improvement system. Once you've pulled charts, once you've identified people, it means that we or some third party can confidentially send letters out to the patients and ask them how they're doing, or ask them about the quality of care, to learn about the process, to figure out how to improve what's being done for these people, rather than just paying for them." This is melding risk adjustment in terms of payment systems with risk adjustment to improve patient care and quality of care and thus improve the process of care.

The third piece that's linked into this is the message is given early on that risk adjustment will force health plans to be efficient and provide high-quality care, and that risk selection will work only for a year or two until competitors figure out how to identify it. Remember, what's the first thing that happens when somebody switches health plans. The new health plan asks for the medical records. They get transferred over, they start noticing who's coming in with what illnesses from what plans. The competitors will start making the system self-policing. The intent is to let people know that this is a game that sharks will not be able to win in the long run, and the best approach is to focus on providing good quality health care efficiently, and the focus of risk adjustment in the future will be to encourage people to do what's right, rather than to do what's just profitable in the short run.

MR. DANIEL J. DUNN: Hal's analysis presented evidence of selection in the market for health insurance. The issues that Hal was discussing would be a key component of any thoughtful approach to try to remedy the problem. However, it's clear that an adequate risk-adjustment method will also be an important part of the solution. Risk adjustment would address the incentives for an impact on subselection, and would insure more equitable compensation for plans, induce plans to compete on efficiency and quality rather than selection, remove the effects of plan's health risk on the premiums that consumers face, and promote access to health insurance.

This need for a suitable risk-adjustment method, and other reasons, led the SOA to fund a research project on comparing different risk-adjustment methods. In particular, you can take real live claims data and apply them to these methods to see how the methods stack up. I was lucky enough to be part of the research team that was chosen for the project and worked with a group at Harvard and Coopers & Lybrand.

I'll bring you up to date on what we've accomplished, present some of our preliminary results, and also discuss some of the implications of them. In the future, we'll be preparing a detailed final report, which will go in-depth into many of these methods, but today I can give you a flavor of where we're going and what we've done.

The objectives of the project are, first of all, to compare the ability of the different risk assessment methods to predict annual health expenditures, so we're looking at total annual health expenditures per enrollee. Second, given that these methods don't seem to work extremely well for high-cost individuals and conditions, we're separating them out as a separate area for research, and we'll explore alternatives for them. Finally, you need to weigh predictive accuracy against the practical issues, including how much it costs to administer a program, whether the data are available, and how easy it is to game a system. What's equally important is what are the incentives for efficient provision of medical care.

The methods that we're going to be looking at include age and sex groups, ambulatory care groups (ACGs), ambulatory diagnostic groups with age and sex, and DCGs with age and sex. Later, I'll describe what each of these are.

For ambulatory diagnostic groups (ADGs) with age and sex, which you're most familiar with, we used 28 different groups, 14 for males and females. Other than the group age zero to one, each group included a range of five years. We're working with a nonelderly population, so we have no enrollees over the age of 65.

In addition to age and sex, ACGs, ADGs, and DCGs include information about the disease patterns of the enrollees. In particular, they work with the diagnostic codes that are included on the hospital or outpatient claims. ACGs work primarily with outpatient information. Each diagnosis is categorized into one of 34 ADGs. Those ADGs are combined with age and sex to produce one of 52 ACGs. A person can receive more than one ADG.

However, one person is assigned only a single ACG. So our ADG model, in a sense, is an intermediate step to ACGs, and given the fact that a person can have multiple ADGs, it may in fact include some additional information that's not incorporated when things are boiled down to a single ACG per person.

The other variant of our risk-adjustment method we're looking at is DCGs. And at the time, there are actually three different working models of DCGs. We're working with what's called the expanded DCG model. This compares with an earlier version of DCGs which relied primarily on, and only on, inpatient information. So expanded DCG moves from being just inpatient to also incorporating the outpatient codings. You start with all the inpatient and outpatient diagnoses for a person. Each diagnosis is categorized into one of 12 DCGs. So, again, if someone has more than one Internal Classification of Diseases—9th Revision Code (ICD–9), they'll have more than one DCG at this point. An important distinction is made between inpatient and other diagnoses, so the same diagnosis, if for an inpatient episode, will carry more weight than if it's only for an outpatient episode. Finally, each person is assigned to a unique DCG based on the one with the highest expected cost.

The data for the project were assembled by the Society, with the help of a number of carriers who contributed information. We worked with a standardized database that's quite large, a national data set covering a good representation of all the nonelderly age groups, as well as the geographic areas. We had claims from eight carriers. We have two years of data for each enrollee, 1991 and 1992. In addition to a large number of claims, we also have quite a range of health plan types, indemnity plans, HMOs, PPOs, and so on, and

different deductible levels. So in this way, we're able to look at the impact of health plan type on the results. For each enrollee, we're working with enrollees rather than households or insurance units. We know: the plan type; enrollee demographics, such as age, sex, zip code; expenditures for 1991 or 1992; use in terms of number of admissions; and clinical diagnoses, both inpatient and outpatient, coded using ICD-9. So in essence, we have all of the information that is required to do age-and-sex risk groups, ACGs, ADGs, and DCGs. So there's a good database for the project.

I will present a brief synopsis of the methods here. We're performing the analyses pool by pool. By the term pool, I'm referring to the intersection of a carrier and a plan type. So pool A may be from Carrier 1, the PPO plan. Pool B could also be Carrier 1, their indemnity business, and so on. So again, this is being done pool by pool. Given that we are working with roughly 30 to 40 pools of data, you can think of this as sort of a repeated test of the methods, and from that we hope to gain some idea of patterns across pools. Each enrollee is assigned to a risk group, according to a risk-assessment method.

An important step we take is to truncate all the information to $25,000, because, as Hal had mentioned, the methods don't seem to work as well for high-cost individuals. We're treating them as a separate analysis, as I talked about before. So we're not throwing out the high-cost individuals; we're only looking at their first $25,000.

To apply the models, we first estimate the risk rates, using half of the sample for a pool, and then we apply the risk rates to the second half of the sample. This way, we're not using the same group to both estimate the rates as well as to apply them. It prevents some overfitting. So for each enrollee in the prediction half of the sample, we have our predicted expenditures, and we have actual expenditures. In essence, that's what we're looking at here—how the actual compares with what's predicted by our model. We're looking at these models in a number of different applications.

One important distinction is between our retrospective and prospective applications. In our retrospective application, we would be using the information for a year to predict expenditures for an enrollee for the same year. So, use 1991 information to predict expenditures for 1991, and the same thing applies for 1992. This approach may be most relevant for a system where you set premiums at the beginning of the year, and maybe there's a settlement process at the end of the year.

We're also looking at prospective applications, where you're using the risk adjustment information for year one to predict year-two expenditures. That would be more in line with a system where premiums are set based on previous information, and that's what you live with. Hal had talked a little about how these models work for individuals as well as for groups. We're also looking at them along these dimensions. We're looking at two types of groups, however. First we are looking at randomly selected groups of 2,500 enrollees, but we're also looking at nonrandom groups, and I'll get back to that later on. There are a number of ways we can measure predictive accuracy. Some of them I'll present include: the adjusted R-squared, the mean absolute percent prediction error, and errors within a particular range, such as errors within plus or minus $500, $1,000, and so on. In terms of the dimensions for assessing accuracy, we have both retrospective and prospective applications, and we're seeing how these models perform for individuals, as well as random

groups. In addition, we look at how the prospective models perform for specific subpopulations or non-random groups (Table 2).

TABLE 2
ASSESSMENT OF PREDICTIVE ACCURACY

Application	Individuals	Random Groups	Non-random Groups
Retrospective	x	x	
Prospective	x	x	x

The prediction error that we referred to in the results is predicted expenditures minus the actual, so it's expected from a model minus the observed. Percent prediction error is straightforward as shown below. R-squared, for those who aren't familiar with it, can best be thought of intuitively as the percentage of variation in annual health expenditures across individuals or groups, explained by a model. So, the higher the prediction error, the worse the model is doing. The higher the R-squared, the better a model is doing. I'll present results from four of the 30 pools we're working with, which are fairly representative. We found quite consistent results across the data we've looked at. The four pools include two indemnity pools, a nongatekeeper PPO pool, and a network HMO pool. Combined, they represent about one million enrollee years.

$$\text{Prediction Error} = \text{Predicted} - \text{Actual}$$

$$\%\ \text{Prediction Error} = \frac{\text{Predicted} - \text{Actual}}{\text{Predicted}} * 100$$

$$\text{R-squared} = \frac{\text{Model Sum of Squares}}{\text{Total Sum of Squares}} \text{ or}$$

Percentage variation in annual health expenditures explained by a model

The pools are not equal in size. The indemnity and PPO pools are each running a couple hundred thousand enrollees. The HMO has about 40,000 enrollees. Some mean expenditures for the data will be presented. The mean was about $1,325. That's after it has been truncated, at $25,000. Without truncation, that mean would be about $1,500. As Hal has shown, expenditures are quite skewed. Thirty percent of the enrollees have zero claim dollars. About 75% of the enrollees, in fact, have less than $1,000 in claims. Less than 1% of the enrollees have greater than $25,000 in claims, so those are the people we're truncating. And only 5% of enrollees have an inpatient admission.

We'll first discuss the results for individuals and random groups, then I'll talk a little about the non-random groups. As you can see in Table 3, we're showing the numbers for both the retrospective and prospective applications. Each column represents a method, so we have age, sex, ACGs, ADGs, and DCGs. Each line for an application represents the pool that we're looking at. So, for example, the adjusted R-squared, under the age/sex column, shows 0.037 and is for the first indemnity pool. So, when looking at the retrospective results, you might imagine that for individuals, more clinically based approaches like

ACGs, ADGs, and DCGs, do better than age and sex, both in terms of R^2, as well as prediction errors within $1,000.

One interesting thing to note is that the results are quite comparable across the different pools, so the models seem to work similarly for HMOs as they do for PPOs as they do for indemnity. As you'd expect, the ability to predict expenditures drops when you go from predicting next year's expenditures using this year's information to predicting this year's expenditures with this year's information. The R^2s go down, and the prediction errors within $1,000 seem to drop a bit. Again, the clinically based models do a bit better, with ADGs probably doing the best in terms of that application.

Chart 5 summarizes the results in Table 3. DCGs seem to do the best, by a hair, on retrospective application, and ADGs are best on the prospective application. Once again they do better than age and sex. Of course, individuals aren't the only dimension to look at. In fact, some would argue that groups are the most important assessment of the methods. We choose 100 groups of 2,500 enrollees randomly from each pool.

Table 3 shows a summary of 2,500 enrollees for each group. We are looking at the predicted expenditures for the entire group, comparing them with the actual expenditures for the entire group, and the numbers in the table are a summary across those hundred groups. Not surprisingly, age and sex do quite well compared to the more clinically based methods, when you start working with large, random groups. In addition, the drop-off in predictive accuracy from retrospective to prospective applications isn't as significant.

Of course, if we had only random distribution of risks across health plans, then there's no need for this whole discussion of risk adjustment. So we thought, we'd try to create some non-random groups. The objective was to think about what we know about an enrollee in year one, and see if we could use that information to tell how well they would do—a model would do for that enrollee in year two. The first approach we used was to look at enrollees with relatively high or low expenditures in 1991 (Table 5). For low expenditures, we grouped people who had costs below one-third of the mean. And for high expenditures, we grouped people with costs greater than three times the mean. This was a relatively simple approach based on information that a plan would have at its disposal.

The second approach (which is perhaps a little more sophisticated) was to look at selected high-cost conditions, and I'll show you what those are in a minute. We tested these non-random groups using the prospective model. The question is, does the model systematically under- or overpredict for these groups? As you can see on the left-hand side of Table 5, the key information here is the predictive ratio, so that's the mean predicted divided by the mean actual. The mean predicted is the amount that's predicted by the age-and-sex model, the ACG, ADG, or DCG model. Of course, the mean actual is the same, because we're talking about the same group of enrollees. And the predictive ratio is the ratio of the first and second columns. All the models overpredict for the low-cost individuals, and underpredict for the high-cost individuals. If the predictive ratio is one, then they'd be doing a perfect job. Some of the models seem to do a bit better than others, in dealing with these non-random groups of enrollees. In fact, ADGs and ACGs seem to do quite a bit better than age and sex, and also do better than the DCG model.

TABLE 3
SUMMARY OF PREDICTIVE ACCURACY
INDIVIDUAL RESULTS FOR SELECTED POOLS

		Adjusted R^2				Percent Prediction Errors Within $1,000			
Method	**Pool**	**Age/Sex**	**ACG**	**ADG**	**DCG**	**Age/Sex**	**ACG**	**ADG**	**DCG**
Retro-spective	Indemnity UR (B1)	0.037	0.285	0.351	0.358	55%	74%	75%	71%
	PPO(C1)	0.039	0.226	0.248	0.324	57	68	72	70
	Network HMO (H1)	0.037	0.255	0.333	0.309	58	72	74	70
	Indemnity (X1)	0.039	0.298	0.366	0.391	55	73	73	71
Prospec-tive	Indemnity UR (B1)	0.034	0.090	0.104	0.089	56	68	64	55
	PPO (C1)	0.039	0.088	0.103	0.074	53	66	62	52
	Network HMO (H1)	0.040	0.078	0.108	0.086	56	67	56	56
	Indemnity (X1)	0.043	0.095	0.116	0.102	53	68	59	52

CHART 5
SUMMARY OF PREDICTIVE ACCURACY INDIVIDUAL RESULTS:
RETROSPECTIVE AND PROSPECTIVE

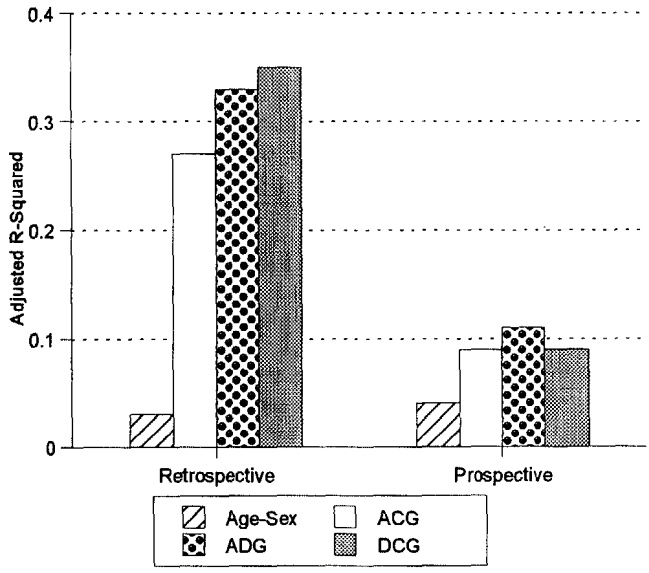


TABLE 4
SUMMARY OF PREDICTIVE ACCURACY
RANDOM GROUP RESULTS FOR SELECTED POOLS
(100 GROUPS OF 2,500 ENROLLEES SELECTED FOR EACH POOL)

Method	Pool	Mean Absolute Percentage Prediction Error				Percentage of Groups with Prediction Errors Within 10%			
		Age/ Sex	ACG	ADG	DCG	Age/ Sex	ACG	ADG	DCG
Retro-spective	Indemnity UR (B1)	4	4	3.5	3	96	87	99	99
	PPO (C1)	4	3.5	3.5	3	93	98	98	99
	Network HMO (H1)	3.5	3.5	3	3.5	96	96	99	99
	Indemnity (X1)	4	4	3.5	3	95	99	99	98
Prospec-tive	Indemnity UR (B1)	4	4	4	4	95	95	93	93
	PPO (C1)	5	5	4	5	92	92	95	99
	Network HMO (H1)	7	7	7	8	80	80	78	74
	Indemnity (X1)	5	4	4	4	94	97	97	97

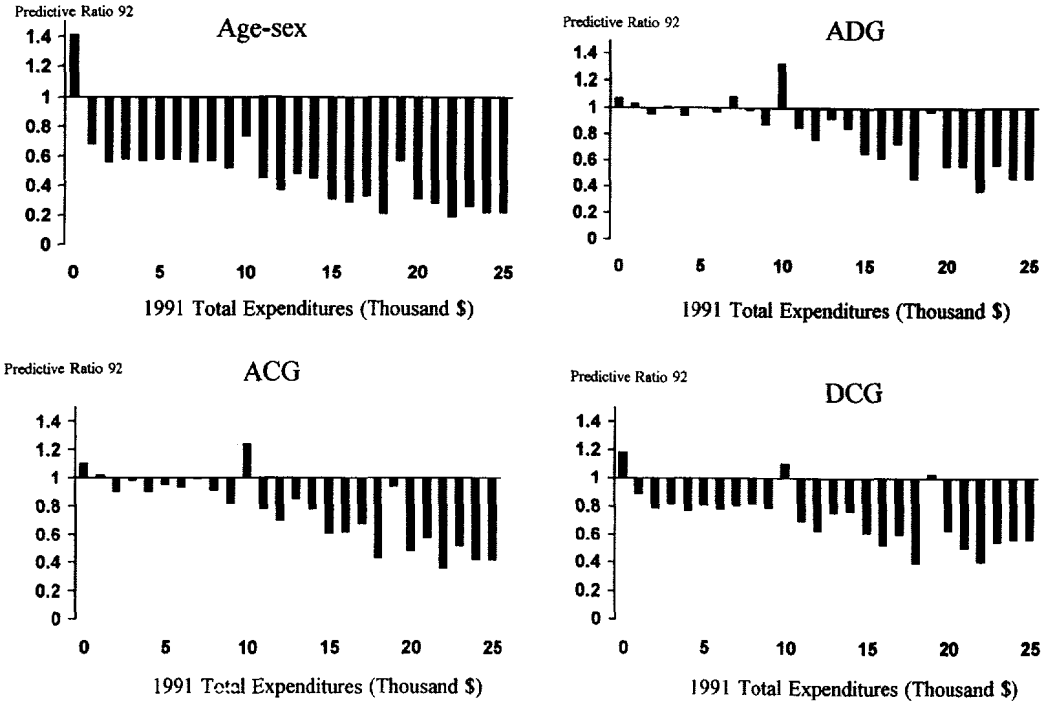
TABLE 5
NON-RANDOM GROUPS
1991 LOW AND HIGH COST GROUPS
PROSPECTIVE MODEL

Method	Low-Cost Group (< 1/3 Mean 1991 Expenditures)			High-Cost Group (> 3 Times Mean 1991 Expenditures)		
	Mean Predicted Expenditures 1992	Mean Actual Expenditures 1992	Predictive Ratio	Mean Predicted Expenditures 1992	Mean Actual Expenditures 1992	Predictive Ratio
Age/Sex	$1,232	$719	1.72	$1,763	$3,688	.48
ACG	837		1.17	2,930		.80
ADG	775		1.08	3,221		.88
DCG	932		1.30	2,732		.74

We decided to look at this result a little more closely. The low-cost groups which were forming in Table 5 are represented by the results of the left part of each of Chart 6, and the high-cost individuals would be to the right of each of the graphs. On the *x*-axis of each graph, there's the 1991 total expenditures for an enrollee, grouped into bands of $1,000. On the *y*-axis, there's the predictive ratios that we saw before. So, a predictive ratio of one would suggest the models are perfectly adjusting for these individuals. A predictive ratio of greater than one means that after risk adjustment, these would be more or less winning cases. And, a predictive ratio of less than one would suggest that you would lose on a case.

One pattern that you can observe is that, for each of the models with individuals with costs less than $1,000 in year one, you win across any of the methods. And that's represented by the bar up to the left-most side of each of the graphs in Chart 6. There is, certainly, a pattern across these methods. For age and sex, you'd be seriously underreimbursed for any of the individuals with expenditures greater than $1,000. That's not necessarily the case for ADGs, where it isn't until $10,000 where the models seem to fall short. ACGs do not perform as well, and DCGs follow.

CHART 6
NON-RANDOM GROUPS BY 1991 EXPENDITURE GROUP
PROSPECTIVE MODEL—MEAN PREDICTIVE RATIO 1992



In addition to looking at enrollee costs in 1991, we also looked at some particular conditions in 1991. We picked two conditions that were defined previously by the Health and Insurance Plan of California grouping system, courtesy of John Bertko. We include cancer patients, so these are patients with diagnoses of different kinds of cancer, who had an inpatient stay in 1991. We also include heart condition patients, again, with an inpatient stay in 1991.

As you can see from Table 6, the model predicts quite low expenditures, relative to actual, for these groups. For example, for age and sex for cancer, the model would predict a premium of $14 for every $100 of actual expenditures. Also, even though they all fall short on this measure, DCGs and some of the more clinically based ones certainly show a little improvement.

TABLE 6
NON-RANDOM GROUPS
FOR SELECTED HIGH-COST CONDITIONS*
PROSPECTIVE MODEL

Condition	Method	Mean Predicted 1992	Mean Actual 1992	Predictive Ratio
Cancer (N = 209)†	Age/Sex	$2,329	$16,727	0.14
	ACG	2,959		0.18
	ADG	3,607		0.22
	DCG	3,960		0.24
Heart Conditions‡ (N = 171)	Age/Sex	$2,484	$12,738	0.20
	ACG	3,195		0.25
	ADG	3,196		0.30
	DCG	4,325		0.34

*High-cost conditions were defined using inpatient ICD9 diagnosis codes and the Health Insurance Plan of California grouping system.
†Cancer includes those with cancer diagnoses of leukemia, multiple myeloma, cancer bone, cancer breast, cancer other sites, cancer prostate, cancer respiratory and digestive, cancer stomach, cancer trachea, bronchus, lungs, and an impatient stay in 1991.
‡Hearth conditions include those with heart condition diagnoses of aortic valve disorders, mitral valve disorders, acute ischemic heart disease, and an inpatient stay in 1991.

I don't have many results to present on our work for high-cost individuals, but I can give you an idea of where we're headed. All the methods I've talked about so far have truncated each enrollee's claims at $25,000. This part of the project will look at those claims over $25,000. We are working to identify them, characterize them both in terms of what conditions are represented, as well as what the expenditures look like for these people. The goal is to potentially come up with some high-cost conditions that can be added to the risk-adjustment methods in dealing with some of the cancer patients or heart conditions we saw. There might be some conditions meeting relevant criteria which might fit into some type of reinsurance mechanism that Hal had touched on.

The high-cost individuals are relatively small in number, but represent a significant part of the expenses. Individuals with expenditures greater than $25,000, made up less than 1% of all the enrollees in our sample, but were responsible for 25% of the expenditures. A significant number of the high-cost individuals we looked at were at the high end of

somewhat medium-cost conditions. This may suggest that they're not very good candidates for some sort of reinsurance mechanism, such as has been used in New York, and it is being explored in Kentucky. Those high-cost procedures, which are truly high in cost, in many cases, are quite infrequent and represent a small percentage of all the expenditures over $25,000. Furthermore, there's significant variation in cost for these conditions. This is just a sample of some of the information we've seen for high-cost conditions.

Table 7 shows four of the procedures in the New York risk-adjustment method, pulled from a sample of one million enrollees. As you can see, there were few enrollees with these conditions. In fact, the dollars over 25,000 that we had truncated for the primary analyses represent less than 5% of the target for a reinsurance mechanism under our approach. The previous slide showed some procedures, these are, in fact, conditions. And again, they're drawn from the market diagnosis groups identified by the Health Insurance Plan of California Risk Assessment Work Group. These are individuals with the diagnoses described in the second column, and they're all individuals with an inpatient admission for their previous year. Here these data are not truncated, so you're looking at true average expenditures. And again, enrollees with large expenditures are somewhat small in number. All together, they represent about 15% of the high-cost dollars in the system. But equally important, the coefficient of variation, in the column labeled CV Percentage, which is the standard deviation as a percentage of the mean, suggests that there's significant variation for these cases.

TABLE 7
SUMMARY OF EXPENDITURES
FOR SELECTED NEW YORK RISK-ADJUSTMENT METHOD
HIGH-COST CONDITIONS/PROCEDURES
(ANALYSIS BASED ON SAMPLE OF APPROXIMATELY 1 MILLION ENROLLEES)

Condition/ Procedure	N	Mean Expenditure	CV Percentage	25th Percentile	Median	75th Percentile
Heart Transplant	10	$295,506	30	$146,471	$185,135	$236,240
Bone Marrow Transplant	19	212,220	47	216,169	328,981	360,890
Liver Transplant	5	196,897	21	185,399	185,635	236,662
Pancreas Transplant	4	156,434	26	121,365	158,748	191,503

In summary, our methods seem to perform well for large, random groups of enrollees. Our methods perform better in retrospective applications. For individual enrollees and non-random groups, the clinically based approaches do quite a bit better than age and sex, and are somewhat comparable to each other. However, no method performed well for non-random groups. In addition, the findings seem to be similar across health plan types. The models work equally well for HMOs indemnities or PPO plans. There are no strong conclusions on high-cost individuals. Certainly, they appear to be a target for future work in this area, and there's more work to be done.

In terms of the implications of what we've found, if plans enroll large random groups, age and sex or a similar risk-adjustment method would be sufficient. However, if plans enrolled disproportionate numbers of high-cost individuals or low-cost individuals, or if risk selection could be done on an individual level, the more clinically based approaches certainly help and would be a significant part of the solution. However, some of the

findings of the non-random groups suggest that, even for the clinically based approaches, they don't quite go far enough. Further research and modeling in this area would be required.

MS. JOAN P. OGDEN: I would like some clarification, Mr. Dunn, on the summary of predictive accuracy of individual results for selected pools—the variation within $1,000 prediction error. Is that prediction error based on truncated claims at the $25,000 level, and is this on a base of the $1,325 mean claim, so what number of standard deviations is represented by the $1,000 error band?

MR. DUNN: I don't have at the tip of my fingers what the standard deviation is, but it is working with the truncated information. Actually, the range of expenditures on which to report that result, depending on what you're trying to look at, could vary. In fact, $1,000 may not be the most interesting range to look at. It has been suggested to us to look at how well the models do for each range of expenditures, including expenditures over $5,000, or areas over $5,000.

MR. TIMOTHY M. ROSS: There is fairly extensive reference to the R^2 approach, and I think a general disappointment with the 4% and 10% R^2, with predictive accuracy from age/sex and also in the diagnostic approaches. What we normally see is R^2 being much higher in predictive models. I think we also see, in more standard predictive models, more normal residuals. Generally, when you look at the individual continuance curves, they're more log normal. And so, I've seen some people use R^2 based on log of the observed variables.

If I'm a health care provider or an HMO, and if I design a very effective way of treating people with heart disease, what's going to happen is all the people with heart disease are going to come to my plan, as opposed to somebody else's plan. And if the risk adjusters that are available in the marketplace don't consider that, then I'm going to take an adverse financial result. From what we see, we must encourage the formation of that sort of a health care system, where people are encouraged to build a better mousetrap. Currently, if they build a better mousetrap, yes, they take care of heart patients more efficiently than anyone else, but their plan costs are higher than anyone else's because they've attracted a disproportionate number of the very sick people.

MS. ROSENBLATT: Thanks for the explanation of risk adjustment. I would add to your explanation that there's an assumption of community rating or a form of rating where your rating can't reflect the type of risk that you have. And so the risk-adjustment mechanism is helping that plan that does do something special for heart disease patients or whatever, because it doesn't want to rate for that. So the risk-adjustment mechanism does it. Hal would like to answer the R^2 question.

DR. LUFT: R^2 is actually a fairly bizarre measure. And when you think about it, it's the wrong measure because it's measuring the percentage of the variance, which is squared dollars. Nobody pays squared dollars, they pay dollars. If you log the dollars, you tend to get a more normal distribution and a somewhat better fit. But the main thing is that none of the models predict very well at the tails. This four equation model was designed to distinguish those people at the very low end and very high end, and it did yield somewhat better results, but there just isn't much information. If you had really good risk models (for

example, if you knew that my father died of a heart attack and that my family history revealed other incidences of heart problems) you would know that I'm at high risk for heart disease. But that doesn't give you a very good prediction of whether I'll have a heart attack this year, and if I have one, it gives you no information on whether I will die in bed cheaply or whether I will have bypass surgery costing $50,000. So there's just enormous random variability that we'll never be able to explain. You can argue about relative improvements in R square, but it doesn't matter. It's the wrong measure. We need to focus more on how to appropriately pay those health plans that take the tough cases.

MR. DUNN: I agree. I'm not sure it came out as well as I intended, but I think these models clearly fall short for the high-cost conditions. They certainly do well for expenditures at the mean or a bit above, but if you remember the graph showing that you only have the winners at the low end of the distribution, it may be that these models can never reach up and deal with those high-cost cases.

FROM THE FLOOR: ACGs are built on ADGs plus an age/sex factor, but they seem to have a lower predictive value than the ADGs themselves. Could you tell us why age/sex did not help?

MR. DUNN: The ACG model we used included ADGs plus age and sex. The ACG model incorporates, in some broad way, age and sex as well as the ADGs. One reason why the ADG model did better is because everyone is assigned a single ACG, which is a combination of all their ADGs, based on some rules, as well as their age and sex information. However, with the ADG model, we're allowing each person to have whatever exact combination of ADGs that they had. So there's some compromise needed in going from the ADG approach to the ACG approach.

FROM THE FLOOR: What do you think would happen to the predictive value of your model if you went over four, five, or six years, as opposed to just the next year, in the prospective model?

MR. DUNN: We actually haven't had the information available to do that, although there is some research going on that is looking at that. John Chapman, at Harvard, is actually looking at that question. Of course, you would not expect the models to do as well, but certainly there is significant evidence that people with high costs may tend to have somewhat high costs throughout their life. You'd expect some information from six years ago could help you predict your expenditures in the current year.

DR. LUFT: Just to show that you can always get at least two answers from two economists, I would argue that, over a longer period of time, the models, particularly the demographic, will tend to work better, because you're averaging those random expenditures over more years. The concern is whether somebody who has, for example, substantial muscular/skeletal problems, and will require much ongoing orthopedic intervention and will tend to be a relatively high-cost person for a long period of time. Also, someone with cancer may drop out of your sample within four or five years. So I think you need to look carefully at what the particular diagnoses are and at the ACGs and DCGs.

MR. DUNN: Using year-one data to look at year two may be unrealistic, given the amount of time it's going to take for the plans to pull together their claims data to do this

type of analysis. So, it may be better to look at year-one expenditures to predict year three.

MS. ANNA M. RAPPAPORT: There's a related social question. The social question is who's going to pay for the high-cost claimants, and is that a concern for all of us? Even with risk adjustment, we're concerned about the high-cost claimants. We can't risk-adjust them out. I'd be interested in your comments, because it seems that, from a policy perspective, finding a way to spread that cost on a social basis is part of what's going to be a satisfactory long-term solution.

DR. LUFT: I think you put your finger on a very crucial question. Back in the old days, when we all had community rating and all the plans were the same, you didn't have to worry about high-cost cases. They were there, but they were bundled together and nobody looked at them separately. We are now in a world in which there are health plans that have gotten very good at identifying and trying to avoid high-cost cases. That was how I entered this area of research. My argument is, we can't go back. We can't make them forget how to identify those kind of cases. The profit rewards that result from identifying and excluding those cases are too great. So we must figure out a way to pay the plans appropriately for those high-cost cases. That does not mean, however, that those individuals with high expenditures need to bear both the financial risk of their illness, as well as the physical and emotional risk, which we can't take away from them.

I think, as a society, we ought to pay for it. I think large employers are usually willing to pay for it, in terms of their contribution. But, I think that is a societal question. I think we have to have the risk-adjustment technology to allow us to give the right incentives to the health plans, so that they don't do the wrong thing.