RECORD, Volume 29, No. 3*

Orlando Annual Meeting
October 26–29, 2003

Session 37TS
A Predictive Modeling Primer

Track: Health

Moderator: JOHN W. C. STARK
Instructor: MICHAEL COUSINS†
JOHN W. C. STARK

Summary: Various predictive models are presented, including a discussion of comparative efficacy among popular models, current and potential uses for predictive modeling and issues surrounding implementation of a predictive modeling process. *Participants have the opportunity to compare and contrast specific predictive model features as well as gain an understanding of challenges in successful implementation.*

MR. JOHN STARK: I've been with Anthem BlueCross and BlueShield for over 18 years. In that time, I've worked in various capacities. I've worked in individual, group and HMO. Currently, I'm the corporate actuary for Anthem Southeast. Some of my responsibilities include keeping an eye on the overall financials, acting as the valuation actuary, developing and improving mathematical models and mergers and acquisitions, which clearly now is not a real significant part of my job.

DR. MICHAEL COUSINS: I am the director of Informatics at Health Management Corporation, a leading disease management company. I have a Ph.D. in neuroscience. I studied the neurochemical basis for depression, Parkinson's disease

*Copyright © 2004, Society of Actuaries
†Dr. Michael Cousins, not a member of the sponsoring organizations, is Director of Informatics for Health Management Corporation in Richmond, VA.

Note: The chart(s) referred to in the text can be found at the end of the manuscript.

and addictions way back when, before doing a post-doctorate fellowship in the Department of Psychiatry at the University of Chicago, where I continued that line of research. My interest is in research and research methodologies. I hope you see some of that expertise here.

I joined Health Management Corporation in 2000, so it's just been about three years. I manage the outcomes reporting division. That's our program evaluation area. John and I will be speaking about that. We will also discuss the research and development (R&D) area where we develop and implement our stratification model. John will be mentioning that later.

MR. STARK: When you think of predictive modeling, you think of a bunch of formulas. We're not going to do a theoretical, mathematical talk. It seemed like it would be better to use some of our studies as a vehicle to show how predictive modeling works and show all the different aspects of it from data preparation to modeling to interpretation of results. That's what we'll be discussing.

As Michael said, his field is disease management. We used predictive modeling to do a study for our company on the value of disease management. Disease management is a little different than utilization management. It's much more focused and takes into account specific diseases. The diseases that we studied were asthma, coronary artery disease, congestive heart failure and diabetes. The goals of utilization management are to improve the health quality, improve health status, optimize utilization and, of course, save money. We'll focus mainly on the cost component and a little bit less on the behavioral piece.

I think one of the things we can all agree on is that the financial outcome is pretty important. Our key message is that if you're going to understand how we arrived at these things, you probably don't have a lot of confidence in the numbers. As actuaries, I think we have probably got more skeptics per square foot than the rest of the country right about now.

One of the things you say is, "Okay, disease management. How hard can it be to measure? It should be fairly simple and straightforward." Nothing could be further from the truth in this case. I'll give you a couple of reasons for that. First, you're dealing with very small populations. Normally, we're used to dealing with hundreds of thousands of lives. Here we're dealing with maybe a couple of thousand. Also, we're not dealing with a broad spectrum of medical conditions. We're dealing with very specific diseases. Finally, with disease management, this is not something you can really put your finger on. One of the things we had to do was try to adjust for everything else and come up with a savings estimate. Once we came up with that, we would attribute that to disease management. So it's more of whatever is left, we'll call savings. Trying to do something like that and trying to make all the adjustments does get pretty dicey.

We have a couple of goals for this presentation. The first is to review evaluation methodologies. We have several of these that we used. We're going to talk about some of the issues surrounding them. We'll talk about some of the strengths and weaknesses of each methodology, as well as describe the use of predictive models in program evaluation.

Predictive models have been around in health care for quite a while. Underwriters use them. These are much more like expert systems, which I think people have probably heard of. Resource allocation for medical management is the case in terms of planning where it's all well and good to know who's sick right now. What you really want to know for planning purposes as you run staff is who is going to be sick next year and what are they going to have. Finally, predictive models are used in program evaluation, which is going to be our focus.

The big question that we had to answer with some of our studies was does it work—does disease management really work? The real question was does it save money? We'll talk about how we did this, how we measured this and the different ways we approached the problem.

Michael and I have done this talk for disease management organizations, where we do have to put an emphasis on numbers. With this group, I think you could all

agree with the statement that the numbers mean nothing unless they're rightly understood, rightly related and rightly interpreted. This really applies to predictive models and the studies that we're going to talk about.

We have a couple of key messages. No one number stands by itself. You want to make sure that you've looked at several different ways to approach the problem. Statements not accompanied by data need to be taken with a boulder of salt. There are multiple ways to measure outcomes, and we ought to use them. What you really like is to do this several different ways and have the results all point in the same direction. Finally, the devil is in the details.

DR. COUSINS: Our presentation is in three main sections. This is a really interesting field, I think. To begin, we're going to briefly cover pre- versus post-group comparisons. This area doesn't involve predictive modeling, but the reason we want to talk about it is that it's one of the most common types of evaluations in the industry today. We'll be describing it and covering some of the issues of this type of methodology, such as regression to the mean. Next, we'll briefly cover several different types of control groups. We'll describe several and then focus on the one that we used, which is called the concurrent control model. We'll also cover some issues such as equivalency. Finally, in the third section, we'll briefly cover the issue of general liability or portability of results from one scenario to another. I know we don't get into a whole lot of detail on this, but it's pretty important, particularly these days when there are contract guarantees in place for particular disease management outcomes, and the level of guarantees from a particular population may or may not generalize to your population.

We'll begin with the cohort and historical control pre- versus post-design. I'll be comparing and contrasting two approaches. When I do this, I'll actually be using data. In general, there are two different types of pre versus post with comparisons. The first, which goes by several names, is somewhat reflective of the fact that in this industry there's very little standardization, and standardization doesn't even tell you the names or the types of design. This methodology goes by the non-cohorts, sometimes for the population base, or the historical control methodology. We'll get into that in just a second. Second will be the cohort methodology. Where possible,

I'll try to use the nomenclature of historical control and pre- versus post-group, but I sometimes slip, so just let me know if you get confused on that.

The big question whenever we're doing the evaluations is who is being compared. One would think that this would be a relatively easy question, but it's really not. There are several different ways that pre versus post evaluations can be designed. I'll barely be scratching the surface today on this topic, but, suffice it to say, there are a lot of details and a lot of building the details that can really influence the results.

The first methodology I'll describe is the cohort methodology. I won't spend a lot of time on that. The numbers are identified after we start the disease management and medical management program from claims typically using combinations of diseases. The costs for these people are added up, then we take the costs for these very same people from the Year 1 period and compare them to the costs from the baseline period for these same people. This is how one can compare and calculate the financial outcomes or savings.

This type of design answers the question what are the costs for, let's say, diabetic pre-program as compared to the post-program? It's the same diabetics. This will be a little bit of a tip-off to something I'll be getting into later with methodologies susceptible regression.

Historical control methodology is one that's currently relatively popular in the industry. It was designed where members were identified just like before but then applied to the post-program period and the cost calculated. Then something a little bit different happened. The exact same identification criteria shifted back twelve months and applied for the base line period using exactly the same criteria and the cost by population were computed. So if we're dealing with a population of diabetics, this methodology answers the question how much for the diabetic, how much do diabetics cost the plan in this period as compared to how much the diabetics cost the plan in this period. That's not necessarily the same diabetics, although there are quite a few that are in both periods. What we'll see in just a second is that these two approaches produce remarkably different results.

I have to explain the results of an analysis where we found that there was a whopping 2400 percent difference between the results when we used the cohort method versus the historical control method. Again, we used the exact same data and everything was held constant: trends, identification criteria, the use of exclusions, et cetera.

Now, this is one place where I could pause for a second to mention these criteria exclusions. The criteria can have huge influences on the results. Just a couple of weeks ago, John and I showed data at the National Managed Health Care Conference in Louisville that demonstrated that even simple changes in identification criteria from looking at the primary diagnosis position to looking at the first pre positions can really change the results and the financial outcomes. Furthermore, those outcomes are dependent upon or interact with how long members are enrolled in the program. This is really some interesting stuff; unfortunately, we don't have the time to get into that now.

There's a huge difference between the historical control method results and the cohort method results. The difference was about a dollar-fifty versus about thirty-five dollars when we used one method versus another. As for benefits, there are large savings with the cohort methodology. This method has actually been changed. Very few people actually used data when they made these claims. The cohort method is susceptible to regression to the mean. That is, people who identify with common events in one period tend to regress toward the average, which is lower cost over time. There can be savings even in the absence of a disease management program or medical management intervention.

In contrast, the historical control method is not susceptible to regression to the mean. This is the approach that we recommend when there's not a lot of research available. This is the approach that we recommend when no one wants to answer the question of how much of the population was really in one period versus how much were they in another period? Although the historical control method is recommended, it's not perfect.

Now we'll turn to the issues. First, there is the regression to the mean. Then there are the contents. It does not adversely affect the historical control method as much as regression to the mean. –It is just referring to the fact that there were relatively higher probabilities that a member who has average claim possibility in a particular year will in the future have lower claims cost. The stock market analogy is a good one. The S&P 500 average is a percent. If the average is 19 percent or 22 percent or whatever, the return will probably be lower the following year as opposed to being higher.

When I talk about regression, I'm not saying on a group basis or on an individual. An example would be health care cost. They give a diabetic an ER room because they have insulin injections. These can be very costly and result in inpatient stays, et cetera. It's unlikely the member who has those high costs was identified as being a diabetic or won't have those high costs in the future. That's where regression to the mean is.

What goes up must come down. I don't ignore the converse of this, which is I'm not sure if they're saying progression to the mean that the members with low cost in one period may not necessarily be low cost next year. Their cost might go up. I don't remember the statistics exactly. Ten percent of the numbers were more extensive in one year, the next year they were less. Those were members who were in one of the disease categories. We've done a lot of studies on that.

All right, moving on: regression to the mean is relevant for this discussion of pre- versus post-group comparisons because the historical control method is, I think, unfairly criticized by the consultants. Actuaries whom I have worked with I find to be susceptible to regression to the mean. I'm fairly certain no one, or at least no study that I've ever seen, has looked at historical control method to determine whether or not regression to the mean is, indeed, a factor driving results. I'm going to show you the results of a little experiment that we conducted to address the question is regression to the mean a factor that is driving the results of the historical control method design.

To test the effects of regression to the mean on the historical control, we looked at all the members who had asthma, diabetes and CHS in pre time periods. So we identified the populations. We first applied the criteria, the principal criteria, four of those were conditions in the program year, and applied it to the trial period (year two). Then we applied it to the period preceding that (year one). The exact same identification criteria were applied for all three periods. All we did was add the cost for these three groups: the cost for approving their claim based on year two and based on one. Then it's simply a matter of comparing the cost.

Before we get to the results, I'd like to pause and show you what the policy was. We typically found program savings when we looked at the difference between base line year two and program year one. The industry wouldn't be as successful without comparisons.

The difference between base line year one and base line year two should also result in non-program savings. The program savings and non-program savings are similar regression to the means, a factor. Now, that's a pretty important thing.

FROM THE FLOOR: What do you mean by similar? If it's one-third of the program savings, it's got significance.

DR. COUSINS: In non-program savings, the program saves $10. But your program savings would be at $30. I would say that the non-program savings is significant. .

This regression to the mean is what is producing the savings when you look at base line year two and program year one. In other words, if we find savings here, we'd expect to find savings on base line one and base line two.

So if this is true, the savings between the two base lines should be essentially the same or similar as between the base line year two and the program year one. And we expect the savings between those two periods, base line one and two, to be the same as the savings between, or similar to the savings between base line year two and program year one.

These are the results that we would have expected. These were not the results that we found. This is what we would have found if regression to the mean was a factor. We found looking at the true results to be somewhat effective. What we found is that there was a financial loss between the base line one and the base line two, but savings between base line two and program year one. In contrast, there was a loss, a small loss, but a loss between base line one and base line two. So at least in this evaluation, regression to the mean is not shown in the results, otherwise, the actual results would look like the previous example, but they don't.

So there's no doubt that regression to the mean occurs on an individual basis; extensive diabetic one year, probably favored not to be extensive next year. Someone has an extensive heart attack this year, probably favored not to be extensive next year. No doubt regression to the mean occurs on an individual basis. But the data do not support that regression to the mean occurs on a population basis with the non-cohort/population base/historical control method. So the data does not support the notion that the results of this historical control design are due to regression to the mean.

We have this one particular experiment we did. We had one group based on year one and based on year two in our program. In this particular experiment it did not appear that the population regression to the mean, was a factor. If you did this a hundred different times, you will come up with a hundred different results.

MR. STARK: I think the point is that when you have the data set and you're preparing to do predictive modeling that you should really do this test. You need to look at this so you don't get any false positives.

Suppose that we had seen savings in the non-program calculations that were about half, then what you have to do is look at your design because, clearly, your savings are going to be affected by regression to the mean. It's going to be very difficult to say how much of your savings is true and how much isn't. That's where the confounding issue comes in when things are maybe a third or a half, something like that. But this is one of the tests you really would want to perform.

DR. COUSINS: This is just one evaluation. We do have one other data set. We've only done it twice. We have another data set in which we've done the same thing. I don't know that we'll find the same pattern 100 out of 100 times. What I suspect and, again, this is a speculation, is that 90 or 95 out of 100 times this is what we'll find. But as we all know, with these evaluations, there's still a 5 percent probability of making a false conclusion. I don't want to overstate the results, but the results are what they are. When we look at this particular data set, this is what we found.

Let's consider groups that covered regression to the mean. As I said earlier, no design is perfect. I mean, in fact, historical control method design is not perfect, and you want to have reasons why it's not. It is because it's susceptible to the confound. One potential problem –is that it is a rather susceptible confound, which just is another way of saying susceptible to events outside of the disease management program, which some of the events may have impacted the results. For example, new treatments, new facilities like private hospitals, which typically have more favorable outcomes than municipal hospitals, coincident programs, for example. It could be disease management programs, medical management programs that the health plan initiates, free sugar strips to diabetics or something like that that coincide with or enhance the disease management program. Then it becomes difficult to choose a part of the program or what should be the other factors. In addition, there are other factors, particularly contacting a specialist, which we ran into a lot when we were looking at the conditions of diabetes and CHS. These aren't general practitioner type diseases.

Some of these confounds can be adapted for adjustments with small chronically ill. However, adjustments for chronically ill populations can't be perfectly calculated. We all know they can't ever be, but even more so now. So that's one thing. It's not possible to have knowledge of and then to appropriately adjust for all confounds. So you need to keep confounds in mind when we look at results from pre- versus post-group evaluations. One of the ways that we in our company attempt to get around this issue is, when we initiate a program with a very large health plan, we attempt to try to start the groups at different times. Sometimes this happens gratuitously if the large lines of business in the health plan aren't already set to have programs begin at the same time. What we like to see is, for example, six months after the

program starts in line of business A, their savings, and six months after the program starts in line of business B, their savings, et cetera, recognizing that they don't all start at the same time. So that's just one little way that we've looked at the plan.

In summary, the cohort approach answers the question what is the cost of a particular cohort over time. This methodology is affected by regression to the mean and confound. It's a relatively weak design and, unfortunately, early on at least, this method was used a lot of times. I honestly don't think it was used consciously or people consciously tried it with themselves or others. I just don't think the knowledge was there.

The second approach was the pre versus post comparison. This answers the question, what would be the cost of a population in one period versus another period. This is, in my opinion, a relatively robust methodology and, you know, coming up with the next best thing when it is not available. It is not affected by regression to the mean. It's important to use utilization and cost trends and then make any other adjustments that would be necessary.

All right, now I'm going to start the last third of today's presentation and talk about some comparisons. I'll describe some designs and identify some issues. John will be getting into those also in just a little bit. There are a few relatively different ways to construct and control the experiment. I won't get into the details of these now. They really go on and it can just be a construct of another that just kind of build on each other. I will briefly cover three control group designs and focus a little bit on the one that we used here.

The first trial is said to be sort of the standard in the industry. The control group is made up of members who are expected to be treated away from the entire population. The assumption is that everything about those members is equivalent to the members who would be in treatment in this study. Now, this is a view of this methodology. It's a methodology favored by pharmaceutical companies. I did it back in my academic phase as well. It's a very rigorous design. But even this method has some pitfalls. There are ethical considerations, and it's pretty expensive

in terms of time and money. Sometimes it is practically hard to implement. Say there's a large health plan that wants to roll up population. Providers have X number of patients in the program and Y number who are not. It's, I think, kind of naïve to think that there wouldn't be some negative effects.

This methodology may be perfect for simple experiments, but even this methodology does not help when you want to avoid some issues that we'll be talking about such as benefit design and normalization. That's practical and in some ways a better alternative. For most of the plans that decision is made after employers purchase. The alternative is made up of benefit managers who do not purchase the disease management program. So again, the section is done at the group level, not individual level. This is the method that we used.

In contrast, the next design is one where the two groups are made up of members who have opted out of the program. Those who remain make up the study and then the two are compared. Keep in mind that the control group is made up of members who do not purchase disease management. It has been shown in many cases that those people are different in many respects from those who remain. There are, of course, no motivations to change, but there are some change behaviors, and there also may be some differences in severity, even more severe disease or less.

In terms of the control group we used, it's a compromise between current control trial and previous control trial, but there is a fairly rigorous methodology. It's not subject to regression to mean. The experience of the control group was the same as the study group.

We'll be focusing in on the current control model. I will briefly cover this design. We have two prints that we could send out if you would like. This study was accepted into general disease management. We have the two groups—the control group and this other group—and they're mainly selected. Actually, John is part of the selection team. What we're trying to do is make the two groups as equivalent as possible, recognizing that the members are not randomly selected.

The control group is made up of self-insured PPO blocks of business that did not purchase the disease management program. There is a study group that some of the managers did purchase the program. One important note, of course, they're underlined here, is the members who are included in the study group would go to the participation level. If we had taken those members out, there would be selection bias due to participation. The exact same identification criteria apply for members for both the control and the study group. Evaluation is included in these two groups—the control group and the study group. We built the model to compare the base line to program year one, bringing in the data from the control group. Then we applied that control group model to the actual group, the calculated expected cost, and compared expected to actual. I'll describe this process in a little more detail shortly.

The current model line is sort of rigorous, but, again, it's not perfect. In the industry though, it's considered to be one of the best. I'm proud to say that we won a special award at this year's Blue Cross-Blue Shield Association, Best Is Blues, Best Practices Conference. That is what motivates us, I guess, to share this with you because we want to make sure everyone knows that there are many different ways to evaluate total and outcomes. This type of design can be used in health plans that have large ASA populations. In fact, we recommended it in one of our proposals to a large employer group, a large health plan, to use this type of a design. It does have some large group applications. I'll be getting into the process in just a minute. Like I said, the study is not perfect and, appropriately so, John will be pulling out some of the imperfections.

MR. STARK: I think one thing you can say to this point is the data selection is critical, and there is one thing we have all seen—garbage in, garbage out. This is one of the points we're trying to drive home: the model itself. There are tons of software packages that for certain studies you need a lot of preprocessing.

One of the things that we tried to do, even though we didn't select the data randomly, is to get these groups as close to equivalent as possible. By equivalent I mean we wanted these groups to have the same characteristics in terms of age, gender, level of benefits, level of disease severity, industry, you name it, just as

close as we could get it. The other thing that happens here is you can get them only so close, and then you have to make some external adjustments. The first step is selection. Then the actuaries come into play and we do benefit adjustment factors and all the black magic that we normally do.

For our study, age-gender distribution looks pretty good until you get to the eighteen-year-old males and then things start to fall apart. . On any other kind of study, we'd probably go back and make some revisions, but, again, we're talking about very small populations. We do not have a huge pool of members to draw from, so we made the decision that this would be sufficient in terms of matching, say, the age-gender characteristics, and we had to make these decisions for quite a few other ones.

Another thing we looked at was the base line claims cost. You'd like to think that if you matched these groups very closely and you've made some adjustments for claim costs, then you should come close. For asthma, we got pretty close, closer than you would expect. For coronary artery disease, we got pretty close, and that's probably more in the range you would expect to see a difference. For diabetes, not close at all. This is one of the problems we faced. This is one of the reasons you can't actually do a group-to-group comparison. This is the reason we use predictive modeling —to adjust for situations like this.

Let's talk about some of these external adjustments. Two of the more obvious ones have to do with the different kind of fee schedules and different benefits. Most, if not all, health plans have agreements with providers. Even within a health plan you don't have identical agreements. You can have them differ by network and region. You can have capitation and incentives. You name it, you can have it. What can happen here is if certain members go to lower cost providers, you can have either savings or losses simply due to the selection of the providers. Of course, you don't want that as part of your study. In our study, we used a normalization process to try to make sure that we charged everybody the same for every service and "reprocessed" claims to get everything normalized.

The next thing we looked at was benefit design. Benefit design has the same effect. If a group changes benefits, you can see higher or lower costs just because of the benefit change, so you need to make an adjustment for that. If you're doing a Medicaid study or a study on a big employer group that has the same benefits, you'll save yourself this step. But for us, again, we were using small populations and drawing from a lot of different groups.

Benefit design affects both cost and utilization. We only made adjustments for cost in our study. Utilization is much more difficult to predict. For big blocks of business, we all have models that can predict the change in benefit design. So you say, "Well, why can't you use those to do utilization adjustments?" We were dealing with small chronically ill populations, so whatever normal methods we use may or may not apply there. So we made the decision to ignore the effects of benefit design on utilization.

I'll digress for a minute and talk about what Michael refers to as cynicism, which I refer to as highly refined questioning and investigative techniques. Be leery of applying conventional wisdom to these studies. A lot of times when you use predictive modeling, it's going to be for things that you can't use standard techniques with. You use typical adjustment factors, for example, hospitals are 30 to 40 percent of claims. For this type of population it can be 50 to 60 percent, you just don't know. So don't fall into the trap of using just the standard actuarial estimates.

Another thing to remember is, in this case, benefit design can have some very strange effects for a chronically ill population. You may find benefits that do not encourage the use of services that may be beneficial down the road. Normally, we use benefit design to steer people, and that's done by the means of co-payments and deductibles. We've got three-tier drug benefits, all kinds of things. If you think about a chronically ill population, those may or may not inhibit these members from getting the services they need. One of the things we decided to look at was, is this true? We did a literature search to look at what happened with co-payments.

The first study said that when they looked at the chronically ill population, they found that under certain conditions higher cost sharing kept members from getting services. I was pretty surprised at this because, personally, I would think that if you had diabetes, no matter what the cost within a reasonable range, you would still get services. In contrast, the final study had a control study, and this group found that it did matter. One of the things to remember about this last study though is that it's ten years old. Managed care has changed quite a bit, but even so, the results are mixed. Since there's really no consistent conclusion, we decided not to make any adjustments for utilization.

Michael will talk about the actual predictive models and some other aspects of the studies that we've done.

DR. COUSINS: What we'll do is share an approach for using the model to adjust for the differences that John described earlier: the base line differences on cost for those diabetics and for the aged. We wouldn't have to use a model if the two groups were equivalent, but data showed they're not equivalent, so we have to make some adjustments. What we'll be doing is opening up the black box system a little bit and, hopefully, you'll see that it's really not all that complicated. You all have probably studied mean of regression and classification trees back in the actuarial sciences program. What we'll do is really just apply these techniques.

There are several different ways to measure financial outcomes using a control group. The way that we measured it was to build a model from the control group and calculate expected costs and then split the comparisons to actual. This is similar to an approach that another disease management company took just a couple of years ago. They also used a control group and modeled those results. But the control group was an opt-out control group, so the opt-out control group had models built on that population and then applied to a savings. There were substantial biases in that design that one has to be careful about. That's certainly a distinction between our study and that study. From a service end, all control groups are not created equal. You will have to remember that. In other words, there are dozens of different ones. They're not all equal.

The predictive modeling tools are used to answer the question, how much were the claims cost in the year 2000? The year 2000 is the target year. That's the dependent variable. I don't think we say that very often. The year 2000 is the dependent variable in this evaluation. As I said earlier, these time periods for the control group correspond to the time periods for the study group, so what's occurring with the control group is also occurring with the study group. That's pretty important, particularly when you start to think about confounds.

Getting back to model building, it's a pretty simple approach. All we did was list out the members who were there in the year 2000 and calculate their normal life costs or medical costs. For each person, we then went back to the previous year, 1999, and looked at their age, their gender, their diagnosis risk, utilization patterns, frequency of medications, et cetera. After doing all that, we got these modeling tools. What we used was one-year regression and classification of relationship between the independent variables or those drivers and the year 2000 member lives cost. By selecting out different variables, diagnosis groups, frequency and utilization, we could compare how well that particular combination is predicting the year two costs. Since the actual year 2000 costs are known, now the tool can adjust itself and select a different set of independent variables and rates. This is artificial intelligence in the vernacular.

There's a cycle of learning that goes on. It goes on thousands, tens of thousands at a time until the optimal drivers and rates are identified. The output of all this is our mathematical formula. This is an example of the formula that's produced by the modeling tool. The modeling tool physically determined these drivers and selected out age, gender, type of medication, A1C blood test, et cetera.

Taking a look at these variables a little bit more closely, there's age, gender, type of medication and A1C test. We looked at these variables closely, not just from a statistical point of view, but also from the point of view of a clinician. We had a clinician on our design team. One of the things we wanted to do was make sure that the independent variables that were being selected made sense clinically. Once we were satisfied that there was a decent looking model, we could then apply it. You need to understand how well the model performed.

The model performance used two different modeling tools. We used one-year regression and a classification tree, and the results I will describe are from the classification tree. The census size is 1548 people. It's not large, but it's not small, not too small. We have a score, which is simply the explanatory value of the model. It essentially measures accuracy. We had at our score 0.45 in four months. The score of 0.45 is pretty good. It ranges anywhere from 0.2-0.9, and similar to football and unlike golf, higher is better. So 0.45 is a pretty good score. We also looked at the group bar score and regardless of the number of groups we group, our scores are consistent and in the 0.65-0.69 range.

As I mentioned, with these two modeling tools, we wanted to be very certain in what we were doing. We're a pretty conservative organization, and we wanted to be sure that the results we were getting were actual. The results of one-year regression were similar. In fact, our score was a little big higher. We're pretty satisfied that our criteria in terms of performance were confirmed. We had conversion results. We knew that the independent variables made sense. We tested something I didn't describe, which was part over fitting, and now we could turn to actually planning the model.

All we do is simply run the claims data from the study group through this control group model. Essentially, what we're doing is scoring the year 2000 expected costs for the study group to the 1999 points from the study group. We're calculating the expected costs using the control group model. In the second stage, we'll be comparing those expected costs to the actual cost. The results showed the costs that we would expect if there were not a disease management program in place and the actual costs. The expected is higher than the actual, so there's savings. Indeed, it was $0.94 net PM/PM savings for when we finished our process.

I have two points I would like to make. One is, again, this method is pretty conservative. We actually calculated the results when we looked at that, and we used a different modeling tool. The results were a penny with the regression. The net savings were $0.95 net PM/PM. This application is good because it validates the approach that we used. The second point is the savings are a decent size. I mean $0.94 is pretty substantial on a population basis.

Given our approach and given the results, we felt that the evaluation is a good basis on which to make a decision. In fact, the decision that we were evaluating in our company was whether or not to expand the disease management program from the HMO into the PPO. This is the way that we made a committed decision to do that.

In summary, there are many types of control groups. I don't think I can emphasize this enough. When you hear the words "control group," don't automatically think that everything is right in the world, it's not. But all of them have strengths and weaknesses. It's important to understand what those are in order to make the correct decisions. Also, the concurrent control model design is adjusted for the population differences that existed. Again, our groups were unequal, so we had to make these adjustments. We couldn't just compare them directly.

MR. STARK: I'd like to summarize a little bit before we head to the last piece. One of the things that Michael has told you is that at each step we look at data, models and results from several different vantage points. The point here is predictive modeling can be presented as this great tool that's going to solve all of your problems; it's not. It's like any other model that we deal with. You got to be careful with your inputs. You can't take the results at face value. Your model has to make sense.

I think you can tell from some of Michael's comments, we had a pretty diverse team that looked at the results of the model. We had clinicians, statisticians, actuaries and data folks. In doing pricing studies or doing any kind of study with a predictive model, you're going to want that same type of group.

Let's talk about the applicability of these results. When we're constructing savings estimates and contract guarantees, we want to find out whether or not we can take the results and transform them without any adjustments. What we found is you can't do that. I don't have to tell any of you that you just don't take some number you calculated and apply it to another block of business

We took a look at our study group, and we looked at the percentage of asthmatics, coronary artery disease patients and diabetics. Suppose we took our average PPO group and we applied the $0.94. What would happen? We would probably save a little bit more. You have a few more asthmatics, a little bit more in terms of CAD, a little bit less in terms of diabetics, but pretty much you probably come out better.

Let's look at another pool of business where the diseases are more across the board. You come out very well and, especially if you were making savings guarantees or applying it to rates, your savings would be greater. The important thing is diabetes. That is where you get the biggest bang for your buck. You really get a lot of savings. Let’s consider another example. For asthmatics, we've got 1.3 percent; for coronary artery disease, 0.5 percent; for diabetes, one percent. Now, this could be a real problem if you just took the $0.94 and slapped it on rates or slapped it in your forecast for this pool of business. These are examples from our block of business. This would be medically underwritten individual business, or maybe medically underwritten small business. The reason is asthma can be treated and can make it through your underwriting process, whereas diabetes probably wouldn't make it through. So as you're looking at the results, be very careful about making adjustments.

We finally get to consistency of results. What you'd like to think when you do a study of this nature, is that everything that you look at points in the same direction. In other words, if you look at some clinical indicators and you predictive model shows a decrease in your costs, this clinical indicator should corroborate this. What can happen is if you don't look at some additional factors, you can have your costs going down. You can have clinical indicators saying that nothing worthwhile happened and, all of a sudden, your results are suspect. Even though you're using a predictive model, it can't take the place of really fully investigating all different aspects including clinical of the group.

Let's go over the key messages again. No one number stands by itself. I think we've beaten to death that you need to look at this from different perspectives and use different techniques to confirm some of the results.

Statements need to be taken with a boulder of salt and, like Michael said, with regression to the mean, which is a very critical piece of information. You need to test for that. People have despaired some of these techniques and, all of a sudden, apply a little science and you find out that may not be the case.

There are multiple ways to measure outcomes. We did this with design. We used pre and post in the control group. Within our control group we used regression and classification trees. So at each step, we took several different approaches.

Finally, clearly, the devil is in the details. We used some evaluation methodologies, talked a lot about strengths and weaknesses and talked about the use of predictive models.

FROM THE FLOOR: I was wondering with your study what were the participation levels that you were seeing in your programs? How did that factor in? What were you seeing in terms of savings rates for those people who were actually participating? What were the costs of the program versus the $0.95 cent savings that you're finding? Did you see that there were any particular features of the programs that you found were the primary drivers for the savings?

MR. STARK: If I remember correctly, the $0.94 was net. So that was a net savings, and Michael can speak to the participation levels.

FROM THE FLOOR: What was the savings? What was the cost? Was it two dollars per member cost?

DR. COUSINS: We did not speak to that but, actually, in the manuscript we do share that. It's roughly around $0.50 back at that time. It stayed around $0.50 PM/PM. You asked about participation, that was about 80 percent. Your third question was about what were the drivers of the savings; that's a fantastic question. In fact, that's one that we raised in our summary section as a place referral research. No one really knows which particular intervention is the one that mostly drives results. That's really our question is that focusing on compliance of A1C medications, you know, for a particular outcome. What's the best way to do it:

telephone, e-mail, by phone, fax, by the doctor, and who knows? But those questions are unanswered at this time.

FROM THE FLOOR: Did you look at the savings at all for those that were actually participating in the program for somebody that's diabetic versus those who were diabetic and not participating? What would be your expected savings in terms of the savings for those people?

DR. COUSINS: We did not, in this control group evaluation, look at the savings of participants versus non-participants just because of the size of the group. We have to do another evaluation and see what the results would be. In every evaluation that I'm aware of that we've done, there have been greater savings for those who participate than those who do not. But I caution you not to draw excessive conclusions from that.

MR. STARK: Yes, I think the point here is to not generalize these results in any way, shape or form. For this study, this is what happened, and that's really about it.

MR. DAVID A. SHEA: Michael, how were the rates, the rating factor K developed? Did you stress test those to see if changes in them produced measurably different results when you calculated expected costs next year?

DR. COUSINS: We did check test the robustness of each of our models. We did it in a way that doesn't directly answer your question, I think. There's a fancy name for what we did, and I can't quite remember what it is. We built the model and had, let's say, variable A through F that came out and those had particular rates. We knew that if we took out any one of those variables that another set of variables would come out. We mainly did this afterward, because what we wanted to make sure of is that no one variable was driving the model performance. That was done pretty extensively. We really wanted to make sure that the results weren't tainted, and so that's one of the ways that we did it. In fact, that's what drove us to use that second model.

MR. STARK: In terms of the calculations, when you can use any regression package. Just to reinforce what Michael said, we knew that our senior management was going to look at this. If they went forward, they would want to have as close to irrefutable proof as possible when they put this on all the business that we were right, so that if anybody came back and asked questions, we could address them.

FROM THE FLOOR: I'm curious if you started to measure subsequent years' results. You said you saw $0.94 savings in year one. Are you starting to see any indicators for whether that trend is continuing and which way it's going, if the gap is growing larger or staying the same?

DR. COUSINS: With this particular design, we have not. We have done a different type of historical control type math evaluation for the past into year three. What we've seen is that there are year over year increases in savings. It doesn't flatten out after one year.

It's actually not surprising to us when we started to look at the savings in the individual conditions. In fact, in year one there isn't savings in most cases with diabetics. The longer a person is enrolled, the less of a loss there is. However, it isn't until early in the second year that the diabetes population tends to show savings and then adds into earlier years. What we've seen is up through year three, and that's the furthest that we've gone so far, is that there will continuously be savings. We're not just seeking out one more and more savings for the same people, we're continually adding as well.