RECORD, Volume 26, No. 1*

Las Vegas Spring Meeting
May 22-24, 2000

Session 125PD
Risk Adjustment Implementation Issues

Track: Health

Moderator: JOHN M. BERTKO
Panelists: MEL INGBER†
CRAIG N. SCHMID
Recorder: JOHN M. BERTKO

Summary: Risk adjusters were first used for the Y2K reimbursements to Medicare+Choice laws. How did plans fare? What was the financial impact to the Medicare program? What changes are expected for the future? What can be learned from the Medicare experience and applied to the use of risk adjusters for other types of health plans?

Participants discuss these topics and share their own experiences with the practical use of the Medicare risk adjusters.

Mr. John M. Bertko: I'm the chief actuary of Humana. We are going to talk about risk-adjustment implementation issues. Let me introduce our two speakers. Mel Ingber is with the Health Care Financing Administration (HCFA) and I have known him for probably five or six years. Generally, he and I are on opposite sides of the table, but he commands a great deal of respect. He has a Ph.D. and is a senior economist at the Office of Strategic Planning. These are the people who do the heavy thinking at HCFA, or at least one of the groups that do. Mel, by training, is an economist. He has been working on capitated payment systems for Medicare for the last six or seven years. He has been really the lead analyst on the development of risk-adjustment systems.

Many of you might not be aware that HCFA has really worked in this area for nearly 20 years in terms of the funding researchers. Mel is responsible for developing, implementing and looking at data collection, rescaling of rate books, and almost anything that has to do with risk adjustment as it affects Medicare today. He also has done some prior work on analysis of hospital outpatient, diagnostic-related groups (DRGs), and health care reform.

Craig Schmid is a former colleague of mine. I moved on and Craig is with Reden and Anders in the San Francisco office. He's a senior consultant and actuary. I can

†Dr. Ingber, not a member of the sponsoring organizations, is Senior Economist of the Office of Strategic Planning at the Health Care Financing Administration in Baltimore, MD.

only give Craig my highest compliment. In fact, he was the smart guy behind all the interesting work that I did for more than three years at Reden and Anders. Craig, prior to that, spent eight years at Blue Shield of California in its actuarial department. Craig has had a lot of work on the other side of the fence, working for health plans and doing Medicare+Choice (M+C) work.

In particular, he has worked on implementation issues, as we begin to move from inpatient-only risk adjusters to those that involve more data, sometimes called the full data models. Mel is going to be talking about what those models are, and what HCFA is thinking about them. Craig, through a project that's called the Joint Purchasers in California, has actually been figuring out**:** Is the data there? How does it work? He has some real life war stories to tell us.

Let me set a context for all of this. The first thing to note is some work on risk adjusters that has been going on for five or six years. There have been various implementations on a small scale until just recently. As of January 1, 2000, HCFA has begun the phase-in of risk adjusters for the M+C program. This is, I think, one of the new and important payment mechanisms of the new decade, if not the new millennium. We've had prospective payment systems for hospital and resource-based relative value schedule (RBRVS) for physician payment. This now is something that I think is potentially of equal importance to health plans for payments. Somewhere down the line this might be of just as much importance at the physician group level. Models in place now are inpatient-only models. We need to move beyond that to models that have greater predictive power, which better explain the risk differences between various populations. Mel is going to talk to us about that.

Dr. Mel Ingber: We are doing risk adjustment right now, which is a small miracle in and of itself in Medicare. I worked on it for many years thinking that this was a little annuity that I would be continuing to work on forever with nothing ever happening. It would just keep me occupied in a corner of the office for a while.

Then the public mood changed. The Congress picked up on this public mood. Suddenly, managed care plans were no longer so wonderful, and it gave an opening for introducing a way to think of many ways for them to cut payments. Risk adjustment is not a way to cut the payment. With all that happened in the Balanced Budget Act of 1997 (BBA), which I won't talk about, the fairest thing to the industry is risk adjustment. That's the one thing that is neutral. If you have unfavorable risk, it pays you more. If you are a favorable risk, it pays you less. But the only way it snuck in was because people were getting a little bit annoyed. People were calling their congressmen. So it happened. They "chickened out" a little bit, and we've had a phase-in that is pathetically slow.

Let me just go over very briefly what it is and what we are doing. The first thing is that everybody thinks we have some anointed next model out there, and we don't. I don't know which model we're going to use, and I don't know if we'll use one of the models. My position is that we can learn from all of them and put together something that maybe incorporates facets of each. We are still funding development of models. If I had to take one I had lying around off the shelf, I

could make it work right now. We're really at the point of flattening out the improvement of these models and just deciding where the tradeoffs are, what information you use, and how to use it. We're ready to go.

We are collecting data starting in October for a new comprehensive model. The plans are supposed to be ready to start sending physician encounter data in the door starting in October, and the hospital outpatient data starting in January. I can't say that those deadlines won't slip. I sure hope they won't, because it will slow things down a lot. HCFA has to actually implement by 2004. The BBA moved all the deadlines way, way up, and so we have to know the methodology in January 2003 and have it pretty much locked up. If we back out of there, you can see we really have to make some decisions fairly quickly.

I want to review the method we're using now. The principal inpatient model (PIP) was developed by the Boston University people and expanded a lot by the people at Health Economics Research (HER). That's one model that has been funded by the government essentially from day one. It was a model that we looked at last time around as we were developing more models. It was a base model from which we wanted to see how far we could go beyond this by using other information. We were not intending to implement this model, but we have because we have it, and it was the easiest thing to do to get inpatient data at the start. There's really nothing wrong with this kind of a model. It just begins to slice, only at the very top, a portion of the top of the quite expensive people who spend a lot of money.

By our estimates, 30% of the money is in the people for whom we are going to pay extra due to the PIP model, and that's nowhere near 30% of the people. It's more like 12% of the people. When it was developed, it initially created the standard grouper if you're going to base things on the International Classification of Diseases-9th Revision-Clinical Modification (ICD-9) codes, putting people into disease "pots." The number of disease "pots" is going to change next week when we get the next report, but it had 172 diagnosis groups. It took all the people and ranked them by what diseases it had in their total expenditures.

We haven't taken these models and tried to predict subsets of the services. You can do it. There's nothing, in principle, wrong with it, but we really are interested in predicting the whole ball of wax, as it were. These diagnosis groups are aggregated. There are groups that say you have either no diagnosis or a trivial diagnosis from a hospitalization. I don't mean to belittle it. but it is trivial based on the ability to predict much in the future.

All these models that we're using are predictive of future utilization. We also have models that are concurrent models. It doesn't make much sense to do a hospital-based model concurrently. It's just like paying the DRG rate in some sense, but all the models we're working with, and we work with both kinds, have an R^2 when we get to the next level, which is around 10%. These models don't get R^2 anywhere near that high when you're predicting the next year.

How were these all determined? You pick a physician panel, and you get a different grouper. They had a group of doctors who were mostly in the Boston area. Many

of them were affiliated with Harvard. They were doing the slicing and dicing. As usual, when you're building these groupers, it's like sausage, you don't want to watch it being made. I've been in some sessions where we have tried to decide what goes into which pot. It can get pretty ugly, but the examples are clearly expensive diseases in which the diagnosis ended up in PIP diagnostic cost group 26, which is metastatic cancer. The highness of that number was reflective of the number of dollars in thousands these people cost in the data that were being used.

The costs were predicted for these people for the year after the diagnosis from the hospitalization. They were predicted to cost at a rate of $26,000 a year. They might not live longer than a year, but their monthly expenditures averaged out to be that much. Congestive heart failure is a very popular disease, and very expensive if you look at the people who were hospitalized in the fee for service (FFS) system, but they're not all that different from what we've observed in managed care.
A much lower one is the ever-popular pelvic fracture, which is one of the fractures that the elderly get. A lot of those will result in future utilization increments. There were some diseases that people get hospitalized for that we did not take account of in terms of paying extra. The ones that are described here are the kinds to raise questions about. You might ask why would a person be hospitalized for this? It might look very suspicious, or it might be an appendicitis, which predicts very little for the next year, even though it's expensive in the same year. The ICD-9 book is not perfect, especially in terms of precision in describing diseases. There are some situations that are vague enough, (particularly a lot of the symptom codes). You really don't want to take every chest pain that comes through the door and pay another $10,000 for it. It's just not appropriate until you have a firmer diagnosis. Here are a couple of examples. Simple diabetes is not really predictive of much as a hospitalization. Only when you have a higher level of diabetes does it mean much. Fever is a symptom.

The one-day stay issue was a bit contentious, but actually, before we made the decision, we had the researchers go through the data and look at the predictive costs of people with one-day, two-day, three-day, and 10-day stays. In principle, like making a threshold, it is monotonically increasing. The predictive costs next year of people with longer stays are higher than people with shorter stays. There is this progression, and we made a decision. I could have said, "Okay, let's do a five-day stay." Everybody would have said, "You're nuts. That's incredible. You can't do such a thing." Ninety-nine percent of our stays are now less than five days, and that would have been ridiculous.

The question is, do you include anything or do you put in a threshold that gets, at a somewhat sicker class of patients, but also has other virtues with respect to converting hospital outpatient procedures into a one-day stay? It seems that there was a little bit of danger in doing that. It would not be very expensive, depending on how the hospital is being paid by capitated plans, to convert an outpatient procedure with observation, into a one-day stay. You can still take any stay and extend it to a two-day stay, but we have to cut it off somewhere.

I did some simulations of the difference, and the results were miniscule when compared with the payments. This is because of people that you move from one pot of getting extra money for having their one-day stay into the other pot of people who don't have any stay at all. They just carry their money with them. It raises the dollar values of everybody in the lower pots because these are slightly more expensive people. The net differences are very small. It was worth it from a program integrity point of view.

The first year we started collecting data was from July 1997– to June 1998, and data flowed in a little bit later than that. We had about 1.6 million stays to work with that came in from the managed care world, and it came in at a number of about 0.25 per person per year. In year two, which was the actual data used for payment after our experiment, it is looking very much like this. We do expect it to be a little bit heavier because we're allowing data to come in well after what we would have normally set as a deadline. We expect things to creep up.

What is the actual impact? The impact at a 10% blend is not significant. Pretend for a moment that we were fully implemented. Our initial impact estimate based on the first year was a 7% aggregate decrease to the industry. Then, when we went in and looked again later on, we found a huge decrease in our estimate and wondered what happened. Certain numbers had not been finalized at that point, when we had to crank out the estimate very quickly. That difference between having a denominator in an equation of 5,186 versus 5,100, which was the final number, accounted for a lot of the difference. We're getting at numbers a little less than 6% and what we estimated was 7%. Little things mean a lot.

The actuaries, in fact, hadn't finished the rate book. I was using a simulated rate book when I did the estimates for the first year. That accounts for a few tenths of a percent as well. That's why our estimates of the impacts have gone down. Also, when we finished getting the data, we found that there are some plans that made large "boo-boos" and did not manage to get their data in. They would tend to make things look even worse. That's where we are.

Let's discuss the comprehensive model. That's where we're going unless everybody suddenly turns around and tells Congress that they love their HMOs and don't think they should be touched. The winds keep blowing in D.C., and you never know which direction they will be blowing in. We're tentatively going to get what we call comprehensive, full data models. What's included in these models is a decision that has to be made. These models definitely bring in physician bills. We've also brought in hospital outpatient bills. We're still experimenting with what can be left in or left out.

I will say that, despite the tendency of trying to exclude data from the models, there's a very good reason for getting in more data, whether or not it's in the model. That's because we would like to partially calibrate these models on what we see as utilization in managed care plans. There's a sense that managed-care plans might be substituting different kinds of care. Things look different, and one disease might have a different cost in managed care than it does in FFS. We'd like

to have enough data from the managed care industry to be able to get a picture of the relative costs of diseases in this hypothetically defined environment.

I'm not sure there is such a thing as managed care in a very narrowly defined way of doing business or treating patients. It's probably quite variable depending on how physicians are involved and how they're being funded. The complications are enormous, but we would get some industry level averages for what the differential ought to be. The other things that can be done, if we get more data, would be to expand to other uses because we would have enough information to build HEDIS measures out of encounter data. There are a number of uses for the data.

There is a lot of tendency to push for quality. We keep hearing about paying on quality. We don't have a clue about how to pay on quality because we don't know quite what it is, but if we have data, we might be able to come up with some of the traditional fallbacks to quality. For example, do you give your diabetics tests? Eye exams? We could get it right out of the data.

The other thing that we could do, based on data, would be to compensate for coding variations that occur over time. In the old days, we had something called a "DRG creep." As soon as we start coding a hospital diagnoses and paying on hospital diagnoses, the pattern of hospital diagnoses begins to take on a different flavor, and we expect we'll have variations in the way people code in managed care, especially when the physicians are involved.

The hospitals have their own ways of coding mainly because of the DRG system. It's harder for them to rethink how to code to get too many systems optimized. They can probably do it with enough computers though. When you start bringing in physicians, you're at a whole different level of coding and coding competence. There are very few coders involved in physician bills. Being able to get a handle on the patterns of coding will be very important by looking at our data.

Our models will probably be prospective. We are using concurrent, same-year models that are very powerful because predicting the cost of somebody in a year that has the disease is a little bit easier. If you have an appendicitis, you know that it's going to cost you a certain amount for that person during that year. However, we didn't want to do that for a couple of reasons, and one of the reasons is the volatility that you get with things like the flu epidemic. We're not really supposed to be paying costs, and the concurrent models tend to follow costs pretty closely. They also won't be known until the end of the year. They're very retrospective in the way they pay. We will be using, initially, inpatient, physician, and outpatient data for making predictions.

I'll show you a few of the clinical algorithms that are out there. Most of them have characteristics that are additive in that more diseases give you more money. Sometimes each disease has a coefficient in the model that adds more money. Even if they put you in one bucket, it may not be additive, but more diseases will get you into higher buckets. There are a lot of ways in which they work the same even though they're structured differently.

All the models include demographic factors at the moment. That is because there are things we can't explain by just looking at diseases, although there are some models that attempt to do this. We still have to pay some predictive amount because somebody's going to get hit by a bus next year who wasn't a user this year. Off we go and pay based on age and sex. We've also been using a new variable, which is, "Were you previously a disabled person?" That's something new to the system that we added because we figured that the disabled are really on a different trajectory of costs. When they hit 65, they would otherwise revert to the lower trajectory of those well-over age 65.

Of the five models that we're playing with is the hierarchical coexisting condition (HCC) model. It was one that evolved out of the DCG work done by Boston and health economics research (HER). That's the model we have most experience with. We've had the software the whole time. It's nonproprietary because we paid for it all. It's been our test-bed model, and we're still working with that one.

Long ago, the people at Hopkins started work on group ACG/ADG modeling. They originally were working on ambulatory data because they were going to work on profiling physicians. They were using a model to project ambulatory expense. Ambulatory experience, **is** much less volatile than hospital use, so it's easier to work with as well. They have two models. We work with the one called ADG, which I'll describe. It basically puts diseases into buckets and pays the bucket-load of disease you have. The ACG model that they commercialized and are selling mainly interacts with all these buckets. The ACG puts people in a unique bucket based on the pattern of these ADGs that are found out there. We found with Medicare that the ADG worked a little bit better.

The chronic illness and disability payment system (CDPS) model by Rick Kronick has been expanded. He did that for the disabled population. He has now expanded to do all the Medicaid, and it's now being calibrated on Medicare data. Putting diseases into categories is similar, in spirit, to the HCC model. Some of the things that have evolved in that model have changed it over the years. One that almost nobody knows about is one from RAND that we funded, chemical-dependency risk, which was the model Grace Carter worked on. That model was done initially for Medicaid. We're also having them do a Medicare calibration on it. It's an interesting variant in the way it puts you into various severity groups.

Last are the newly popular and heavily marketed CRGs by 3M. I have a very nice four-page detailed brochure with nice pictures in it, that describes it. It is like the ACGs, but on steroids. It puts people into 1,082 buckets, based on the patterns of all kinds of things. It eats all kinds of data. The company has designed it to use home health and procedure data and dates of visits in order to do the optimal version of their model, but they also have many more consolidated versions of it.

Of course, there's a reason to go to all this trouble, and one of them is the notion that we are more site-of-service neutral. By picking up the physician bills, we get most sites of service included. We're not saying we're only going to get people hospitalized. This has some advantages and some disadvantages. If you pick up a person hospitalized for congestive heart failure, the extra payment for that is going

to be quite high because it's picking up a severity level implicit in there. When you start picking up every Tom, Dick, and Harry whose doctor reports show congestive heart failure, you really average a lot of people, and they lower the actual increment for such a disease. Given that you don't have a completely selected population of people who are only hospitalized, it works reasonably well across the board.

It does take a lot of money out of the demographic factors and puts it more into health status. They all do that because there's much less in the base bucket than there is when you're using hospital-based models. Diabetes, for instance, will garner you some dollars in a physician's office in most models. The payment accuracy is definitely improved. All our R^2s go up to this wonderful vaunted 10% number by using the models that use diagnoses only. If you're willing to use procedures and other kinds of information, you can get the R^2s up pretty high. It depends on whether you're having R^2 rates or whether you really want to be very neutral about practice patterns. We've been worried about that for years. We don't want to say, "If you use this procedure, we'll give you more. It gets a little bit sticky. This is a number I generated that has been unseen by mankind yet.

Why do we bother going to fancy risk adjusters? Here's a little motivation. This is not done on plan data. It's done on future service data. I ran some correlations of the risk adjusters that I had lying around with per-capita costs in counties. I took the actuary's rate books and backed out the per-capita costs in each county. I happen to have two sets of counties lying around. One set was 124 large counties that I was researching a few years ago, and the other set was 105 counties involved in our "choices" demonstration that do risk adjustment.

Those 105 counties are really idiosyncratic. For instance, at least 15 of them are from Georgia, and a lot of them are rural. I have only one or two real urban counties in there. They're the tough nuts, and if we look at what the correlation is between per-capita costs, and in a purely demographic model on the large counties, it's very respectable at 0.61%, because you have a lot of people, and the law of averages tends to work.

If you use the PIP, you get a much better correlation. You're not getting much more. It's a little bit better, but where you see the real differences occurring is the small group, and these counties are effectively small groups. Many of them have fewer than 10,000 people. The demographic correlation was a negative 0.15 for these 105 strange counties, showing that, indeed, there's plenty of variance out there. Age and sex just don't take it very far. The PIP is a major improvement at 0.43, and the HCC is holding its own at 0.70 using the physician data.

This is really an important reason for making progress in this direction and is our motivation to continue in this direction. As for the beneficiaries that we're dealing with right now, 18.6% in our data set were hospitalized. Those who had diseases that were counted in giving bumps in the PIP model amounted to about 12%. If we look at the people who are accounted for by the comprehensive models, you get different numbers, but they're definitely different by orders of magnitude.

In the HCC model, which still discounts a lot of vague codes, 57% of the people are in some elevated bucket that are not, strictly speaking, just demographic. The ADG model in the last version actually had 77% in some bucket, for better or worse. More people in an elevated bucket doesn't necessarily make it lighter, but it does mean an awful lot of people are now being measured as opposed to just "mooshed" in by age and sex. There's something about them that we know are being used clinically.

To show something about those age/sex factors and for varying of age groups, we could pick one of them. Let's pick the 70–74-year-old female. If we use a basic demographic model, the dollars involved in that demographic cell were $3,600 roughly. The PIP DCG model in the age bucket, having taken out some of the more expensive people, leaves about $3,000. In the HCC model, $1,650 is left because we've detected diseases that are put into other buckets, and the ADG-HOSDAM puts so many people into buckets that there's only $767 left in the demographic cell. These have interesting patterns because the ADG starts out lower than anybody else, but by the time you get to the oldest people, it's actually higher than the HCC model. There are some dynamics in here that we're investigating as to what is going on with these various models.

What's making them behave the way they are? We are taking people out of the base pot. Some of the characteristics of the models that are interesting are the predictive ratios for various groups of people in which we look at the predicted expenditure divided by the actual expenditure. This is just one of the zillions of predictive ratios that we've done.

I can't remember which version of HCC this is. You can see what it does to take the adjusting average per-capita cost-like (AAPCC) demographic models, which for the lowest 20% in the base year, we're not looking at the prediction year itself. We want to know who was cheap last year? We are paying two-and-a-half times what they ought to be paid according to this. The PIP model brings it down a significant amount, but it's still double. With the HCC, it's still overpaying by 21% in the version of it that we had here, but as we get up higher, we have for the middle 20%, 1.35.

The highest 20% were severely underpaid if you look at them this way. PIP, because it addresses this high group, brought it up to 0.75. The HCC, even though it's not picking out people who were hospitalized, is still better, because it's getting a lot of the ambulatory expensive people in there. We're up to 0.88 on this version of the HCC. Again, none of these are final.

When we're looking at the highest 5%, you can see that PIP is far better than the old models when you start looking at particularly expensive people. It's a question in our minds of not so much perfect as better. We're always looking for better.

These were some disease-specific predictive ratios. We just picked out people who had particular diseases on an AAPCC-like model. The numbers are probably best for the breast cancer group for some reason I don't really understand. The PIP tends to raise all those predictive ratios. The HCC is doing incredibly well because

there tends to be a lot of overlap between the people who have a disease and the groups that they put together, which overlaps the code. I put together sets of codes that they were supposed to use as a target, but they weren't supposed to know which exact ICD-9 codes I was using.

Some of the variance is just because there's a validation group. Some of it is in there merely because picking out certain code groups and saying, "How do you pay for them?" overlapped with the way HCC is structured. They tend to get bigger R^2s on that basis.

Mr. Bertko: You mentioned the importance of the physician part of the HCC method by the amount that remains in the demographic-only component. I was wondering whether HCFA was going to do any research into the effect of changes in physician coding, maybe across fiscal intermediaries. You have different fiscal intermediaries in different parts of the country, within different health plans, or even at the different medical group level.

Dr. Ingber: We're certainly going to look at coding patterns. Right now, we've been working with the aggregate 5% file, which pretty much takes into account all the intermediaries and carriers, and we haven't taken a big enough sample of any one subset to see what's going on. It is of interest to know where there were any systematic coding differences going on. I'm in correspondence with one of the folks at Kaiser who is working on setting up guidelines for their company's physicians.

How does one code this disease and that disease? The coders don't like this, but the physicians are the ones actually using these check-off lists. I think it's a worthy thing to do. What they're doing is going to overlap with what we've seen in FFS. Right now we're going to have to assume that it's in the noise.

That's why this recalibration effort is going to be so important. We need to get enough data to be able to see systematic differences in the managed care plans. For instance, in the detail let's just say the incidence of diabetes was about the same. The spread between the fifth and the fourth digits in FFS might be different than what we see in managed care, and that's an important issue.

Mr. Ronald G. Betz: Mel, can you comment on how the model reacts to the different practice patterns of physicians, both regional variations and possibly the type of managed care model? Does it come from a staff model, group model, individual practice association (IPA) or a combination?

Dr. Ingber: If I could define what a model of an HMO was, I'd be ahead of most of the industry now. It used to be group or staff. That kind of thing used to work, but everything is a hybrid mish-mosh. There are a few pretty well classified objects out there. As far as I know, within one plan, the risk arrangements with the physicians might vary so much that it's very hard. It's really the risk arrangements that matter.

It is not so much a matter of whether you're called a group or an IPA. It's a question of how the doctors are getting paid. There's just too much variation, and

we haven't been able to get in the door to find out enough information about plans to do the research that you're suggesting might be done. I'd love to know the answer to that question.

Mr. Bertko: That's actually almost a perfect prelude to some of the war stories that Craig is going to talk about.

Mr. Craig N. Schmid: I probably won't talk much about Medicare. Instead I'm going to talk about the joint purchasers project, which is a California project.

The joint purchasers are four large purchasers of health care in California: California Public Employee's Retirement System (Calpers), MediCal, University of California, and the Pacific Business Group on Health (PBGH). They all have a similar interest in implementing a risk-adjusted payment system for their employees.

This is about a project they're doing to look at that. These are Calpers' goals. The other three purchasers would have similar goals, but maybe they are not identical. Calpers' primary goals were to create financial incentives for health plans and providers to attract members regardless of health status. It wanted to change its focus from selecting good risks to providing good care and to compensate health plans according to the risk associated with the members it serves.

In the design of the project, ultimately it's looking to get to a risk-adjusted payment system, but it decided on a three-phase approach. The first phase of that is the data study or gap analysis. Find out what data is out there, what's being collected from providers, what is sent to health plans, and whether it's suitable for risk adjustment. The second phase would be to try to fill the gap and to work with the health plans and the providers and to try to improve the situation. The third phase would be to actually implement something. I'm just going to talk about the first phase.

We've been working on this since last September, and we're approaching the end of the project. The Calpers' specific goal for phase one is to determine whether health plans and provider systems in the Sacramento area can provide data elements needed for various risk-adjustment models.

Although they're ultimately interested in implementing these things statewide, we focused on a four-county area around Sacramento for a couple of reasons. There's a high HMO penetration and Calpers is heavily concentrated in Sacramento. As I mentioned, there were four purchasers involved. We also have six health plans, four large provider systems, and a multitude of medical groups involved. The California Health Care Foundation is funding it, and the consultant team consists of William Mercer, Reden & Anders, and DxCG. DxCG is a firm put together to market the various DCG models that Mel talked about earlier.

As I said, the goal of this is not to come up with risk-adjustment results. It is to find out where the gaps in the data are. To get at that, we set up three phases to

the project. One is bringing in data, analyzing it, looking for gaps, seeing what kind of data is there, and what's not there. The second phase is site visits, where we send a team out to various provider sites to talk to them about how they collect data, how they entered it into the system, what sort of edits they do, and how they pass it on to the health plan.

In the third phase, which is really not the focus, but certainly a lot of the participants are looking at this phase carefully, is to run some risk-adjustment models and get some results. I'm not going to be able to say too much about that. I'll have a few comments on it at the end, but these are the models we are looking at. Most of them are DCG models. In the beginning, Calpers, in particular, had a strong interest in the DCG family of models, particularly the HCC model, which uses comprehensive diagnoses from all sites of service or most sites.

The joint purchasers also had an interest in drug models. In addition to the HCC model, we're looking at an inpatient model. We're running a new commercial inpatient model that looks at all inpatient diagnoses and not just the principal diagnosis. DxCG is working on some pharmacy models. It is at an experimental stage right now. We ran two of those—the inpatient pharmacy combined and the pharmacy-only model. Because MediCal is one of our purchasers and we have a lot of MediCal people, we ran their Medicaid model.

We also ran the Rx Risk model, formerly the chronic disease score (CDS) model, which was developed at Group Health of Puget Sound. The reason we did that is because most of the participants in this project are volunteers from Calpers, PBGH, University of California, and MediCal. Some employers say they want your data, and it might be hard to say no, but there's no contractual obligation here.

Some of this turned out to be a problem when we got down to errors in the data. We found out a lot of the tests that would have been ideal to do couldn't be done because we didn't have the data. I think we've learned a lot and we'll get to that later.

Here's what we did collect: enrollment, member IDs, coverage dates, and affiliation with the purchaser, health plan, and medical group. We got a three-digit zip code, and we collected a National Drug Code (NDC) for the drugs. On the claims and encounters, we collected hospital inpatient and outpatient professional data. We collected the ICD-9 diagnosis as the focus of most of these models. We also collected date of service on drugs, and a member ID, which is encrypted. The encryption process was a big focus.

We spent several months on this, discussed various approaches to it, and ultimately decided, for this data study phase, to just let each health plan choose their own encryption method and send it to us. The disadvantage of this was that we couldn't track the same member if he or she switched from one health plan to another. The advantage is it was easy to implement, or we thought it was, but even negotiating an agreement with each health plan was fairly difficult. I guess the moral is if you ever want to do this sort of thing, you should allow a lot of time for confidentiality agreements.

Two things we did not collect were financial information and procedure codes. Financial information is usually some measure of allowed amount or paid amount. Measure of the services used is the basis for calibrating any sort of risk model. That's the thing you're trying to predict. Once you've calibrated the model, you can apply it to a population without having this information. We didn't need to collect it, but then it makes it impossible, for us to come up with any R^2s. It makes it hard to say which of our models performed best in predicting a population.

We also didn't collect procedure codes, which would have made it easier to come up with some visit rates. We had a lot of problems in the data acquisition. I have a long list that I've presented a couple of times. I'm going to spare you that list, but I'll tell you a short story. A friend of mine likes to tell the story of a Society of Actuaries convention he went to many years ago. Two researchers came up to present their findings. One of them talked for about 50 minutes describing their research. At the end, the other researcher came up and his only remark was, "You wouldn't believe how hard it was to turn data from multiple sources into a usable data set." My friend tells the story because he thinks that was a silly comment, but I tend to think it's a good comment.

Most actuaries are going to be familiar with the problems of doing that. The good news is most of the problems were correctable, but we had to get most of the data submitted two or even three times. Things like this always crop up, but we have been carefully tracking them, so that we can hopefully avoid them when we go on to other phases of projects.

Two problems weren't correctable. One was that was most of these models don't use lab and radiology data. The diagnoses from those sources tend to be rule out diagnoses. If somebody comes in that you think might have cancer, you do a test. Even if the test comes out negative, the diagnosis still shows up on the claim. We don't want those. They're excluded. One of our health plans was not able to exclude those. We think they might be able to exclude them going forward, but we weren't able to pin that down. In any case, the results of this study include those data for that one carrier.

In another big problem, although one plan collected drug data, it didn't keep the NDC codes that tell you everything about the drug dispensed. All of the drug models we were using rely upon these codes for identifying the drugs, so we couldn't use that. We don't know how we'll adjust that plan going forward.

Before I tell you what we did find looking at the data, I want to just give you an idea of what we expected to find. California has a lot of capitation. We were expecting to see a lot of delegated claims processing, little incentive for plans or for providers to submit their diagnoses to health plans. We expected to see big gaps. Also, these managed care organizations are under a lot of financial stress. We expected to see a lot of systems problems that reduced the quality of the data. What we wanted to find out was: Were there other gaps in the data? Were these gaps of the encounters being reported? Were the diagnoses being reported? Were

they of good quality? Was it accurate and consistent? Were there gaps in the drug data? If so, where?

As I mentioned earlier, we collected the bare minimum of data elements. We had some fairly weak statistics to measure, but I think they told us a lot. One statistic we looked at was the percentage of nonusers. If somebody had an encounter, they were a user. If they had one encounter, they were a user. If they had 15, they were a user. If they had zero, they were a nonuser.

These numbers varied from 19% to 26% for the five health plans. By comparison, in a study that had been done with the Washington Health Care Authority, they saw a lot of plans in which the initial number was more than 35%. However, these percentages improve rapidly if the plan actually implements risk assessment. This is a particularly good result in the first phase.

What this means is that there doesn't seem to be any systematic problem with any of these health plans. To the extent that there are problems, they're small enough not to have a big impact on the top numbers. That doesn't mean that there aren't problems with some components of the health plans. We'll talk a little bit more about that. We found similar ranges for the health systems and the purchasers. The health systems are the large hospital systems in our study.

For the medical group, on the other hand, there's a very wide range, 17-62%, and that's not for the full range for our study. It's just for nine of the large ones that we looked at. Sixty-two percent is very poor data quality. It's highly unlikely that it is because of healthy people, so that's just something you need to adjust. Health plan data was consistent.

Our preconception was that the data would not look very good because of the high-delegated claims processing. In fact, it looked reasonably good. The health system and purchasers' numbers also looked good. Medical groups definitely have a problem.

Then we did the same thing for distinct diagnoses. An example of a distinct diagnosis is, if you had diabetes, you see the doctor 12 times for this and it shows up on 12 different claims, but it's counted as only one distinct diagnosis. In contrast, if you had diabetes and asthma, then that's two distinct diagnoses. We see the same pattern here as we saw with the nonusers. Health plans have a fairly narrow range—3.5–4.7. Those struck us as reasonably good numbers.

Health systems had a similar range: 3.4–4.3; purchasers had a range of 3.1–4.1. Then you get to the medical group. On this statistic, a low number is the suspicious number. Medical groups ranged from 0.9 to 4.4, and 0.9 were quite low. Again, the conclusion seems to be that things are okay at the plan/system/purchaser level. At the medical group level, there are some problems that need to be addressed. We looked at the similar statistics for pharmacy. These are the percentages of nonusers. They had a percentage of people who did not have, during the one-year period, any pharmacy claim. The health plan was 25–36%, the health system was 24–28%, and purchasers were 26–32%. The difference for the medical group, 24–

34% doesn't really stand out as being different from those other three groupings, whereas the medical group looks good.

As for encounter data, some of them were pretty bad on the diagnoses. They still look good in the drug. The reason is the medical groups aren't involved in collecting the drug data. There is a pharmacy benefits manager that would almost always collect these data, and so any problems tended to be smaller.

From the Floor: Are the data for those under age 65 or Medicare-aged members?

Mr. Bertko: It is probably for active employees under age 65. The results are probably reasonable because drug use seems to be pretty consistent in terms of its frequency across the country.

From the Floor: Would prescription drugs provide better risk-adjustment data for seniors?
Mr. Bertko: Yes. I believe it's 82%.. Over 80% of seniors take drugs.

Mr. Schmid: I think it's also worth saying these are Calpers populations, and these numbers are based on the commercial members. MediCal is not included. The results are based on Calpers and PBGH. At least the Calpers population is older than the average commercial population. That would tend to increase usage.

The following are more statistics on diagnosis coding. We took the diagnosis data, ran it through a DCG model, and looked at people through the HCC model, specifically, to see which people didn't get a score. There are several reasons they might not get a score. One reason would be that they don't have a diagnosis. The second reason would be they have a diagnosis, but it's invalid. The third reason would be they had a valid diagnosis, but it was a lab or radiology, and so it didn't get scored. With health plans, 19-30% of the people are unscored. The four health systems were 20%, 20%, 28%, and 49%. The 49% number is very poor. The medical group is 18-73%.

Again, as always with diagnoses, we see problems with some of the medical groups. We determined that the 49% number reflected a plan that had a problem with the source code. The source code is used by the HCC model to determine whether the diagnosis is inpatient. If so is the diagnosis primary or secondary inpatient? If it's outpatient, is it hospital, physician, or something else?

Codes 7-9 get topped out; 7 is radiology and lab, which, as I mentioned earlier, has the rule out diagnoses. A medical group that was a big component of this one health system omitted everything with a source code of 7 and this plan popped out. We didn't catch that until right near the end of the project when we ran these models. It's something we'll definitely look for as we go forward with the project. We concluded, from the data analysis, that the problems we had in collecting the data were mostly overcome and probably can be avoided in the future. The overall data quantity was better than we expected. Certain medical groups looked very bad in terms of their ability to collect diagnoses, but on the drug side, those problems don't show up. The diagnosis coding quality is somewhat variable.

Let's go on to the site visits. Why did we do site visits? We wanted to know how each site receives data. We wanted to know how they verified eligibility, what their systems looked like, how they used patient IDs, how they record encounter data, how they submit it onto the health plan, the marketing service organization (MSO), the TPA, or the individual practice association (IPA). The bottom line is get a better understanding of how data are recorded and transferred.

We also were looking for opportunities where data might fall off the truck and get lost. We visited providers from four systems, three to four locations per system. We visited hospitals, medical groups, MSOs, TPAs, IPAs, and corporate headquarters. We spent two or three hours at each site, from mid-February to early April. As far as eligibility data, we saw certain similarities at the various sites. Many of them used something called SpotChex to verify eligibility, which is an internet-based system where physicians sign on, answer information, and get a verification from the health plan. Some of them used fax-back systems, where they would fax information into a health plan and get verification back by fax.

Most of our participants did not use any sort of universal ID. There are a lot of problems matching the ID with the system used from the ID that the health plan used. Within a system, different parts of that system might use different IDs for the same number. Many of them reported errors in the eligibility data that they get from the health plan. One of the differences between the sites was the extent to which they used electronic versus paper. Overall, we seemed to find the data were being sent from the plans to the providers.

Opportunities for improvement included more electronic transfer from the plans, electronic distribution within the health systems, increased use of internet/fax-back eligibility systems, and something that is critically important for our project, which is movement towards one ID per member.

On the hospital side, we saw similarities where almost all of them, if they got a claim with no diagnosis, rejected it and sent it back. They had trained professional coders with fairly consistent diagnosis coding. The penalties for physicians who didn't submit diagnoses were sometimes interesting, like loss of vacation days. Most of them kept their data warehoused with data for their own systems, but we didn't have access to that in our study, and they all used UB92 to make claims.

Some had to do double data entry because they had multiple systems that didn't talk to each other. In some cases, the same staff that did the inpatient coding did not do the outpatient and ER codings. Perhaps this was left to physicians in those departments and their staff.

They didn't have integrated billing and registration systems. In summary, they seemed to have processes for recording and transferring the data. Point two is very important. No one said to us, "Since it's a capitated service, we don't collect the data." That was good. We actually expected to hear that. Once they sent it to the health plans, then the health plans entered it. Generally, I think things are pretty good on the hospital side.

Opportunities for improvement include increased use of electronic data transfer, more meaningful error reports, and more advanced claim processing edits. Mel talked a little bit about Superbills earlier, but in almost all cases they use a Superbill, which is a form that has preprinted lists of procedures they might do and the diagnoses. They generally check off which one. Encounters without diagnosis information were consistently returned to the provider for more detail; however, the degree of follow-up on that might have varied.

They all had some procedures for training coders and auditing the results and some sort of initiative to improve physician coding. They all generated reports for appointments that didn't have accompanying encounter detail.

The Superbills varied a lot across providers. One way they varied was by specialty. This makes a good deal of sense, but even within specialty, they varied a lot in the degree to which physicians could customize them for their patients. Some just had check-off boxes. In terms of diagnosis, in particular, you can code a diagnosis at a three-, four-, or five-digit level in some cases. In at least one system, they just had a diabetes indicator, and it would be marked off but you'd have no way to differentiate what sort of diabetes it was. Others would leave a blank for you to fill in the last two digits.

There was variation on whether the diagnosis was done by trained staff or by the physicians, and a big variation in whether they submitted data electronically or on paper. We saw a lot of systems where you might enter it into a system at the physician level, send a print out, on to the MSO where it is again entered. That certainly increases the possibility for errors. There were a lot of incompatible billing systems and variations in auditing practices.

An outstanding issue that we didn't really get to investigate is whether any data recorded by the providers is lost between the providers and the systems or between the systems and the health plans. We didn't do any sort of audit of this data. We just talked to providers, but we didn't look at operations at the health plans at all. A big question is, even if the providers do a good job of sending data to the health plans, what happens to it after it gets there?

What's left to do? In addition to these things we described, our project calls for delivering a road map, which is a description of what we do to get from what we are to what they need to do to implement risk assessment. That's going to be in May or June.

We also will be testing risk models within the same time frame. We actually have some preliminary results. We presented them to an advisory committee on Monday. I can only tell you a few things. We saw results that varied greatly. In one case, 1.00 is an average population from 0.9-1.4. That is a big range. On the other hand, the age/sex factors varied over roughly the same range.

The general range of results wasn't any broader than we would have expected by just looking at age/sex models. However, in many cases, there were deviations of

as much as 10% between the age/sex prediction and the predictions of some of these models. What we saw is that the drug models and the diagnosis models gave similar results for each population. They seem to be predicting the same thing. There were a few exceptions to that, but by and large, that was true.

There were medical groups where we knew they had a diagnosis data problem. The diagnosis scores were ridiculously low, but their prescription drug scores looked very reasonable. I have a feeling going in that the drug might address that problem. I want to thank the Joint Purchasers and the other participants, most of whom are financing their volunteers in this process.

Mr. Bertko: Mel, one of the things that I saw in Craig's presentation was that some of the medical group results were a little alarming, mainly because of their variance. In terms of public policy, I drew the inference that, on a health plan level, a lot of these data were homogenized together, but on a medical group by medical group level, there were some really poor performers.

Dr. Ingber: We are hoping, from the Medicare point of view, that we don't have a lot of control of what goes on beyond the point of contact by the insurer, the M+C organization. We are in fact encouraging the organizations to talk to their provider groups and to modify their contracts to assure a better stream of data. When there is an economic incentive to do so, it's not that bad a deal. It works out better for everybody.

Right now, I had this strange experience in one of our meetings in which we were presenting risk adjustment. Then somebody said, "We capitate everybody, and they don't send us any information. We save a lot of money that way." The notion was they were getting a better deal from the physicians because the physicians didn't have to send them information.

What were they buying? Did they have any idea? All they knew is they were able to get X dollars in from Medicare and pay out X minus 10%. They were getting a good deal. That kind of an attitude is not good for anybody, and I think the notion it's not only Medicare, but the employer groups; everybody wants more information at this point to be prudent purchasers. In order to do this, we probably get everybody on line on submitting information.

There is a lot of talk about electronic medical records. Maybe the physicians can start entering things, like diagnosis codes in the office, or the procedures codes, or prescriptions. Kaiser Portland has phased in some of that. This is the kind of thing that's going on. It's not going to be very hard to get it out. The only mistake it in designing that one was not to make it easy to get it out. It was all going in, but they didn't have to push the button and give me all the visits designed into the system. Computing is cheap. these days. Telecommunications is becoming cheaper. If we don't give an incentive to use this, we'd be nuts. We're willing to put up with some noise at the beginning, but we think it's definitely the right direction to go.

Mr. Thomas J. Stoiber: Mel, I have a concern. Maybe it's borne out by what you said here today. I want to make sure that I've got my conclusions correct. Let's take an urban area where we have a decent population. Let's set up a hypothesis. Split that group evenly in half. In one case, the organization is doing a good job with their incentives to get people out of the hospital and to have all these services done on an inpatient basis. The others are just operating willy nilly. They're saying they are managed care companies, but they're really not.

They submit the data. The HCC fund comes back with the inpatient base PIP scores. Is it possible that you might see significant risk score differences here and maybe under the comprehensive model, you would improve that dramatically? Is that the whole goal? Am I reading your conclusions right?

Dr. Ingber: Well, if I accept your hypothesis, then the answer is yes. We would definitely improve the prediction if we were to use a nonsite-specific method. That is important. However, the hypothesis is that the managed care industry is having huge successes in getting these hospitalizations down, but it doesn't seem to be supported by any evidence we've seen, so we're a little less worried about this first stage. First, it's phased in at 10% one year, 10% another year, and 20% next year. The actual financial effect of the PIP is very, very small, but conceptually it's not even that corrupt. There was just another paper published in *Health Affairs* showing the same thing that we see all the time. There is not a real systematic difference between hospitalization rates in managed care and FFS anymore. I think there might have been at one time, but things are tight all over now. I think it seems that it's gone down all over.

Mr. Bertko: Mel and I might have a modest difference of opinion on some of that.

Mr. Kevin Kenny: I have a question on the 6% figure you mentioned as a recent estimate of the amounts that HMO payments might end up getting reduced by in fully implementing a risk adjuster. I'm probably oversimplifying, but does that suggest that all of the Medicare managed care work that has been going on for the last few years has probably increased the taxpayers cost for the program by something like 6%. Does that mean that when this is fully adjusted, there will be positive or negative impact from the managed care program?

Dr. Ingber: I think that there have been a couple of things going on. First, the actual dollars of cash flow have been much more affected by things totally extraneous, but we have to subtract from that. What would happen if we just look at the risk adjustment piece over the years every time we did a study that showed that there was a clear selection bias? Somebody said that was old data. All data are old data. By the time you get them together, they are old data, but we've had a very consistent pattern of the old data telling us the same thing over time.

I think the notion is that the people in managed care have probably cost the American people 6% more than they should have. No risk adjuster is going to get the exact number. That's the cut that would be taken by a hospital inpatient risk adjuster.

When we go to the full risk-adjustment method, I expect the variance that we will see among plans will be much greater. I think there will be many more plans that are on the positive end than the 6% we see now. The consistency of overpayment, as perceived by the model, is remarkably high. There is a distribution, but maybe we'll see some finally move into the positive.

Mr. Kenny: I didn't quite understand the comment on how extraneous factors affect the test scores.

Mr. Bertko: That can be summarized in three letters—BBA.

Dr. Ingber: The base payments have been distorted in ways that have nothing to do with selection.