MCH EPI Conference

Using Population Attributable Fractions to Assess MCH Population Outcomes

December 8 – 11, 2008

DEBORORAH ROSENBERG: …University of Illinois at Chicago School of Public Health and we know this is going to be an intense two days. We've hopefully tried to pace what we're going to do here so you all will have a chance to really understand what we think is a really potentially new tool for all of us in terms of in fact things like the needs assessment and the other program planning and evaluation tasks that we do as maternal and child health professionals in addition to anything we might do as researchers. So, I am going to introduce now my colleague, Kristin Rankin and she will introduce herself a little bit more and then we'll just get going. So, welcome.

KRISTIN RANKIN: As Deb said I'm Kristin Rankin. I just actually in the summer got my Ph.D. at UIC in the Maternal and Child Health EPI program there and I will be the co-trainer today and I apologize for my voice and if I start coughing. I have a little bit of a cold due to sleep deprivation because I have a 6-week old who I did bring along with me with my mom. I'll have her here tomorrow morning probably after breakfast if you want to meet her. If there are any clinicians in the room and you want to come to me at break and tell me what I can take while I'm breastfeeding because I'm not taking any medications right now because I wasn't sure what to take. So, if you can tell me that at the break that would be great and then I'll limit the coughing I promise.

So, without further adieu, I think we're just going to get started. As Deb said we're not quite sure about the pace of this, so just communicate with us. We're going to make this really informal. If you have any questions let me know and if you think we're going to fast or anything like that just let us know. We want to make this best possible experience for you guys. So, what I'm going to start with today is a background and overview. I'm going to do kind of a real high level overview of what we're going to be covering in the next two days because I want to kind of show you where we're going, why we're going this and then we're going to step back and do it in more detail and actually go through the method step by step. So, the impetus for this is as you probably know we usually use ratio measures in our work. So, a lot of ratios, relative risks just to describe the magnitude of association between the risk factor and an outcome but impact measure such as population attributable fraction sense they account for both the magnitude of association and the prevalence of risk in the population they can be really helpful in describing the impact of a particular risk factor in an outcome. We think that PAF's are underused or misused right now because there are the methodological concerns right now about how to appropriately account for multiple factors at one time. So, often if you see PAF's used in work it's maybe in the last paragraph like in the discussion section of a paper they'll mention the accrued population attributable fraction for one risk factor of interest but it's not very often that you'll see PAF's presented for multiple risk factors and maybe even ranked to show which risk factor if you eliminated it from the population would give the most impact on a particular outcome. So that's what we're really trying to get to today and I've probably said most of this already but in a multi-variable context the goal is to generate an individual PAF for several risk factors

so you can prioritize them for program planning and decision making. But, generating mutually exclusive and mutually adjusted PAF's is not straightforward because there is overlapping distributions of exposure in the population. So, methods that go beyond the normal adjustment procedures that we usually use such as regression modeling, methods are required beyond that.

So, I just want to start with a few historical highlights. I used the population attributable fraction for my dissertation, so I did a huge literature review. So, I had to share a little bit of this with you. So, Levin was responsible for introducing this measure in 1953. It was a paper about lung cancer and he described it as the indicated maximum proportion of disease attributable to a specific exposure. Never really called it attributable risk or the population attributable fraction but this was really the birth of this measure. The idea was that if an exposure is completely eliminated that the disease experience of all individuals would be the same as the unexposed. So that's kind of the theory behind it and you'll see this is the formula for the crude PAF and there are a couple of different ways to look at it. There's either the prevalence of the exposure and the population as a whole times relative risk minus one over the prevalence of exposure time relative risk minus one plus one and this is a typical crude PAF equation that you're probably used to seeing. This is really the same as a more intuitive formula, which is the prevalence of outcome in the population as a whole minus the prevalence of the outcome in the unexposed over the prevalence of the outcome in the population as a whole. So, those are some formulas we're going to start with but we're really going to be actually using formulas that are more appropriate for adjusted measures later. So then, really nothing

happened for about twenty years in the literature. Nobody else kind of tried to use this method until Metyan picked it up in 1974 and developed a formula for an adjusted PAF and what the adjusted does is it's a proportion of disease that could be reduced by eliminating one risk factor after controlling for other factors and accounting for any potential effect modification. So, these methods are used both adjusting like I said for confounding or if there was an affect modifier. So, there's two different formulas that Metyan suggested for that. Then, Brussey and Brelin and Dressler came up with a summary PAF, which is actually the proportion of the disease that can be reduced by simultaneously eliminating multiple risk factors in the population. So, just an example here and one of the examples we will be using a lot. For low birth weight two risk factors would be smoking or cocaine use and the adjusted PAF would give you a PAF for smoking controlling for cocaine or a PAF for cocaine controlling for smoking. The summary PAF would give you the PAF for eliminating both of those smoking and cocaine simultaneously and we'll explain to you cannot actually add the adjusted PAF for smoking and the adjusted PAF for cocaine and get the summary PAF for eliminating both exposures and that's going to be like a core piece of why we're using the methods we are using right now. Like I said we'll go back over all of this in detail. Brussey also gave us a method for using regression modeling to generate these adjusted PAF's, so we'll be also using regression models later in the course.

Then Benetel and Gayle were responsible for developing variance estimates for the adjustment and summary PAF. We really aren't going to be discussing much about the variance today because the average PAF, which we're actually going to be focused on

doesn’t have variance estimates generated yet. So, we’re not going to hear a lot about 95% confidence intervals today and we’re hoping that as this method starts getting used more we’ll have those formulas available for you in the future. So, the idea of this is if you think of this pie chart as the whole of the disease. So, this is everybody who has a low birth weight infant and you want to calculate the summary PAF. The summary PAF is the total of all of these colored slices of the pie. So, it’s going to be about .46 or .457. So, 45.7% of low birth weight is explained in this particular model or is attributed to factors A, B and C but you can see that these are component PAF’s. That’s what we’re going to call component PAF’s, where you say because a woman could both be a smoker and a cocaine user or just a smoker or just a cocaine user you have all of these mutually exclusive categories. So, you’re exposed to factors A, B and C and that accounts for this small slice of the pie here. You’re exposed to factor A alone. There’s a high prevalence of factor A here. You’re exposed to factors A and B, A and C, B alone and you get the idea. There’s all of these mutually exclusive pieces of the pie but why this isn’t satisfying is that we don’t have a PAF just for factor A that we can compare to the PAF just for factor B and actually say, which would get us the bigger bang for the buck. If we did an intervention program for factor A how much of the disease would be reduced if we had an intervention program for factor B and then we can compare them at an equal scale. Right here, we don’t know, which portion of this blue pie to give to A, B or C in order to compare them. Is this making sense so far? So that’s why we’re trying to get to methods that would allow you to compare factor A to factor B to factor C. So, we want three slices of pie basically, one for A, one for B, one for C and not all of these mutually exclusive factors. So, methodologic work has been and is still being carried out

to develop approaches that apportion that summary PAF in a way that yields estimates of impact for each of a set of risk factors. So, Ivan and Gafeller, they are a couple of Scandinavians who came up with a method in 1995 and 1998. There were a couple of papers about this. They came up with the idea of the sequential PAF. What the sequential PAF is, is the portion of the disease that could be reduced by eliminating one risk factor from the population after some other factors have already been eliminated. It's giving a sequence to the factors saying, okay, we're going to completely eliminate smoking from the population and then completely eliminate cocaine from the population of pregnant women and what is leftover for cocaine after we've already eliminated smoking? In the first sequential PAF, it's actually the same as the adjusted PAF and the first sequential PAF is the PAF where that risk factor is eliminated before any other risk factors. So, an ordering is imposed here. We're saying that we're going to have a risk factor prevention program that's going to completely eliminate one factor and then we can start on the next factor. So, an example of the sequence would be to eliminate factor A, then factor B and then factor C. So this first sequential PAF here for factor A is the PAF for A adjusted for factors B and C. So, you can see how the first sequential PAF is the same as the adjusted PAF because it's A adjusted for B and C. The sequential PAF for B after you've already eliminated factor A is the summary PAF for A and B minus the adjusted PAF for A because you've already eliminated A and it's so, what's leftover for B after you've already eliminated A from the population. Then the sequential PAF for C would be the PAF for the union of A, B and C, so the summary PAF for A, B and C minus the summary PAF for A and B controlling for C and then you have what's leftover for C. So, if you eliminate A from the population then what's left for

B, then what's left for C at the very end. So, you remember that pie that we had that had all of the different mutually exclusive factors. Well, now we've colored them in so that you see if you eliminate A first. It takes all of these four slices of the pie because this is A, B and C, A alone, A and B and A and C. So, any slice of the pie that had A in it was eliminated first and then you have what's leftover for factor B and then just a small slice is leftover for factor C, since we've already eliminated the other factors. So, this is how you have a sequential PAF where A would be eliminated first, then B and then C.

Now because that isn't really practical in the real world to say if we had a risk factor reduction program for smoking that we could completely eliminate smoking from the population and then start on the next factor and then start on the next factor. Either some of our intervention programs are multifaceted or some of our intervention programs are going to have an impact on other exposures. So, if you're eliminating cocaine you might have some people who stop using or stop smoking as well or the other way around. We know it's not going to happen in a specific order like that. So, Ivan and Gafeller suggested using an average PAF, which is a simple average of all of the sequential PAF's. What this tries to do is get an equal apportionment of that summary PAF that whole PAF over every possible sequence because the order that risk factors will be eliminated in the real world is unknown and this is based on the Shapley Solution in game theory, which is a method of fairly distributing the total profit gained by team members working in coalitions. So, if you have five people on a team working for a common goal and you want to reward them for that how are you going to apportion the reward to those five members of the team? This has been shown by

Shapley that this is the most fair way to do it. This simple average over every possible sequence. So for three risk factors, we're back to factor A, B and C, there's six sequences in which they could be removed in a population, A, then B then C, A then C then B, B then A then C and etcetera. So, you see the A's are all circled. So, you see there's going to be a value for this sequential PAF, sequence one. For sequence two it will be the same value but it's still a second sequence, so we want to count that. Then, if you remove it second after B, there's going to be a value for the sequential PAF here. Third after B and C and so on. You'll also notice in sequence four and sequence six are going to have the same value too because it's always A coming last but there's still two unique sequences. So, there's a total of six sequential PAF's for each of the three risk factors and the average PAF is going to be the average of all of those sequential PAF's for A. Then, we'll be able to say if you eliminated factor A from the population you could reduce the disease by whatever the value for the average PAF is. So, then we get to what we really want, which is three pieces of the pie that can be compared. This is a little more realistic than that sequential PAF where A took almost the whole portion of the summary PAF because you're thinking you're eliminating that first. In this one, you have an actual average for A, B and C and then you probably are always going to have a proportion of the disease that's unknown or unexplained by what we have available to us.

Question: (Inaudible)

KRISTIN RANKIN: Would it be the same as what?

Question: (Inaudible)

KRISTIN RANKIN: Right and that's always going to happen. What you'll see when we start doing the methods, we're always going to start from the summary PAF. So, even if we have five factors, we want to know what is the summary PAF for eliminating all five of those factors simultaneously. We're almost going to be working backwards from that number because we want to make sure that we're not going above that because what you'll find out is that adjusted, if people are explaining the adjusted PAF for A, B and C, if you added the adjusted PAF for A, then B, then C, really what those are, the first sequential PAF's. You're going to have a much larger number. Well, not necessarily much larger, it depends on your data but you're going to have a larger number than the summary PAF's. So, we want to make sure that we're only using the summary PAF as the maximum potential disease that can be reduced in the population. Does that answer your question? Anybody else before we move on? So, in summary, the summary PAF can be a portion into component PAF's reflecting every possible combination of the factors being considered. We saw that that's not very useful. The sequential PAF's, which reflect pieces of one particular sequence in which risk factors may be eliminated. So, that's only describing one sequence. And, the average PAF, which reflects an average over all of the sequences regardless of the order in which the factors are eliminated from the population. I just want to mention, I'm saying the word eliminated. Now, we know we're not going to completely eliminate smoking from the population but the value that we're going to get is going to assume that we're completely eliminating it

from the population and we'll discuss just in brief towards the end some methods that can be used to make that a more realistic estimate and say, "We know we have a risk factor reduction program that is 75% efficacious." I mean that's huge but if there's 25% efficacious and then we can apply that to the attributable fraction we get for completely eliminating it from the population and get a more realistic estimate.

PAF's can be generated from data from different study designs. Using cross sectional data we can use the prevalence and the measure of effect from the same data source. Then the interpretation for the PAF is the proportion of prevalent cases that can be attributed to exposure. From a cohort study we can also get both the prevalence and the measure of effect for the risk factor from the same data source and then the interpretation is the proportion of incident cases that can be attributed to exposure. You can also estimate PAF from case control data where the prevalence of exposure among the cases is used and the odd ratio is used in place of the relative risk and the rare disease assumption is assumed. The interpretation is the proportion of incident cases that can be attributed to exposure. So, it's similar to cohort studies. So, just so you know you can use these methods across all different types of data sources. So, in addition to the different computational approaches that we talked about decisions about how variables will be considered may be different when focusing on the PAF compared to ratio measures such as the odd ratio relative risk. So, we're going to be differentiating un-modifiable and modifiable factors. So, we're going to be thinking a lot in our conceptual model in that stage of our analysis about what is going to be modifiable and what's going to be un-modifiable. You'll see that that definition might change depending

on your perspective and where you're sitting. Confounding an effect modification is handled slightly differently with the PAF and causal pathway issues could be handle a little bit differently than when you're thinking about modeling for just the relative risk odds ratio.

So, a little more about modifiability. Un-modifiable factors end up being used only as potential confounders or effect modifiers and we don't calculate PAF's for them. So, a good example of this would be race or ethnicity. To calculate a PAF for Hispanic ethnicity really doesn't make a lot of sense. I've already said the word eliminate so many times this morning. You're talking about risk factors that you can eliminate from the population like factors that you can really use to reduce the disease in the population and a lot of people call this…I know Art Enhandler always thinks of this as a risk marker. So, Hispanic ethnicity is a marker for risk for a potential outcome potentially but it's not a factor that we can develop a program around and it's nothing that we want to eliminate in the population. Right? Do, that would be a un-modifiable factor but if we know that it is related to the disease outcome we want to make sure that we're controlling for that. So, we're going to use it as a confounder and an effect modifier in our analysis but not calculate a PAF for it. Then the modifiable factors are factors that can possibly be altered with clear intervention strategies. Like I said it depends on your perspective and it may alter the results if you choose to use un-modifiable and modifiable. You'll see that there are some factors that are a little less clear than race ethnicity. Then model building and this is something that I explored a lot in my dissertation and your perspective changes a little bit when you're building a model for

the PAF as well because first of all you have these un-modifiable/modifiable factors and you might handle them differently in terms of the level of measurement and coding choices and if you find effect modification either within modifiable factors, across modifiable/un-modifiable and within un-modifiable factors you might decide to handle them differently. So, are you going to run stratified models? You run a separate model for Hispanics versus African Americans versus whites and then compare the PAF's for the modifiable factors across races or are you going to use an interaction term of the model just to adjust your beta estimates for calculating the PAF? There's a lot of different decisions that need to be made and you're thinking of it from a different perspective than if you were just generating a odds ratio or relative risk and you'll see how that plays out. We're actually going to do an exercise. It's kind of experimental tomorrow at the end of the day where we're going to walk through building a model with real data and talk about all these decisions at each point. I kind of call it like a "Choose Your Own Adventure" type. You know those books where we'll get to one point where we'll discuss what will we do next and then we'll get to the next point and really walk through an analysis and discuss how all of these issues come up when you're building a model for generating PAF's. So, you'll see how selection of the final model may not be based on statistical significance of the ratio measure of effect and this is important because you might have a factor that has non-significant but elevated relative risk but is very highly prevalent in the population. You're going to see that's going to affect the PAF. If there's huge prevalence of that in the population there's going to be a larger PAF. You might want to keep that variable instead of thinking of it as something that's not significant and leaving it out of your model. Then we'll have a little discussion about

statistical significance of the PAF. We're not really sure how to define that yet and we'll talk about that. Yes?

Question: (Inaudible)

KRISTIN RANKIN: Oh, I'm sorry. If you're thinking of the variable as maybe as dichotomist or as a continuous variable or as an ordinal variable that might change depending on whether it is un-modifiable or modifiable and then you'll see there's a lot of importance placed on your reference category for the modifiable factors because you want to have a reference group that makes sense in terms of developing a PAF. That's what we're talking about.

So like we discussed the average PAF's allow for the sorting of modifiable factors according to the potential impact of risk factor reduction strategies on an outcome in the population where the ratio measures only provide the magnitude of the association but we have to remember that the PAF like I said is a proportion of an outcome that can be reduced if it was completely eliminated from the population. So, we have to be careful when we're interpreting these measures to make sure that our audience understands that is the highest possible impact that we can have and it's really quite unlikely unless we really depending on the risk factor really think we can eliminate it from the population but most risk factors probably won't completely do so.

So, I hope I've made a little bit of a case for why this could be useful in your work, in these assessments, in research but why isn't it used more commonly. I think it has to do with there are no known statistical packages to complete all the steps of generating average PAF's. So, you'll see we're going to be using a combination of some of the statistical packages in Excel in this particular training to develop average PAF's. We're going to be using a lot of different softwares together to do this. Also, the variance estimate as I mentioned for the average PAF are not yet available either for random samples or for samples from complex designs like (inaudible) or any other sample surveys. We can get 95% confidence levels around the crude PAF, the summary PAF and the first sequential PAF, which is the same as the adjusted PAF but for the average PAF, right now, we don't have 95% confidence in those. So, you can see how we kind of use the 95% confidence level from those other measures to kind of estimate what they might be for the average PAF but I should mention that we actually have a colleague in the bio statistics department at UIC who decided to do this as her dissertation. I think she's almost done. So, we should have 95% confidence in those very soon for the average PAF and (inaudible) job is going to be I think to figure out how to make it practical to calculate those. Yes, go ahead.

Question: (Inaudible)

KRISTIN RANKIN: I wish we could tell you. We have from my dissertation, we actually have a manuscript that's almost ready to be submitted. We did have two abstracts accepted at SER and SPER in June about this without variance estimates. Now,

abstracts are easier, of course, to get approved than manuscripts. I did talk to Alan Lohaus when I was there because I said I have this paper I would really like to submit to Journal but I'm not presenting 95% confidence in it because we don't have them yet but I really want to start pushing this idea, you know pushing the boundaries but unfortunately I had a baby in the meantime, so we didn't quite get the manuscript to him yet. So, I don't know how the reviewers are going to look at it but we can definitely let you know once we've gotten those first probably rejection letters but like I said hopefully we'll have estimates very soon and we can distribute those to all of you hopefully in an easy to use way.

The other really important thing I think to bring up here is that the interpretation of average PAF is strengthened by the evidence of causality. A lot of people have said in the literature how can you even estimate a PAF and report a PAF for a factor that you don't know is causally related to the outcome. So, what we're saying is 6% of low birth weight is attributed to smoking. Now, if you don't know that there's a causal relationship there is that an okay measure to report? Deb and I tend to not be as formal. I don't know how to explain it. I don't think it's a reason not to use the measure because how often do we have a factor we know is causally related to an outcome? So, I think as long as we're cautious in interpreting findings I think we can use this method. We can discuss this more as the training goes on but what we do know is an average PAF cannot help to establish causality. Now, we know what are there the seven different…I'm going back to EPI 101 here but the seven different ways you can start establishing causality. One of them is magnitude of association. So, relative risk on an

odds ratio that's elevated is one piece of evidence that a risk factor is causally related to an outcome. Right? Well, an average PAF really doesn't add to that evidence about the causal relationship but we also don't think that you necessarily have to wait until causality is determined before you can start generating PAF's. So, it's just in being cautious about interpretation and we'll talk more about that as the training goes on.

As always having an explicit conceptual model or framework before you start analysis is important. That's an important if you're estimating relative risks or odds ratios but this is particularly important when you're producing PAF's because of how you handle the factors and model buildings. We talked about the un-modifiable and modifiable factors. You really want to give a lot of thought to that before you start your process because it's going to actually change your computation. Now, if you have un-modifiable factors and modifiable factors in the same model and you're estimating relative risks and odds ratios you're not going to handle them differently in the model. You are for average PAF or any kind of PAF estimations. So, you really want to think that through before you get started. That is the end of my broad overview and we'll go back and handle each of these topics in more detail and actually get into the number crunching, if you will. Do you have any questions at this point before we go on? Do you have a good sense at this point of where we're going in the next two days?

Question: (Inaudible)

KRISTIN RANKIN: I think I mean that's definitely true but even if you're not doing the PAF you discuss things as being modifiable or un-modifiable. I think why we've made it such a point here is because it actually changes the way you use the formulas and I see what you mean. I think if I'm answering your question right there are populations, which you could target to help reduce disease but what we're trying to get at is what exposures in that population do you want to modify to reduce disease. So, it becomes tricky. You'll see later in our example for tomorrow in the model building exercise I've used teens or teenage, so less than twenty as a modifiable factor thinking about not necessarily…and low birth weight as the outcome, not necessarily thinking about reducing…I mean it doesn't make sense to eliminate teens from the population. We don't mean that but what we're thinking about is potentially a teen pregnancy program. So, you have your proportion of the disease that is in teenagers and could that be eliminated with a teen pregnancy program. So, modifiable and un-modifiability kind of gets tricky to think through. We'll be doing that. We'll take you through that more so later. Does that make sense? Did I answer your question?

Question: (Inaudible)

KRISTIN RANKIN: Right. So, you're saying you wouldn't want to eliminate a particular group from the population at all but it might be a marker for other factors. Like I said it's a marker for other factors we can't measure right now. That's a really good point and if you don't calculate a path for that racial group then how do you know what the proportion of the disease can be attributed to what that's a marker for, the modifiable

factors that the marker are for. The other way to look at it is to look at the modifiable factors that you can measure and compare them between different race/ethnicity groups, which is one method. Actually that's what I used in my dissertation. I compared African Americans to whites. I did stratified models and looked at the modifiable factors for childhood obesity within those two racial groups and compared when the slices of the pie were different or when the factors in the pie charts were even different between the two racial groups. So, there's a couple of different ways that you can do that. That's right. I would like you to continue to challenge that as we go because the way that we've looked at it so far has only been that you can stratify by these factors but we wouldn't calculate a PAF for them and you might have a good argument for why you might want to calculate PAF's for those types of markers for other reasons and how we can do that.

Question: (Inaudible)

DEBORORAH ROSENBERG: Can I make a comment about this, which is just I think here's what you get if you think about how you would interpret. So, if for instance if we calculated an average PAF for Hispanic ethnicity, what you can say is maybe thus and such percent of the outcome can be attributed to those unknown, yet to be measured factors that being of Hispanic ethnicity is a marker for but, see it wouldn't get you though to I think more what traditionally the PAF is saying. So, we're going to develop a program to address X. We don't know what the X is yet and that's where the problem arises but I was actually thinking Kristin as Wanda was asking this you can think about a new kind of pie chart that somehow had a different kind of shading or something

indicating, we don't know what these factors are yet but we know that they are associated with a given, un-modifiable risk marker or something. It's a really interesting thought to come up with some creative way to do that. One thing I would say too is hold that thought because we actually are later this morning going to get to an example. So, just start thinking about where we add. We talk about poverty and should we consider poverty as un-modifiable or modifiable in an analysis, which is different enough from race/ethnicity that you could think about a eliminating it but then how realistic is that? So, hold that thought because we are going to talk a lot more about making these kinds of choices.

KRISTIN RANKIN: I really like that perspective too because if we were to generate a PAF for a racial or ethnic group it would really utilize…it's why the average PAF would be so helpful because you could have all of these modifiable factors in the model in addition to race/ethnicity and be accounting for that overlap in the distribution because we know that there is some Hispanic women who are smokers but we have a slice of the pie already for smokers and this is what's left for Hispanic ethnicity. So, these are the factors that we don't know about that are impacting Hispanics. So, I really like that and I think that's a different way of looking at things than I have ever looked at it so far. So, we can think about that when we're doing the model building exercise tomorrow.

DEBORORAH ROSENBERG: One thing, Kristin, also in your work I wish we had those pie charts because when you stratified what happened sometimes though even with stratification so not directly calculating a PAF for race/ethnicity but if I'm remembering

this right it was really obvious that I think it was for whites you could identify a lot of pieces of the pie. Am I saying this backwards? But, amongst blacks and you also stratified by gender, boys and girls but you could tell within one strata you knew a lot about maybe what you could address with programs and another strata where you didn't know a lot. Was that male/female? Was that boys and girls?

KRISTIN RANKIN: It was both. Actually, the African American males I think my summary PAF and this was so just to give you a little background this is using the national survey of children's health to look at childhood overweight, which isn't the best resource for that just because we don't have the diet information from the National Institution of Health but it's a population based survey with a lot of great information and a lot of great factors for childhood obesity. So, I think that my summary PAF in the strata of African American boys and this was only among adolescents, my summary PAF was about 13%, which it seems really small and for the factor within that I mean I had one factor that accounted for 2% of the obesity amongst African American adolescent boys. Whereas, I think for white girls I was able to explain closer to 35% in my summary PAF. So, it's interesting when you put the pie chart side by side. You could see a larger proportion of the pie for childhood obesity was explained for white girls than for African American boys and that was only by the factors that I had available to me from NSCH but it was a first look at that. So that's the interesting part about stratification but you could look at it both ways. That's really interesting. I think that would be great and that's why I love to come and do these trainings because usually the best ideas come from you guys. So, I learn a lot from you.

DEBORORAH ROSENBERG: Okay, well to follow-up. I'm glad actually your work came up because I wanted to make a comment about how really pleased I am I mean with my faculty hat on that this was a real example of a collaboration so that working on Kristin's work when we started to involve a bio statistician one of the things that was really problematic was as Kristin said here we were using a data set, really important, the NSCH, the National Survey of Children's Health and so we have doing a PAF and we have a complex sampling design and what do we do about variance estimation and all that and there was our now colleague Ph.D. student in bio stats who found this interesting and so we almost had like a joint project and because of us I think in her dissertation she first has done work establishing the variance estimates for the average PAF and random samples but she's going on and doing it for complex samples. I don't think she would have done that had it not been for us saying in our work we have all of these…you know, we have PRAMS , we have NSCH, we have NSCSHCN, etcetera. We need it for conflict sampling and designs. So, she's doing that and she actually just told me that for her dissertation she did it in R I think but she's going to do it in SAS and hopefully develop either some kind of user defined SAS macro that can be shared. So, we're really hoping that we make it into a relatively user-friendly approach and will be able to do 95% confidence levels around these measures. So, we are going to go on and go back as Kristin said and really start at the beginning and depending on how you were trained and when you were trained and all of that lets really start back at the beginning and talk about all of the attributable risk measures that are based on differences as opposed to ratios and everything here is being done in terms of

proportions as though it's either cross-sectional or cohort data but there are tweaks to these, if you have case control data under certain circumstances but in the simplest case just think of cross-sectional or cohort data and we've put up here the typical two by two table as a way to start thinking about this. As we move on you see we have to sort of adjust our, at least for me it was an adjustment of my kind of ingrained epidemiologic way of thinking about this but starting from a simple two by two table with two dichotomous risk factor and a dichotomous outcome, what are the standard attributable risk measures. So, we first have what's most often probably called the attributable risk, which is just the difference between the risk of disease in the exposed or those with the risk factor minus the risk of disease or outcome in the unexposed and I've put in the two by two table here we're going to go through all the notation, all the words… Half the battle here I think is getting through the notation that of course varies from author to author and the sort of nomenclature that varies from author to author. We are going to try and resolve most of that. So for right now p1 is the Roe percent or the risk if you had a cohort data. So, A over n1 is p1. C over n2 is p2. Then, the overall risk of the outcome regardless of this risk factor is p naught m1 over N. So, the attributable risk is the p1 minus p2 and then the attributable fraction is that attributable risk as a percentage of the risk in the exposed group. Then, we move to the population level. So, the population attributable risk is now the difference between the overall risk and the risk in the unexposed and the population attributable fraction or the PAF, which is what we're focusing on here is that difference between the marginal risk of the outcome minus the risk amongst the unexposed as a proportion over the total amount of risk in the population. You'll notice that the attributable risk and the attributable fraction is all about

what's happening just in the group that's exposed or the group with that particular risk factor. So, the attributable risk. The risk of an outcome attributed to a given risk factor among those with that risk factor and the attributable fraction is the proportion of cases of an outcome attributable to a risk factor also just with those with a risk factor and then these which I think are always more interesting at the population level, the population attributable risk. The risk of an outcome attributable to a given risk factor in the whole population. Then the PAF is the proportion of cases of an outcome attributable to a risk factor in the population as a whole. Another way this is sometimes talked about just going back here. You can see, it's really about excess risk and in the attributable risk measures the baseline is the risk amongst the unexposed. And, you're saying how much excess risk among the exposed is due to that exposure? And, down here the difference is how much of the overall risk in the population? What's the excess risk in the population compared to the risk in the unexposed? So, the benchmark is really always the risk to the outcome in the unexposed. Right? That's the baseline. They don't have the exposure. What's their level of risk? So, nomenclature. We've just put up here all the various, if you read articles and the literature you can drive yourself nuts because everybody's saying different things and many different things. So, these are just some of the terms that have been used. I know I was trained to think of the population attributable risk percent. That's the PAF, population attributable fraction and again, depending on what you read and where you were trained you might have called these things slightly different things. So, we're just making sure that everybody's clear on the differences in these four basic measures.

Back to formulas. Kristin showed two of the most common formulas and again thinking about this two by two table, all of these formulas are algebraically equivalent and again when you're reading the literature you need to just figure out what that author is talking about. One thing in particular you might want to notice the difference in…I realize I didn't put the footnote again on this slide of the PE at the top of the slide, the first formula that Kristin showed that's the equivalence of the exposure in the population. That is, in the two by two table that's n1 over n. So, the marginal. What's the prevalence of the risk factor overall in the population as opposed to down here this is just in column one. What's the prevalence of exposure in the cases? So, a over m1 and that changes so you notice that down here we just have no relative risk minus 1 over relative risk as opposed to that slightly more complex thing but if you're there too you can work this out and prove to yourself that the algebra is identical. It obviously gets more complicated as we move into the multi-variable world. We're not quite there yet but just in the crude sense these are all algebraically equivalent and they substitute for each other. Yes?

Question: (Inaudible)

DEBORORAH ROSENBERG: Yes but wait, wait, wait, wait. This is wrong. Right? Is this wrong? Oh, yes, yes, you're right. It's right. It's the same as up there . It's the PE. Right? Yes. This formula is identical to that one up there. It's just got this spelled out as opposed to the PE in parenthesis. That's what that is. Yes, make sure you because especially with that light on I can hardly see any of you back there. So, if you have questions, please feel free to interrupt and just yell at me.

Okay, so we're going to go through a simple example and we're going to start out looking at smoking and low birth weight and then you'll see cocaine and low birth weight as Kristin already mentioned and just building up to look at…begin to look at these different kinds of PAF's that Kristin was talking about. So, what you have here is a simple two by two table. I've only obviously produced in SAS. I'm only showing the frequency in the Roe percents because for this example we're going to use relative risks. So, I've got the crude relative risk over here and now the crude PAF, which you can see I've used the formula that just looks at exposure amongst those with the outcome, amongst the cases, if you will. So, in column one, I've got 200 smokers who have delivered low birth weight infants and then O divided by my total of 700 who delivered low birth weight infants. That's the prevalence measure of exposure and then times the relative risk minus one divided by the relative risk and I get a crude population attributable fraction of 10.7% or 0.107. Yes?

Question: (Inaudible)

DEBORORAH ROSENBERG: Not the risk of the population, the low birth weight percent in the population, right, is 10% higher than it would have been if there were no smokers.

Question: (Inaudible)

DEBORORAH ROSENBERG: We can do the calculation and I think you'll see it in a minute. We'll do that in a minute but yes you're right in the sense that if you think again or this is the low birth weight percent, 6.25 amongst the non-smoking women. And so, the expectation is that if you can eliminate smoking from the population everyone would look like the non-smokers and they would have a rate of 6.25 as opposed to a low birth weight percent of ten. We will go through that.

Question: (Inaudible)

DEBORORAH ROSENBERG: Yes. So, just comparing the risk interpretation to the PAF interpretation, so women who smoke are at 1.6 times the risk of delivering a low birth weight infant compared to women who don't smoke and 10.7% of low birth weight births can be attributed to smoking. If smoking were eliminated we would expect 75 fewer low birth weight births and the low birth weight rate would be reduced from 10% to 6.25%.

Question: (Inaudible)

DEBORORAH ROSENBERG: Yes, I am right. So, if you think about… If you do the zero…I don't know if you can see this at all from there but 0.107 x 700, I think that it comes out to approximately this is the 75 cases that would be reduced. We know we have 200 cases amongst smokers. So, if you do this the 200 cases minus 75 equals 125 and then if you do 125 over whatever it was like 10,000 or something like

that…2,000. So, if you do 125 divided by 2,000, you'll now be at 6.25, just like the unexposed group. So that's the expectation if we could eliminate smoking completely from the population. So, now we're going to do the same thing with cocaine and these are made up data by the way. This is not real data but I think you can agree that relatively, realistic so for the low birth weight, you know, those are fairly typical low birth weight percents and I think also fairly typical prevalence of smoking amongst pregnant women. So, it's reasonable even thought its made up. The same thing with cocaine. These are not from any real data source. These are made up and you can see we have a really high relative risk here. Lots of big magnitude of risk if you're a cocaine user during pregnancy of delivering a low birth weight infant, 4.77 but we have cocaine use a lot rarer than smoking during pregnancy, thanks goodness. And so, what happens when we calculate the PAF? We have a lower prevalence. We have a higher relative risk and therefore we end up with PAF of 10.2%. Again, we can go through the same set of interpretations. Women who use cocaine are at 4.77 times the risk of delivering a low birth weight infant compared to their counterparts who don't use cocaine and 10.2% of low birth weight births can be attributable to cocaine use and if that were eliminated from the population, we would expect 71. We had 75 over here, fewer low birth weight births and a low birth weight rate would thus be reduced from 7% to 6.25% in the population and we could do that same kind of calculation where we end up having the same low birth weight percent. There's no exposure group anymore. I mean really I said for low birth weight you changed row one but what it really is there is no row one anymore. There are no cocaine users anymore. Everybody's now is going to be in row two. We've made everybody unexposed. Then, you talked about this. So, this is why we

think this is a kind of instructive example because here we have smoking with high prevalence and definitely 1.6 is nothing to…