RECORD, Volume 30, No. 1*

Spring Meeting, Anaheim, CA
May 19-21, 2004

Session 64PD
Predictive Modeling – Current Practices and Future Applications

Track: Health

Moderator: Emma Hoo

Panelists: R. Adams Dudley
Lori Weyuker

Summary: Predictive modeling continues to show great promise as new applications are uncovered and approaches are refined. A panel of actuaries and other experts provides an assessment of current and future applications of predictive modeling techniques.

MS. EMMA HOO: Many of you use risk adjustment and predictive modeling tools, both in underwriting practices as well as for member identification in disease-management programs. This morning we have two speakers who will discuss the practical application of these tools, as well as opportunities for their expanded use and development.

Speaking first will be Lori Weyuker, who has 15 years of experience as a health care actuary, both for health plans and consultants. She has focused on the application of risk adjustment to health insurance payments, as well as provider payments. She is currently a member of the Society of Actuaries Health Section Council.

Following Lori will be Adams Dudley from the University of California-San Francisco. He is an associate professor of medicine for health policy epidemiology and biostatistics. His work includes the application of risk adjustment for measuring quality and use in provider pay-for-performance programs. His current efforts include advising the Centers for Medicare and Medicaid Services (CMS) on its

applications, as well as looking at the measurement of hospital quality and efficiency in California.

MS. LORI WEYUKER: This is a case study that addresses a business problem by using risk-adjustment tools. I see these as tools to solve business problems. After this, I'll specifically lay out whom the players were, why these players were willing to participate in this scenario, what the incentives were for participation from all the different players, including the employer, and why we would risk-adjust at all. I'll discuss the results of the case study along with the public policy and financial implications. Then we'll talk about implementation challenges. This may seem really crisp and clean, but when you're actually doing it and dealing with the data, it can get very messy.

I want to give some background information. As many of you know, there's been a big focus on so-called chronic conditions and health-care consumption. There are many statistics about individuals with chronic conditions consuming a disproportionate share of health-care dollars. For example, 50 percent to 60 percent of inpatient stays are due to chronic conditions; 33 percent of total health-care expenditures are due to preventable conditions. Sixty-six percent of patients with diabetes, asthma, congestive health failure or clinical depression do not receive adequate care. If a person has one of these chronic conditions, we know he will have health-care costs due to these conditions year after year.

Some carriers feel they are not being adequately compensated for the parts of their population with chronic conditions. In addition, they would like to find some way to manage risk selection in a multiple plan option environment to meet targeted medical loss ratios, and ultimately, targeted margins.

Here are some of the long-term implications. If you have a carrier that has a disproportionate share of the population with chronic conditions, it could end up in a death spiral and go out of business. This could lead to lack of competition by health insurers and HMOs. You can end up with a scenario in which you don't have enough carriers, if the risk selection is too disparate between two or among three or four carriers that are providing insurance coverage to a large-employer market.

The first objective in this case study of a large group employer is to find a method that allows each carrier to be paid adequately for these chronic conditions for its portion of the employer population. In addition, we wanted to find a method to encourage competition based on efficiencies, not on the ability to select risk. This is a large group and there really is not underwriting per se, but there still are games that one can play to select risk in the large-employer arena. The third objective is to find a method that does not reward superior underwriting. Again, the focus is to reward efficiency and not the ability to do a good job at what I guess I would call "pseudo-underwriting."

The solution we decided to test in this case was risk-adjusted premiums. The premiums are adjusted to compensate each carrier for its respective so-called disease burden. Let's say, for example, if you have a large employer with two carriers. The carriers will have different disease burdens, depending on enrollment. You could end up with a disproportionate share of pregnancies. You could end up with a disproportionate share of AIDS patients or diabetes patients or what have you, so the purpose of the risk-adjusted premiums is to make sure that the carriers are adequately compensated for their part of the disease burden.

I'll go through some of the players and explain why they're necessary. To do this type of risk-adjustment exercise, one needs at least two carriers by definition because money will flow between them. When I say carriers that also includes HMOs. We need an employer group, obviously, which must be of reasonable size. In this particular case, this employer size was not a jumbo group. There were about 1,500 employees in the region that we decided to focus on, and we felt this was large enough to have credibility, but the minimum size is a gray area. There are other issues relating to credibility besides the size of the employer. There is the credibility of the data, but we'll talk about that a little bit more later on.

In this particular case we also use an unbiased third party to administer the risk-adjustment process. This can be done in one of several ways. You can use a consulting firm or a data warehouse that knows something about risk adjustment. We also used a data-encryption vendor. Although there are alternative approaches, we wanted to make sure that all of the i's were dotted and the t's were crossed with respect to the Health Insurance & Portability Act of 1996 (HIPAA) compliance. At the time that this project was coming to fruition, HIPAA was very new, and we wanted to make sure that none of us went to jail.

There are a couple of things one needs to obtain before going through the process. One thing, obviously, is to select a risk-adjustment model. There are pros and cons about the various models, but the main point is to have a risk-adjustment model that is mutually agreed upon by the employer and the participating carriers. That's a process that can be done pretty quickly unless the employer and/or the other carriers would really like to dig into the details about what makes one model better than another model and what kind of data are required.

In terms of data, I'll mention HIPAA quite a few times throughout the talk because I think it's very important to have HIPAA compliance. I've been talking to some folks who have been doing risk-adjustment projects, and I think people are getting a little sloppy with HIPAA. In terms of the HIPAA considerations, I divide this into two parts: data in and data out. First, you must get the raw claims data itself, or encounter data if it's from an HMO. This data must be obtained in a way that's HIPAA compliant. Then once you have the data and all the legal agreements, there are HIPAA considerations regarding what you do with the data after it's risk-adjusted.

Why would you want to risk-adjust? There are incentives depending on which of the participants we're talking about. From the carrier side, risk-adjusted premiums compensate the carriers who have a higher-risk part of the group. Let's say, for example, one carrier believes that it has adverse risk. This is a way for that carrier to feel that it will get a fair shake. The carrier that believes it always gets a fair shake or always does a great job at risk selection may not be interested in participating in this project.

Risk-adjusted premiums allow the carriers to compete on the basis of health plan efficiencies. I think there has been more focus on the efficiency of medical groups, hospitals, the delivery system and efficiency on the health-plan side as well. Employers are very much in favor of looking at the health plan and seeing some demonstration that it's efficient.

We talked about the death spiral. If a carrier is compensated adequately, there's less likelihood of a death spiral. If this works according to the theory, it should allow continued competition among health plans that may be on the edge in terms of anti-selection. Similarly, it increases health-plan financial stability. I talked to some employers who feel that knowing that their carriers will be there next year is very important. They don't want to have to change carriers every other year because they're financially unstable or they're taken over by the state or what have you.

In addition, risk-adjusted premiums allow some carriers to gain exposure to new groups, and this was true in this case study. One of the carriers was being offered to the group for the first time and that was one of its incentives for participating. It wanted to have additional exposure in the large-group market.

Finally, risk-adjusted data can be a better predictor of expected health-care consumption.

Why would you risk-adjust from the employer's point of view? The employer may be interested in being able to offer more plan choices for the employees. For this type of exercise, by definition you have to have at least two carriers. Second, financial stability of the carriers, as I mentioned earlier, is desirable to the employers. Third, and I think a lot of employers have voiced this; reward goes to efficient health plans. Reward does not go to the health plan that does the best job of avoiding risk. In this case we're not paying the carriers to avoid risk. We're paying the carriers to take care of specific parts of the population.

I'd like to talk a little bit about why health-based risk adjustment/risk assessment is being discussed now. I believe that this is an innovation whose time has come, and I believe that this technology is in its very early stages. I think it will evolve to be a lot more impressive than it is now. A lot of the early research was funded by CMS in the mid-1980s, and therefore, is in the public domain if you want to look at it. CMS finally implemented risk-adjustment to Medicare Plus Choice beginning in January

2000, and they have stepped it up and implemented it in a more serious, robust way since then.

I believe that when CMS or especially any federal agency implements this type of technology, it is soon accepted by the health insurance and medical worlds. We've seen this with Diagnostic Related Groups (DRGs) and Resource-Based Relative Value Scale (RBRVS) and so on, and this is another example of that.

Risk adjustment is a more accurate predictor of health costs. The Society of Actuaries' study looks at R^2 statistics or predictive ratios. I won't go into all that right now, but unless you have a population of 1 million, risk adjustment is a superior predictor of health care consumption.

Risk adjustment is coming into fruition now, I believe, almost entirely because of the big improvement in electronic data. Fifteen years ago, much of the research was based on survey data, which I think has some bias problems. Now the quality of electronic medical data has improved to the point that relatively stable models can be created.

I'd like to talk quickly about the steps that we took because this project took 11/2 years, about six months of which was spent with lawyers. Attorneys from four different firms were involved. First, we had to get agreement from the employer and all of its relevant carriers to risk-adjust the premiums. We had to agree jointly on the risk model and obtain census data from the entire group. This can be complicated if part of the group is self-insured. In my experience in working with self-insured groups, it is still not that easy to get a census from the dependents if they do not have any healthcare claims. But for this type of exercise, you need a list of the entire census — all of the dependents.

We had to agree jointly on a data encryption company. Data confidentiality must be ensured, and not only from a HIPAA point of view. Regulations in various states have additional data confidentiality implications. We had to ensure proper use of risk-adjusted data. Data probably was 80 percent of this project, and it's not insignificant. We had to define the data set for all the carriers to use in providing the data to risk-adjust. Each carrier had to extract its respective data and clean it. I'm putting so much emphasis on the data because if it is sloppy, the results of the risk-adjustment exercise will not be accurate. They won't accurately reflect the disease burden of that part of the population.

Each carrier sent its data to a third-party data-encryption company. The risk-adjustment calculation then took place. We had some conversations back and forth to define the algorithm for the risk adjustment among the carriers because this was a pilot project. Then we met to talk about the results of the calculation and to see if the carriers and employer agreed on the algorithm. This wasn't that bad.

We did this project in California. Northern and Southern California are often quoted separately, as if they were two different states, so that is how we did this project as well. In Northern California, only one carrier was previously offered. As a result of the risk-adjustment exercise, we have three carriers. The risk is not really that even. For example, Carrier A had a 0.85. This was the first time Carrier A had ever been offered. I believe this is part of the reason their risk was much lower than the other preexisting Carrier C. After a couple of cycles of open enrollment, I believe that this will become more stable.

Similarly, in Southern California, Carrier A was already previously offered and Carrier C was not. I think it's often the case that when employees have the decision to switch carriers, the healthy ones switch, especially in this case. Carriers A and C are different medical provider arrangements. By definition, people who wanted to switch between Carriers A and C had to change doctors. I believe there's a lot of data that shows that people who are willing to change doctors are typically healthier than an average person, and I believe this explains some of the different risk results.

Again, the objective was to encourage participants to choose the most efficient health plan. The carriers gave a large-group quote, and then we went through the risk-adjustment exercise based on the open enrollment results. These risk factors were used in an algorithm multiplied by the original quote on the premiums. So, for example, Carrier C had some anti-selection, so they received an additional 7 percent premium in Northern California. In Southern California, the reverse situation occurred. Carrier A actually lost 15 percent of its revenue because it had favorable selection. Carrier C had a sicker population, and was 7 percent sicker, on the average. In Southern California, the reverse occurred. Carrier C, which has a sicker population in Northern California, will be paid more to take care of its sicker population. This goes along with our objectives to pay the health plans for having a slightly sicker population, and this is what we did in this exercise.

What are some of the pitfalls? It required the use of individual-level health care data, which is protected by HIPAA. I can't emphasize this enough. This is a very important consideration. Detailed data requirements add to the administrative burden, so this kind of exercise is not for everybody. Additionally, some health plans are very good at cherry picking, and they would prefer just to cherry-pick. They are not interested in trying to become an efficient health plan, but they are interested in continuing on as a good cherry picker. At this point, some HMOs and capitated office-based physicians do not collect all the necessary data to do this type of exercise, depending on what kind of risk model is used. If you can't get the data, you can't do a good job with the risk adjustment. Again, some self-insured health plans do not collect all the necessary census data. If you don't have the census, you can't really do this exercise.

In conclusion, risk adjustment can provide incentives that improve efficiency in health plans. If properly applied, it's superior to many techniques currently used in

the health sector. For example, it's superior in many cases to age/sex or even just to using claims data and adding trends. If it is applied improperly, incorrect conclusions may be made, and this can be quite dangerous. An in-depth knowledge of risk adjustment and how to apply it is needed to avoid incorrect conclusions.

Risk-adjustment models are likely to improve further as data quality continues to improve. I think that we'll see a lot of really great innovations in the next 10 years on risk adjustment. To implement risk-adjusted premiums on a big scale, you need buy-in from many parts of the health sector, and I don't think this has really taken place in a big way yet. You need buy-in from physicians' groups, employers, insurers and HMOs, and from the government.

R. ADAMS DUDLEY, MD: I'll talk about where we might go further with risk adjustment. This is an area that desperately needs the involvement of actuaries. One of the things Lori said that I want to address is if you only have 1,500 people, then you might need risk adjustment. If you have 1 million, age/sex would work fine. That's only true — and this is very important as you think about this — if there isn't biased selection. No amount of sample size increase will fix a bias. I think and I hope that the future of health care is biased selection that optimizes health-care specialization.

An extrapolation of that idea that might work is that we clearly need health plans to learn how to do more to manage care properly. An extrapolation of that might be the establishment of boutique health plans. There might be health plans that focus on certain conditions. Lori mentioned that there already are health plans that focus on cherry picking. We're hoping that they will find more socially acceptable things to focus on in the future. There might be plans that integrate disease management and focus on folks with multiple chronic diseases. They will want to be that kind of a health plan and market themselves that way. You can only do that if you have some consideration for the clinical status of the patients.

I believe Gallup just did a survey of Fortune 500 executives that asked them, "What's your biggest concern right now?" The stock market is down 30 percent from its highs and Iraq is happening, and they said health care cost was their No. 1 concern. The next time — which I predict will be sometime within the next year — that a health plan comes to them and says, "You're facing a big increase," they'll be very interested in learning why that increase is happening. There are a number of reasons you can imagine why an increase is happening. Among them is that the underlying costs of care rose. That might not be acceptable, but others are even less acceptable, such as your costs are going up overall because you have a situation like the numbers that Lori found in the risk assessment before the risk adjustment in the health plans. Think about the numbers she showed you. Somebody has 1.07, and somebody else has a 0.85. That's roughly a 22 percent difference in the underlying cost. You know that in the absence of risk adjustment, the plan with the 22 percent healthier patients would not say, "I have lots healthier people. Here's an extra 22 percent."

As a way of explaining to employers what's going on, or at least being able to talk to them about what's going on — because their human resources (HR) people might have some sense of what's happening as well — undertaking risk-assessment and then talking to them about potential solutions could be very important. I think this has now for the first time reached the level that CEOs care. Before, it was hard to get something complicated that might require a little push, such as better data from the health plans or from the medical groups. It was hard to get that because you wanted the CEO or that someone at a senior level to do the pushing, and they weren't really interested in this topic yet. Now I think they're finally interested. They finally got to the point that they understand. So I think the future for risk adjustment is very bright, and it needs very bright people to understand the methods, lead the way and help people understand how they can use it.

I'll talk a little bit now about the current state of risk assessment/risk adjustment. I'll describe some of the perceived limitations and what people think is the ceiling of maximum performance. I'll show you that the perceived ceiling isn't correct. Then, I'll suggest some future directions.

In terms of the current state, there's been talk of using patient surveys and asking them about health status, but I think use of diagnosis data and pharmacy data has become more prevalent. Diagnosis data is the most widely used, in no small part because of Centers for Medicare and Medicaid Services (CMS). When the "big dog" picks something, that becomes the dominant method. CMS has said, "Tell us whether or not you have people with these diagnoses." I'll tell you a little bit more about the evolution of CMS's methodology in a second.

Primarily people use prospective risk adjustment — that is, they take last year's diagnoses or last year's information, and they say, "We'll predict costs for next year." You can get pretty decent predictive ratios in select conditions. If the prospective model predicts the average cost for breast cancer will be X, then you say, "What was it?" You can get it within a few percentage points of X. Overall, the relative predictive power is low because the breast cancer patients don't allocate randomly among the plans. If they did, you'd be OK. So, the predictive ratios give you a misleading picture for how good risk adjustment is right now, using just diagnoses.

For most models, the R^2 — that is, the summary of how much of the variance do we explain — ranges from 0.10 to 0.18 for prospective risk adjustment. There hasn't been much research on the underlying behavior that leads to these big differences that we mostly can't explain. Lori showed you big differences when she looked, and I tell you that the models don't explain them very well. We should also identify the underlying behavior that causes this situation. Should that behavior — or what we think of the types of behavior that might be happening — change our response?

I want to talk about perceptions and perceived limitations. One of CMS's main consultants, an economist named Joe Newhouse, has argued that only about 20 percent of the variance in total expenditures is predictable, which may leave us feeling a bit impotent, faced with the amount of variation that we actually see. But I think that's probably not correct, and I'll show you at the end why. One of the contributors to that low number is the preference to use prospective versus concurrent models. Concurrent models would say, "What's your data for this year? Who had what this year? And we'll make adjustments for that." It's probably intuitive that if I'm allowed to say somebody had breast cancer this year, I can get a better prediction for what their average costs will be than by saying the incident of the breast cancer happened last year, and now I have to predict their follow-on costs. That's a more variable number than the cost of work-up in initial treatment for breast cancer. On top of that, the concurrent method captures acute events. For example, I had cholecystitis, so I had to have my gall bladder removed. Next year, that's won't cost anything, but correcting this year allows me to get better predictive power.

There are reasons that concurrent isn't used. The first is that the cholecystitis example I just gave you creates a bounty. If we've used concurrent, for which prices are higher for any given thing, and then the sad person goes from a little sad to depressed and there's bounty for depressed — it's bigger than it would otherwise be. Also, health plans or anyone else who is being paid or receiving a risk-adjusted payment generally prefers to have a stable pot. They want to know what that pot will be. I've heard people say, "I can manage risk. Give me an amount of money, and I'll manage the distribution of those dollars. But don't give me the uncertainty of what could happen clinically and uncertainty about what could happen in terms of the size of my pot over the course of a year." What we need then is better predictive power, some way to get the pot size to be maybe not perfectly predictable, but pretty predictable and some way to deal with not inviting bounties — not inviting people to find things that aren't there that pay a large amount of money.

How can we do that? The alternative to these two separate directions — viewing prospective and concurrent as two separate directions — is to mix them. You could start each year and say, "We'll take last year's data and give you a prediction for what we expect you to get." Then on top of that, for a few conditions for which we have a sense of how often they happen, we'll tell you ahead of time that we'll do concurrent risk adjustments.

For example, even at 1,500 people, if you know their ages and sexes, you have a good estimate of how much heart disease they will have. You could say, "We understand how much heart disease they're likely to have. What we'd like to do is make sure that the dollars are allocated. If there happen to be more heart attacks, we'll make an adjustment to reflect the fact that you have more heart attacks or more bypass procedures or whatever it is that we're adjusting for." On heart disease there might be a little variation, but another common health risk is cancer.

We know there are a lot of cancers, and we can get a pretty reasonable estimate of the risk of cancer in a population of a decent size. If we can mix the concurrent and the prospective approaches, maybe we can do a little bit better.

To get away from the bounties, where people are looking for things that might not be there because they have big price tags, we proposed — and then did a project that I'll show you the numbers from — that those things that are concurrent should be verifiable. So, mild depression would not be included. That doesn't make sense because it's too gameable. But let's return to our breast cancer example. If there isn't a slide with cancer cells on it from someone's breast, then someone committed fraud and that's a much higher level. People are less likely to be willing to do the kind of gaming of claiming someone had HIV but being unable to produce a positive HIV test than they are to say these allergies are quite severe and need treatment or this sadness is depression and warrants a diagnosis and a risk adjustment with it. So, make them verifiable. Of course, they must be expensive or it isn't worth the effort, and they have to predict expenses in subsequent time periods. For lack of a better acronym, we call these the VEP conditions – verifiable, expensive and predictable.

To show this is a viable method — and we shared this with CMS — We selected 135 candidate conditions from the entire *International Classification of Diseases, Ninth Revision,* (ICD-9) codebook. We took data from 21 health plans, and we said, "For this year — because we're looking for things we can do concurrently — what are the 100 most expensive?" We took those 100 just as a round number to demonstrate it. It doesn't have to be 100, and in the end, CMS would pick 68. Then we took an entirely new data set that had patients' data from two years from seven health plans, and we said, "How would it work if we applied the prospective method that's most commonly used, the concurrent that has better predictive power but is almost never used, and the hybrid?"

We have 320,000 patients. The remaining population consists of people who don't have one of these VEP 100 conditions – again these are determined to be verifiable, expensive and predictable. You can see their relative cost weights of 0.57 versus 5.23. What you can't see is how many patients they represent. That's important because only 9 percent of the patients had one of the VEP 100 conditions. So, that's a relatively small number of patients. The important thing is that they had half of the total cost and most important of all, they had 83 percent of the total variation in cost. So, 9 percent of the people who have these conditions that are verifiable explain the huge majority of the variation in cost, which is what risk selection and risk adjustment are all about. Now, instead of doing what both DCGs and all of the other software that's out there for risk adjustment does, which is to deal with every ICD-9 condition, we're focused on the 100 most important. That makes life a lot easier.

For patients with one VEP 100 condition, their average cost rate was 3.58. If they had more than one, their average cost was 12.33.

We did several models. The first model is just age/sex prospective. For the second model, we took what Lori was using — the DCG system has the hierarchical co-existing conditions (HCCs), and it's what Medicare is currently using — and did it prospectively. We got an R^2 of 0.08. When we did it concurrently, we got 0.37. This is exactly what we expected — the difference in improved predictions with prospective versus concurrent. When we did the VEP 100 dummy variable — I make a binary variable for each of the 100 conditions — in the VEP 100 patients only, we get an R^2 of 0.21. That's because we're only predicting within those patients. When we use HCCs within those patients concurrently, we get 0.33. Those are about numbers we expected. The VEP 100 dummies have less information. The HCCs use all of the ICD-9 codes, so that's why they can do better. We knew they would do better than the VEP 100. We're OK with that.

Let's now look at combining these approaches. If we run the HCCs prospectively for the people who are in the remaining group — that's the 90 percent of patients who don't have a VEP 100 — we get an R^2 for them of 0.07. If we take them and the predictions for the VEP people and mix them together, we get an R^2 that is better than either one. The hybrid is 0.26. The reason for that is, now I have this population in which, on average, the remainder are low risk and the VEP are very high risk. When I was trying to predict just within the VEP, I got a 0.21. But now I've added the knowledge of the difference between them as part of my prediction, and so I have a better R^2 . So I've added the fact that they are these two populations with a very large difference in mean between them, and end up with a better prediction.

If I do the same thing for hybrid with the HCCs — which contain more information — for the VEP 100 patients, then I get a 0.36. So interestingly, the hybrid has almost exactly the same predictive power as the HCCs applied to everybody with the conditions. The R^2s are almost exactly the same. All we did was identify the 100 important conditions, and they're verifiable. On those we'll use this concurrent approach because it's not as gameable as it used to be, and everybody else will still do the prospective. We have a pretty stable pot of money that we can offer people because for 91 percent of their patients, they know ahead of time what it will be. They have some sense of how often they'll get people in this other pot, and the overall predictive power is almost the same as just having applied HCCs all the way across. In terms of fixing the various incentive problems that we have, these represent big steps forward.

And it's complicated, which means they need actuaries. They will not just say, "Why don't we just do a hybrid of prospective and concurrent?" They need help. They need people to explain to them in ways that they can understand the incentives that their plans face and how we can fix them.

By doing the hybrid approach, we address the weakness of prospective by getting a better predictive power and by reducing the gaming and providing stability.

To complete the story about the evolution with CMS, we presented this idea and there was concern about the data burden. So, we went back and forth, and they ended up with 68 — but their 68 were from our 100, except for trauma. They added trauma. We thought that you don't need to risk adjust for trauma. Nobody knows that it will happen. But they found that when trauma happens in an old person, the costs go into the next year. Someone who is 78 years old and involved in a bad car accident may have problems for a long time, so they were right as far as I can tell.

Now we have a list that we can focus on that represents 84 percent of the variation. To do more for the future, we need to learn more about exactly how the variation within each condition can be explained. If I'm a pulmonary doctor, a lung doc, I know that how I treat someone with lung cancer is a function of the stage. There's 1A, 1B, 2A, 2B, 3A, 3B and 4. For 1A, a small cancer in the lung that doesn't touch anything else, I'd recommend that we remove that and you go on your way. If they have a 3B cancer, that requires preoperative chemo to shrink the tumor, surgery and then postoperative radiation. It's very expensive. But if I know it's 1A versus 1B, 2A, 2B, 3, I have a treatment plan – they vary a lot in cost, but I know what those costs are.

ICD-9 codes do not capture any of that information about lung cancer staging. Instead of just treating lung cancer as one big lump, the next step is to say, "There are categories of lung cancer, and we could do a much better job of allocating the dollars by going within them." As an example, Medicare tried to do this with end-stage renal disease (ESRD), which on our list of 100 is the worst candidate for doing that. Here's why. If you get ESRD, your kidneys don't work. The definition of it is that you need dialysis. Almost all of your costs are associated with dialysis, so the potential to do much risk adjustment is actually pretty low. There are no stages. There's no numeric number of how bad your kidney disease is. It's just as bad as it can get. Most people start at $50,000. There's some variation around that, but it's a little bit difficult to predict that variation. Medicare found that they could get an R^2 of 0.25 by using "how long have I been on dialysis" and "did I have diabetes that caused me to get here" and a couple of other rare conditions.

For other conditions, I think we can do better. But within the 100, if we can get to an R^2 of 0.25, how would we do? This is what your R^2 would be from zero to one, as a function of the percentage of your predictive power within the VEP 100 patients. If we're down to zero — we just use the VEP 100 dummy — we get the R^2 that we saw, which was 0.26. If we could predict everything for them — if we project 84 percent of the variation plus a little bit more because we do a little bit in the rest of the people — we're up to 0.85 or 0.86 or something like that. If we just do what Medicare showed they could do in ESRD, we get our R^2 all the way up to 0.42. These are huge steps forward in our ability to tell people how much this will cost or how much the patients that you have will cost to manage. If it's the employer, here's why this is happening to you. Here's why they need more resources. Yes, that price increase is actually probably fair.

I want to jump to behaviors. You probably talk to HR people who grouse sometimes about health plans and about the behavior of health plans — trying to get the healthiest patients or to dump the very sickest or in general just being difficult. You can imagine how any of these would lead to risk selection potentially. But it turns out that the risk selection it leads to might be different, so if you just are out there cream-skimming, what happens to your cost next year might be different than if your plan's main modus operandi is, when somebody gets really sick, to dump them.

No one has ever looked at this before. We took data from Medicare. We said, "How much of a risk selection benefit can plans get?" Medicare plans were found over and over again to have about two-thirds of the expected cost for the fee-for-service population. Just as an example, we said, let's imagine there are a set of plans that have a risk selection of 0.8 and a set of plans that have a risk score of 1.2. What would happen to them? Let's look at how they got there, and whether it matters in subsequent years.

Using health plan data, we created dumping algorithms. If I wanted to dump, how would I do it? Who would I get rid of using this year one data. If I wanted to cream-skim, whom would I be getting? We sampled as if I was a pretend cream-skimmer and I know that on average I'll get to 0.8. Or I'm the person on the other side from the cream skimming, and I know that on average I'll get to 1.2. We did 1,000 simulations of each of the different kinds of behavior using 1991-92 data from seven health plans. The difference is that the prices I am about to give you or the losses and gains are probably about double now. If you use the demographic risk adjustment model — just age/sex — and you have dumping, skimming or stinting going on, the losses to the plan that wasn't engaged in that activity are in the random behavior column. Everything is close to zero. If there's dumping, the dollar losses to the plans in the early 1990s would have been about $75 in the next year from dumping, but $130 in the next year from skimming. If you dump that patient — they're really expensive this year — to get to my 0.8 to 1.2, next year their expense difference goes down because often what they had is getting better by then. If you get there by skimming, your younger people will still be younger the next year, so that difference persists a little longer.

What if I try to fix it? If I use prospective HCCs, that actually works pretty well because the correction that I'm making is for relatively minor things. It's a bunch of young, healthy people who are moving around, and the proportion of young healthy people that you have is what's changed. However, if you're dumping or stinting — that is, dumping the very highest cost or in general being difficult with people — those differences persist more with prospective HCCs.

If I use concurrent HCCs, I fix some of the differences, and if I use hybrid HCCs, I also fix it some. In fact, in this particular instance, it turns out that you can fix it even more. You want these things to be zero. You don't want to be paying people extra or too little. It turns out that the underlying behaviors that lead to the risk

selection do matter. What does this mean? This means that employers and health plans need actuaries to explain that to them and to show them how it could work, especially if we ever get to a world in which we actually want skimming or we want reverse dumping, where people are trying to get certain types of patients. You have to explain to them how that will matter and, therefore, the best way to fix that. Employers want the resources to be there for fair care, for real need, and then they don't want to pay anything extra.

The only way you can give them that is to be able to understand the forces causing people to move. If you talk to HR people, often they have an idea. This is an example from real life here in the University of California. They'll say, "Health Net offers health club memberships." One time at UC we had a zero contribution plan, and then you pay the dollar amounts above that plan. Health Net passed PacifiCare. PacifiCare had been 69 cents less expensive than Health Net, and they switched by $1.01 per month. For $1 a month, 11 percent of the UC population changed plans. You can imagine the cost of doing all those switches was so much more than $1 a month. The HR people often will have some sense of what's going on. That was a cream skimming. That will have implications for how I should go about risk adjustment.

We made software so you could do simulations on predictive modeling and alternative approaches. You can actually say how this would play out for them. You can give them different predictions. These are 1,000 dots on each graph, but it shows different predictions for where the costs are likely to end up falling out, so you can explain to people in ways they can see what behavior matters. You get a different graph for each risk selection behavior and for each risk adjustment approach. You can't give them an absolute number, but you're trying to give them some sense. Obviously the more that you get lines where costs and predicted selection add up, the better you're doing. There are ways that you can make this visible to people, make it make sense to people, make them understand why they should care about it and then tell them what to do about it. I think what makes it exciting to be doing this now is that we're making big steps forward, and there's an awful lot of need for your help. There is also an awful lot of potential help for you to offer.

MS. HOO: We have about a half-hour for questions. If you could, please announce your name and organization before you present your question.

MR. CHARLES S. FUHRER: Chuck Fuhrer, The Segal Company. I have a question for Lori and a follow-up comment, if I might. You said that under a slide that said "predictive power," "Risk adjustment is more accurate than what is currently used." Can you explain to me your understanding of what's currently used?

MS. WEYUKER: Yes, I was referring to a couple of different studies that I have done using health plan data and also to literature. These studies compare actual to predicted claims using some commonly used methods — for example, age/sex. You

can get age/sex factors from your Milliman book or what have you, or you can print your own age/sex factors. So I was talking about the case of using age/sex factors as opposed to using risk adjusted factors, which are the factors that come out of using the risk model or even using prior claims data. When you look at R^2 statistics or predictive ratios, which are predicted claims divided by actual claims, the risk-adjusted data actually comes closer to predicting next year than the other methods.

MR. FUHRER: Let me tell you what most people in this room would answer as to what method currently is used. The first thing they do is to look at prior claims. The next thing they do is to cap those claims at some point, which is often called the pooling point, to cut down on the variance. The next thing they do is take a credibility average or weighted average between that and the age/sex predictor.

In 1988, I published a paper on how to get those credibility factors. My criteria was to maximize the R^2. In addition, though, it was somewhat innovative and somewhat a precursor to risk adjustment in that I looked at the correlations between individuals within the group in coming up with my R^2s and therefore was optimum in the sense that it took into account the effect of health conditions, at least in terms of what the prior incidence was. The selection of the pooling point also was done in that paper in a way to optimize the R^2. Furthermore, at a couple of meetings I pointed out that the method could be further improved by giving greater weight to those prior claims that occurred more recently, as well as deleting those claims that were resulting from people who had left the group or had died, which I guess is really the same thing. I didn't do the R^2s for that last modification, but I have to tell you that I got some pretty high R^2s in that 1988 paper. I don't know who is using that particular method; it never made it to the Society's syllabus, but informally I think some people are using it.

MS. WEYUKER: I just wanted to make one more comment. There are some risk-adjustment vendors out there who use the timing of the claims as part of their prediction calculations. For example, the claims that are six months old or newer have a higher weight than the older claims. So maybe they're combining these two different technologies.

DR. DUDLEY: And Chuck, what happens to the dollars above the pooling point?

MR. FUHRER: The usual procedure, of course, is to ignore them. But in order not to have bias in your answer, they're replaced with an average expected cost over that point.

DR. DUDLEY: The R^2s you're looking at don't lop off the dollars because somebody has to pay them. In the real world, what often happens is that reinsurance is purchased. Often the reinsurance has a pretty significant load on it. To the extent that you can get the actual predictions without lopping off any dollars — without having a point at which you stop things — closer to the actual events, the less reinsurance you have to buy and the less load you pay on it. The more that we can

get the actual predicteds right to the number, which means we have got to have more spread, the less that everyone else will be paying to people who are in the reinsurance business.

MR. FUHRER: Actually, you don't need to reinsure those amounts. You only need to increase your predictive power and increase your R^2s by using the unbiased capping at that point. The cost of the reinsurance is really not relevant to the discussion.

DR. DUDLEY: It may not be relevant to the discussion of the R^2. It's very relevant to the discussion of how I as an employer allocate my dollars and what total I will pay. The more that we can get to a place where health plans believe that they'll get what they need without having to play games, the lower the costs will be for employers and probably for health plans. Right now they must spend a lot of money doing advertising to help them get certain kinds of patients that probably has little to do with what actually happens on the ground with care. That's mostly just wasted money — not for them because they get profits from it, but for society and certainly for the employer.

MS. HOO: There are a lot of different types of uses beyond predicting cost and setting premium. A lot of plans are also using these tools to support their disease-management programs and there may be a different focus around identifying thresholds levels over which members can be stratified into specific interventions. Why don't we move on so additional folks have an opportunity to ask questions.

MR. FUHRER: One comment more. I would like to see a real comparison between the R^2s of the predictive modeling versus the actual method that's used.

MS. HOO: I'd like to make just one other note. Bill Thomas has a paper coming out in *HSR* — I think this summer — that compares a number of different tools and their predictive power.

FROM THE FLOOR: I'm Marilyn Kramer, and I'm from DxCG in Boston. I want to echo your point about using the tools for medical management and also to reference the paper that I wrote with my colleagues last year on looking at using Monte Carlo simulations, as Adams described, to look at the pricing and how you price various groups using the prior costs with some pooling as well as age/sex versus the diagnosis-based models. We hope to expand on this study this year with information on real-world employer groups because the reality I wanted to highlight was, as Adams said, that the world is full of biased groups. People make selections as they choose employers, and as they choose physicians, as they choose hospitals. Our world is surrounded by biased groups that we need to take into consideration.

But I did have a question for Adams. You used primarily Medicare data, and our experience says that the commercial population — under 65, privately insured — is very different than the Medicare population, both in terms of the number of

conditions they have and the variation between the healthy and the unhealthy. I wanted to see if you could speculate on what your results might be with a commercially insured, privately insured population.

DR. DUDLEY: As I'm thinking back, I was unclear on that. The data all came from the commercial insured population. I was talking about in the context of choices made by policy-makers, using CMS as an example because they're the big dog and because I do think that now that they've committed to working with Maryland, risk adjustment is a much more real phenomenon. I was talking about CMS and how it's gotten to a pretty similar kind of place conceptually, but the data are from commercial health plans that I presented.

FROM THE FLOOR: I was wondering if either presenter had any knowledge of people using the prospective model in a different twist to look at what cost is potentially avoidable. Are people starting to intervene with providers to try and get the cost down instead of trying to disperse the pool?

DR. DUDLEY: I'm sure that the disease-management people would say that's what they do, that's their whole business. In addition, I know that behind the trend toward using episode-based assessments of quality and cost of care is the rationale that at some point we want to intervene. Whether or not people have successfully gotten to the point where they can take the data and turn it around fast enough to be able to intervene is not clear. I think even the disease-management people would say that probably there is some time lag between when they identify someone and they actually start saving you money.

MS. HOO: At the Pacific Business Group, we've been engaged in work to assess physician performance at the individual level and looking at both quality and efficiency. In your mind, how might a plan think about marrying risk assessment data with evaluative measures that could differentiate providers on quality and efficiencies and the effectiveness of their interventions?

MS. WEYUKER: I'm not sure if this answers your question, but I have been thinking for the past couple of years that some commonly accepted metrics need to be created, sort of vis-à-vis National Committee for Quality Assurance (NCQA), and I haven't seen that happening. How to create these metrics so they would be widely accepted seems complex because there is variation among risk-adjustment models. How that actually will happen, I'm not sure.

DR. DUDLEY: One of the things that's happening is that episode evaluations are focusing on things for the most part for which quality is assumed not to matter or where use is bad quality. For example, there's been a lot of focus on the treatment and the use of antibiotics in acute bronchitis or acute pharyngitis and so forth. The thought there is that the main problem is overuse, everyone will get better and we can just sort of drive utilization down. What has not yet happened because we don't yet have the data — and it's a data problem — is to go to some of the examples I

used, such as lung cancer. Adams says it has seven stages, and don't just take his word for it. Go to the American Thoracic Society and ask them what should be done for the different stages. What we then need is to look at the costs of patients treated in those stages and the quality of care.

Unfortunately, that currently requires more data, namely data from different buckets from which we usually get data. It requires the kind of clinical data that we mostly haven't gotten. That's exactly why I wanted to come up with a VEP 100 — to say, "We won't do that for everything." We won't do that for bugga-bugga disease, but we can get started on priority conditions if we can identify what those conditions are, and it's reasonable to identify them by their clinical impact and their costs. For the most part, unfortunately, right now we're not to a point where we can really marry data for significant and expensive conditions.

FROM THE FLOOR: Can you comment on the fact that you might hear from the physician groups or the insurance carriers, "Look, we did a good job. We have a good method of managing some of the chronic illnesses. As a result, our risk sort of goes down in the following year, and then we're being penalized in getting a lower premium for that."

MS. WEYUKER: I view this whole scenario as something that's fluid. Let's say, hypothetically speaking, that the risk adjustment was implemented in a situation you're talking about. It's not necessarily done in a neutral way. I believe the risk adjustment has to be fluid and follow these decreases that you're talking about. It can be done in a way that it's budget neutral.

FROM THE FLOOR: Especially when you look at the chronic conditions and you try to stage them, aren't you essentially punishing people for having better efficiency and showing a lower stage of a chronic condition?

DR. DUDLEY: Risk adjustment is part of the fix. It's not the whole fix to the incentive problem. If we thought that we as people who are interested in actuarial numbers would fix this thing, we need to lower our opinion of ourselves. An important addition to that is to make some other incentive systems that give people a reason to control chronic conditions. This is a marriage for you, Emma. Marry risk adjustment with performance-based payment, which may be something with which you folks are a little less familiar. Say, "If you control diabetes, then the incidence of complications and such will fall, and therefore your risk-adjusted payment next year will fall." But that will be accompanied by payments that are higher for health plans that do a better job of controlling diabetes and prevent the long-term complications. Then you must figure out through negotiations who shares how much of the gain, but you want to create the gain in the first place. You absolutely have to get diabetes controlled to have gains to be shared.

FROM THE FLOOR: We take both into account. Obviously, the issue is that to some degree you are still — even if it's 30 percent to 50 percent — recognizing the risk-adjustment score, and they don't want to hear that at all.

DR. DUDLEY: In general, people are very resistant to changing the way they are paid. But I think we're at a point now where there's enough conviction at very high levels of society that we need to change the way people in health care are paid and change what data we require them to collect and present to the world. I think there's an opportunity to move this in the right direction. I sometimes use the "Star Trek" analogy. You know that 100 years from now, we won't have any health plans that don't know who has diabetes and how severe it is. We don't know when, between now and 100 years from now, we'll get to that point. I think the main determinant is when we start. It's technically doable right now. It's a matter of will and presenting to people whenever they run into a problem, "We can provide you with a solution. We can figure that out." That's what you folks can do.

MS. HOO: I would add also from a purchaser perspective that to the extent that the premium base is lower, it allows a plan to be in a more competitive position in a commercial marketplace to gain new enrollment that sustains better risk mix over time. I think there are opportunities as well as looking at the investment that you make on the front end to support overall population health improvement.

MR. ROBERT B. HARDIN: Bob Hardin from Gen Re. If you're ultimately successful and you're able to accurately predict the costs of various plans that are offered by an employer to an employee group, what are the consequences of that in terms of how the employer should think about sharing the cost? Would we use any of your data or information to help the employer figure out what employee contributions should be?

MS. WEYUKER: Yes, I think that's a really interesting point. The other half of the case study is showing so-called risk adjusted contributions. I think through risk-adjusted contributions, an employer can really use a specific strategy to channel their employees to a specific health plan. They can do that for many reasons, such as having better contractual arrangements with hospitals in a certain region. I've seen some health plans have discussions about picking and choosing certain provider-group specialties that they want to really funnel employees to specifically, and I think that some of this can be done with risk-adjusted contributions.

FROM THE FLOOR: If you're not careful with a risk-adjusted contribution strategy, you could have a high-risk group of some sort that would be penalized.

DR. DUDLEY: Actually, risk adjustment is the only way that you can be fair to those people. I mentioned the big stampede of people from one health plan to another, and you see it during the death spiral — actually now we've had several. We started out 10 years ago with nine choices and one by one, one of them gets identified and spirals down. During the end of the death spiral for one of them

recently, the per-month premium was more than $300, and still 453 people signed up. So we called them all because we just thought they didn't notice. We thought that maybe they thought it was $4.53 or it was $453 for the year or whatever. That small number of people, that tiny slice, wanted to pay it. You can never set something up that's fair, at least if you consider "fair" to mean that people shouldn't be penalized for being sick, until you first do the risk assessment and get the dollars allocated to the plans to reflect sickness. With what's left over, then you can talk to the employees about what they should have to pay. Think about medical savings accounts without risk adjustment. Basically it's a place for healthy people to store money, and it's not very fair to sick people.

MR. KEVIN M. DOLSKY: Hi, Kevin Dolsky from Actuarial & Health Care Solutions. With regard to future applications of this technology, one of the issues in the area of disease management is defining ROI calculations, proving whether the stuff is working. Do either of you have any insight into the application of this technology to ROI calculations for disease management programs?

MS. HOO: There are a number of vendors who are using these tools to look at changes in their population status and establishing better measures of baseline experience. It is difficult to quantify the component that's associated with regression to the mean, as opposed to the effect of the intervention. But the Disease Management Association of America and others are working on better defining methodologies. There's a paper that Tom Wilson and others published as part of a consensus process in December in *Disease Management* that speaks to some common methodologies and use of tools to this end.

DR. DUDLEY: Part of the reason that I'm hoping that we get more detailed clinical data in the future for specific conditions is that it will help you deal with regression to the mean. The reason we can't tell who among the diabetics is actually chronically sick and who was just having a bad time and will get better — or asthma or any of the things that we're doing — is that we don't have enough clinical data about them to know what's wrong with them and whether that's a fixable thing. Heart failure is a big thing for disease management. Some people have just had a myocardial infarction and their ejection fraction is low — they have heart failure — but you should expect it to get better clinically on its own. For other people who have had, for example, viral myocarditis, and it's been around for a little while, it won't get better. Or if it's from valve disease, it won't get better. But we can't pull those people apart right now. They're all just listed as heart failure in the ICD-9 code.

Getting more clinical detail about people would help us deal with making sure that disease-management programs take the right people in and then with helping them demonstrate their ROI. Then you could look at people who didn't have it according to their severity rather than their selection and see what actually happened with people who are in the disease-management programs.