RECORD, Volume 27, No. 3*

New Orleans Annual Meeting
October 21-24, 2001

Session 99PD
Using Prescription Drug Data For Risk Adjustment And Underwriting/Rating

Track:	Health
Moderator:	MR. ROBERT BRUCE CUMMING
Panelists:	MS. ARLENE S. ASH†
	MR. ROBERT BRUCE CUMMING
	MR. J. FRANKLIN ROSE

Summary: Insurers today are looking for new ways to use currently available experience data to better predict future costs. Both the federal and state governments are looking at the possibility of using prescription drug data because it is more available and often of higher quality than other experience data.

MR. ROBERT B. CUMMING: Our topic today is the use of pharmacy data for risk adjustment and underwriting/rating. We're going to start off with Arlene Ash. Arlene is research professor at Boston University School of Medicine. Her doctorate is in mathematics. She is a founder and senor scientist at DxCG Incorporated, which is a company that has been developing and selling various risk adjustment packages since about 1996. Arlene is going to talk about the use of pharmacy data and risk adjustment, including a risk assessment model that DxCG recently developed.

Our next speaker will be Frank Rose. Frank is a consulting actuary with Milliman U.S.A. in Minneapolis. He has over 25 years of experience in the group health arena. His practice focuses on individual and small group underwriting and ratings. He has helped a number of large small group carriers with underwriting operational reviews, rating projects, and implementing underwriting guidelines. Frank is going

*Copyright © 2002, Society of Actuaries

†Ms. Arlene Ash, not a member of the sponsoring organizations, is founder and senior scientist at DxCG, Inc. in Boston, MA.

to talk about using pharmacy data in the day-to-day underwriting and rating operations at a medical insurance company.

I'm going to talk about a research study that we're doing for the Society of Actuaries on risk adjusters. It's a study that compares a number of pharmacy-based risk adjusters along with a number of diagnosis-based risk adjusters.

MS. ARLENE ASH: I'm going to talk about a couple of things, but the research that I'll be discussing has been a collaboration across a variety of disciplines and organizations. Several of us at DxCG and at Boston University have been involved in developing these new pharmacy models. We collaborated with the Kaiser Foundation, and there we worked with some health economists, actuaries and MDs. Also, our company in Boston has collaboration with Care Group Provider Service Network, which is the Beth Israel Deaconess Harvard Hospital. We borrowed pharmacists from them, and also working with the company is another Harvard general internist. We have a whole team with a lot of different skills that has been looking at the question of how you can use pharmacy data to predict health care cost.

First, we'll talk about the strengths and weaknesses of pharmacy data in assessing actuarial risk. We're going to take a population base approach. We're going to talk about looking at a population and this year's data that comes in, and using something you can find in the automated data sets that you have during routine care, and use it to predict next year's total health care cost.

Then I'll describe how we developed our models and how these models were reformed. I'll describe some findings from the CalPERS (California Public Employees' Retirement System) joint purchases study of the availability of data and their thoughts about the feasibility of using either diagnoses that you scoop up off these administrative files or pharmacy indicators for predicting health care costs to move money around. To figure out in different plans whether or not you have systematically sicker or healthier people and therefore, to move money around and see that there's more money to the health plans that are dealing with sicker people.

The health care data that we want to look at is provided in a computerized society, such as diagnoses from hospitalizations, diagnoses from other sites of services, from the doctors' offices, and pharmacy claims. I'll talk about the strengths and weaknesses of each. Depending upon you and your client, one or another of these might be completely unavailable, might be available more rapidly then something else, might be more incomplete, or less incomplete. We are not going to talk about data that is not generally available in a computerized form, such as medical chart reviews and the actual values that come back from lab test surveys.

What are the strengths of pharmacy data? It's extremely timely if you are able to capture it on computer. The national drug codes (NDCs) that are used are recorded with standardized protocols that are disseminated and enforced. It's often available across very different kinds of providers and may be comparable across those groups in ways that other data might not be. It is directly tied to care given, which often isn't true of diagnoses, where someone might write down a code in order to get some tests done. The drugs that people get are real and are tied to real care. Finally, the data is predictive and we'll talk about how predictive.

One limitation of pharmacy data is that knowing the set of drugs that a person has taken during a year is not as informative, it's not as predictive as knowing the set of medical problems that the person had during the year. To offset the limitation, we use hierarchical condition categories (HCCs), which are models that are built based on diagnostic information. If you can build a model that extrapolates from the diagnosis to determine which medical problems are present, you can do a better job at predicting what cost those people will have next year than if all you had was their drug profile.

It's hard to map individual drugs to medical problems. Many drugs are used for a very wide range of problems. There are rapidly changing implications of what it means for somebody to be taking a drug. For example, certain drugs might have been introduced to be used exclusively for HIV-positive patients, but then they get used for people with hepatitis. Those are very different conditions with the same drug, which you thought had a very specific purpose and did last year, but doesn't necessarily have that purpose this year.

You have to be careful about data that seems to be comparable across groups. In benefit plans, you have to be careful about how much incentive there is for a person to get all of his or her drugs through a particular pharmacy benefit manager (PBM). It's both a plus and a minus that the pharmacy data are directly tied to the care given. There's also a concern about gaming. If people understand that judgments are made about whether they're treating sicker or healthier people based on the drugs given, they might shift their prescription behaviors. There's also concern that drugs given by one provider in one part of the country might have a different meaning than the same drugs given somewhere else.

Table 1

Comparing Data Sources

	Timing	Predictiveness	Incentives	Penetration
Rx	+++	++	--	++
Inpt Dx	-	++	-	-
All-E Dx	-	+++	++	+++

Let's compare the three data sources (Table 1): the pharmacy data—these NDC codes, the inpatient diagnosis, and the all-encounter diagnosis. There should be a column called availability. Because if you don't have the data, you can't use it. And as I said before, I think that the source that is most available to you is going to depend on who you are and where you fit. People are excited about using pharmacy data to predict cost because of the timing. It's there and it's there immediately.

How long it takes you to collect and aggregate information about the diagnosis that shows up during hospitalization or the diagnosis that come from dispersed sites of care will depend very much on the particular application that you have.

In terms of predictiveness, your gold standard is these HCC all-encounter diagnostic models. The predictions from that data are just about the best there is. On the other hand, either pharmacy data or models built on inpatient diagnosis are pretty predictive, and in combination they're actually good competitors for all-encounter data.

In terms of incentives, the concern is about sensitivity to practice pattern variations. I'm more concerned with the incentives associated with putting dollar amounts on a prescription drug than I am with putting dollar amounts on a diagnosis. There's a negative associated with inpatient diagnoses and incentives because these models only see that a person is sick if he or she is in the hospital, so we get a chance to see that diagnoses were written down during the hospitalization. A person could be very sick, but not go the hospital. The inpatient model would show him or her to look just like a perfectly healthy person.

In terms of penetration, this is the medical use prevalent in an under-65 commercially insured population. About 4 percent of the people go to a hospital. So that top 4 percent can be pulled off as a more expensive group. You can get good discrimination with those hospital models, but the rest of the people are left to be priced by age and sex. You have very poor penetration with inpatient diagnoses. Typically, about a third of the people don't have any pharmacy data, so you're left to look at a third of the people and not really know much about what's going on with them. For all-encounter diagnoses, somewhere between a quarter and a fifth of the people have no diagnoses during a year, if you're looking at a population under 65.

We started out with our pharmacy model wanting to map drugs to diseases, but we decided that we could not sensibly map all drugs to diseases. Our model instead collects classes of drugs that have the same sort of therapeutic purpose. You will see later a contrast between what our models do and what another pharmacy model called the chronic disease score does. This is a different conceptionalization of how to use drugs. The idea is to attempt to say that if we see a certain pattern of drugs, we think that the person has diabetes. We see this pattern of drugs, we think that they have congestive heart failure. Those are examples of two different ways in which to try to organize the drug information.

We have a combination of empirical and theoretical inputs to try to build models that make clinical sense. The model tries to be relatively robust against artifacts of what's in a drug formulary and different practice patterns. So right now we're talking about models where you use this year's drug to predict next year's total cost: inpatient/outpatient and drug costs for next year in a commercial under age 65 population.

In developing our models, we actually used two kinds of data. We had a development data set of about a million people. It's the Medstat MarketScan data. It's nationwide; it's about a million people who had drug coverage during the two-year period, which we needed to calibrate our models. We were working with Kaiser, and we were using their dataset with our model to compare to our results. When we felt we had a pretty good model, we would apply the same classifications to the Kaiser dataset to see if we got similar results. If there was a number that we thought was funny in our data, we would see if it looked equally funny in the other dataset. In any event, we have two years of information, and we were able to build all the different models because both datasets were complete with respect to pharmacy NDC , inpatient diagnoses, and outpatient diagnoses.

The MarketScan data is on one million people. Half the people are women, and none of them were seniors. We actually had a few seniors in the Kaiser data set, but our development data didn't actually look at the experience of those 65 and over. The average age was 33.

Table 2

Demographics (N~1 million)

%Female	50.2
% Children (0-17)	26.2
%Young Adults (18-44)	41.6
%Older Adults (45-64)	32.3
% Seniors	0
Mean Age	32.8

In the Medstat development data set, we found that 36 percent of the people had neither inpatient nor pharmacy data. Only about one-fifth of the people had nothing. There were people who had diagnoses, but diagnoses that showed up outside the hospital. We're using the presence of two kinds of data to split the population into four groups. And it's a yes/no on whether there were any inpatient diagnoses reported, and that was only on 4.3 percent of the people. A yes/no on whether they had any pharmacy data reported. It's very rare to find a person who goes to the hospital but isn't taking any drugs. That's 0.5 percent. You have pretty broad penetration of the population when you have both kinds of information. Of course, as you would expect, the people who didn't go to the hospital and didn't use any drugs are less expensive than average, and the people who use drugs and went into the hospital are about four times more expensive than average.

How does our classification system work? We start with the NDC. At this point, we have about 69,000, but when we began the project there were about 58,000. We collapsed them into 125 Rx groups. Then, just for reporting purposes, not for modeling purposes, these Rx groups were rolled up into 18 larger categories of drugs. A larger category of drugs might be the diabetes drugs, which has the two classifications of insulin or oral diabetic agents. Pulmonary drugs break down into three categories: asthma, chronic obstructive pulmonary disease (COPD); cystic fibrosis; and methylxanthines.

There's a small amount of hierarchy building in these models. If a person is taking both oral agents during the year and insulin, we classify that person only as taking insulin. Thus, the category of taking oral diabetic means only oral diabetic agents not insulin. We try to separate out the people with the more severe disease in the upper category; in the lower category, these people are not taking insulin. The model included a small number of drug group interactions. An example was a person taking insulin and loop diuretics. The doctors and the pharmacists felt comfortable that those two things together would probably suggest that you have somebody sicker, more expensive, than the sum of the two combined. Deciding which drug interactions got included had technical, clinical, and empirical contributions.

Table 3

Pharmacy Vignette	
Year-2 Prediction	Year-1 Prediction
$924	43-year old female
$534	**Rxgroup 40: Beta-adrenergic blocking agents** *NDC: 00378004701 (Metoprolol tartrate)*
Rxgroup 41: Calcium channel blocking agents	
	NDC: 00069154066 (Norvasc)
$2,332	**Rxgroup 66: Insulin** *NDC: 00002871501 (Humulin 70/30)*
$5,009	Final Prediction

These models are regression models. You start with an age/sex category, and you have a 43-year-old female with nothing else going on. You predict a $900 payment for one year for this one member. The various drugs that might show up would add amounts. In our example, (Table 3) she has heart problems and diabetes, and she takes the two drugs for heart problems and the one for diabetes. Once they are added, you end up with a prediction of about $5,000 for this person based on a profile of what drugs she's taking.

We also built the model to account for pharmacy data and inpatient diagnoses because we found that clients had the most trouble getting diagnoses from non-hospital events. We thought it was useful, but if you couldn't get diagnoses from non-hospital events, you could get both the inpatient diagnoses and the pharmacy data. If we've seen that same person in the hospital for diabetes with chronic complications, that's not the person who happens to have diabetes that might be controlled by insulin, that's the person who's out of control and been to the hospital, and that person is likely to be more expensive. In the combined model with pharmacy and inpatient data, using the same 43-year-old woman except that this time she had a hospitalization for diabetes with chronic complications, we would slightly reduce the predictions on the drug use, but the hospitalization would cause the final prediction to jump up to just over $18,000. Basically, if a person goes to the hospital, your sense of how expensive that person is likely going to be next year really goes up.

How well do the models predict overall? We're going to compare the all-encounter HCC, which is the gold standard, to what you would get if you used the inpatient HCC where you build the same profile that you would if you had all the data, but you only have data that comes from hospitalization. The two new models are the ones with the drugs alone and the drugs plus what you can learn by pulling diagnoses off the inpatient form. Let's compare the four models.

An age/sex model has as ability to predict with an R^2 of 1.7 percent. If you have the inpatient diagnoses alone or if you have the pharmacy data alone, you are basically able to predict equally well with an R^2 of 8.4 percent. The all-encounter HCC model, based on getting your diagnoses from everywhere, is at 11.3 percent, while your pharmacy plus inpatient is right up there at 11.8 percent. Thus, if you do combine pharmacy and inpatient diagnoses, you can be as equally predictive overall as the gold standard—all-encounter HCC.

One of my major endeavors over the years has been to imagine pictures that we could draw that would help us to understand how well our model works. One picture that I like to draw is to use each model to divide the population into deciles of predicted risk. So I take a model, for instance the HCC model, and apply it to the data, which gives an expected cost next year to everybody. I then sort people into 10 groups by costs, so that the first group is the 10 percent with the highest costs, the next is the 10 percent with the next highest costs, down to the 10 percent with the lowest costs. Using this approach, the 10 percent of the people that the all-encounter HCC model thought were least expensive had average costs of $292. The 10 percent of the people that the all-encounter HCC model thought were the most expensive came in at just under $7500.00. So you really get the ability to spread that population out.

When you use this approach with all four models, you learn that all the other models have similar results, except for that model that looks only at diagnoses that come from inpatient hospitalization. It doesn't get the high end as well because, by the time you drop off below the top 4 percent, you haven't seen any hospitalization. You really can't make distinctions there.

Another way of looking at the same question is to sort the population based on year-one costs (Chart 1), the same year that we build our profile. We sort them into 100 groups, starting with the lowest cost, then the next-lowest cost, up to the top percentile of cost based on year one. Actually you have to do a little fudging here because about 22 percent of these people had no cost in year one. We just randomly assigned them to 22 bottom groups. Each dot represents a group that was sorted based on how much it cost last year. If you look at the dots that are over on the far right-hand side of the graphic, they represent four predictions for the same group of people—those who were the top 1 percent most expensive last year. A dot that was on the 45-degree line would represent a perfect match between the model's prediction and what actually happened. Well, none of these models know that the person was the most expensive person last year. These models key off of diagnostic information or pharmacy data. They don't know that that person is going to be that expensive. As a result, they're all too low. But in fact, the pharmacy plus inpatient HCC model actually is the best at predicting costs. It produces the highest predicted cost for that group of all the other models. The one that's hanging out at the bottom is the pharmacy predictive. Pharmacy models are least effective at finding that really high-end expensive group. On the other hand, if I had added another set of dots here for age/sex you would see that the pharmacy models are a heck of a lot better than that. If you only have pharmacy data, you've been using age/sex models, and you want to do some underwriting, don't say that the pharmacy data is bad; it's just bad compared to a model that uses data that you might not have available to you.

Yet another way of asking the same question about how these models perform, and what we've done here is we've used this gold standard all-encounter HCC model to lay out people according to its perception of how inexpensive or expensive people are. So you're to the left on this graph, if the HCC model thought that you were going to be very inexpensive. You're at the far right if the HCC model thought that you were going to very expensive. If you look at the HCC line and the actual line, they track very well all the way across. So the actual cost of people in each of these bins as determined by the HCC model, the actual cost and predicted cost are very, very close all the way out to a small group of people who you think will cost $70,000 or more. And in fact, you pretty much predicted how expensive they are going to be.

Again, the pharmacy-only model tails off at the bottom. It looks like the bad outlier, but don't go away with too bleak a picture here. The people who you predict to be $20,000 or over represent only about 1 percent of the total and these models still know that those people are going to be expensive. They just haven't made predictions as high as those people actually are. The model doesn't know what medical problems are present; it only knows what drugs are used.

Summarizing this information on predicted performance, the Rx-only and the inpatient-only models have similar R^2 at 8 percent. But the all-encounter HCC and the Rx plus inpatient HCC are better and are pretty much equal, 11 percent to 12 percent prediction.

The California Joint Purchases Study is looking at the question of what kind of data are out there? Is there adequate data to do good purchasing based on risk adjustment models, risk assessment models that would come from the kinds of data that are out there? So this was a data quality study, and it was an assessment of the current state of the art a couple of years ago on what data were out there. It's very frustrating if they weren't able to get the real cost. This was just looking at how much data was there. If you look at the same population using different models, to what extent might you get a different sense of whether this was a high risk or a low risk population?

Five plans participated in the study. And they differed a little bit in the percentage of non-users. We found the same kind of findings in this data set that we found in our Medstat data. There were substantial numbers of drugs per person, of diagnoses per patient, and the invalid codes were not so great. The plans didn't differ so much in terms of how much data they were supplying per person, which was somewhat comforting because we feared that some plans would have a lot more data loss, while others would be much more complete in their data capture.

Table 4 shows the relative risk scores for the different plans. The age/sex models showed the plans to be quite different. Plan 2 is an older group, the age/sex model has 1.35 risk based on age and sex. If you look at Plan 2, you'll see that the six different models came in pretty much the same. For Plan 2 you had a pretty consistent sense that this was just an older group and that age alone explained how expensive it was going to be. But we saw some things that were a little bit disturbing. On the 4th and 5th lines, the chronic disease score (CDS) and RxGroups, were split on whether or not a plan was more expensive or less expensive than average. They differed very much from what the diagnostic cost group (DCG)/HCC found, so Plan 1 really is a disturbing case where the HCC model is saying that this group is only 90 percent as expensive as average, while the CDS (Pharmacy) model is saying that it's 5 percent above. Well, there's a big difference between a .89 and a 1.05 risk factor. I think people don't exactly know what to make of the fact that two different teams of well-respected researchers had produced pharmacy models that were giving very different answers on the same population. It makes us a little nervous. We think that these models are good, but we want more convergence before we're really comfortable using them.

Table 4
Relative Risk Scores By Plan And Model

	1	2	3	4	5
Age/Sex	0.99	1.35	0.98	0.93	1.04
IHPCC	0.96	1.31	1.00	0.93	1.02
DCG/HCC	0.89	1.34	1.06	0.88	1.10
CDS (Pharmacy)	1.05	1.32	0.93	0.91	1.14
RxGroup	0.99	1.35	0.98	0.83	1.11
Rx + IHPCC	0.99	1.33	0.99	0.84	1.09

The findings of the study were that we sometimes have different answers, but we don't really know why. We didn't know what the real risk was, so we couldn't exactly say which one came up with a better answer. The pharmacy models are tempting because they're definitely better then nothing, but they can be a little confusing as to what you're seeing. There was also a great deal of surprise and pleasure in the fact that the diagnostic data was much better than expected. And to the extent that the data was not so wonderful, often the data had actually been recorded at the site of the care giving, but it was lost in the data transfer, which is a more solvable problem than if it's not being recorded in the first place. We made visits to the site to see what problems were destroying the quality of the data, and we found that a lot of the problems were fixable.

The decision of the joint study was to proceed with a diagnoses-based risk assessment and ultimately to move toward risk adjustment and to have implementation via phase-in. We wanted to go slowly before you started moving real money around so that you could actually have more time for people to get used to what was going on and be convinced that what was happening was fair and appropriate.

On the question of using pharmacy models as management tools, there are limitations. They are just not as clinically informative or as well validated as diagnostic models are. They're still in their first or second generation of wide use. They're there, they're useful, but there are concerns. The data is timely. It's informative. Especially if you combine them with inpatient diagnoses, you really can get a very informative picture of how sick the population is.

MR. J. FRANKLIN ROSE: My topic is how prescription drug data is being used in underwriting. Most of the information that I'm providing was taken from a survey we did in 2001 of individual medical carriers. It was a mixture of commercial carriers, HMOs, and Blue plans. Although it was exclusively individual medical carriers, a lot of the applications apply to small group.

Drugs have become much more important in the underwriting process. The primary reason is that prescription drug costs as a slice of the claim dollar have gotten much, much larger. Claim cost trends on drugs have been very high. Let's use rheumatoid arthritis as an example. There are about 30 drugs that are commonly used to treat rheumatoid arthritis. Of those 30, about two-thirds of them are non-steroidal, anti-inflammatory drugs. Among those are aspirin, acetaminophens like Tylenol, ibuprofen, which costs you maybe a couple of dollars a month. Those are actually used to treat very mild forms of rheumatoid arthritis. But the two leading prescription drugs to treat arthritis today by sales are Celebrex and Vioxx. Those two drugs were introduced both in 1999, and they each cost about a $1,000 a year. Those have sort of become the gold standard for treating rheumatoid arthritis. But there are other drugs out there that are even more expensive than Celebrex and Vioxx, such as Enbrel, which was introduced in 1998 and costs about $15,000 per year per treatment. For those who suffer very serious forms of rheumatoid arthritis, Enbrel has been very successful where Celebrex and Vioxx clearly have not been able to do the job. Thus, when you get somebody in on an application and they write that they have rheumatoid arthritis, it becomes very, very important for you to know what kind of drug is being used to treat that applicant's rheumatoid arthritis.

Drug information is being used in three out of four different underwriting possibilities: new individual business, new small group business, and small group renewal business. For the most part, you cannot do it at individual renewal time. For small group, you can actually use drug information to re-rate and underwrite at new business time and certainly to re-rate at renewal time in states where you can adjust for a health status factor. There are some limitations and I will go into those a little bit more. On new business, where do you get your drug information? You get it off of the application, and you get it from some supplemental sources like an attending physician's statement (APS). I have seen some companies actually go to the pharmacy, and the pharmacies have been providing the drug history for a particular applicant. In the small group, those are the main sources. When you're doing a renewal, such as on a small group, you also can use your own drug claim database.

What kind of information is asked on the application about prescription drugs? Typically the carriers are asking what drugs have been taken within a certain time period. The most common time period is five years, but some are asking for two years, and some are not putting any limit on it. We're seeing more applications asking for the actual dosage. The reason is that the dosage will often give you an idea of what the cost is, and the dosage can significantly affect the cost. For an ace inhibitor with 10 or 20 milligrams of the drug and versus 40 and the 40 may be twice as expensive as say a 10 or a 20. So it becomes even more important to know the dosage, while for others it doesn't seem to make any difference. The applications are also asking for medical condition, and they usually ask for drug name and date of use, while some are asking for the prescribing doctor because they like to get an APS if their underwriting guidelines call for it.

You have four underwriting options on individual medical at the time of new business: accept, reject, rate up, and exclusion writers. In small group, you can either accept or rate up for the most part. Can you use drug information solely to make a rejection or do you have to go beyond that? Typically most of the carriers said that they would base rejections only on the medical condition being treated by the particular prescription drug. There were some companies that actually have a list of drugs, and if you're taking any of the drugs on the list, you will be rejected. Others also would reject based on the cost of the drugs. If, for example, a maintenance drug for a chronic condition exceeded some threshold, the application would be automatically declined. So there were some companies that base the rejection solely on the drugs, not necessarily the condition being treated.

Those companies that are using exclusion riders are typically using them in all of the states where they're allowed. About half are using exclusion riders. Of those that don't use exclusion riders, pretty much all of them used the other three options, the accept, reject, or rate up, but we did find two or three that were only doing accept or rejects. Those companies were not even doing rate ups, and they were not using exclusion riders.

As for the scope of exclusion riders, about half of the companies using an exclusion rider excluded not only the condition being treated by the drug, but also the drug. In some instances when they did that, they would provide the PBM with a list of the drugs to be excluded under that particular exclusion rider for that particular applicant. In other instances, they would provide a therapeutic class of drugs for the PBM to exclude. In both situations, they are struggling still with this approach because often a drug can be used to treat more then one medical condition. If they have excluded a drug for a particular condition, but that person happens to get another drug in that same therapeutic class, but it's really to be treated for a different condition, the PBM is going to not cover it. In addition, many of the PBMs have not had, until recently, the capability of excluding either certain drugs or certain therapeutic classes. That was the reason often cited on why they don't exclude drugs as well as the condition when they're doing exclusion riders.

As for the rate-up structure, typically the rate ups are done in 20 percent or 25 percent increments. They will go up to 50 percent to 100 percent, where the 100 percent is double the standard premium rate. That's pretty standard. We found that very consistent among the carriers that we surveyed. About half of them based the rate up both on the cost of the medical condition and the drugs; the other half would base the rate up solely on the medical condition. Those that based it solely on the medical condition will often implicitly include the drug cost within that medical condition. But some of them were just looking at the medical condition. If they're just doing a rate up solely based on that, they could run into problems. To return to my example on the rheumatoid arthritis, if they based that strictly on an average for rheumatoid arthritis, they would have inadequately rated up a policy for someone taking Enbrel. It is best on the rate up if you have drug-specific information so that you can get a fairly accurate estimate of the cost of the drugs that a person is taking.

Some carriers, as part of their underwriting process, will not necessarily make the full portfolio of products available to people. If you're taking an expensive drug, or you have a particular medical condition, they may not write you with, say, their lower deductible, they may not write you with a co-pay deductible plan. . They may say that if you want drug coverage, you're going to get, say, a $1,000 drug deductible. About half of the carriers had no limitation like that in their underwriting guidelines, while the other half had something where they would eliminate certain plan benefits from being offered under certain circumstances. Some of those would offer only the higher medical deductible and/or a higher drug deductible, while others would offer a medical plan, but they would not cover any prescription drugs.

My final point is about recent developments. One of the concerns about an application is often how complete it is. There are several companies around now that are working with the PBMs and are essentially aggregating PBM data into a single consolidated database. Their goal is to try and get all of the PBMs into this database. If they succeed, you could get a very complete history of a person's prescription drug history over the most recent years. One I'm familiar with has contracts with a half a dozen PBMs. Their database is about two billion drug claims, and they have 140 million lives in their database. They get about a 45 percent hit rate, which means that about 45 percent of the time they're finding prescription drug claims on a person.

The second trend is that there are several fairly new companies out there that are in the process of developing what will be essentially a group renewal rating system. They are using drug databases that these aggregators are developing for them. To my knowledge, none of them have anything available yet, but they either are in the process of doing it or have plans to do it.

The third trend is a company that has developed a product that will take your drug claims and run them through their little black box to determine the probability of a person having a particular condition. For example, it might say that based on a person's drug claims, there's a 5 percent probability that he has this condition and a 75 percent probability that he has that one. This information can be used by the underwriters to do the renewal rating for small groups.

We've seen a tremendous interest in putting very specific drug information in the underwriting manual. We've done it for a couple of clients, and now we have incorporated it into our individual and small group medical underwriting guidelines. The guidelines have extensive drug information on quite a few conditions, including the common conditions, to help the underwriters realize the differences in expected claim costs when conditions are being treated by various different drugs.

MR. ROBERT CUMMING: I'm going to talk about a research project for the Society of Actuaries. Its purpose is to compare various risk assessment or risk adjustment models. The first goal is to analyze and compare the predictive accuracy of some recently developed pharmacy-based risk adjustment models. Our second goal is to compare how well those pharmacy models do versus some of the commonly used diagnoses-based models, which have been out for a little while longer.

One of the reasons we're interested in looking at the pharmacy models is there has been a frenzy of activity in that area in the past couple years. Although some pharmacy-based models have been around at least since the early 1990s, several new pharmacy-based models have come out in 2001 or in 2000, in terms or risk adjusters and predictive models. There is a lack of independent comparisons of these various risk adjusters. Most of the developers of the risk adjusters will provide numbers on how well their models do and predict. But there hasn't been a comprehensive study done since probably the last SOA study in the mid-1990s.

The first goal for the project is to promote the actuarial profession, especially in the risk adjustment area. The actuarial profession did a fair amount of analysis and work in this area in the mid-1990s, but we haven't done much since then. It is a hot topic, and it's very important for insurance, ratings, and health-care reform. The second goal is to produce some useful research in a very timely fashion. Sometimes these research projects tend to drag on, so one of our goals is to produce something fairly quickly. We now have some preliminary results on the study and we hope to be able to publish the final results, the final report in probably another month or two.

The research team consists of various support people at Milliman and me. We also added Dave Knutson, who is a well-known researcher in the risk adjustment area with the Park Nicollet Institute Health Research Center. We added Dave because some people thought the research report would get a better review generally if we added some non-actuaries. The project oversight group, set up to provide some input and review, is headed by John Bertko.

We are looking at six different risk adjusters in our study. Three are pharmacy based, by which I mean that they predict costs by mapping the NDC to various kinds of therapeutic classes or medical conditions. The other three are all the traditional diagnosis-based risk assessment models.

MedicaidRx, the first pharmacy-based model, was released in late 2000 or early 2001. It was produced by the same set of people who produced the chronic illness and disability payment system, which is a fairly widely used system for state Medicaid agencies to determine payments for HMOs. Rx Groups and RxRisk also have come out either in 2000 or 2001. RX Groups, Arlene talked about that from DxCG Incorporated. RxRisk is another pharmacy-only based risk adjuster by Paul Fishmen at Group Health. Arlene had mentioned the CDS in her study, and the CDS is used as a basis behind both MedicaidRx and RxRisk; both have been built by expanding and updating the CDS model.

One of the traditional diagnosis-based models, Adjusted Clinical Groups (ACGs), used to be known as Ambulatory Care Groups. That's the most recent version of the John Hopkins model. The second traditional model uses a chronic illness and disability payment system, which is an updated version of the disability payment system. It's fairly widely used for state Medicaid agencies for paying HMOs. The final one is the all-encountered model or HCC model that was done by DxCG. The intent was to get a handful of the newer pharmacy models and a handful of the most commonly used diagnosis models.

A lot of other risk adjusters or predictive models exist. We heard from many of those organizations for purposes of being included in the study. But we had a somewhat limited scope, budget, and timeframe, so we had to limit it in some way. But there is a potential for doing some follow up studies on others.

Our study intends to compare how well these different models predict. What we're trying to predict for is a commercial group population: group employer/employee business. We have a mix with HMO and PPO business in our datasets, although about 80 percent are PPO business. In developing the dataset, we went through a number of data quality checks to look at the completeness of the codes, the percentage of non-users, and so forth. There were some problems with some of the HMO data. Some of the HMO data was capitated, they had encounter records, and it looked like that data was not complete, So there were chunks of data that we threw out of our dataset.

We're focusing on how well we can predict cost for certain people. We identified all the people who were enrolled for all of 1998 and 1999, for which we have both their medical and their pharmacy data, and for which we also know some demographics such as their age and gender. We have about 800,000 people in this dataset.

Our process is to take the first bucket of people and use that to calibrate the model. We develop the parameters, or the weights, or whatever you might want to call them. We run our regression, and we use that first bucket to calibrate the models. We take that model with that set of parameters and apply it to the other bucket of people and use it to validate, to see how well it predicts. What we're talking about today is based on how well those models predict on that other subset of people. The reason for doing it is because there is some potential to overfix the data and bump up or artificially increase your performance measures, the R^2 scores. You have to use the same data set to calibrate and validate.

We also looked at results under a number of different scenarios in terms of stop-loss limits or truncating claims. We truncated claims at three levels: $50,000 per person/per year in total claims, $100,000, and no truncating at all. If a person has $150,000 of claims, we're going to limit that person's claim to $100,000. We cut them back to $100,000. That is fairly common in these types of studies. The original Society of Actuaries study of five years ago used thresholds of $25,000 and

$50,000, but their data was eight or nine years old at that point. Thus, $25,000 and $50,000 back then are roughly equivalent to $50,000 and $100,000 today.

Another reason to look at results on a truncated basis is that some of these very large claim people can't be predicted anyway, so why leave them in the data? The other is that common measures of how well these models predict is the R^2. The R^2 is overly sensitive to how well you're predicting these very large claims. By truncating, you're taking off that problem when you're dealing with that issue.

In looking at these six models, we analyzed them under a number of different scenarios. A number of these risk assessment models come with their own set of parameters or risk scores. I think of the six that we have in the scope here, five of the six do come with their own set of weights. The adjusted clinical groups did not come with their own set of weights, but the other ones all come with their own set of weights. During our analysis, we looked at how well they predicted using the weights they came with. We also looked at how well they would predict when we recalibrate, by which I mean that we're taking our data and we're calculating our own set of parameters. We do that in both a prospective and a concurrent approach.

One reason to recalibrate is that often, if you have a larger plan or if you're working with a state Medicaid agency, it's common for them to recalibrate their own models. Another reason is to put everything on a level playing field, to make everything more of an apples to apples comparison. We looked at how well the models predict individual person by individual person, by looking at the R^2, which is commonly described as determining what percentage of the overall variation and cost among individuals is explained by your model. We also looked at what we call the mean absolute error. It's the average miss that you have when you compare what the model is predicting for cost for a given person versus what actually happened. If you just added those pieces together, you'd have some positives and negatives and you don't want that. So we just take the absolute value of the difference and then calculate the average value of those absolute misses.

We also looked at how well the models predict for various non-random groups. To put people into groups, we used some different criteria. One is to just look at what kind of medical condition they have. For example, we looked in 1999 to identify all the people that have some type of problem, such as depression, to see how well the models predict the cost for those people in 1999. We also will sort people into buckets based on their claim dollars and see how well the models predict for that.

We have some preliminary results on how well the models are predicting on an individual level using the commonly used R^2 measure. Remember, we also truncated claims at the three different levels. When you don't truncate claims, the best predictor, one of the diagnosis-based models explained about 15 percent of the overall variation and cost among individual people. The worst predictor, another diagnosis-based model, was explained about 10 percent of the variation.

A couple things to note were that there seemed to be a little bit more variation in results of performance across the diagnosis-based models. The pharmacy-based models tended to be a little tighter together. Also, it's interesting to note that when you start to truncate the claim, the pharmacy models start to perform a little better than the diagnosis-based models. When you truncate at $50,000, the best performing pharmacy model is almost the same as the best performing diagnosis-based model.

Compared to a given scenario of how the models rank, you look at R^2 as a measure of how well they're doing, versus looking at the mean absolute error. The mean absolute error is kind of an average prediction miss on an individual-by-individual basis. In general they perform similarly, but again when you push from an R^2 measure to looking at the mean absolute error, some of the pharmacy models start to perform better. For example, looking at R^2 as a measure of how well it's doing, one of the diagnosis-based models performed the best. But to try to figure out which one is doing the best, one of the pharmacy models that has the best results.

On a concurrent basis, these diagnosis-based models do much better. It's much easier to predict the same year's model claims using that year's diagnoses then next year's claims. One of the diagnosis-based models is about 50 percent in this case. The other note to take away from this data is that on a concurrent basis, the diagnosis-based models improved in their performance greater than did the pharmacy models. The diagnosis-based models ranked one, two, and three when you're looking at things on a concurrent basis as opposed to a prospective basis.

For another measure of how well the different models are doing, we looked at how the actual claims compared to what the models predicted. We take every single person in the data set, add up their total claim dollars, and put them into a bucket based on the amount of claim dollars they have. For example, we took 20 percent of the people in 1999 that had the most claim dollars to look at how well the model predicts for those people. We see that it's difficult to predict those high-cost people. For the highest-cost people in 1999, their actual claims were twice what the model predicted. When you look at results for the least expensive 80 percent of the population, most of the models perform fairly similarly in terms of how well their predicted values match up with what actually happened. But the difference comes into play more on the most expensive people. The most expensive people tend to correlate pretty well with other measures of how well the models perform.

We also looked at how well the models predict for people to have a given type of medical condition. We're looking to 1999 and trying to find all the people that have a given problem, such as asthma, depression, or HIV, and seeing how well the models predict the claims for those people. When you do this, you're aggregating together larger groups of people and the model starts to do better. Again, the diagnosis-based models tend to perform a little better here than the pharmacy-based models.

There are a number of areas in which there's a potential for some follow-up studies. Some want to apply these models to Medicaid and Medicare populations on a federal level. Medicare is moving in that direction for risk adjustments. About eight to ten states are basing their Medicaid payments on risk adjustment at this point.

We would like to look at the impact of claim lag and population turnover on the results. We would like to incorporate in our study at some point the fact that pharmacy data is much more timely. That's going to add to your predictive value in the real world, and it may overcome to some degree some of the differences between the pharmacy models and the diagnosis-based models. Perhaps the pharmacy models would be better when you incorporate timeliness.

We'd also like to look at results for real groups of people. How much does it really matter what risk adjuster you use when you start to look at an entire HMO or a larger group of people? Do they get very similar results? Or are there vast differences in terms of how well they predict across much larger groups of real people? We're looking at what happens if you try to tweak the weights or the parameters behind these models. We run the regressions, we get a set of parameters out, we stick it in the model, and we see how well it predicts. But there's a lot of judgment and art that can go into setting those parameters and trying to maximize the performance of the model. We haven't really looked at that issue because we were trying to keep everything as objective as possible.

The last thing that we're looking at for follow up is comparing how well the model predicts versus actual claim dollars. What happens if you combine claim dollars and diagnoses to project? We've done a little bit of just playing around with it. Some of the claim dollars seem to predict just a little bit better then some of the best-performing risk adjustment models we have. It would be interesting to combine some of those things together and see how well they do.

FROM THE FLOOR: Based on some discussions over the years with a medical director in a private company, changes in drug therapy are great predictors as opposed to point in time observations of drug therapy. Sometimes patients go to a higher dose or to a stronger drug. To what extent do any of the models that we're looking at try to take that kind of information into account?

MR. CUMMING: The three pharmacy-based models that we looked at do not take that particular factor into account. In other words, they wouldn't add something extra just because they change a drug therapy. Of course, it's possible that two drugs might somehow fall into different classes. Some of these models might result in a higher score for someone who took both of the drugs versus just one drug.

Some of the Ingenix predictive models start to look at some of those types of factors. They may look at whether the person switched physicians, which may be an indicator that the person is going to be more expensive than someone who has not. Some of those other models start to use utilization-type measures and treatment-type measures to improve the predictive accuracy. But at this point I haven't seen them involved other than in these other models. The other concern is that if you're intending to use the model more for payment purposes, you don't want it to be susceptible to bad incentives or gaming. So if you start to improve utilization measures or other types of measures related to the multiple drugs versus just one it may lead to some bad incentives for some unfair payment approaches.

MR. ANTHONY J. WITTMANN: Arlene, you about convinced me that the combination of the Rx and inpatients was a new gold standard when compared to the HCC/all-encounter model. Then you showed that the HCC model was closely correlated with the actual results. Considering the complexity, do you really think that using all-encounters are better then using Rx plus inpatients?

MS. ASH: Yes. The reason that I'm so enthusiastic about the HCC model is that it is a management tool. It is well vetted in terms of being robust across systems. I really like models that build on what problems people have rather than the way they happen to be treated in one system or another. The drug models are just too new; they haven't converged well enough.

MR ROSE: Bob, you mentioned that combining claims with these models might produce some interesting results. I think that would be a good follow-up study. It might be a better predictor than anything else.

MR. CUMMING: Yes, we haven't combined the claims and the diagnosis yet. We combined diagnoses, an all-encounter diagnoses with the drugs using the chronic illness and disability payment system (CDPS) model and MedicaidRx to bring in the drugs. You do get a significant increase in the R^2. When we did it on that basis, it improved a couple of percentage points. You can play around with these models and combine some different components and pick up things that aren't picked up in any single model and improve your predictions.

MR.TODD C. GALLOWAY: It seems like the adjustment models do a pretty good job of predicting when you have a population that has a higher percentage of people who have a higher cost diagnosis. But it doesn't seem to measure the acuity of differences within a diagnostic category. I suspect that a lot of the forces that created a plan to have a higher percentage of higher cost categories would also cause it to have higher cost within a given category.

MS. ASH: What basically makes one person with diabetes more expensive than another is the set of multiple co-morbities that they have. These diagnostic models don't have, because it's not available in diagnosis data, the hemoglobin A1C, some specific measure of the severity of the diabetes. Nonetheless they do a spectacular job of predicting the people who have diabetes whose expected cost might be $2,000 next year, and those whose expected cost might be $45,000 next year without knowing a lot about how severe the diabetes is, but knowing a great deal about what other medical co-morbities are present.

FROM THE FLOOR: I'm wondering if you thought at all about having an even lower claim truncation level on studies. If you're using a model in underwriting, particularly for small group underwriting, you are limited in what kind of rate action you can take. If somebody is above $10,000, even $1,000,that's going to put him in the highest category.

MR. CUMMING: We haven't at this point, for underwriting purposes. You might certainly be interested in doing that and also in perhaps making some other adjustments to the way the models work. Our focus to this point has been in how well the models predict in terms of a payment approach to a health plan. You'd have a much larger population involved. If you're going to use these models for underwriting of small group business, you're right about that. There are probably a lot of other things you could bring into play as well. You can start to bring in utilization measures, actual claim dollars, and combine some of those features and get a much better predictor of the future.

MS. ASH: At DxCG we also do a truncation at $25,000.

MR. DANIEL L. WOLAK: I want to follow up on the prior question on how satisfied you are at this point on predictability of the models. It doesn't look like they're predicting all that well. What are your thoughts on where we are right now with the models?

MS. ASH: Relative to the age/sex models, these models are significantly better.

MR. CUMMING: Some people have theorized that on a prospective approach the highest R^2 you can get to is maybe 20 percent on an untruncated basis. Some of these models are getting in the 14 percent or 15 percent range. Using that rule of thumb, the models are doing pretty well. Using diagnostic codes alone, I'm not sure there's much more you can do in terms of improving the predictive ability of the model. There are other issues that may be more important such as making the models less assessable to gaming or incentives to up-code. Some of these models have actually taken out some diagnostic codes that are ill defined or that might be used differently by different physicians. They are trying to get away from the potential for gaming and also get away from measuring differences in practice patterns or coding patterns and not real differences in health status. The focus of a lot of models now is how to best protect against gameability and bad incentives. When you look at how well the models do for a collection of people with certain diagnoses, they do pretty well. If the health plan had few HIV people, most of these models would very well correct for that. If your plan didn't have a lot of people with cancer, these models adjust very well for not having that set of people. The models are very useful for a fair payment system, especially in Medicaid, where you're paying different HMOs based on who they enroll.

MS. ASH: Whenever we see ourselves in a horse race like this, it's tough because you have to sacrifice predictive power to have a model that is more robust, that is less game-able, and that is better as a general management tool. We built two kinds of models. Ones that we call payment models and ones that we call explanation models. We might be willing to throw in more stuff and gain a little extra predictive ability in a situation we weren't worried about gaming.

FROM THE FLOOR: In our PPO, individual HMO products, we offer the old shoebox prescription drug after the deductible and co-insurance. You send it in, we send you a check. I'm getting tremendous pressure from our product management people to the effect that that benefit has been outdated and should be replaced with a co-pay benefit. My concern is that the HMO products have a much higher overall prescription drug use. In your survey, can you give me any ammunition to convince my marketing people that I can maintain an affordable product?

MR. ROSE: When you go to the electronic submission with co-pays and working with PBMs, you clearly are going to have an increase in a number of drug claims. That's been proven. But you don't have to go to a co-pay plan. There are carriers out there that are covering drugs the same as any other sickness that is subject to the standard deductibles and co-insurance. I'm seeing it more often where people have decided that the co-pay plan is too expensive, they're going to stand-alone drug plans with an independent drug deductible of $250 or $500. And I see even higher.

FROM THE FLOOR: You would have to require some pharmacy coverage inside your product. It may be that if you don't have drug usage right now, you get the standard $50.00 deductible, but if you have high drug usage, you get a $300.00 deductible. But you do have to have some pharmacy coverage, wouldn't you agree?

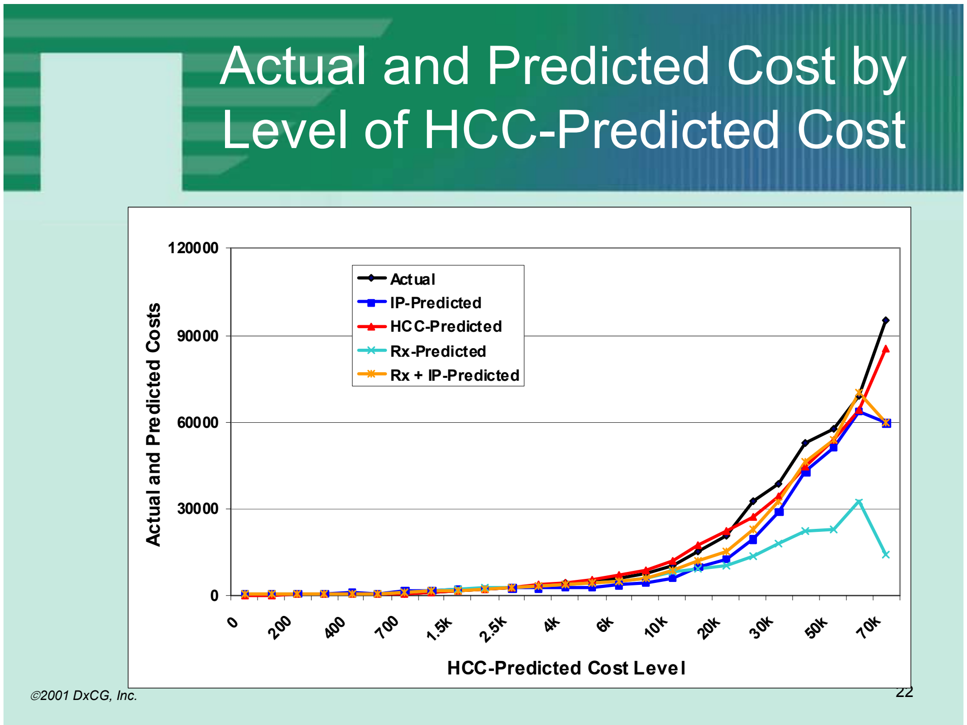
MR. ROSE: I agree, yes. You need some pharmacy coverage. But you can still offer a competitive product with a stand-alone benefit that's not co-pay but is subject to some deductible.

MR. KEVIN DOLSKY: My question relates to using this data for rating purposes and for use in underwriting shops. Along that line, it would seem to me that implementation of this approach would involve selling underwriters, who are a cynical bunch by definition, on the idea that whatever they're using is not as good as this. Generally an underwriting shop would use other things, particularly in small group they might use age/sex, some experience, they might have some tiering where they have some kind of guidelines that they use, certain disabilities and certain experience, and so forth. Do you have any comments or guidance as to how you might go about taking these technologies and comparing them to what an underwriting shop might use so they may have an idea whether or not it's an implementable solution for them.

MS. ASH: It's a good question. We do have some experience with a client of DxCG's who actually did some simulations on what if we did purchasing using this tool. We bid on certain products using this tool versus what we would do otherwise. We wouldn't get as many jobs was the answer, but we would make more profits. When you have maybe 200 people in a plan, the data suggests that you have a credible number when you look at this average risk core. The major implementation question would be, can you get the data? Can you get the data at the time that we need it to be there? If you have the data, it's a terrific tool.

MR. CUMMING: We had one client where we went through and, for small group ratings, took measures based on claim dollars or loss ratios and combined those with measures based on a risk adjustment. Basically, the risk adjustment score becomes another parameter in your credibility formula. We're looking at loss ratios, risk adjustments score and a manual rate, and blending those together. Using both the claim dollars and risk adjustments score did result in a little improvement in terms of the predictive accuracy of that rate and formula. So there have been situations where we've combined those two together, but that's an area where a lot more work could be done to improve the performance.

Chart 1



Actual and Predicted Cost by Level of HCC-Predicted Cost

©2001 DxCG, Inc.

22