RECORD OF SOCIETY OF ACTUARIES
1995 VOL. 21 NO. 4B

RISK ADJUSTERS: AN UPDATE

Moderator: JAMES N. ROBERTS
Panelists: DAN DUNN*
GARY M. KOSCIELNY
ALICE ROSENBLATT
Recorder: JAMES N. ROBERTS

This panel will present the results of risk adjuster research and discuss the potential use of risk adjusters in healthcare reform efforts.

MR. JAMES N. ROBERTS: We have two topics that we're going to be discussing. One of them is an update on the Society-sponsored research study on risk adjusters. To bring us up to date on that, we have Alice Rosenblatt, who's a principal at Coopers & Lybrand and is also the chairperson of the risk adjusters work group.

With Alice, we have Dan Dunn. Dan is a senior research economist at the Department of Health, Policy, and Management at the Harvard School of Public Health. Dan has been working over the past year as the principal investigator on the Society-sponsored study that compared different methods of risk assessment. Previous to that, Dan worked for an eight-year period as the technical director of the resource based relative value schedule (RBRVS) project at Harvard, which is probably of interest to many of us. In addition to his work at Harvard, Dan is the director of research at Cambridge Health Economics Group.

On a second topic, Gary Koscielny is going to give us an update on the experience of the risk adjustment technique being used in the New York small-group reform statutes. Gary is vice president and chief actuary at Amalgamated Life, and his company is the administrator of that program in New York state.

MS. ALICE ROSENBLATT: I'm going to be talking about the research paper as Jim mentioned. The name of the research paper is "A Comparative Analysis of Methods of Health Care Assessment." This was the project that followed a path that we're hoping will continue to be used on many of these types of projects where there's a public policy issue. In this case, the public policy issue is the subject of risk adjustment and its use in health care reform. I'll be talking a little bit more about that later.

The American Academy of Actuaries (AAA) published a monograph on the subject of risk adjustment. I chaired the work group as Jim mentioned, and one of the things that was done in that monograph was an identification of the need for actuaries to do more research. That led to a request for proposal from the SOA, and selection of a joint team from Harvard University and Coopers & Lybrand. Dan and I jointly are going to walk you through a very brief summary of the report, which has just been completed.

The research team was a joint team consisting of Harvard University and Coopers & Lybrand people. The oversight was provided by the SOA risk adjustment task force. That

*Mr. Dunn, not a member of the sponsoring organizations, is Senior Research Economist at the Department of Health, Policy, and Management at the Harvard School of Public Health in Cambridge, MA.

was one of the task forces created as a result of the Academy work group determining that this is an area that needs some research. We also created an advisory committee that consisted of health economists, as well as actuaries. The team studied data that were provided by insurance carriers.

The research team from Harvard University included Dan Dunn, who will be speaking, Debra Tarra, Eric Lattermer, Peter Braun, and Susan Bush. And just to delineate the areas of responsibility, it was Harvard that dealt with all of the data, and all of the data problems, which I think was the real hard part. I'd like to give a great deal of credit to Dan and the team who I think worked incredibly hard over the past year on this project.

On the Coopers & Lybrand side, in addition to myself, John Bertko was involved as was Tom Stoiber. We provided much of the linkage in terms of the public policy implications and how actuaries normally look at this. We hope the audience will be both actuaries and healthcare researchers, and of course, policymakers.

We had a very illustrious advisory committee. We had Jim Hickman, whom many of you may know from his activities in the Society. Jim provided much of the statistical knowledge that we needed in terms of how to measure the results and how to set up the different modeling techniques that could be used to test the methods. Bill Hsiao, in connection with Harvard, was on the advisory committee. Hal Luft, a speaker here, is from the University of California at San Francisco. He spoke at the Vancouver meeting on the subject of risk adjustment. He's one of the healthcare researchers who has done a great deal in the study of risk adjustment. From having worked with Dan, he let us know where some of the "glitches" were in the data. Joe Newhouse, who is also at Harvard University, has also done a large amount of work on risk adjustment, particularly in connection with the Medicare program and the adjusted average per capita cost (AAPCC). The SOA risk adjustment task force included Bill Lane as Chairperson.

I'm starting out with basic definitions. The definition of risk assessment is measuring the deviation of each individual's expected cost from the average cost. We're defining risk adjustment in a healthcare reform type of scenario as methods used to compensate for difference in risk. If one imagines some kind of alliance structure in a healthcare reform environment, carriers with the higher risk individuals would receive payments from an overall pool, and the carriers with the lower risk individuals would make payments into the pool.

Risk adjustment was actually mentioned in the Clinton plan, and some of the other versions of healthcare reform. It's also being discussed at the state level. But what leads to the need for risk adjustment are a couple of factors. First of all, much of the alliance structures involve individual selection of plans in a multiple choice environment, as opposed to the employer selecting a plan, and all the employees being insured with one plan, that leads to selection bias. Also, in the small-group market and in the individual market we're seeing compression of premiums through some form of community rating. If you think about community rating in its strictest sense, we're only varying rates for benefit design, and maybe for individual versus family coverage. You can think about the fact that, if one carrier in the marketplace had all of the bad risks, it would have very high rates even if it was low cost from the point of view of how it reimbursed providers, and how it performed utilization review and things like that. It might be the most efficient, but the public would see a very high rate because of the high risk. The intent is for a risk

adjustment mechanism to be found that would make community rating work in an ideal world. It must be one that would promote competition based on efficiency and quality rather than on which carrier is the best cherry picker or which carrier can select the best risk.

The primary research objective of the study compares the predictive accuracy of the different risk assessment methods. Dan will be talking about the different risk assessment methods. Some of them, like age/sex, are what we, as actuaries, are very familiar with. We will also throw some new initials on the table. Also, the study compares different risk assessment methods based on other criteria, such as, is it practical? Can it be done? Does it do what we just said, encourage efficiency? We all think about the distribution of claims cost. We know that it's the tail of that distribution that causes the problem in terms of healthcare. We know that 4% of the claimants drive 50% of the claim cost, so it's very important to look at those high-cost individuals and claimants.

Some other research objectives compare a prospective versus retrospective approach to risk assessment—I'll define that shortly. To compare the predictive accuracy for a nonrandom subpopulation, the study asks a question. When you have a carrier in the market with all the high risks and another carrier with all the low risks, do any of these risk assessment methods just blow up because of the nonrandomness of the population that we're dealing with? The study looks also at the sensitivity of study findings to the type of health plan, and the mix of enrollees. This is where this research project differs tremendously from some of the previous research projects. A good deal of the previous research was done on one HMO, for example, so that everybody in the population belonged to a given HMO. In this study, we received HMO data, indemnity data, PPO data, and we had data from many different carriers who were doing claim recording in different ways. Some of our findings are first-time findings because we had that wealth of data that we were working with.

I'm going to briefly go over some of the state applications of risk adjustment. I'm not going to spend much time on New York because Gary will be talking about New York. Basically, there's a demographic adjustment that relies on age, gender, and family size. There's also a high-cost procedure and condition pool. Connecticut adopted a mechanism that's similar to New York's high-cost pool. Washington State does not have risk adjustment in place, but it is studying risk adjustment on the state employee group of 300,000. Florida has just taken risk adjustment under study.

California has three forms of risk adjustment. California has a Health Insurance Purchasing Corporation (HIPC) for small groups. Because in California the small groups can be age rated, the demographic adjustment there depends on gender and family size. There's also a set of high-cost marker conditions. There's a much longer list of conditions than the New York approach, and it uses a weighted average to develop the risk adjustment transfers. There's also California Public Employees' Retirement System (CALPERS) which is the plan that insures the employees of the state, and demographic risk factors are used to establish targets for rate negotiations with the carriers. There's also a group of large employers called the Pacific Business Group on Health. They are studying risk adjustment for use in negotiations with insurers and HMOs.

I promised a definition of retrospective-versus prospective-risk adjustment. You're probably familiar with those terms in connection with experience rating. This idea is

similar here. You can do both assessment and adjustment prospectively and retrospectively. On the assessment side just to define it by example; you would take the 1995 attributes to predict 1996 experience—for example, use diagnostic conditions a given individual had in 1995 to predict claim costs in 1996. Retrospectively, at the end of 1996, you look back and say, which diagnostic conditions were experienced? And that's going to derive an average claim cost.

On the adjustment side, where we're actually transferring money between carriers, prospectively we could make an estimate of what those transfers are going to be and build that estimate into the premium. That would be a prospective form of adjustment. Retrospectively we might transfer based on the actual usage and the risk. For example, stop loss, a reinsurance mechanism that you're all familiar with, could be considered a retrospective adjustment.

Some of the types of health risk assessment methods, and Dan's going to get into this in more detail, would include demographic models. It's just what we're familiar with, age/sex. Then the AAPCC also uses location and institutional status. Health status models are generally based on actual diagnosis. There's also a questionnaire called the SF36 that asks questions such as, Can you walk up stairs? Can you feed yourself? Can you dress yourself? Is your health condition now better then it was a year ago? That would be self reported. And then there are also prior use models based on actual prior usage.

I am at this point going to turn it over to Dan who will talk about some of the risk assessment methods tested.

MR. DAN DUNN: I'll fill you in on some of the methods and data we use, and talk about our results for the predictive accuracy part. Alice will then come back and talk about some of the practical considerations, which are certainly as important as the predictive accuracy findings. I then will summarize.

Note there are a number of different ways to assess risk. We're only looking at basically two classes of those types. Of course, there are demographic models, which would include simple age and sex. Then there are a group of models that are based on previous diagnosis, like the health-status and clinical based models. All these models are based on information that could be found on administrative claim forms that insurance carriers typically collect.

We looked at two subcategories within the groupings of our clinical based models. One is ambulatory care groups, which includes the ambulatory care group (ACG) model as well as a subset ambulatory diagnostics groups (ADGs). I'll describe what these are. We looked at two diagnostic cost group (DCG) models, one which is based on inpatient information, and the second DCG model includes additional information.

I won't go over the methodology in detail. It can be quite complicated, just grouping different people into different categories at the end. The grouping is based on the diagnosis, age, and sex. After applying the logic, you'll end up in an ACG or DCG. It may seem complex, but one of the nice things about these models is that their developers have been quite open about all the assumptions that are used, and are usually helpful in helping you to understand exactly how things work.

For the purposes here, the key information for these models is what's used in assigning the individuals to each ACG or DCG. ACGs are based on outpatient diagnosis. Coded using ICD-9 diagnosis coding system, each diagnosis is first categorized into one of the 34 ADGs based on what clinicians and statisticians argue would be homogenous clinically as well as resource intensive groupings. These individual ADGs are actually one of our models, so it's an intermediate step of the ACG model. Finally, based on the ADGs, age and sex, patients are grouped into one of 52 ACGs. Someone can have more then one ADG if he or she has multiple diagnoses that are somewhat different, but everyone is assigned to one and only one ACG.

The first of the two DCG models we looked at is called the principal inpatient DCG model, which we termed the PIPDCG. All the PIP diagnoses are recorded for an individual for a year. Each diagnosis is categorized into one of 12 DCGs, which are groupings based on their expected cost related to that diagnosis. Finally, each individual is assigned to the highest expected cost DCG.

The initial PIP model had tried to distinguish between discretionary and nondiscretionary hospital admissions. This would get around the potential incentives for admitting a patient to get a higher risk grouping, and thus a higher payment; however the model that we used removed this qualification. The way they did it before was they got together a panel of physicians and had them assign the relative discretionary areas. In the end, this didn't help all that much in terms of predictive accuracy. The current PIP model doesn't account for that.

The second DCG model might be considered something of a kitchen sink relative to the other. We call it the expanded DCG (EDCGDX) model with high-cost conditions. It uses both inpatient and outpatient diagnoses. Like the PIP model, it categorizes each diagnosis into one of 12 DCGs. However, with this model, a distinction is made between whether they're inpatient or outpatient diagnoses. It gives more weight to the inpatient ones with the rationale being that they'll probably be related to higher costs. Again, these are assigned to the highest expected cost DCG. In addition, they identify 25 high-cost coexisting conditions. These are conditions other than the diagnosis that was related to the DCG assignment, to try to pick up co-morbidities. Examples of these are diabetes, some infectious diseases and so on. These would be expected to take care of the higher cost individuals in the modeling.

This is a summary of the five models we talked about here. The age and sex model had 28 age and sex groups with 14 each males and females, and with five-year age bands approximately for each one. We had no elderly individuals in the database since we're only talking about individuals under 65. ACGs had 52 groups, based on ambulatory diagnostic information. For ADGs, we actually added the age and sex groupings since ADGs, unlike ACGs, do not account for age and sex. The PIP model, again, uses only inpatient information. The EDCGDX model includes quite a number of risk factors, namely 12 EDCGDXs, 28 age and sex groups, and 25 or 50 high-cost coexisting conditions and includes both inpatient and outpatient information.

As Alice had mentioned, we were lucky enough to receive quite a rich database from a number of national insurance carriers. We ended up with the final analysis using data from seven carriers, and it described enrollees for 1991–92. Thus, we had two years to work with. Approximately four-and-a-half million nonelderly enrollees were included.

These are both the claimants and the nonclaimants that we're picking up with these data. Again, as Alice had mentioned, we had a range of health plan types: indemnity, HMOs, and PPOs, plus we had a range of deductible levels to look at.

The databases were actually designed to test the models that we evaluated. They included, in addition to the plan type, enrollee demographics including their locations, based on zip code, age, and sex. A complete record of their expenditures and utilization was obtained, and the clinical diagnosis code using ICD9. These are the key variables to group the patients with these models.

The analytic methods are pretty straightforward and consist of a number of discrete steps. First, we assigned each enrollee to a risk group according to the risk assessment methods. So for age/sex, each enrollee is placed in an age/sex group. The ACGs are based on their diagnosis, age, and sex. Each individual is assigned to only one group for each model. The risk adjustment formula is tested using the expenditure data, and we're using total expenditures for the year for each individual for these models. Examples of the risk weights employed are the average cost for males between 35 and 39, or the average cost for ACG5.

Next, predicted expenditures are calculated for each enrollee for each method. So based on which risk group they're in, and based on the risk rate assigned to that group, we can come up with a predicted value. Finally, we also have their actual expenditures, and can compare those with the amounts predicted and that becomes the essence for our measures of predictive accuracy.

We use both prospective and retrospective applications. For the prospective application we're using 1991 information and trying to predict their 1992 costs. For the retrospective application, we're using 1992 data to predict 1992 cost. Thus, we can not only compare the methods within prospective applications but we can also compare between prospective and retrospective.

A final important point on our methods is that, in order to narrow down the potential noise that was related to the extreme skewness that you usually observe, we truncated expenditures for the primary diagnosis at $25,000. When we looked at the high-cost conditions, we relaxed that assumption and looked at every dollar.

FROM THE FLOOR: How did you handle geographic differences?

MR. DUNN: Given we were working with data from across the country, we were left with the problem of controlling for things such as price differences and differences in intensity of care. We explored a number of different adjustment methods. There are very few that can deal with cost difference and also with the problem that some parts of the country have more hospital days per person or more doctor visits or whatever. The differences are related to a number of different things. Here we're most interested in netting the differences out, so we could be looking at a level of XXXX. We ended up using what was the AAPCC that Medicare uses. It basically is a measure of the relative expenditures per enrollee for across the country. There are certain problems using that method. However, given its uniformity across the country as well as its availability, it was probably the best approach we could use.

We did some sensitivity testing to see how the models would change if we had used different approaches. It ends up that the methods weren't sensitive to the approach.

And also, there will be a fairly lengthy and fairly detailed final report that will describe in great detail all the assumptions and results. We're just giving you a synopsis in this session.

There has been some debate over what is the best measure of predictive accuracy for these models. Some people argue that all the models need to do is predict well for groups of enrollees, because differences across individuals would tend to cancel themselves out. Given that insurance is the pooling of risk across individuals, that should be the target for the models. However, given the potential issues and incentives for risk selection, this argument ignores the incentives which plans may have to try to seek out the lower risk individuals or avoid the higher cost individuals. If that's the case, then you also need to consider predictive accuracy for individuals, as well as for nonrandom groups. Rather than pick individuals versus groups as the one measure of predictive accuracy, we'll look at both individuals and groups and also look at nonrandom groups. I'll come back to these shortly.

In terms of predictive accuracy, we used a number of measures. Three were the individual measures. I'll come back and talk about what the r^2 is—the mean absolute percentage prediction error. It is just the percent difference between the predicted value and the actual value observed for an individual. We determined the percent errors within a particular range, such as within $500, $1,000, greater than $5,000, and so on.

FROM THE FLOOR: Did you test the sensitivity to truncating claims at $25,000?

MR. DUNN: We have done a little of that. We actually had two reasons for using our approach. One, our approach had been used previously in other studies on risk assessment, and it gave us a chance to link that with these studies. This maybe isn't the perfect reason, but it was nice. Second, our approach was looking at what the tail of expenditures looked like, and $25,000 seemed somewhat reasonable based on that.

But to answer your question, we actually reran a number of the analyses after the fact, at $50,000 and with no truncation, and all those results are in the final report.

I'm giving you very much of an overview of the results. The report itself has more than you'd ever want in terms of all the measures. In fact, it may be too much, but in this session, we're just pulling out some of the key things to give you a flavor.

It turns out that all the other individual measures we looked at showed a very high correlation as measured by r^2 so this is a very representative measure. For the retrospective model, the age and sex didn't do nearly as well as the clinical based approaches. The PIPDCG model, which is based on only inpatient information, performs better. The method really picks out those patients who have the greatest cost of those who were hospitalized. The ADG and EDCGDX models follow somewhat close behind. Again, these are retrospective models based on what actually happened for that year.

For the prospective models, age and sex again fall short of the clinical-based approaches. However, as expected, the clinical-based approach comes a little closer. The ADG model,

which is based only on outpatient information, performs slightly better then the EDCGDX model, which really has quite a few clinical variables involved. Again, as expected, the prospective models were not able to predict cost as well as those applied retrospectively. Those are the results for individuals.

For random groups, we randomly selected 2,500 enrollees from each insurance plan and compared their predicted amounts for the entire group with the actual amounts for the entire group. We did this 100 times. So we're repeatedly selecting 2,500 enrollees randomly again and again. These measures of prediction are basically a summary across those 100 random groups.

As shown, there's actually very little difference between the models, in terms of how well they predict for random groups. All of them do pretty well. Interestingly, there aren't many differences between retrospective and prospective models. So even something as simple as age and sex does quite well for random groups of 2,500 individuals. We also looked at different sized groups, and results didn't really change that much. But certainly with smaller groups the r^2 is increasing and vice versa.

Probably one of the more interesting analyses we did was to look at nonrandom groups. These were groups of individuals of expected high or low risk, where we might think the models would fall short. They're also groupings that you might expect if the carriers are going to perform selection based on information that they have at their disposal. We really used two different types of groups. One is nonrandom groups based on previous expenditures. We looked at the 1991 expenditures for an enrollee and how large or small they were relative to the average for all enrollees in that year. We then saw how the model did in forecasting expenditures for 1992 for those individuals.

The second approach for nonrandom groups uses clinical conditions, which I'll get to shortly. A ratio greater than one means an overprediction and less than one is an underprediction. In terms of risk adjustment, it would result in an over- or underpayment for those groups of individuals.

Age and sex actually does not do very well for these groupings. It greatly overpredicts the expenditures for those with previously low claim amounts and greatly underpredicts those for individuals with higher claim amounts individually. For the individuals with the expenditures greater than, say, $7,000 in the previous year, age and sex is predicting only $0.38 on the dollar. Some of the other models do better. But, again, there's a very systematic relationship between previous expenditures and the inability of the models to predict. The ADG model seems to do best.

We created three different nonrandom groups for high-cost conditions. One was for individuals with heart disease in the previous year. A second, for individuals with cancer, and a third for all other individuals with hospitalization. Again, some of the models are better, but overall they don't do that well. Age and sex for cancer patients pays roughly $0.22 on the dollar under risk adjustment. For heart disease, $0.38 and for all other people with an inpatient admission in the previous year, it still falls short. The best model here is the EDCGDX model, which pays about right for those with previous inpatient admissions and underpays, but not by as much as the others, for cancer and heart disease patients.

If they had worse illness the next year and the models were able to predict that, which is what their goal is, then it would be a problem. But this is saying that even after we account for everything in the models, there's still noise left over and it's systematic in this direction. It says, they're sick, sicker in the following year, and they're even sicker than the model was able to tell us to expect.

And these are probably the most problematic results for risk adjustment, because these are all based on information that's at the plan's disposal in the previous year. And also, these are the types of conditions that create the greatest equity problems in a risk transfer process. So a plan that ends up with most of the cancer and heart disease patients under an age/sex approach faces great underpayments.

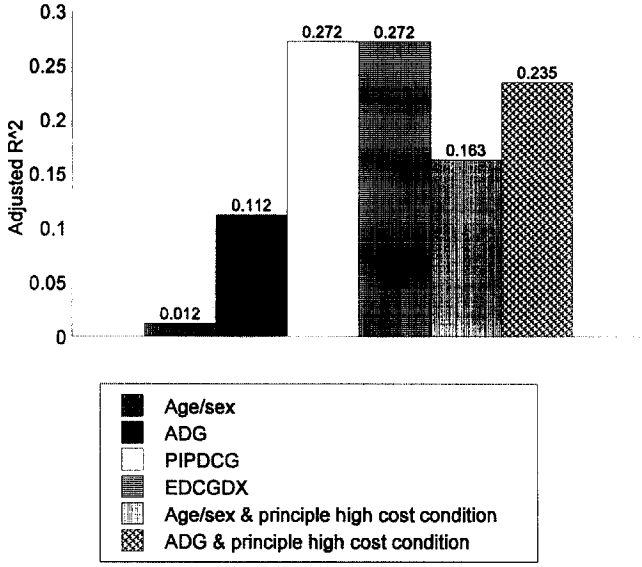
Briefly, one of the things Alice had mentioned is that we had a number of different health plan types at our disposal. And these are the findings for the individuals for r^2 for some selected pools. We actually had 19 of what we call "pools" or health plans, in the data. Those that begin with an "I" are indemnity plans; plans with "P" are PPO plans; and "H" are HMO plans. But, in general, any way we sliced it, the models all did about as well, or as poorly, on predicting the cost. So they apply in the same way across plan types.

As Alice had mentioned, we also looked at some of the issues and results for high-cost patients. We did this for two reasons. First, we had truncated expenditures at $25,000 for each analysis group that I just showed you. Second, a number of states, including New York, Connecticut, and California, have been looking at high-cost pools as a way to deal with the truly high-risk patients. We wanted to see if we could identify some high-cost conditions, add them to the models and see how much it increased the predictive accuracy. We ended up identifying 43 high-cost conditions. One obvious point is that, on average, they are high cost. Another point is that they're actually quite infrequent, with a few exceptions such as some of the heart disease conditions there. There are not too many of these folks in the typical health plan enrollment. However, and not surprisingly, if you look at the distribution of enrollees, all those over $25,000, which is less then 1% of all individuals, are responsible for 25% of the total cost. So, in terms of dollars, these are certainly important; patients are also important for risk adjustment.

We modeled these high-cost groups by basically saying, if an enrollee was in one of the groups with really high prospective rates, we're going to reimburse the plan at the average cost for all individuals in that condition and then add these to the models (Chart 1). Since the PIP and EDCGDX model are actually doing somewhat of a high-cost condition approach, we didn't add them to those models. Instead, we added them to a simple age and sex model, and the ADG model, in looking at outpatient events, is more looking at a high cost event.

The first four columns are the results without high-cost conditions. Of the two columns on the right, the first column is what you get if you add the high conditions to simple age and sex; so obviously it helps quite a bit. The far right column is what you get if you add the high-cost condition to the ADG model, which, again, was only outpatient on diagnosis. All of these are run without any truncations. Alice is going to talk a little bit about some of the practical considerations we looked at.

CHART 1
ANALYSIS OF HIGH-COST CONDITIONS,
SUMMARY OF PREDICTIVE ACCURACY, RETROSPECTIVE MODEL
NO TRUCULATION
INDIVIDUAL RESULTS—ADJUSTED R^2



MS. ROSENBLATT: One of the unique things that we did in this study is that many times we actually went on to test the concept of risk adjustment. Most of the other research that has been done has stopped at testing risk assessments. But we went on to actually test the monetary transfer that would occur if all of these data had represented risk pools within an alliance structure. Gary is going to talk more about the New York mechanism. But, very simply, what we're trying to do is use these different measurements to develop risk assessment. If you had a group and you had an age/sex/weight for that group or for a particular carrier's pool within an alliance, and compared it to the other carriers within that alliance, those with the low age/sex factor would pay into the pool, those with the high age/sex factor would get money from the pool.

That's what we're doing in these simulations. The first one actually uses the age/sex factors. The weights were developed from the actual data. We actually went through the data and asked how males age 40–44 compare to the total risk pools of those under age 65. Then we developed different weights and applied those weights for the entire pool. You can see in Table 1 that for all 7 pools you get a weight of 1. For example, pool IA had a weight of 0.932, and pool IH had a weight of 1.185. The New York method uses a similar type of demographic adjustment. Here, we're using a simplified version of it. We're calculating the average claims cost per person for each of the pools. So, based on the 0.932 example, the model would predict an average of $1,271 per person. However, the cost actually came in at $1,330. The transfer ratio is just the difference between one and the age/sex factor of 0.932. And if we apply that 0.068 to the average, we get a transfer amount of $93.

Consider the situation where within an alliance structure, transfer is going to be built into the premium. Keep in mind we're dealing only with the claims cost portion of the premium. Of course, in a real world scenario, you need to account for things like retention. But we're just dealing with claims here. We developed an adjusted premium, which because we're dealing with claims, we called a pure premium. We get an adjusted pure premium rate, which is the sum of that transfer amount plus what we're saying the premium started out to be. Then we calculate the total amount of premium based on the number of enrollees, which is not quite the total claims. You get an adjustment amount that you need to then spread across the pool. In this case, it was $259,000 that we spread. We got a final adjustment just to give you a sense of what kind of transfer amounts we're talking about. Our pool IH with the high age/sex factor of 1.185 would get money in from everybody else at $43 million dollars. The other carriers who had the age/sex factors less than one, would pay in amounts like $14 million or $18 million and so on. Across the whole pool it's zero sum total.

You can obviously do the same thing for all of the different methods like ADG, the PIPDCG and so on. Then you can simulate that, and we would end up with different results. The carrier that received the approximately $43 million under one method, may have to pay under another method. That's the real issue from a public policy point of view. The really scary thing is that that did, in fact, happen. For pool HD, an HMO, we had a situation where we had an age/sex factor of 0.953 and the ADG value was 1.638. We did all sorts of checking for quality and that is a real result. For that HD pool other methods produced factors of 1.361 and 0.933. In some instances, pool HD would receive payments of $16 million or $9 million, and in other cases HD would pay minor amounts like $1 million or $2 million. You can imagine the carriers sitting around the conference table having done these simulations and trying to agree on which method to use to do risk adjustment. We're talking about pretty major monetary transfers here, so it's a tremendous public policy issue.

Let me show you some of the reasons that are causing those results. Basically, Dan and his team did a great deal of testing. Why are we coming up with a factor like 1.63? What was occurring were differences in reporting and coding of diagnoses, particularly on ambulatory encounters. The HMO pools had a greater percentage of enrollees with claims and greater intensity of ambulatory diagnosis per enrollee.

Table 2 is just a sample. It shows some of the ADGs and gives you some of the descriptions of the ADGs. For example, consider a condition that's time limited. The 9.6% of the people in the pools fell into that particular ADG description. In an indemnity pool, it was 5.4%, but if you look at that HMO pool that I was emphasizing, it is 16.7%. If you look at malignancy, there is not as much of a difference, 1.9% and 4%. But administrative prevention was a real big difference: all pools, 17.9%; the indemnity pool, 7.6%; and the HMO pool, 50.3%. You can also see for no ADGs at all, you had across all pools 38%, in the indemnity pool, 47%, and in the HMO pool, 14%. There was a big difference.

Our database included not only claims, but also enrollees who did not have claims. Across all pools, 34% of the population had no claim; in the indemnity pool, 32%; and in the HMO pool, 14%. One of the reasons, obviously, for that is the plan design. The indemnity plans had deductibles, and there's probably a great deal of underreporting that occurred. For the HMOs with copayments like $5 and $10, we're probably getting all claims, even minor ones.

TABLE 1
SIMULATION OF RISK ADJUSTMENT TRANSFER PROCESS USING DIFFERENT ASSESSMENT METHODS
EXAMPLE OF CALCULATING OF TRANSFER AMOUNTS USING AGE/SEX METHOD

	Pool IA	Pool PA	Pool HA	Pool IH	Pool HD	Pool PC	Pool HB	All 7 Pools
(a) Ratio of each pool predicted to average	0.932	0.919	0.923	1.185	0.953	0.979	0.990	1.000
(b) Predicted Average claims cost per person	$1,271	$1,254	$1,259	$1,617	$1,300	$1,335	$1,351	$1,364
(c) Actual average claims cost per person	$1,330	$1,339	$1,400	$1,806	$1,375	$1,102	$1,246	$1,364
(d) Transfer ratio	0.068	0.081	0.077	(0.185)	0.047	0.021	0.010	0.000
(e) Transfer based on "average cost" per capita	93	110	105	(253)	64	29	13	
(f) Adjusted pure premium (c) + (e)	$1,423	$1,449	$1,505	$1,553	$1,439	$1,131	$1,259	$1,364
(g) Number of enrollees	153,526	169,513	20,098	171,764	18,498	204,351	120,965	858,715
(h) Total Transfer (e) x (g) ('000s)	$14,278	$18,646	$2,110	($43,456)	$1,184	$5,926	$1,573	$259
(I) Adjustments to achieve zero as sum of amounts in (h) ('000s)	($58)	($64)	($8)	$0	($7)	($78)	($46)	($259)
(j) Transfer after adjustment ('000s)	$14,220	$18,582	$2,103	($43,456)	$1,177	$5,849	$1,527	$0

TABLE 2
DISTRIBUTION OF AMBULATORY DIAGNOSTIC GROUPS (ADGs) BY POOL

ADG	Description	All Pools	Pool IH	Pool HD
1	Time Limited: Minor	9.6%	5.4%	16.7%
2	Time Limited: Minor-Primary Infections	1.6	8.0	31.8
3	Time Limited: Major	3.8	2.6	9.0
10	Chronic Medical: Stable	10.8	8.9	24.1
28	Signs/Symptoms: Major	1.3	7.9	23.7
31	Prevention/Administrative	17.9	7.6	50.3
32	Malignancy	2.1	1.9	4.0
33	Pregnancy	2.2	0.9	3.3
	No ADGs	38.4	47.5	13.7
	No Claims	34.0%	31.7%	13.7%

Let me get into some of the less technical areas of comparison. First consider practicality and administrative cost. The age/sex model is the most practical. The PIPDCG model, which requires only inpatient data collection, probably also runs lower cost than the other diagnosis based methods. Because there are fewer inpatient claims, you need to collect less data. But it would be extremely difficult to obtain consistent and high quality information on ambulatory diagnoses. Plus, these encounters may be missing completely. For example, staff model HMOs may not be coding the ambulatory data at all. Similar situations might exist in some HMOs that are capitating.

One of the other big issues is the ability to restrict manipulation. The age/sex model would be least likely to be manipulated. Also, it could be subject to easy audits. For the methods that depend on diagnosis like ACG, ADG, and DCG, we could get upcoding. Upcoding means that the providers of care would actually learn to code a higher diagnosis, because there would be greater payment from the risk adjustment mechanism. And again, this probably is more difficult to game on the inpatient side than on the ambulatory side.

Timeliness and predictability is one of the big issues. You notice we're using 1991–92 data, and this is 1995. We have similar problems if we think about an alliance. How long does it take to get the data collected? How long does it take to go through cleansing of the data and reasonableness checking of the data? You can end up applying the data with a sizable time lag. From that point of view, our methodology was not realistic, because we were using 1991–92. In a real world situation, you might end up with more of a two-year lag. The longer that lag time is, the less predictive the models are going to be. We weren't able to test for that because we only had the two years of data. But it would seem reasonable that if you ended up with a longer time lag, all the ratios that Dan was showing you in terms of predictive accuracy would drop.

On the retrospective side, there are a number of practical issues you come up with as you think about a given state having this type of mechanism. People who are in small groups may join large groups, or move out of state, and so on. By the time you make the adjustments for this type of activity, you might have a risk pool that looks very different. So you get into equity issues and then, of course, there are cash-flow issues. If there's one carrier that is supposed to receive a sizable payment, and you have it adjusted on a cash-flow basis, with a $42 million payment, for example, you could even end up with solvency concerns.

I mentioned up-front that one of the reasons for doing this is to get away from measuring carriers based on their ability to select the best risk. The objective is to get toward the system that actually has people selecting carriers based on efficiency and quality of care. Do these methods provide that incentive? The age/sex model really doesn't do much about that. The retrospective models might provide a greater incentive to actually increase cost and utilization because a higher payment is allowed. Some of the models may also, if they're looking at just inpatient services, provide incentives to treat on an inpatient basis rather than on an outpatient basis.

Aside from predictability, age and sex is the best, in terms of being practical, and we can audit it. But as you saw, the numbers are pretty low, in terms of the predictive accuracy. On the diagnosis based models, the prospective models provide greater incentives for efficiency. But as was shown by that HMO that I highlighted, there's a big issue in terms of the data on the ambulatory care side.

MR. DUNN: In terms of predictive accuracy, all methods as you saw perform with all random groups for reasons that we talked about. For individuals, retrospective models outperform prospective applications. For individuals in nonrandom groups, age and sex is certainly better than no adjustment. However, the diagnosis-based models have much to add to the age and sex approaches. Of course, you need to consider the practical problems with them, but there's some promise there.

Finally, we concluded that there's room for improvement since even the best models still provide some incentives for risk selection. Also, they could hold some inequities for plans with the higher risk patients.

What does this all mean? We certainly need risk adjustment. It's going to be a crucial part of anything we do, in terms of reforming the way health insurance is organized in the U.S. But alone, our results say it's not enough. Are you going to need some of the reforms related to the rules of the game such as guaranteed issue? Not having exclusion for pre-existing conditions and so on? Certainly these reforms of rules of the game have been thrown out before. Practical issues must be resolved. Data systems, auditing, and timeliness are certainly problems for some types of models. Methods using only inpatient diagnosis, such as the PIPDCG model may hold the greatest promise as an interim approach, all things considered. Further research, based on our findings, is unlikely to provide a major breakthrough or go much beyond the predictive accuracy of the models we tested. The alternative then would be to try to work with things like the high cost condition pools on top of these models to deal with the truly higher risk patients who aren't able to be modeled in the ACG or DCG models. Another alternative is some sort of blended payment systems, which are combinations of the good qualities of a prospective approach, and the extra predictive accuracy of a retrospective model.

MR. GARY M. KOSCIELNY: I will be talking about the risk adjustment system currently used in New York. I'll discuss what we're currently doing, and I have some initial results to share with you. Also I'll talk briefly about where this is going in New York.

The laws were passed and the risk adjustment was set up in 1992. Starting in 1993, the mechanism covers the individual, small group, and Medicare supplement markets. The laws implemented community rating and open enrollment, and set up the risk adjustment

pool. The individual small group market has a specified medical condition pool and a demographic pool. The Medicare supplement market just has a demographic pool.

The risk adjustment mechanism recognizes the geographic variation by carving the state up into seven different regions, and then operating the pools independently within each region. There are actually 21 independent pools operating in New York State.

The specified condition pool is a reinsurance type pool. There is a quarterly premium required. The premium is $5 for a single contract and $10 for a family contract. There are two categories that are recognized: the hospital only and the major medical only. Hospital is $3.75 and major medical is $1.25, and for a family contract, the rates are twice that.

There is a very short list of specified medical conditions that are recognized. These are called the table one conditions, including the transplants and the neonates. The amount shown in this table are the maximum amounts that will be reimbursed to a carrier from a pool. If a carrier does not pay more than these amounts for the patient, then only the amount paid will get transferred to the carrier. The idea behind this was to encourage carriers to keep their costs down by using managed care. We would get a dollar for dollar savings for any dollars that they can save.

The second type of conditions contains only two categories. Table two conditions are HIV and ventilator dependency. They're reimbursed on a monthly basis. Anytime a carrier has one of these conditions, it would submit claims and get up to these amounts: $2,000 for HIV, and $13,000 for ventilator dependency.

There's a two-step claims process. First there is something called a prenotification, which is essentially a one-page claim form, that identifies the patient, the carrier, and the condition. There are two reasons for doing this. One was we were hoping that it would help with some money management. Some of the conditions are relatively infrequent and high cost, and so we felt that, if we had some prenotification of these, that would help us. That turned out not to be the case. Virtually everyone submitted their claims right before the deadline.

The second reason relates to another problem. The deadline for 1993 claims was December 1994; for 1994, it's December 1995. There's about a year-and-a-half to two-year window to submit. Amounts paid are on a per-patient basis. If two carriers cover the same individual, we have to split up the payment, and so this gives us a warning that we have to look out for that.

The second step is a formal claim. This involves a longer claim form requiring patient's name, social security number, address, and things like that. There's something called the proof of payment form, and this includes actual copies of bills for bills over $1,000, a simplified claims listing or explanation of benefits for smaller amounts. There are also selected medical records submitted, including doctor notes or hospital records of some sort. Such documents would just simply identify things that we would need to know to be sure that the condition did actually occur, usually not more than two or three pages per condition.

Let's turn to some initial results for 1993, which is for a nine-month period. Claims of $6 million were paid to carriers for neonates, $7.2 million for transplants, ventilators for under $69,000, and probably the real surprise was HIV at only $97,000.

Then the number of claims was 77 each for the neonates and transplants; 3 ventilator cases, and 14 HIVs. Probably the surprise there is the low number of HIV cases. The reason has to do with the criteria in the regulation for an HIV case. The criteria is CD4 count below 50 on 2 consecutive tests taken a month apart. The problem with that is that the CD4 count is a result of a blood test, and it's not ordinarily collected as part of the normal claims processing. So, in order to get that, the carrier has to ask the patient, the provider, or a lab to send in that information. None of those are normally all that willing to do it without prodding. Their concern is confidentiality because this is HIV. A number of carriers are reluctant to even ask the question because of concerns of confidentiality. That's why there are so few of those.

Where is the money going? Again, this is nine months of 1993. I've broken the state into three types of carriers. First are the Article 43s (the Blues), the traditional large players in these markets. They've put in the most money and also are getting the most out of this pool. The other types are the commercials and HMOs. The big thing here is, everyone seems to be losing from this pool. No one is collecting more money than they're paying in.

There are a couple of reasons for that. One is that HIV was originally anticipated to be a more sizable condition than it was, and a large source of distributions to carriers. This turned out not to be the case. Second is that, when these pools were getting started, the pools were hit with lawsuits from the HMOs and later by the commercial carriers. I suspect that some of the carriers felt that these pools might not be around. In fact, the HMOs won their initial lawsuit, and so the carriers may have felt that this legislation might not be around, and it wasn't worth making the extra effort. It was surprising that a number of carriers did not submit any claims at all. Two of the largest Article 43s, two of the largest blues, did not submit any claims, and the single largest commercial carrier didn't submit any claims for 1993. In fact, to this date, through mid 1995, none of those three groups have submitted claims. Two of them actually called me within the past couple of weeks asking for the instructions on how to submit a claim, so I suspect that they'll follow through with the submission.

An interesting question to ask is, how much risk adjustment is being done with this? For nine months of 1993, $22.1 million was collected. Annualized premium for this block of business was $3.6 billion, that works out to 0.8% of premium. That is probably not a good indication of how much risk is being adjusted. Suppose that everyone collected exactly what they paid in to the pool. How much risk adjustment is being done in that case? Well, if everyone gets out exactly what they paid in, there's no risk adjustment. So another way of looking at it, I think a better way than just looking at the annualized premium, is to take the difference between what's paid in and what's taken out for each carrier, and add those up. When you do that, you come up with $10.5 million, or 0.4% of premium. An odd year was 1993, because it was a start-up year, and we didn't have that many claims, and there's the HIV problem, and so forth. We'll have to watch that later, to see how that works out.

Let's turn to the demographic pool. First, we'll go over some of the simpler concepts. There's an average demographic factor that is determined for each carrier in each region. It's just a weighted average demographic factor. It attempts to assess the risk that a carrier is assuming in that region.

Next is a regional demographic factor that's simply the weighted average of all the average demographic factors in a given region. The weighting is done on the annualized premium. The factors are based on age, sex and family status. They're similar to factors you might see in a rate manual. The payment disbursement formula is very simple. Claims are the first factor, and then there's the adjustment factor, which is simply the RDF minus the ADF or the difference between the regional average and the carrier's average for the given region, divided by the ADF.

Now how are we administering this? There are two cycles that go on. One is a quarterly cycle where we collect data, analyze premium, expected loss ratios, and the average demographic factors. We also do preliminary contributions and distributions. In this case, what we use for the claims portion is the annualized premium times the expected loss ratio. We start out using projected claims. On top of that is the second cycle, which is the annual cycle. It's called the annual reconciliation. For that, we collect actual claims incurred for each region, for each carrier, and substitute that into the formula. We ultimately use actual claims times the risk adjustment factors. Since we've already collected the preliminary amounts on the quarterly cycle, we compare what the carrier would have paid or collected on the annual cycle with what it has done so far. Then we make further payments, require further payments, or make the carrier's distributions to get to the annual cycle. We do an annual reconciliation twice for each year. In fact, we're still finishing off 1993 at this point. The first reconciliation requires a larger estimate of claim lag, and the idea is that by the second reconciliation there isn't much of an estimate to deal with.

Let's take a quick look at some of the regional demographic factors and how they behave (Chart 2). Most of them I think are pretty stable. New York has been very stable. Buffalo on the bottom has not been stable, but the reason for the big jump you see is that the largest carrier in that region decided that it wasn't doing the calculations right in the first place. So it discovered its mistake and adjusted it.

I don't recall whether this came about because of our audits of them, or whether they just figured the problem out on their own. I think it does emphasize the point that you have to do audits with these. In New York, everyone is audited once every three years.

Where's the money going? On the left side of Chart 3 is 1993, and that's nine months. On the right side you have 1994, and I've broken it out by the three types of carriers. As you can see, the Article 43s are pulling out the vast majority of the money, $83.4 million in 1993, and it's expected to be about $99.9 million in 1994. Again, 1994 is 12 months, so that should be much larger. The commercials and HMOs are putting in most of the money.

How much risk adjustment does that involve? Quite a lot compared to the specified medical condition (SMC) pool; a total of $160 million dollars. I'm measuring risk adjustment the same way that I did before. I look at each carrier and take the difference between what they paid in and what they're pulling out, and then add it up for each carrier.

And it comes out to $164 million in 1993, which is 6.0% of annualized premium. In 1994 it will be about $200.6 million, which is 5.7% of annualized premium.

CHART 2
I/SG DEMOGRAPHIC POOL—RDFS
APRIL 1993 TO JULY 1995

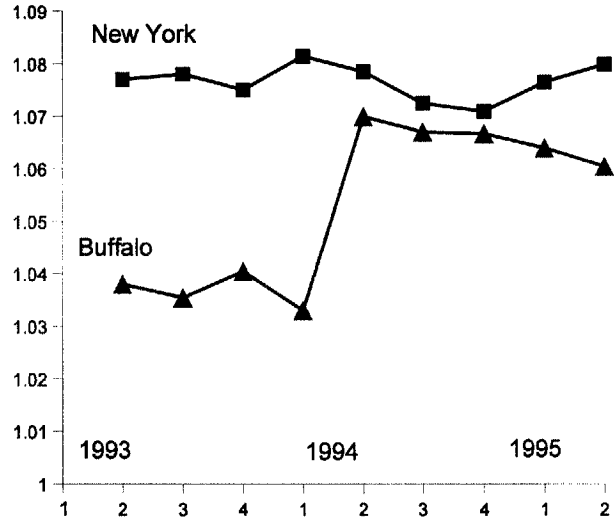
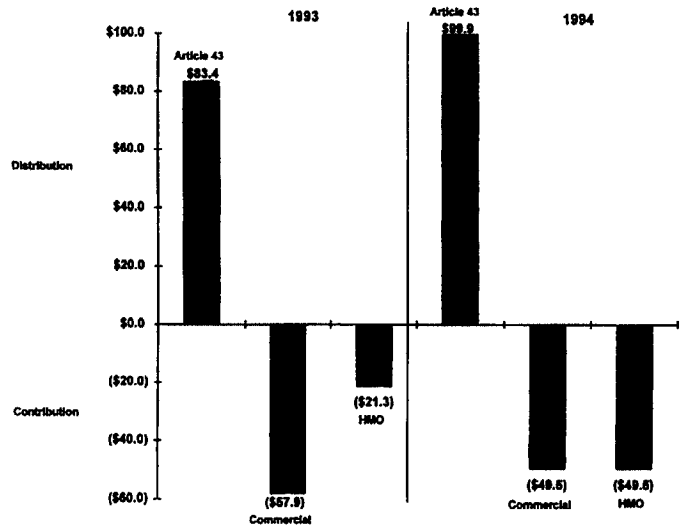


CHART 3
1993 (9 MONTHS) AND 1994 I/SG DEMOGRAPHIC POOL
CONTRIBUTION/DISTRIBUTION ($MILLION) BY CARRIER



Where is all this going? It's going away, at least the demographic pools are. Senate Bill 5469A is going to phase out the demographic pools by 2000. There will be a gradual reduction in the contributions step down over the next four years, and it's going to expand the SMC pool to take the place of the demographic pool. Based on the risk adjustment being done, you can see that it's going to take a great deal of expansion to get the same amount of risk adjustment. Why is that being done? I guess I would characterize it as a deal with the HMOs to offer point of service plans and drug coverage, which is not something that they have done universally.

MR. STEPHEN G. BUTZ: I have a question regarding the ratio of predicted to actual expenditures in the prospective model. You mentioned something interesting—the claims were actually higher than the model predicted for certain diagnoses, like cancer and heart disease. I wonder if the model that predicted that was based on the individual predictions or was it based on the pools? Do you take into account the possibility that the rates are different before movement of people from one pool to another?

MR. DUNN: I think I can answer the first part of the question. Alice will have to help me on the second part. You're right in that there are certain models that were providing significantly lower payments than actual claims. For certain conditions, we had shown you the results for heart and cancer conditions. Was this result somewhat sensitive to which pool we were looking at? In fact, this result was quite robust. We found it for both the HMO plans, PPOs plans as well as the indemnity plans.

MS. ROSENBLATT: We couldn't test for a spiraling effect. It would be very unlikely. We were using national data sets, and no area would have enough weight for one carrier to sort of dominate the data set to that extent. I would say it's pretty unlikely we were getting any sort of adverse selection spiral in the two years of data we had.

FROM THE FLOOR: Was the model based on previous extrapolations from one year to another year?

MR. DUNN: No, we had two years of data—year one to predict year two, as well as using year one to identify who has the cancer condition, a heart condition, and so on.

MS. ROSENBLATT: Let me try to answer your question. This is not experience rating; instead, it's closer to age/sex weighting factors and relating age/sex weights with an average claim cost. When Dan is saying the models predict, we're saying that based on whether it's an age/sex category or an ACG category, you're using the experience of either the previous year or that same year. Based on people who fall into that category and the model which you say, the average for that category is what the prediction is.

FROM THE FLOOR: So the model was based on the group of individuals regardless of which pool they were in, right?

MR. DUNN: We performed analysis pool by pool except for the transfer analysis that Alice presented. Within each pool the weights were particular to that pool.

MS. ROSENBLATT: I might add that the weights also differed. One of the things that we didn't get into in this synopsis but that is in the report, is that, if you look at the weights across the different pools, there are differences.

MR. W. RANDOLPH ADAMS: I'd like to know if you included claims for mental health treatment.

MR. DUNN: We did not exclude those claims.

MR. ADAMS: I didn't see them in your table of high cost procedures, and I know that mental health claims account for a very high percentage of premiums that are collected for health insurance.

MR. DUNN: Right. It wasn't an exhaustive list of the high cost conditions. One of the problems with mental health diagnoses is that we're working with plans of potentially different benefit packages, limits, and so on. This is one of the reasons why we didn't. Much of our analysis is plan by plan. We did not explicitly throw out those types of claims—they were in if they were reported, and if they weren't, then we didn't have them.

MR. ADAMS: Did you get ICD9 codes that would identify specific conditions such as schizophrenia, manic depression, and so on, or were they all lumped together into one code?

MR. DUNN: If the ICD9 psychiatric codes we had were the principal reasons for the encounter, either inpatient or ambulatory, then that's what we had reported.

MR. ADAMS: When will the report be available, and how may one obtain the report?

MS. ROSENBLATT: Tom Edwalds has told me that the research committee needs to approve the final report, and that will probably occur soon. [The completed report, also titled, "A Comparative Analysis of Methods of Health Risk Assessment" was published in 1996 in monograph form. It is available for $35.00 through the SOA books department.]

MR. PAUL J. DONAHUE: I wonder if you reached any conclusions. You mentioned the possibility of the need for market reforms, in addition to predictors. Did you conclude whether reform worked for state programs? Would there have to be market reform on, for example, controls on marketing, in addition to the use of risk adjusters? Is perhaps one of the reasons the HIV pool had so few claims that there has been an improvement in the average monthly costs so that fewer claims qualify for the $2,000 threshold?

MR. DUNN: I can give you a quick answer. Our conclusion was that, based on the individual nonrandom group results, these models alone would not remove all incentives for risk selection or with the outcome of such behavior. Therefore, unless there is some type of measure being put into place that limits risk selection behavior, then the models themselves are not enough.

MR. KOSCIELNY: With regard to the HIV results, I think it's the criteria in the CD4 concept. Of the 14 claims that we received, I think 10 or 11 of them came from one company. It was a small company actually, but one that was very active in getting those claims. I think it's the criteria. We don't have any information to judge whether the $2,000 is a barrier or not because we just didn't get any claims.