MCH EPI Conference

Using Population Attributable Fractions to Assess MCH Population Outcomes

December 8 – 11, 2008

DEBORORAH ROSENBERG: Percents, right, going the other directions for row one, row two, row three in the reference booth. And then the relative risks and what do you notice again, about these relative risks? They're all, except for using the Bruisey Method, they all have the same denominator, right? It's all the last row percent. The risks in the reference group and in the sense of those are the Bruisey Stratum specific relative risks. Stratum, being defined differently than we are used to thinking about it. We can do this both ways. One thing that I think is very intuitive about the Rothman approach is, you can actually see that in the Rothman approach, the summary PAF is indeed a sum of the component PAF's. Remember those component PAF's are the little, small pieces of the pie that represent every possible combination of risk factors in the population. So this one is for both smoking and cocaine, this one is for cocaine only, this one is for smoking only, and actually we have this is the funny one, this one actually drops out of the equation, zero, but it is there. I put it there only because moving on to Bruisey, you need to consider it. The fourth PAF is the PAF for the reference group, which is the reference group over itself, so it turns out to be zero, right? There is nothing to be eliminated. These women are neither smokers nor cocaine users, but it is actually there. The summary PAF is the sum of, it is a four by two table, there is four rows in the four by two tables. Therefore there are four component PAF's. We get back to this point

one six. So this is almost crude PAF's, if you will for this is a good question actually, this is if you are both a smoker and a cocaine user. Six percent, this is interesting actually, how do you interpret this? Can we get Kristen, because it is not eliminating both from the population simultaneously because you still have all women who did both

KRISTIN RANKIN: Smoking and cocaine, women who were both smokers and cocaine users, you would reduce the birth weight by six percent. You'd reduce low birth weight in the population by six percent if you got that specific group to quit both exposures. So that is why we don't find it very useful.

DEBORAH ROSENBERG :Then you'd have to go on and say yeah, we've eliminated it for those women, but here now we have women who are only cocaine users, and if we had a program that was for them and we could eliminate their cocaine use in the population we'd reduce low birth weight by another four percent. Then we'd still have smokers in the population, who don't do cocaine…these are just the pieces part, these are the components of the summary PAF, but they, themselves, I don' think we would report these directly. In fact, when you use the Bruisey formula, his approach doesn't lend itself to that. I've rewritten, this is the Bruisey formula, here we have the PJ's, those column percents. Here over the relative risks, this is the fourth one, the reference group, so it's dividing by one, but the point six six is still in the calculation so you need it there. In the Rothman formula, the fourth component PAF drops out because it ends up being zero, but here in the Bruisey formula, you've got all four in there. It is just showing you again that just regardless of the way that you do the formula. These are not interpreted, this is not anything. These pieces of the calculation, these are not component PAF's or

anything. There is nothing to report here except this, the final result and it is the same. We've got these component PAF's, we've got a summary PAF, but again we haven't disentangled smoking from cocaine use. We still have this piece of the pie, that women who do both and we haven't managed to disentangle it at all, which is why it was hard to interpret those component PAF's. We are not there yet, we have not reached our goal of getting a separate…being able to report a PAF for smoking and separately a PAF for cocaine that is meaningful. So typical in Epidemiology, how are we going to start thinking about adjustments and moving towards disentangling those two, in our case, two risk factors? Again here are the formulas; we've looked at these before. The difference here, I wanted to point out to you is, as usual in Epidemiology we also start thinking of when we move to multi variable confounding and affect modification and …so now this is the same formula that we've looked at, at least two or three times now, the difference is up here, this is the way that we've been looking at it, so you Relative Risk J minus one, Relative Risk J, those are again stratum specific, depending on how you define your strata. Here in if you think of the Rothman approach, what we end up doing is…this is the number of all exposed cases in all the strata and over all the cases, over all strata, and we are using the adjusted measure. So if we stratify, we are used to stratifying, we are thinking about the problem in that way. Rothman has provided us with two different ways we could calculate PAF. One saying we are not making any assumption about affect modification, so we are not assuming homogeneity, we are not assuming heterogeneity, either, but we're stratifying, let's do stratum specific estimates and if we can assume homogeneity, we can use this formula to actually take advantage of the fact that there is no affect modification going on, and we can make it slightly

simpler and use the adjusted measure. Here you are summing over how many strata you have. Here you are using all the cases across all the strata and the adjusted relative risk. We'll talk more about this when we get to modeling. As you start thinking about the differences and how we are going to handle confounding and affect modification, now that we are talking about PAF as opposed to thinking about just the odds ratio over relative risk. This is as usual, if there is heterogeneity, we probably don't want to use an adjusted relative risk. Even as input to the PAF calculation, we again will talk a lot more about that particularly when we get to modeling. For the PAF though, this is probably a little bit different than what we usually do, even if there is not affect modification, at least if you have enough sample size, so that you don't have sparse data in you strata, there is nothing wrong with using the strata specific estimates, if you want. You'll see that that is actually what I have done. I'm going to reorganize the risk system data that we have for smoking and cocaine, and right now I'm reorganizing them in a traditional sense. This is the way we think of stratification. I'm looking in each of these tables, I'm looking at the smoking, low birth weight relationship, stratified by whether the woman is also a cocaine user or not. This is our way of stratifying. When I do that, what do I get? I get two stratum specific relative risks. As you can see in this case, it doesn't look like there is affect modification. We've go one point three seven, one point three six. That is about as homogeneous as you can get, that is just the way that it worked out. Using stratum specific estimates, I'm going to calculate now the adjusted PAF, which is the sum of the two stratum specific PAF's, if you want to think about it that way. We've got the prevalence estimate from stratum one, the prevalence estimate from stratum two, and the relative risk from stratum one, and the relative risk

from stratum two, and when I do that, adjusting from cocaine, I now have a population attributable fraction for smoking of seven point six percent. Again, this is the Rothman approach. I certainly feel most grounded, because this is the way that we typically think about doing the stratification. So here it is the other way around, now I've stratified by smoking and each two by two table is for the cocaine by low birth weight relationship, again I'm going to sum…no affect modification, which it wasn't the other way around either, I'm again summing the two stratum specific PAF's to get an adjusted PAF for cocaine use of nine point nine percent.

Now let's do it the other way, the Bruisey approach; here again, now we have a different way of thinking about what are the stratum specific risks that we are talking about here. For the PAF for smoking, controlling cocaine use, the four P's, that is the prevalence estimates, the four column percents, and the four stratum specific relative risks. Look at this, RP1 over RP2, so that is row one over row 2. Then we've got row two over row two, which is nothing. Row three over row four, and row four over row four, which we know is not really meaningful. These two equal one, these two relative risks equal one. We know that when we use the Rothman formula that meant we didn't consider them in out calculations at all because they drop out of the equation.

For Bruisey, I realize that there is an error on your hard copies. This one is wrong because it is over the…before I say this I want to make sure that this is right…yeah it should be row three over row four. You want to move this down to here, correct? In fact, we can go check this out. I looked at it this morning. So here you are ending up with the same stratum specific two by two tables. So it is the cocaine yes over cocaine no, within

the smokers. Then what I tried to show here is that this one is the meaningless one over neither. That one is meaningless too, cocaine over cocaine only. So it is a collapsing of those. I'm trying to replicate the Bruisey method and impose my way of thinking, about the way that we usually stratify things and it gets a little hairy. I just put the box around the wrong…I think it's this row over this row…is that right? See its fifty two over one forty eight and see it's the thirty eight over sixty two. I think what is happening is it's this over this. Its this over this, twenty five over…let's think about this…that is right, eight over five, so it's three over four. I'm sorry that was an error, obviously. So there is four, in our way of thinking, strata here, but you can see that two of them are the meaningless ones, when you think about it in the Bruisey sense. Here again, I tried, because it was hard when I read Bruisey, figuring out RJ star and RJ Tilda. The RJ Tilda's are really the important ones in a way because they are the relative risks for the factor being adjusted, conditioned on combinations of the other covariates. This example, these are the two relative risks for smoking, since of the other variables. These, the RRJ Tilda's are what correspond to what we think all the stratum specific relative risks, and the RRJ's are just combinations of the covariates in the absence of the factor being adjusted. So what is happening in the nonsmokers, what is happening in the non cocaine users, this gets more complicated once you start adding variable three, variable four, and variable five? It's always going to be like in the nonsmokers, and the various combinations of the cocaine users, and they have hypertension. They are cocaine users and they have diabetes. They are cocaine users and they have hypertension and they have diabetes, whatever your other risk factors are in the risk system, but it is always in the nonsmokers, if it is smoking that you are adjusting.

For the PAF for smoking controlling cocaine use, PAF for cocaine controlling for smoking, now using the Bruisey formula instead of Rothman, again we end up with…we have two factors, we have two funky strata, if you will, where those relative risks are one because it is just amongst the nonsmokers comparing them to the nonsmokers, so you get a one there, in each case, or the non cocaine users compared to the non cocaine users amongst the smokers. Here we get zero point seven eight, point zero seven zero, that is just rounding error, believe me, they are equivalent. We've got zero point nine nine, when we did it the Rothman method here, I think point zero seven six, or something like that up there, it's the same though. We've ended up with the same result a slightly different way to calculate.

This was important as I worked through this, and as you think about the Bruisey method, which is more generalizable to the modeling case, which is why I want you to learn that one. I really…all of this was so that you could imprint on that Bruisey strata are slightly different than the Rothman strata, the way that we think about it. Actually we will go through an example in an Excel file in a little bit that shows you this even more clearly. You just have to start wrapping your brain around what are the components needed using Bruisey's approach, how do you organize variables into a risk system, thinking of that one reference group, the group where women have non of the risk factors in that risk system as the reference group for… it becomes the reference group for all of the strata, for all of the component PAF's.

You can see that with adjustment we've gotten a diminished, slightly lower PAF. So we've gone down from ten point seven to seven point six for smoking. And down from

ten point two, not as much down, to nine point nine for cocaine use, so slightly different numbers. No we've adjusted and the important thing to understand is, why can't we stop here and just use the adjustment technique, and know that we can calculate ninety fiver percent confidence intervals for the adjusted PAF using STATA, or by hand, but we can certainly do it by using STATA. Why don't we just stop here? The reason we don't stop here is because we still, although we have a separate, this is the adjusted PAF that we just calculated, smoking adjusted for cocaine, this is cocaine adjusted for smoking, but if you add those up, you've over shot the summary PAF for the two factors organized into a risk system. This means that you still haven't quite managed to disentangle all that is going on with the overlap between smoking and cocaine use. In fact, this is whey these methods, traditionally, even with adjustment procedures, haven't been used. You can understand this, so people, one of the classic cases, one of the people wanting to look at cancer incidents and attributes. How much is attributed to the environment, how much is attributed to genetics, how much is attributed to...Just using these methods you are going to totally misestimate what those pieces are because you haven't made it mutually exclusive, yet. That is why there has been lots of criticism when people try to do that. The impotents to do it is correct, that is what we want. We like to be able to say that this amount of cancer is due to the environment versus genetics versus whatever. Until we get more refined methods, the start of which we hope we are getting to, by moving towards getting an average PAF is going to allow us to do that. When you do simple adjustments they are not yet independent. The goal is to get mutually exclusive estimates. That is what we are moving towards, but we are not quite there yet. You just asked about assumptions, we'll get to that later. Hold that because even with the

average, we are making certain assumptions, we will talk about…let's put it this way, there is a lot more work to be done in this arena. We think that it is worth it to go down this path though because of the ultimate utility and what we will end up with. Even with the average PAF, we know that there are limitations too and we will talk about that, but this has really been why people don't even…speaking of publications, people still do publish this kind of stuff every once in a while, but for the most part in the literature there is an understanding of this. That when you try to total these you are really getting…these are not mutually exclusive, mutually adjusted estimates that can add to that piece of the pie that is the summary for the whole risk system, so people aren't using these methods. What drives what, and why aren't there software procedures to do these, because there hasn't been the theoretical until recently, groundwork laid for writing the software that could do it? I think that we are moving in that direction. I always say to…why did we always do logistic regression for years and years when we even had cohort data and it was because the software didn't exist. There is a give and a take here in terms of us pushing methodologies that we think are useful and have a utility and then the software developers developing the software that is going to allow us to do it, including variance estimation. So we are out there trying to push for these because we think that in terms of our work for priority setting and better understanding how we are going to think about program development, that this could be really important for us at MCH. I'm thinking that it is break time and this is a good place to break. We ran a little late, so could we be back here at about twenty five to.

KRISTIN RANKIN: That's it we are going to extend this case to about two. As we go on in the training you will see that we are going to add more and more variables. You will

see how it gets more complex as we do that. I'm a SPFS user and I know that in the online course that we teach, I'm always getting can we get a code for STATA, can we get a code for SPFS, and I'm sorry I always go back to SPFS, I always provide a little bit of SPFS code, in case it will help SPFS users, and I'm sorry for people who use other programs. This is just a little…it's a strange way to start this, but it is a little bit of a code on how to get an individual level of data set, an aggregate data set that you can use as we are using the data in the slides for this. If you have individual level data, for say, the ten thousand women that we've been talking about in our fake dataset, we want to get an aggregate dataset so that you know how many people are in each of these cross classifications of the variables that we are interested in. If you sorted the data by all of the variables, low birth weight is their outcome and smoking cocaine and the third variable that we are adding as Deb mentioned is poverty. So think of this dichotomous over one hundred percent federal poverty level over one hundred percent of every level. Then you do PROCFRK. I always like to show this because I think it is a helpful way to format your data in SAS. You use a PROCFRK with a table statement, you cross all of your variables that you are interested in and you use the option afterwards with a slash. You use the option list then you get every cross classification of your variables in a really easy to read format instead of if you use PROCFRK. If you didn't use this list statement, you would get a bunch of two by two tables and it is really hard to… this is kind of just a neat piece of code that I like to show people, if you don't know it already. It is a good way to summarize your data. It is a great way, whatever analysis you are doing to get an overview of what you are working with. You see here that you have low birth weight yes and low birth weight no. These are all the cases at the top, these are all

the controls at the bottom, this isn't case controlled data, but these are all the diseased and these are all the undiseased. Then you have every possible combination of these three variables, smoking, cocaine, and poverty. In this fake dataset we end up have people in every stratum, but what you would see if you didn't have…if you didn't have anybody with a combination of…they are nonsmokers, but they are exposed to cocaine and poverty, then this strata wouldn't show up at all. In this particular case we have people represented in every possible cross classification of these variables. This is the format of the data that we want to be looking at, and actually, we want to simplify it even further and count up the number of cases. We know that for smoking, cocaine and poverty that yes, yes, yes, we have twenty four cases, or twenty four diseased, low birth weights. And for smoking, cocaine, and poverty, yes, yes, yes, we have thirty five undiseased. So we are going to collapse this even further so that we have the cases and the controls, or the diseased and the undiseased in the same strata with the smoking, cocaine and poverty, yes, yes, yes. So really we only have eight different strata of possible combinations of smoking, cocaine and poverty. Yes, yes, yes, yes, yes, no, yes, no, yes, etcetera, and then the number of diseased and undiseased in each one. So really all of it is, is just reorganizing the data. I gave you a little code to show you how you would do that with individual level data of your own. The reason we want to reorganize it this way is because we want to get the prevalence and the relative risk for each one of the strata because as you see Deb laid the groundwork for the formula for calculating the summary PAF. You want to get the prevalence and the relative risk for each strata. You can see the prevalence here is the twenty four cases over the total. I don't have the total shown here, but we know that the total is seven

hundred. It is the same fake dataset; we are just adding new variables. There should be seven hundred with low birth weight. So twenty four over seven hundred is point zero three four, which is our PJ right there. This is the total proportion of the cases that are exposed to smoking, cocaine, and to poverty. Then the relative risk for this is going to be twenty four over fifty nine, so the risk in this strata, for the yes, yes, yes, over the referent group, those exposed to none of these, smoking, cocaine, or poverty, and you see that that is one, so the ten point seven over one. So it is twenty four over fifty nine divided by one seventy five over forty six oh five, that is ten point seven. That is how we get all of those values. Basically, you will see when I move into Excel. You could have programmed it, I guess, in SAS and then added these to your SAS dataset, once you get and aggregate version of the dataset, you could do it that way to, but I start working in Excel at this point just because I think it is a little more flexible and you can see what you are doing. It is up to you.

Just to generalize how many strata that you are going to get, so you know that you are working with the right cross classifications here, for binary variables or dichotomies variable, the number of strata, which we are calling K, it comes up in all the formulas as K, it is two to the N, where N is the number of the variables. So if you have three variables two to the third is eight, eight different cross classifications, and that is what we saw here. When we get into…but in general K is the product of the number of levels for each variable. If you have variable, an ordinal variable with three or four different levels, you are just going to multiply the different number of levels. In the Bruisey data that we are going to be using for our exercises, you will see that there is a dichotomies variable, there is two ordinal variables with three levels each, and there is a four level

variable. We are going to have seventy two different strata here, different cross classifications. You will see how it gets a little more interesting then.

Going back this is our three level variables, you've seen this pie chart before, but we just put some names to the variables, now. This is the pie chart with the component PAF's in the very beginning when we were talking about Factor A, B, and C. Now we are talking about smoking, cocaine, and poverty. I think we've kind of beat this to the ground. These are the different component PAF's they add to the summary PAF, what we are trying to do is get the summary PAF for this particular system is point four six. So Deb do you want to discuss…we've loaded on your machines and Excel file and it is called the Supplementary.

DEBORORAH ROSENBERG: You should have now on your desktop, there is a folder called desktop, if you click on that you will see a bunch of files. Do you all see that? The file that you want is called Supplementary…

KRISTIN RANKIN: Supplementary Excel file, PAF Training Supplementary Excel File, I think. You also have this printed out in your packet, the first worksheet.

DEBORORAH ROSENBERG: That is what we are talking about now anyways. So you should be looking at something like this. This is worksheet, the first worksheet; this is all we are going to go over right now. This is exactly what you've been looking at in the PowerPoint presentation, so it was perfect to ask have we moved into Excel, because this is what you've been looking at. I first want to point out to you, that now you see that when we say K equals two to the, in this case, three for eight strata. Those strata are

each row in this risk system. So again, it is not strata defined how we might define them more classically. I'm not going to go over the whole worksheet, but we'll go over a few things. What I have done here, is just for one instance, I've calculated all of the pieces, the PJ's, and the RRJ's, the Rothman formula and the Bruisey formula. You can, if I click you will be able to see, to go at your leisure and look at all the formulas in the formula bar, because this is the kind of…not exactly like this, but you will be able to refer back to this if you want as you think through these issues. Remember right now we are just using stratified analysis here; these are just the actual prevalence estimates, the PJ's, and the relative risks. What I want to point out to you is, if you look at this column of relative risks here, in this context, these stratum specific relative risks all have row eight, or the reference group, as the reference group for that stratum. In here you will see that everything is over row G, I'm sorry row fourteen, which is down here, the eighth group. As you look at each of these relative risks, it is all over the reference group that is in row fourteen.

Then what I have done down here is taken those PJ's and those RRJ Tilda's, that we have, then I have calculated with both Rothman and Bruisey, the adjusted PAF. What I want you to notice here is that I just chose to use the PAF for smoking, adjusting for cocaine and poverty. We could do it…at some point you might want to try to do it for the other variables, PAF for cocaine and adjusting for smoking and poverty. What you need to see here is which of the relative risks here turn out to be similar to the same stratum specific estimates that we are used to and you can see here that I have got a relative risk in the first four rows and then the relative risks in the last four strata all turn into one. If we go back here, let's think about why that is. So the relative risks that we are

interested in have to do with comparing smokers against nonsmokers. So you see here is the yes, yes, yes row, smoking, coke, and poverty, who do we want to compare them to? We want to compare them to the no, yes, yes, that is the strata where all the other variants are the same, yes for cocaine, yes for poverty, and what is changing is smoking. You are going from yes to no. So in row two you've got yes, yes, no, we are going to compare this row to this row, which is the yes and the no for the adjustment variables stay the same, but we are getting the comparison for smoking yes, no. Those would be the regular two by two tables if you think of them that way. That we would form if we were asking SAS and PROCFRK to do a regular stratified analysis and getting an adjusted relative risk. In this configuration we could go on and think about it for the other variables to, though. So for example, if we were doing cocaine adjusting for smoke and for poverty, we've got yes, yes, yes, now we want to go to yes, no, yes, which is the comparison to this row, right? Now we want the cocaine group to change, but we want the smoking strata to stay the same. You could do that for poverty as well. Those then are the relative risks that actually are going to have real values, the other ones are going to turn into ones, because you are going to be comparing Bruisey…when you think about the relative risks that turn into ones, those are all about…remember we said in Bruisey's terms, they are the one variable that you are adjusting for isn't there, so it is like…all the no, no, yes versus all the no, yes, yes. All the smoking no's in this case, if we were going to get an adjusted PAF for smoking, adjusting for cocaine and poverty, these last four become the relative risks of one, and the other four are the way that we think of the strata that are of interest. Remember for Bruisey all of them are strata of interest, they all have values. So you can see in this column here, when I do the

calculation, here I'm doing the Relative Risk J minus one over Relative Risk J, I end up with these zeros using the Rothman version. Over here because I am incorporating the PJ divided by these Relative Risk J's, I'm dividing by one, I am going to have values in all of these rows to do the calculation. This is just showing you for three variables, how to compare the Rothman approach to the Bruisey approach. They're both really the same and we keep wanting to move ourselves towards thinking of these eight strata and this formula, the Bruisey formula, we are going to keep using because it makes more sense as we move into more complicated modeling.

Let's just look at my results here. After I sum across for both formulas here, what do we have for smoking, we got seven percent. Remember we started out with a crude PAF for smoking of ten point seven percent. We had an adjusted PAF for smoking, only adjusting for cocaine, so in a two variable risk system, it went down to I believe, it was seven point six percent. Now we've added poverty and we are adjusting for poverty as a dichotomies poverty variable. Now we are down to, a slightly diminished PAF adjusted, first sequential, what would happen if we eliminated smoking first from this risk system? Cocaine and poverty still haven't been tackled and we are going to eliminate smoking first, we are down to seven percent for the population attributable fraction for smoking for looking at low birth weight. So how much of low birth weight is attributable to smoking? Then I'm not going to spend time now, but I tried to as much as I can because I need this for myself as well, given you as much detail about how we get to these calculations. What are the RRJ's and the RRJ Tilda's and all of that? I've worked it out in each case down here for if I was looking at smoking, controlling for cocaine and poverty, looking at cocaine controlling for smoking and poverty, etcetera. I don't want to

go over this all right now, but I want you to know that it is here. Kristin and I have been trying to internalize this stuff over the past three or four years, we know you are not going to walk away from here in two days time, but we hope that you have enough and enough resources that you can work it through completely as you move forward and it is going to make you want to use it. So I'm not going to take our time now doing this, but I want you to know that it is there, and you could of course ask us questions about it too later, but I'm going to turn it back to Kristin now to go on. Unless you have any questions right off the bat, is there anything that you need to ask right now?

KRISTIN RANKIN: Just to let you know that before you leave tomorrow we are going to have CD's for you that have electronic versions, but it is also going to have answer keys for the exercises so that is why you don't get the CD's until tomorrow. What was the slide number we left off on? We are going to really quickly move into modeling because as Deb said you can see that we were generating the relative risks based on the fully stratified data, but what you can do is run a regression model and use your beta estimates to calculate the relative risks. So that you can generate a model it is most appropriate for your data rather than doing the fully stratified version, especially as you get more and more variables, you are not going to be able to fully stratify because as I mentioned before in this fake dataset, we have data in every strata, but you are not going to always have that. When you start having zero cells, is when you are going to run into trouble generating relative risks, you are going to want to move to modeling, just as you would want to move to modeling from the stratified method when you are estimating relative risks and odds ratios. So some of the advantages, like I just said modeling is not sensitive, sparse data in individual cells. Then if you chose to consider

confounding affect modification in the same model, you can generate estimates more easily that way. You may have an interaction term between two of your variables and control for a third or a fourth variable. You can do that more easily in models more than you can with the stratified method.

I wanted to first show that using this assumption free approach that we were talking about, if you treat all variables as affect modifiers, basically you are running a fully specified model, which is the same as doing it in the stratified version, so I will show you how that works. Like I said, this method breaks down quickly because you start getting zero cells. Just sharing some SAS code about your fully specified model, what I mean by that is that your model is low birth weight equals smoking plus cocaine plus poverty plus every possible combination of them. What you are going to estimate here, your beta estimates here are going to give you the exact same results that we saw with the fully stratified data.

This is just to show you how to do it with binomial regression and logistic regression in SAS. Ideally, binomial regression you can generate relative risks directly from, that is in PROCGENMAD and SAS, with this distribution equals binomial and link equals log. Because you usually want to use the relative risk with a PAF, this is ideal.

I'm also showing you the logistic regression though because the program that we are going to use in STATA to generate adjusted PAF's is going to be using the beta estimates from logistic regression instead of binomial regression. We have that kind of limitation, so I just wanted to show you both for now so that when we move to using STATA to generate our adjusted PAF, our first sequential PAF's , you will see why we

are using odds ratio to estimate the relative risk instead of directly using relative risk. Ideally, we would be using the results from the binomial regression to get the relative risks directly. If you have any questions about code or anything later, feel free to email me.

These are just the results from the binomial regression model; I just wanted to show you the SAS results. It is the same data, seven hundred low birth weight cases, and ninety three hundred normal birth weight. You always want to check this line in your results, so you can make sure that your modeling low birth weight equals one, which is yes. All the dichotomies variables are coded one and zero, by the way, in all of these data. These are typically what comes out of SAS, but here is the main thing that we want to see, these are our beta estimates right here. These are the results from the regression model, where this is the fully specified model, you see that you get a beta estimate for each of the factors by themselves and then a beta estimate for each of the interaction terms. You see that we have the interaction term for smoking cocaine, and poverty, that three level interaction term, too. This is fully specified.

We are going to use these beta estimates now to calculate the same PAF's that we just did, same summary, same adjusted PAF's that we just did. We are back to the Excel file. I apologize for the back and forth, but I think it is really hard on PowerPoint to show you what we are trying to do in Excel with all of the formulas. If you want to go into Worksheet B now. This is a little bit busy, but basically what I've done in this worksheet, is taken the beta estimates that you just saw on the slide. I basically inserted them into every row. You can see that these are repeated; the beta for smoke all the way down it

is repeated. The beta for cocaine, the beta for poverty, then you see the beta estimates for all the different combinations, smoking times cocaine, smoking times, poverty, cocaine times poverty, and smoking times cocaine times poverty. Now I'm using these beta estimates to calculate this Relative Risk J, in the last column. How I've done that, if you've seen the column before, the natural log of the relative risks, what I've done now with these is coded one and zero for yes and no. A9 times E9, smoking is one, times E9, column E right here is the beta for smoke. Basically I'm applying the code for the variable to the beta estimate for that variable to get the natural log for the relative risk. I'm doing it for each of the variables. When smoking is one, cocaine is one, and poverty is one, when you are exposed to all three variables this is the calculation. The calculation is always going to stay the same. I just dragged this, if you are not familiar with Excel and formulas and that, stop me and ask me about that. I'm assuming that, but I might be wrong about that. This formula, I just dragged this formula down, so that it is the same for all rows, but you can see it changes because sometimes you are multiplying the beta estimates by zero. For the top row you have smoking times beta for smoke, cocaine times the beta for cocaine, poverty times the beta for poverty, smoking times cocaine, which is still one times one times the beta for smoking times the cocaine. Smoking times poverty, one times one times the beta for smoking times poverty and so on. That is how I got the natural log of the relative risks. In the next column if you exponentiate you get the relative risks. You see that it is the same ten point seven that you got with the fully stratified data that Deb showed you in Worksheet A, that is because basically what you have done in this model is just replicated your stratified

results. You can see that the relative risk changes because you are multiplying these betas by different values in the first three columns. Are there any questions about that?

There might be a better way of doing this; you might find a better way to mechanically do this. This just seemed easiest to me. It's a little clumsy, but you generate the beta estimates is SAS or in STATA or in SPSS, you just fill them in to your Excel file here, where you have your aggregated data, and then you can determine your relative risks.

I neglected the PJ column. PJ column is calculated the same way as Deb showed. Al you are doing here is this is D9 over D30, and you mayall know this. It took me a while to get this in Excel and I think it is so helpful, but if you use the dollar sign in front of the column and in front of the row, so for dollar sign D, dollar sign thirty, you get the same value and then you drag this formula. Here, let me delete all these. You might all be thinking I know this, but I didn't, I think this is very helpful. If you want to drag this same formula, but you always want to keep the denominator the same, you want to do that because, let's see where D30 is; this is your total cases. This is your seven hundred. You want your denominator to stay the same the whole time because you are getting the proportion of cases that are in these strata. If you just drag this down, the D9 is going to change to D10 in this column because you want twenty eight over seven hundred, you want eighty over seven hundred, but the seven hundred stays the same. That was a long explanation of a little Excel trick that I think is helpful.

So then when you apply the Bruisey formula, and actually, I'm sorry I've used the Rothman formula here. So we are using these kind of interchangeably. These are the component PAF's. So you see how the Rothman formula is basically PJ times relative

risk minus one over relative risk, what you have is a component PAF of three percent for this group that is exposed to all three variables. You add all the component PAF's and you get a summary PAF of point four five seven one. This is the same that you saw in the previous worksheet. Then I just showed it to you with…

DEBORORAH ROSENBERG: Because it is really interesting getting back to keeping us thinking about do we want to consider poverty modifiable or unmodifiable. In this case, if you look at the component PAF, in the second to the last row, so it is nonsmokers, non cocaine users, but they have poverty, how big the component PAF is as a proportion of the summary PAF, it is very big.

KRISTIN RANKIN: This is kind of a preview, we are going to go, we are going to use these same data to go through it both ways, with poverty as modifiable and poverty as unmodifiable when we are calculating sequential and average PAF's later, but this is for more of the mechanics of getting the summary PAF, because that is really our starting point, we want to make sure that we can get that down before we cans tart doing the sequential and average.

I wanted to show you the fully specified model using logistic regression because you can see the betas are going to be a little bit different because low birth weight isn't a rare disease in the sense of how we think of rare disease. Usually when you are using to odds ratio to estimate relative risk, a lot of times that happens in cancer where there is case controlled data, and cancer is quite rare, depending on the cancer, it is considered pretty rare. When you are thinking of a prevalence of seven percent, like we do in low birth weight, it's not quite as rare, so you are going to see some differences

when you are using the odds ratio from logistic regression to estimate the relative risks. I just wanted to show this as a limitation when we go on to using logistic regression, when we are calculating our PAF. I used the beta estimates from the logistic regression here, and you can see that the summary PAF is point four seven five then. So it is slightly…you would expect it to be elevated because usually if you are comparing odds ratios to relative risks, the odds ratios are a little bit of an over estimate of the relative risk. So in the same way the summary PAF is a little bit of an over estimate when using beta estimates from logistic regression. Any questions about that? Any questions about using the beta estimates Debbie?

UNKNOWN SPEAKER: (Inaudible)…

KRISTIN RANKIN: Exactly. So it would bring everybody who is under the poverty line to over the poverty line and eliminating smoking and cocaine. Exactly, you could reduce low birth weight in the population by forty five percent. This is fake data, but yes, exactly how you interpret it.

UNKNOWN SPEAKER:(Inaudible)…

KRISTIN RANKIN: Right, that is what the summary PAF means. Great, Beth.

BETH: (Inaudible)…