RECORD, Volume 28, No. 2*

San Francisco Spring Meeting
June 24–26, 2002

Session 42PD
Show Me the Utility: Practical Applications of Credibility

Track: Health/Long-Term Care

Moderator: CHARLES S. FUHRER
Panelists: THOMAS JACOB LEIBOWITZ
                   CHARLES S. FUHRER

Summary: How should credibility vary by size of groups for various coverages? When can a block of business be considered fully credible? How does credibility factor into a rate filing? To what extent is current experience more credible than past experience? At the conclusion of this session, participants gain a broader viewpoint of the uses of credibility and various credibility applications.

MR. CHARLES S. FUHRER: Welcome to Session 42, "Show Me the Utility: Practical Applications of Credibility." First, Tom Leibowitz will give an analysis of a situation in which there is a small group renewal rating. Then I will discuss some practical considerations in renewal underwriting, for almost any size group.

Tom Leibowitz became an FSA in 1999 and has 11 years experience in health care actuarial work. He started his actuarial work career at Blue Cross/Blue Shield of Massachusetts. From there he went to Private Health Care Systems, also in Massachusetts. Then he was with Actuarial Management Corporation, and now he's with Reden & Anders in the San Francisco office.

MR. LEIBOWITZ: Good afternoon. I put together a presentation on a practical application of credibility, although this session is really about the opposite of this. My presentation focuses on one practical application of credibility, which is renewal rating for small group medical insurance.

_____

I have a few goals in terms of what I'd like people to get out of this presentation. I'd like people to get a good understanding of what the small group health renewal process entails, as well as the phenomenon of changing results by duration.

Small groups are credible to some extent, so past experience does impact future experience. Group size is actually important. And, although the credibility formula should vary based on the loss ratio, the way you approach credibility or the way it is put into effect should not be completely constant.

The last thing I will discuss is a practice that comes up pretty frequently. This is what's called the "Cherry Pick" strategy, where another carrier has what looks like a very positive block of business. You take it over, and a year down the road, that 40 percent loss ratio becomes an 80 percent loss ratio. Then, you lose a ton of money on what was presumably a great block of business.

Although there may be one standard definition of credibility, as far as I can tell, there are a variety of them. Here is my definition of credibility, which is pretty loose: it is the idea that the credibility that is assigned to something is the weight given, relative to that given to everything. So, credibility becomes something that is very specific, and occasionally pretty subjective.

The renewal rating process is that of determining what an appropriate rate should be for a group in a given rating period. The way that you're going to come up with an appropriate rate should theoretically be based on the group's original rates and the experience of the block, overall. Also, if everything looks good, presumably there will not be nearly as much of a change in an individual group's rate as there would have been if everything looked lousy. The amount of time you have in terms of experience for that group, and the number of lives, are considered.

The basic concept behind duration is that the phenomenon that occurs in life insurance, in terms of select and ultimate mortality, also applies to health insurance—select and ultimate morbidity. A number of years ago, Bill Bluhm came out with the Cumulative Anti-Selection Theory. It was the notion that this takes place in individual health insurance. It also takes place with group health insurance, but not to the same extent. It is a lot faster than in life insurance, and it takes a number of months, not years, before the underwriting starts to wear off. It is also a lot more significant. The bottom line, though, is medical loss ratios increase as time from inception increases above and beyond normal rates.

Here is an example of what loss ratios look like by duration (Chart 1). I took a sample from one of my clients and looked at a pretty big number of cases to see what their loss ratios were like. This first dot, where it is around 35 percent, is the month of inception. So a whole bunch of groups that came in over a couple-year period had that loss ratio in the first month that they were in the program.

It increases pretty steadily, then there is a drop in the 13th month. This makes sense, because that is when the rate increase went into effect. Then it goes back up.

If we show it going out a few years, it steadily tapers off, but there is a small increase on a month-by-month basis, even above and beyond what you would see in terms of a standard change in the new business rates. Everything that I am showing so far is background to get to the application of credibility in the rating.

Table 1

Sample Case Study for Analysis

- **124 groups, 1968 lives**
- **Min = 2 lives, max = 140 lives**
- **3 cohorts: period 1 loss ratio under 50%, 50 – 80%, 80%+**
- **Groups per cohort = 60, 37, 26**
  - **Meets 80 / 20 rule**
- **Lives per cohort = 785, 773, 410**

I have a case study in which I took 124 groups—about 2,000 lives—and I put them into different cohorts based on experience in their first year (Table 1). Actually, when I am talking about periods one and two, the first period is going back to the first 10 months. Typically, you will not have the full year's worth of experience when applying the rate increase.

So, in our first year, the average loss ratio was 62 percent. For the second year, there was an enormous increase of about 20 percentage points to 82 percent. This is very typical. It is very consistent with what you will see. I have seen this with at least a dozen different programs that I have worked with. It is more pronounced on individual, but it is certainly very significant with small group.

Table 2

Sample Case Study for Analysis

‣ **124 groups, 1968 lives**

‣ **Min = 2 lives, max = 140 lives**

‣ **3 cohorts: period 1 loss ratio under 50%, 50 – 80%, 80%+**

‣ **Groups per cohort = 60, 37, 26**
  - **Meets 80 / 20 rule**

‣ **Lives per cohort = 785, 773, 410**

So I split the groups into cohorts based on their loss ratios in the first 10 months (Table 2). If applying traditional credibility and renewal rating methods, how do they work? How would using an alternative method work in terms of the accuracy? Presumably, the ultimate objective is to come up with an appropriate renewal rate.

We have roughly the same number of people in the cohort that has, as expected; very low loss ratios, and the same number in a loss ratio range, which is around the target. Then, there are significantly fewer cases and fewer lives with really bad experience. That is what you would tend to expect—the notion of 20 percent of the cases producing 70 or 80 percent of claims experience. So this cohort does a pretty good job of replicating what you would see under a standard small group portfolio.

There is often a lot of debate in terms of how much past experience impacts future experience. It definitely has an impact. A couple of years ago I was at a Society meeting where an actuary said, "I don't believe in giving credibility to small groups, but I do give credibility to good groups."

There are a couple of flaws to that statement. One is if you're calling groups "good" versus "bad," presumably you're giving them credit. If you've flipped a whole bunch of coins three times each and some of them came up with three heads, presumably you would not call those good coins. But if groups come out that way

you're going to look at them and say there is probably something to this favorable experience.

The other thing is that smaller groups have a significantly greater change in their experience from the first year to the second year for groups that are doing okay than for groups that are doing lousy. So good groups will get worse. Groups that are doing okay will get worse, but not nearly so much. Bad groups will stay bad, but get a little better. There is a mean reversion that one would expect with basic traditional credibility methods where you're looking at the average and comparing something to your average. Everything does revert toward that, but the degree to which that happens varies by loss ratio.

So, in this case, we can look at the difference between first and second year loss ratios. We see that the loss ratio in the first cohort nearly doubles. If someone were to offer to give all of his or her groups at some sort of a reduced rate, when there is an average loss ratio of 35 percent, do not cut their rates in half, expecting to get a 70 percent loss ratio the next year. I would leave them exactly the same. As we can see, the changes in the second and third tiers are not nearly so pronounced, but you can also see that each of these three cohorts has a worse average loss ratio than the entire block in the first year.

Here we have the experience on a case-by-case basis (Chart 2). The dark blue line shows all the groups from the lowest loss ratio to the highest in the first year, as well as the same group's experience in the second year. So, there is pretty strong volatility the whole way. This is why, on the lower end, they presumably cannot get much lower. They have got no experience. The potential for downward fluctuation is pretty small. The potential for increased fluctuation and worse claims experience is pretty high.

Chart 3 is what the loss ratios look like by duration. When we split that up into cohorts we can see the green one is what we would call the bad groups that have pretty lousy experience, but then it tapers off. We also have the good groups down at the bottom, which have steady increases in their loss ratios as time goes on.

Table 3

Results Using Traditional Credible Method

Goals – use credibility to predict accurately the future costs of each case, rate appropriately

Variance: Actual vs. Expected, based on

1. **Basic Credibility formula $C = [N / (N + k)]^{1/2}$**
2. **k = number of member months**
3. **Period 2 expected LR = C x year 1 LR $_{group}$**
   + (1 – C) x overall expected year 2 LR $_{all\ groups\ combined}$

K ➔	0	100	250	500
Under 50	-$373,283	-$10,815	$62,759	$105,898
50 - 80	-$40,479	-$71,255	-$80,522	-$86,604
80 +	$456,119	$132,951	$65,802	$27,147
Total	$42,358	$50,881	$48,038	$46,440

© 2002 Ingenix, Inc.

So here I am using what I would think of as a traditional credibility method (Table 3). This is when you take current experience, and apply credibility to it, making some sort of an increase for the fact that it is going from year one to year two, then balancing that with the non-credible piece, which is the overall anticipated loss ratio for the block.

We have got a comparison of actual experience versus expected for each of the cohorts from year one to year two. It shows that if you apply traditional credibility methods to the cases with really good loss ratios going forward and looking at that experience in the second year, it may seem like the company is losing a ton of money.

Whereas, you could give very high credibility to the bad groups. In this case, you might say, this is really great, because the experience is looking lousy, but it is not looking nearly as bad as anticipated.

So what does this tell us about the traditional method? To make money, do not give any credibility whatsoever to the good groups. Instead, just give them the same normal rate increase as for the block as a whole. This is because if you were to drop a client's rates because they are really good, then you'd get a big surprise at the end of the day. On the other end of the scale, if there are bad groups that

are likely to be really bad again, an insurer will end up giving them much more of a rate increase than they really need.

We can garner a couple of things from this that are pretty useful. One is that using the overall anticipated future loss ratio for a cohort does not seem to make a lot of sense. It does not make sense to balance out groups that have a 20 percent loss ratio with an expected 80 percent loss ratio; then to balance out groups that have a 120 percent loss ratio with that same 80 percent. That is not the point to where they are going to revert.

It also indicates that using an unadjusted loss ratio based on the experience of the first year does not make a lot of sense because of the dramatic changes from year one to year two.

So, I recommend three major changes to the way renewal underwriting takes place. One change is to do what we did here in terms of splitting the groups into loss ratios by cohort. This does not necessarily have to be done by loss ratios. It can be done by claims costs per member per month, for example, then use average expected loss ratio, by cohort, as the non-credible piece.

One of the biggest keys to properly working with credibility is determining what to use as that non-credible piece to balance with the credible piece. The way to do this is to look at your historical experience. Then, after determining which are the optimal cohorts, figure out what sort of a change is expected for each of those cohorts from one period to the next.

The other key is to modify what you're going to use as the credible piece. An example would be if an insurer has a 35 percent loss ratio for a case in the block with the good groups. If the average loss ratios go up 20 percent, the insurer should not switch from using a 35 percent loss ratio to using a 42 percent loss ratio for the expected credible piece for that group. Or, even worse, the company should not decide that since it was at 35 percent this year, it will use 35 percent again next year.

Instead, if any group is in that cohort because the loss ratios are doubling for that cohort, the credible piece should be a doubling of the loss ratio. Whereas, on the other side if there is a cohort with claims that drop by 10 percent, with the highest cohort, you would probably want to reduce the expected claims for this case by 10 percent and use that as the credible piece. Then, again, balance that with the same change to the overall loss ratio shift for the cohort as a whole.

If this makes a lot of sense, you may decide to vary the K. The K is the modifier in the credibility formula, based on whether you think groups of certain loss ratios are more or less credible. If in doing this review, you will probably see that the amount of fluctuation varies more on the low end than it does on the high end. Therefore, it may be best to give more credibility to groups with worse loss ratios.

Table 4

Results Using Alternative Credibility Method

Variance: Actual vs. Expected

K ➔	0	100	250	500
Under 50	$21,318	$43,835	$37,935	$33,770
50 - 80	-$11,876	-$13,023	-$12,868	-$12,648
80 +	$22,810	$12,279	$14,901	$16,840
Total	$32,252	$43,090	$39,968	$37,962

So here we do just that (Table 4). We have taken those same cases and have changed what we are calling the credible piece and the non-credible piece under the same basic parameters, the same modeling, only we have said we are not going to treat every case the same. We are going to treat small cohorts of cases as though they have a specific pattern and apply it to all the cases in that cohort.

It does not really matter what is used in terms of the overall credibility level in order to go from first to second year in terms of the accuracy of the rates. It would really depend on the goal. If the goal is to make as much money as possible, then a company may want to do this.

But, if the goal is to get as accurate a rate as possible, then this is a better way to come up with second year rates. One of the reasons that a company probably wants to come up with an accurate rate as possible is that if the company comes up with inappropriate rate increases for certain groups, then they may leave. This may be desired sometimes, and sometimes it may not.

Table 5

Impact of Size on Credibility

‣ **Compared variations in results period 1 vs. period 2 for groups under 25 lives with groups 25+**

‣ **Used simple formula:**

$$\frac{\Sigma \textbf{ Premium x (actual period 2 claims – expected)}^2}{\text{Total premium}}$$

Where expected period 2 claims = period 1 claims x $\frac{\text{(average LR period 1)}}{\text{average LR period 2}}$

Results: weighted average squared variation for under 25 was triple average for 26+

Regarding the impact of size on credibility, we have looked at variations in results from the first to the second year, to see what sort of validity there was to the idea that increased size does mean increased credibility (Table 5). I looked at it just for groups of under and over 25 lives. It looked like the variation was roughly triple on the small groups versus the larger groups. This is what one would expect, because when there are groups of two to 25, then one or two people having lots of claims or nobody having any claims is going to create a lot of variation.

What are some of the implications of looking at the size? One is that there is less fluctuation for larger groups, in terms of experience. It also confirms one of the biggest myths—when somebody comes along with a cohort, a new block of business to pass on. The myth is that if you have a bunch of small groups is not the same as one large group.

A selection of small groups is not going to be the same as having a large group with a 35 percent loss ratio, because it is not. Again, because of that, avoid any sort of a "cherry picking" strategy or something where you're looking at an individual cohort and thinking that the experience of that cohort is going to remain the same.

There are some issues that come into play with doing something like this. One is that in doing this analysis I assumed that I knew what the second year experience was going to be. Obviously, I did not know this information, but I really did not have

anything else to use for the second year experience. Presumably, though, when doing this sort of study, the patterns will stay pretty stable over time in terms of the differences in experience between first and second year.

In theory, I have only split these into three cohorts. In actual practice, you should split it into a greater number of cohorts to avoid the inconsistencies. This can happen when one group is at the top of one band and a second group is at the bottom of another, projecting them forward, then the one with the lower loss ratio looks like it is going to have a higher projected loss ratio in the second year.

Another thing is that when I did this analysis I did not have the rate increases for each case. Presumably, loss ratios will increase more, traditionally, when looking at a block of business, because those groups with the better loss ratios are going to get smaller rate increases. So, when looking at this sort of thing, you need to back out the rate increases that each case had, because on a go-forward basis you're probably not going to be applying your renewal rating in the same way.

The next factor is that I have only looked at two years worth of data. In theory, this should be a much broader study, looking at three and four years out. Then, also, look at the credibility that you would give to each of the historical periods when projecting data for years three and four.

State laws restrict rate increases pretty significantly. None of this matters in California, because almost all of your groups will get what you think of as a 1.1. In Massachusetts, everybody gets the same rate, no matter what, and there really is not any underwriting, so it does not matter. There is the whole notion of new business rates, plus 15 percent, so it limits the amount to which you can raise rates anyway.

It is necessary to come up with your own credibility formulas. I just threw out a formula that I used at a previous company. Actuaries need to go out and use their own formulas, and see what makes sense in terms of applicability to your own cohort.

It may vary based on the experience of the groups. There are other ways to predict future costs. One way is to look at the individual experience of the people in these groups, such as drug use and what sort of claims caused their experience in the first year. That data will be a very good predictor, probably better than anything else of their future experience.

It is necessary to have lots of data to do this. Also, you will need to pool high cost non-recurring claims. Do not just pool any claim over $20,000. Look at any claim over a given threshold and decide if it should be pooled or not, or if you should expect even higher claims, because somebody has a medical need that will lead to something more costly.

So what are the lessons that I'd like everybody to get out of this? The first is that you really need to take historical experience into account when rating renewals. It's not always put into practice. Duration, and the fact that experience varies significantly over time, should inhibit a person from taking unadjusted experience from one year and trending it to the next.

Both the credible and non-credible pieces should vary based on experience. Again, do not take the historical experience for the block as a whole. Split it up into cohorts to get a more accurate reading of how things should be going in the next year.

All groups become bad groups. Companies do well at holding onto groups and keeping a high retention on their business. This is not necessarily a positive thing, because with groups that have high turnover, you're going to get very good experience. If you have groups that stay around for years and years and years, then you're going to need significantly more of a subsidy of first year experience, which is much better than second year and subsequent experience.

The application of credibility is limited. One of the things that it said in the brochure about this session is that you will learn at what level a group becomes fully credible. I do not think groups really ever become fully credible even if sometimes a company runs out of things to balance it with. If you take 5,000 life cases with the same loss ratios, their experience would vary dramatically from the first year to the second or differ from each other from one year to the next.

I think you should experiment and figure out the way this sort of thing works best. I am providing some examples of ways to get a little more refinement into the renewal rating methodology. Again, there is a reason that "Cherry Pick" strategies never work and that is because small groups are small groups. They are a bunch of groups with very few people who have a highly random distribution of claims. You can not just look at the groups that happen to roll the dice in a certain way and come out very favorable, then assume that it is going to happen going forward.

Lastly, since this was supposed to be about practical applications of credibility and I have only given you one practical application, here are other places where credibility can come into play. One is in employer stop loss, both specific and aggregate. Even with specific claims you can use some credibility. The same applies with provider excess of loss.

In creating area factors, certainly, if you're doing an analysis of experience, you do not give every area 100 percent credibility, but you definitely have some sort of variation.

Regarding reserving, for the most part, health care reserving is done the same way it was done before people had calculators, and used methods such as four out of the last six months worth of factors. Then the non-credible piece is a made-up loss

ratio. There are all sorts of applications for credibility. The same goes with forecasting. With large group experience rating, credibility is very important.

Then there is initial underwriting. You can use credibility to weight different potential outcomes based on what is known about somebody's medical history. With pricing of benefit differentials, if you have a bunch of experience from different plans, and then there are actuarial assumptions about how those plans should be, then you want to use credibility to come up with a new set of benefit differential factors. The same thing goes with age rates. Seasonality factors are the same general idea.

Also, look at the experience of every agent a company has on an individual basis as well as general agents. Then look at credibility in the terms that an agent has only one case and it is either bad or good. You cannot really say that much about the agent. Whereas, if an agent has 40 cases, they can be lined up next to each other and the agent can figure out which ones to fire. That is the end of my presentation and now Chuck will talk about credibility in general.

MR. FUHRER: Thank you, Tom. I am celebrating my 25th anniversary as an FSA. I have spent 14 of those years talking about credibility. In 1988 I published a paper in the *Transactions* on group credibility. Since that time, I have been at a half-dozen Society of Actuaries meetings, discussing it.

In addition, I have written a bunch of small articles on credibility and have used it in my experience at both insurance companies, where I worked for a stock company, a mutual company, and one Blue Cross/Blue Shield plan that was a mutual company, and another one that was not a mutual company. Now I am in consulting.

My comments on the analysis will be in three areas. One is the treatment of the traditional model and how we might go about setting values within it. Then I am going to discuss the model that I used in a paper I wrote.

The basic credibility formula was N over M, plus K raised to the half, where K is the number of months. The actual formula did not have the square root in it. Today, Tom brought up the point that there is some value in putting that square root up there. But the true traditional formula, N over M, plus K, was probably developed in the 1970s by a number of authors who found different ways of justifying it. It essentially results from regressing one year on the next. You get a formula of that type. The constant K is the point where 50 percent credibility is reached.

There was a tradition of using another formula, in which a different constant, which I also have called K, was set at the full credibility. Then the partial credibilities were assigned by using the ratio of that, then raising it to the half power or taking the square root. I do not think there was any good theoretical justification for that formula, except that it looked good. But this was beneficial, because sometimes when things look good, it means that they fit our pre-conceived notions of what we

should be doing. Sometimes those ideas are more accurate. K is the 100 percent credibility point.

FROM THE FLOOR: In member months.

MR. FUHRER: Either in member months and N is the member months or in member years and N is in years. It does not matter.

FROM THE FLOOR: You can get a C greater than one there.

MR. FUHRER: It is capped at one.

FROM THE FLOOR: What is N in the top formula there?

MR. FUHRER: It is an exposure.

FROM THE FLOOR: But what about K?

MR. FUHRER: K is the constant.

FROM THE FLOOR: In other words, it should be N for number of member months in K.

MR. FUHRER: Yes. So how do we go about setting the values in the traditional model? One way is to use cohorts in a method that Tom used. I took his actual data and calculated the resulting credibility. The first thing I did was relate the loss ratios for each cohort to the average loss ratio for the whole group. Cohort one regressed two-thirds of the way back to the mean of 100 percent and cohort three regressed even further.

So the effect was that cohort one had 33 percent credibility and cohort three had 19 percent. Cohort two was selected deliberately to be close to the mean and therefore did not give a very useful denominator, so I do not think that it would be a useful value.

If you go back and solve for K, then in cohorts one and three there are relatively close answers, which belies the concept that we should be using a different credibility level for the good groups versus the bad. Now, K is in member years, which strikes me as being way too high, and partially illustrates the problem with the cohort method. By selecting groups to put into a cohort, you're losing a good part of your data by putting all of the claims from all those groups together.

The best method to use, which appeared in the original papers, was that K was actually equal to the process variance divided by the variance of the means. And those are not well-defined terms. They vary depending on the application in question.

In this case, take the process variance to be merely the variance of each individual's claims. By variance, I mean the statistical sample variance, the mean squared variance. Then the variance of the means, one could take as the variance that is still seen between each group's loss ratio.

This is pretty easy to understand, because if there are greater and greater differences between the groups, then K should come down. For example, you would have more credibility, because any claim experience will help to decide between them. On the other hand, if the groups do not vary very much, then it would be very hard to determine where they stand within that.

The process variance is the thing that clouds up the whole issue. It is that each person's claims are bound to jump around. Therefore, to the extent that it is relatively high, we are going to have less credibility. For example, K will be higher because K is the denominator.

Another method I have dealt with for estimating K includes the variance method. For example, where we only use one year's worth of data; whereas, we could use successive years and look at correlations between them.

It is worth pointing out when we look at cohort two we do not want to give it much weight. Inherent in the variance method is the fact that the groups that are further away from the mean in the first year will be given much more weight. In fact, the weight is equal to the squares of those deviations.

That is both good and bad. It is good because that is where the action is. The groups that are far from the mean are the ones we are most interested in. On the other hand, it puts so much weight on those that are very far, but sometimes only one or two values can actually dominate the experience. That is why it is useful, as Tom pointed out, to cap the claims at a maximum amount or pooling point. This will lower the process variance and generally give us more credibility. The price that is paid is that we will have less information. If we use a maximum claim of zero, then we will have zero credibility, so there is some place that is optimal.

What is wrong with this traditional model? On a practical level, we normally get K greater than 100 life-years. If we do that, as the formula indicates, a one-life group for one year gets less than one percent credibility.

This is not enough when we look at the fact that there are certainly repeatable claims from year-to-year. If one person in that one-person group is very sick, then that person is much more likely to be sick in the following year. In fact, that is why we do individual underwriting on one-person groups, if we are allowed to. On the other hand, if that person had no claims, that is good evidence that he or she is pretty healthy and is likely to stay that way in the following year.

I want to point out that using the "N over M, plus K" formula for group insurance was wrong from the very beginning. "N", in those formulas, was always the amount of time for a single risk. It was never intended to be a formula for group insurance. The constant "K" was even called a time constant because it related to the time of exposure. Casualty actuaries have been using this formula for many years, but once again, "N" is the length of the exposure, not the volume.

I came up with a fix for the problem in my 1988 paper, which is in the Society of Actuaries Library. You can easily find all the references by merely going to the Society Web site and doing a search on the library. I urge everybody to use library—it is a tremendous step forward compared to the old days when we had to look through indexes.

My new formula is "K1, plus N, minus 1 over K2, over 1, plus N, minus 1, times K2." So K1 is the credibility for a one-person group. I have looked at data many times and invariably, I get an answer that is around 25 percent. This is a little higher if we cap the claim at some pooling point and a little lower if we do not. Sometimes it has been as high as almost 30 percent. It is sometimes as low as 20 percent.

However, it has always been in that ballpark. Surprisingly, it has always been that a person who has no claims in a particular year on the average will be about 25 percent better than the average insured person. This is looking at, generally, non-underwritten business in group insurance.

Then, when looking at the shape of this formula, the square root that was used by Tom actually looks a lot closer to the shape of this one. So, there is something to be said for that as a simple cure for this instead of having to estimate two constants.

Estimating K2 is a lot harder. A lot of groups are necessary, because of the covariance between the groups, causing it to jump around a lot. The variation is high. I have gotten answers for K2 between about two percent and a half percent per member, per year.

The paper also dealt with a number of other topics. One was finding the optimum pooling point. The conclusion was that although there was an optimum point somewhere, it tended not to be so optimal that it mattered much if you used a lot higher or a lot lower point. But, eventually, using none or a very low one that was far from optimum.

In addition, I thought there was a possibility that good groups ought to get different credibility than bad ones. However, I did not like the resulting formula that had a bend in it. The rate of change with experience of the final rate when it reaches the mean suddenly changed.

So, I thought it was better to try to fit a smooth curve to that. I gave the formulas for using a quadratic, being the one with the least amount of parameters. This is a little poor, because eventually, the quadratic comes around and changes direction. So it might require a certain amount of manual fixing, but it seemed to be one reasonable way of handling that problem.

I also dealt with figuring out how to change the credibility formula when there was competition. In addition, I was concerned with one way of rating specific stop loss. I assumed that the problem was whether we were going to rate the specific based as a percentage of the experience claims, trended of the individual group or based on the manual, which presumably adjusted for things like age and sex, but not for the experience of the group.

The stop-loss carriers use the manual rate largely because they do not think the experience of the small claims says much about the high experience. When you put it together, use somewhere in-between. Those are the factors that came out in the paper.

That concludes my comments on what Tom presented. If anybody has any questions or comments, please ask them now.

MR. TIMOTHY DILELLIO: My question for either of the panelists is, in looking at the experience by duration, first of all, you mentioned how the experience gets worse over time. Why does that occur? What causes that to show up? I ask because that is what the data says, and I have seen some of my own data suggest that as well.

The second question is, does that apply equally to medical and pharmacy claims?

MR. LEIBOWITZ: I can answer the first part. I probably cannot answer the second part in any great detail, so I will talk in general terms. Some of the reasons for the worsening experience include the fact that a company is underwriting the groups. Theoretically, it would be less pronounced in places where you're not underwriting.

Another reason is the inherent positive selection, which comes from the fact that a case can go out into the marketplace because it thinks that the rates it is getting are too high, relative to its experience. If its experience is really good, the group thinks it is possible to go somewhere else to get a good rate.

Another reason that is probably true to some extent, but I think has less merit, is that people have to sign up. They have to find new doctors with a new plan. There is actually some merit to that because there appears to be some pent-up demand in the later months. Look at months of duration 10, 11 and 12 and see loss ratios that may hit 105–115 percent of the average second year loss ratio adjusted for rate increases.

So, the fact that there is this pent-up demand is an indicator that there is something going on above and beyond the wearing off of underwriting, because people may be putting off services as they get used to the new plan. As far as pharmacy goes, I do not know to what extent it may vary, but I assume it is fairly consistent.

MR. GORDON TRAPNELL: What kind of insurer was this? What kind of underwriting did he conduct? Was this a market-dominant insurer or a small player in its market? Was it in multiple states or concentrated in one state? What kind of a benefit package was being offered?

I was more struck by how small the select effect appeared to be, because I am used to seeing much larger select effects on small group cases.

MR. LEIBOWITZ: It was a fully insured product, more or less first dollar medical traditional PPO plan. Deductibles ranged from $250–1,000, with 90 percent or 80 percent co-insurance.

MR. TRAPNELL: Were drugs covered?

MR. LEIBOWITZ: Drugs were covered. There were three-tiered drug cards. Typically it was a 10/20/30 type of a plan with $15–30 office co-pays. What else?

MR. TRAPNELL: Level one...?

MR. LEIBOWITZ: There were standard underwriting forms.

MR. TRAPNELL: Did you pay any attention to it?

MR. LEIBOWITZ: Yes. Every case is underwritten using one of our competitors' underwriting manuals and not our underwriting manual. Average underwriting load at issue is in the 20–25 percent range, predominantly in Texas, which has the rating band of plus or minus 25 percent, so it is 1–1.67.

That is about as standard as you will get in terms of a first dollar medical program. As far as presence in the market, I do not know exactly what their market share is. It is certainly below 20 percent, but it is not so small that nobody has ever heard of them and that they would somehow be getting a market advantage relative to other carriers.

MR. THOMAS BUTZEN: Chuck, when you did your modeling, I noticed you used two concurrent time periods in the study that you had published. When you do your renewal underwriting, you usually do this three or four months ahead of time. What is your opinion on how much lower the credibility would be if you had put a four-month gap between the two time periods? When I have done this, it reduces it.

MR. FUHRER: I am hesitating because I looked at the way the credibility tends to fall off over time, and I wanted to come up with a formula that continuously does that. Presumably, the older the history, the less relevant it is. This is because as a measure of the individual, or that 25 percent, it tends to wear off, since the people's conditions change and some of them leave the group.

But, in addition, whatever group characteristics that are measured from the high credibilities of the larger groups, they presumably change over time. Also, there is probably more randomness during the periods. So, when putting all that together I expect that the older the experience, the less credible.

On the other hand, I have never gotten enough data to peg how quickly that goes. When I have done this sort of work, both at the insurance companies and for the factors that my consulting firm uses, I had to use the little experience I had and fill it in with some guesswork. But, I am sure that whether it is a gap in experience or if we're trying to give credibility to the prior year's experience, it would tend to go down.

In addition, the larger the group, the less amount of time that you want to include, if any at all. At least give the lesser amount of weight to the older experience. However, I do not have enough data on groups to completely be sure of that.

Then, there is the question of incomplete experience, such as not having the runoff yet so when you're looking at incurred you don't yet know what they are. I gave some answers to that in a talk a number of years ago. Basically, increasing what amounts are analogous to the process variance by the amount of uncertainty you have about the actual claims lowers the credibility.

MR. ERIC BEST: I have one question for Tom. What was the typical turnover within a group for those that you observed? In other words, it is not going to be the same employees year after year, so what was the turnover like?

MR. LEIBOWITZ: That is a good question. The only reference point that I have is from doing their compliance statements and seeing when groups fall in and out of what is considered small group, large group or below five, and such. However, monthly fluctuations were fairly small—they would be from anywhere between zero and five employees. I do not remember ever seeing a dramatic fluctuation in covered lives. It was not as though any companies went from five to 40, then back down to three.

MR. BEST: It is not size that I am talking about. Rather, did exactly the same people need to be covered in two successive periods to be dealing with it?

Because, presumably, a group with 50 percent turnover from one year to the next is going to have lower credibility than a group with exactly the same people insured in two successive periods.

MR. LEIBOWITZ: That is definitely true. I can say that I did not do any analysis to see what sort of turnover there was within specific individuals in the groups. The one thing that I did hold constant is that every case that is shown was only there for the full two years that came in within a specific timeframe, then stayed for 22 successive months. I did not do any analysis in terms of individuals.

MR. BEST: My second question has to do with how you developed the loss ratios. Were those based on rates that were actually sold to the cases or were you comparing the cases against some sort of normal manual?

MR. LEIBOWITZ: They were actual historical loss ratios. The big adjustment, that in theory would have made the most sense, would have been dividing the second year premium by the premium rate increase for each case for months 13 and up. This would normalize or adjust for the fact that not all rate increases were the same.

MR. BEST: Thank you.

MR. FUHRER: I want to point out that in the '88 paper that I wrote, I included a formula that adjusted for a known turnover rate. That is, to multiply K1 by (one minus the turnover rate, or the persistency). Then, in order to keep the formula at the right level, instead of N minus one and the numerator and denominator, N minus the persistency.

In any case, the effect that it would have is to directly lower that 25 percent. That could be based on the probability that each person turns over as opposed to what actually occurs, since that is presumably not known.

MR. WES EDWARDS: Chuck, something you said with regard to a slightly different application was about the decreased credibility given to older periods of data when forecasting next period data. You were saying that is in part because a medical condition that carries over from the most recent period is the best indicator of the next period.

Take, for example, an application where one forecasts claim costs per capita five years out. Say it is for a retiree medical evaluation with claim costs that with trend, will go up significantly in the future.

In that situation, wouldn't more data or even older data get slightly more credibility, since the next period claims alone are not that important? This would not even be, necessarily, the next group of retirees, since that group will change. There is also a concern about not just one future period, but many future periods.

MR. FUHRER: That is an interesting question. I had not done much work on retiree health until recently, when my current employer and I spent a lot of time on it. I

completely agree with your conclusion that next period claims alone are less important when predicting many years in the future; whether using experience from last year, the year before, or the year before that.

In fact, when we do retiree health valuations at my company, we use a formula that is very flat for the older years of experience. We used one for active predictions of next year's experience, choosing one that goes down much more sharply than for the older experience. So we have, indeed, gone along with what you just said.

A related question, of which I have a less positive understanding, is, should you assume that the group tends to regress to the mean as time goes on?

There is also the question about whether FASB 106 actually requires a company to look at it. It could be that if the group is currently running quite a bit higher, then maybe you would use a lower short-term trend until it came back into line to what you would expect.

Currently, we are not doing that. We have taken the opinion that it is more in line with what FASB suggests to not include that. But, certainly, there have been some debates about that, so I do not know the right answer.

The problems with the theoretical model were dealt with in my earlier talk. However, there are a few more situations that I would like to mention.

One problem is that, to some extent, you will regress to the mean. The question is, what does that mean? Presumably, it is some sort of manual rates, but now it is very dependent on how sure you are and how good your manual is. There can be real problems with collecting enough data. Heath insurance costs vary considerably from area to area, for example, and you have to have a tremendous amount of data to be able to accurately predict that effect. Not only that, it would have to be very up-to-date data, because local trends can be much different than national ones.

The other problem when working with an insurance company is to try to come up with a good credibility table, perhaps the rest of the insurance company does not agree with this wonderful model that you've created. On the other hand, sometimes these people may know something that is helpful. An example is that the medium-sized group very often is not given a lot of credibility, at least by a lot of the models. This is perhaps because of the "N over M, plus K" problem.

The sales department generally wants higher credibility, because it is a lot easier to present to the client the fact that we took its experience and trended it up. Our trend rate is way too high, but this is better than the black dots that the manual rates become to the outside world.

Chart 4 is the model. It is a nice, smooth curve, with the 25 percent. The sales department will have a different one (Chart 5). They feel like a squaring of the curve occurs. They realize that a curve ratio can not be used with small groups, but then why not just go up to 100 percent and be done with it?

Are there some more questions or discussions?

MR. HOWARD MAYBERRY: I would like to know if it is practical to separate claims between those caused by chronic conditions and those caused by acute conditions when determining the credibility.

MR. LEIBOWITZ: It is imperative, not only for looking at credibility, but for doing any sort of rating or initial underwriting, to have a predictive model to take the claims and model them forward. If somebody breaks a leg, for example, I am not a fan of removing the entire claim from the person's experience, because stuff like that always happens. It seems that the same cases run into that more often than if it were just random.

Not only should you separate the two, but you also need to look at the individual chronic condition and sometimes even inflate it. Or, typically, when you think of pooling, you think of an amount to subtract, and I think sometimes you'd want to come up with an amount that you would want to add as well.

MR. FUHRER: I will point out that we do not want to throw away that data. We have lots of data on a group's health conditions. And, under the rubric of doing risk adjustment, certain firms will actually predict claims based on the conditions.

My only comment is that we do not have enough data yet to figure out the credibility of the total dollars of experience. We do not know exactly where to cut off pool claims and we do not generally even look to see which claims are ongoing and which are not. Also, we do not have enough data to say whether we should go back in our claims data two years or three years, but now we will get to the detailed condition analysis?

It seems like we are jumping over the more obvious things, to go into a lot of detail. I still have not seen where doing risk adjustment for the purpose of projecting claims is necessarily any more accurate than using manual experience.

MR. GORDON TRAPNELL: In either of your experiences, has the relatively small select effect that the set of data indicates been the general experience in recent years? Also, does anybody in the audience have comments on that? I am still struck by how small the select effect was. I am used to seeing much larger effects for medical underwriting of small group health insurance.

MR. LEIBOWITZ: Mr. Trapnell, what sort of a year-one to year-two shift have you seen?

MR. TRAPNELL: I have seen ratios like 35–40 percent in the first year; something in the 60s–70s in the second year, jumping up close to 100 percent the third, then going well over 100 percent in subsequent years.

But, buried in that are so many effects of rate increases, selection on the rate increase, competition, which is why I asked all those questions.

MR. LEIBOWITZ: Lapse and all those things? At the company where I worked before, Reden & Anders, we specialized in small group health, and every single client that we had looked like this. They had roughly 20–35 percent of an increase from year one to year two.

Since I have been here, besides this case, I have also worked with another small group carrier with tens of thousands of lives and the same pattern. Interestingly, that client had very high persistency. I would not say very good persistency, but it was very high, with a percentage in the high 80s to low 90s. Year after year its average duration was close to five years (60 months), which is much higher than typically seen in small group. The pattern that is held there is the same one that I see here.

Typically, I have seen 30–40 percent for the first few months, but then it starts to get up into the 60s. It all depends. I have seen it very stable once it gets three or four years out, and I have not seen it continuously rising at a rapid rate after, say, 24 months.

Somebody asked earlier if we saw this happen with prescription drugs. I can say that it does not necessarily happen with drugs, but that the same phenomenon does happen with mental health claims.

MR. BEST: I will respond to the question about the selection curve. One of the things that we did to try and throw out anything that was going on with rating increases was to look at just patterns of claim costs per member, per month, as time went on. Our actual deterioration in the first couple of years was something more like only 12–15 percent, ignoring what we were doing with rate increases and just looking at the deterioration of claim costs. In fact, beyond the second or third year there was very little deterioration at all.

MR. LEIBOWITZ: One thing that this indicates is that there needs to be a subsidization—even at 10, 12 or 15 percent of future experience by first year cases. If you were to rate your first year case at the proper rate, on average, the experience is going to be 15 percent worse after trend. Then, you would need to give every case the maximum rate increase permitted by law in most states to merely break even with where you were in your first year.

So, as a result, when doing initial small group rating, think in terms of a subsidy or an active life reserve type of thing, or an active group reserve, in this case, because you're going to get this sort of subsidization. So, if there is very high persistency, in theory, your rates will be higher than those of a competitor who has the exact same experience and much higher lapse rates.

Any other questions or comments? I would like to thank the two of you for taking the initiative to have this dialogue.

FROM THE FLOOR: I have one question and one comment. One comment, first. Chuck, regarding the risk adjusters and the study, *A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations,* published on May 24 by the Society of Actuaries.

I am confused about artificial medical intelligence. Everybody is claiming everything that looks like R-squared is still not in an acceptable range. R squared is very low, so I do not think it will work.

The second question is not only for you. I have questions for Chuck as well as Mr. Leibowitz. Why are we looking at credibility on the effect of trend? For pricing, in group insurance, we look at the previous year's trend, and apply that same one for next year's rate increases. What does it mean?

We are giving a 100 percent credit. Looking at the past, since 1965, we have seen that for three years we can be successful, and at the end of the third year, we make a lot of money, because our projected high trend turns out to be a low trend.

So, again, after three years of low trends, we have seen high trends coming back. It looks like with three years, this follows some kind of trend, but since 1992, it does not follow one. So why are we not looking at the credibility of using the pricing trend?

MR. LEIBOWITZ: We are actually not doing that. As an actuarial consulting firm, in looking toward future trends, we are actually looking at changes in provider contracts. We are working with health plans, but we are also working a great deal with medical groups and with hospital organizations to find out what trends they are seeing. Both on the commercial and the Medicare sides we are looking at things like the distribution of the mix of services and the changes in contracts. We are looking at changes in Medicare and trying to figure out how that is going to be reflected in changes in physician fee schedules.

We are predicting, for the most part, 15–18 percent trends on everything but drugs, which should be a little higher. We certainly did not see those trends from 1999–2000 or 2000–2001. This is the historical period on which we are trending forward. We are trying to take into account as much information as we can.

So, your question was rhetorical in terms of what we should be looking at in order to better predict future claims, other than simply past claims experience. You are right that you would fall into that trap. So, I think there are a number of things to look at if the information is available.

How would a traditional insurance carrier do that, when they do not have the luxury that we have, in terms of available sources of information? The key is internal communication among different departments to find out what is going on with each area.

I personally spent five years with a Blue Cross plan and knew nothing about hospital contracting. Then I worked for a network where I learned very quickly that the best way to find out what future hospital costs will be like is to deal with the people who are doing the contracting.

FROM THE FLOOR: My comment is about what I'm doing with my company. We are looking at the last five years of trends and are throwing away two outliers and looking at the average of three, and we will use that trend for pricing. So, we don't know whether the State Insurance Department will buy into it or not.

This is because, in the high trend area, we might be able to provide some low trend pricing, so we might have some losses. But on the other side, when low trend time will be stable, we will not incur high losses. So, whether the State Insurance Department will be bought into it or not, we do not know. That is the way, I think, that we can apply credibility theory and bring stability into the marketplace.

MR. FUHRER: I would like to respond to your first question. Consistent with what Tom said, in my opinion, the selection of a future trend is much more in the realm of the economic sciences, or at least an application of time series analysis. I think the kind of credibility models that we look at here tend to break down because of outside influences. Without looking at those influences and understanding them, I do not think that you can get a very reliable answer. Even then, predicting the future for this type of thing is very difficult.

MR. LEIBOWITZ: One of the risks when doing long-term, stable averages of trends like that, in terms of rating, is that you will outpace the market sometimes. This can put you at a big competitive disadvantage and you may not sell anything for years. Then you may be under the market trend for a number of years in which you are increasing your rates less than legitimately possible to remain competitive.

The other issue that you brought up was the idea of dropping the high and the low from a five-year period. I have never been a fan of dropping a high and low anything for any type of application. I have always disliked the idea of taking, say, the middle four of six months, then giving the superhighway to these four months, and none to the other two months just because they happen to be on the ends. In

that case, seven months back would be worth nothing, but six months back would be worth a great deal.

So I think you are right. It is definitely another application of credibility where you have to look at the value you want to give to each previous time period. But I also do not think you want to throw anything out, because a lot of times the high and the low can cause a distinct pattern to the overall impact.

For instance, in reserving, if you throw out the high and the low factors, you're always going to be understating your reserves. In theory, this is because one is going to be greater than the other, in general, in terms of the fluctuation. If you were to do that with a rating trend, I suspect that you would more than likely throw out a high that is much higher, relative to the mean, than the low that would be dropped. Any other questions or comments?

MR. FUHRER: I thank everybody for coming.
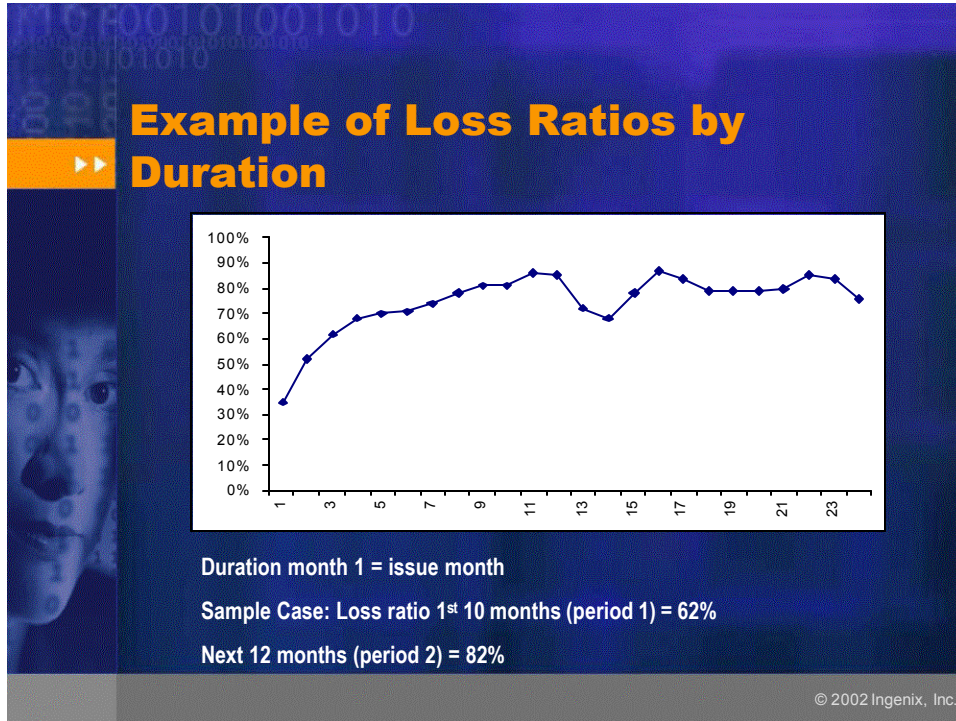
MR. LEIBOWITZ: Thank you very much.
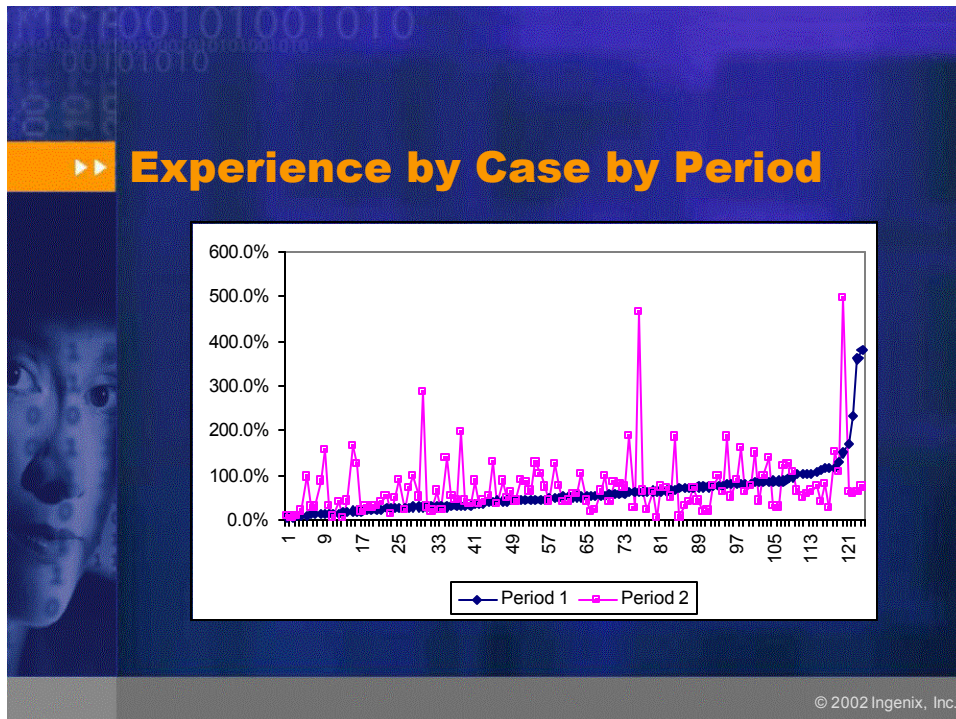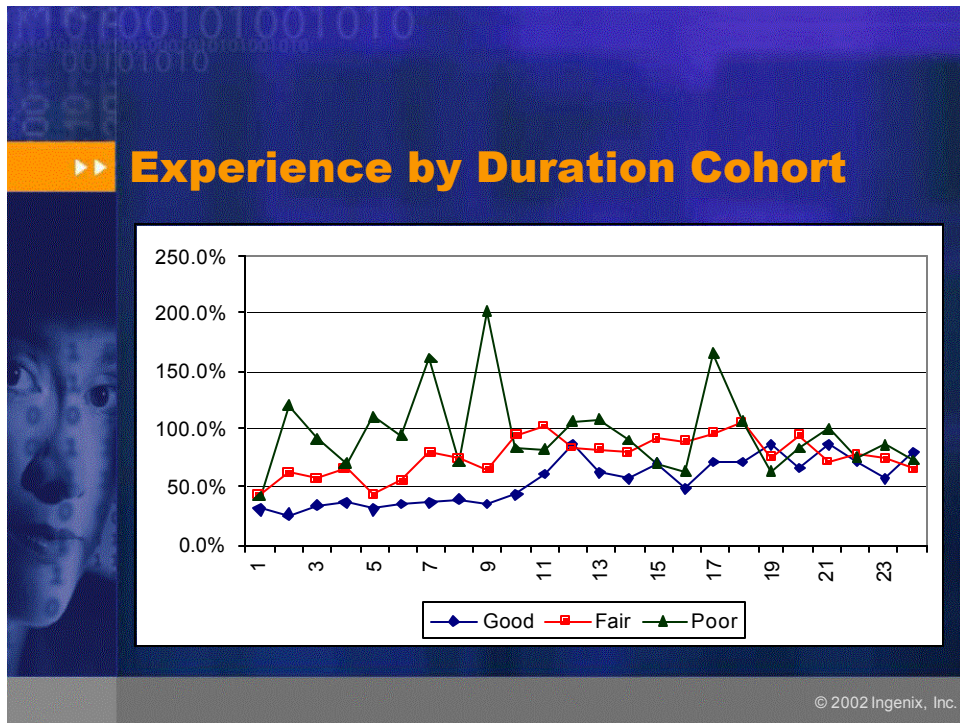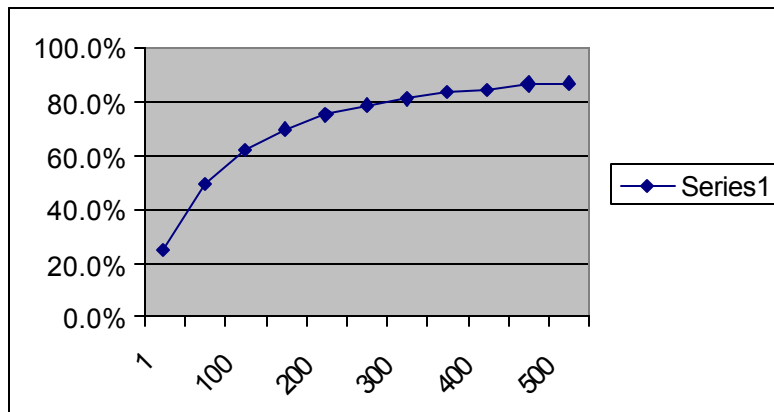
Chart 1

Chart 2

Chart 3

Chart 4

Your Model

Chart 5

Sales Department Formula