MCHB/EPI Atlanta GA Conference

December 10 - 14, 2005

Sample Size Issues and Determination in Maternal and Child Health (Part I)

JUAN ACUNA: Okay, thank you. Okay, let's see. Well, as you see this course is a Latino party. So Brian is the weird one, okay? So you're going to get very familiar not only with sample size (inaudible), which is really good. We're really glad that you are here and that you could be together for this is what has become a kind of not only an exercise from the perspective of statistics and sample size, but from the perspective of our program, The Maternal Child Health (inaudible) Program. And how can we do better to provide training venues for you guys to come to the conference and have the opportunity to high--intermediate to high level trainings, and at the same time, attend the conference, and at the same time network all in just one single week a year. So you could keep this in mind so you could let us know for later how can we improve and if this is really a good idea or not. As you see or you have seen, this is the first time within this couple of days conference that we have had three training activities going at the same time. So you have the idea that you will have several venues and opportunities to seek further training activities.

I'm going to review basic issues. These guys are much—I mean like way better than me. So, I am just here to introduce the perspective of our program in the training activity as I just did, and add the perspective of the epidemiologist. Not

the statistician that understands these like way, way nice and nicely, and they see the formulas and they just chew them up and say, “Oh, my God, this is so nice.” We see them; we get scared with the formulas and say like, “Oh, my God, I don’t get the phi and the theta. Why can’t they cannot keep the same letter for the same thing? And they just keep changing the letters to the things that I have in my mind. And then this is a whole mess, so that’s why I just keep the formula thing.” But they’re going to do a great job, so I’m just going to introduce basic concepts that will keep us within the same grounds. And I hope that I will be able to do so within the next 40 minutes? One hour?

RUBEN: You have one hour.

JUAN ACUNA: Yeah, one hour? Oh, okay. You should have said 45 minutes.

RUBEN: Okay.

JUAN ACUNA: Yeah, you know me better than that, Ruben.

RUBEN: Okay.

JUAN ACUNA: Okay, so this is--okay, good. So, the outline for my section is going to be just some basics, an introduction to the issue of sample sizing, basic epidemiologic concepts that are related to sample size and its consequences,

some basic concepts of sample size, where I am not going to go in-depth on those because this is not the right time. They are going to go in-depth in those, so I will be the ice breaker. They will be your headache. So, into some of these items that we need as part of sample sizing that makes sample size in such as relatively simple concept, but truly later becoming a very methodological issue that creates or damages results, conclusions, study science, et cetera et cetera, and then some epidemiological consequences of sample size.

The first thing that we will see—and Brian will go further into his topic—is the sampling, and sample size are totally different issues. Sample sampling is based on the fact that we seldom need to study whole populations, and we could use just pieces of bits of those populations in order to draw our conclusions. Now, what we have to understand is what do we expect our conclusions to be. And that is one key problem that I find quite often while traveling and talking to people at the state level, at the local level, at agencies' level, that people do not often separate and compartmentalize well enough. So the problem is when you ask people what do you expect about this sample, they say, "Well, what do you mean?" And the right answer should be what they really truly expect from that sample. Not what do you mean because the questions are really simple and straightforward. A sample size will assure that you—we have the statistical power to actually draw the conclusions that you want to draw. So you need to start by figuring out what it is that you expect to get out of your analysis and your studies.

The goals are to maximize the value of research, do the most that I can with the least number of subjects that I can. To the value is measured in resources, both human resources, both physical resources, and economic resources. Maximize the efficiency of the research process, meaning that the research process is going to be as short, as inexpensive, and as efficient as possible but still is going to give me the opportunity to draw upon the conclusions that I really want to draw. Based on what? Of course, based on starting on a hypothesis. So for instance, this is one of the most recent papers that I could find and it was just yesterday--so this is like brand new, hot and steaming, a randomized controlled trial to determine treatment for malaria in pregnant women. So, what would be your questions if you would be the methodologist behind such a problematic question. Do you think that this is an important problematic question? Well, for some places in the world such as Africa, such as many others where malaria is extremely prevalent, these are the questions, and are the questions that are different between life and death. Now what is the point? The point is what would be your research question here. Any volunteer? This is an MCH topic. So what would be your first question, in order to start drawing conclusions that would lead you to the proper sample size and sample sizing or sampling frames?

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: From what population? Okay, so we need a population and we need to sample. Is that true for this particular problem? What field of

epidemiology am I addressing or would these questions be included upon?

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: Excuse me?

UNKNOWN SPEAKER: Clinical trial.

JUAN ACUNA: Is that a design or a topic? That's a design. So that was going to be my next question. What is the design? So, if we're interested in a clinical trial, do we or are we extremely interested in knowing all the population that we can sample from? Probably not. So you see, the first thing is we are not very used to ask ourselves questions. So then again, what is the first question for this health problem?

UNNOWN SPEAKER: What's the (inaudible) population?

JUAN ACUNA: That could be the third, maybe the second question? What is the field? I mean, is it the same thing to calculate the sample size or try to do the study if I would be asking for diagnosis instead of treatment? It would be different. So what is the first question?

UNKNOWN SPEAKER: I think you have to clarify how do you define treatment.

JUAN ACUNA: I think that you have to realize that this is a treatment or intervention research question. Because it could be a diagnosis, which is another field, or it could be a prognosis, which this is not. But it could also be a risk factor study. I mean, what are the risk factors for pregnant women to get malaria, what is the prognosis of them, et cetera et cetera. So all of those are studies and they lie in different fields in epidemiology, which triggers the next questions including sample size. But we're not used to those kind of questions. Why? Because we seldom work outside the risk factor arena. Am I right? How many of you have done clinical trials for MCH? See, a few. How many of you have researched on the risk factors for MCH outcomes? Like all of you. Risk factors for low-birth rate, risk factors for mortality, risk factors for breast-feeding, et cetera, et cetera. So it is important to understand that we need to start with certain questions that address the epidemiologic field that we want to address with our research and designs.

The purpose of sampling among others are to improve the efficiency of research process, as I have already explained, to assure the adequacy of the research process, to address methodological adequacy to respond to research questions, so we have to start with questions. And sometimes, a research question might be a good one or might be a bad one. And we are going to talk about these later. And to address the consequences related to testing, for instance, interventions in more or less needed subject. For instance, what would be the problem if I

establish an intervention study and I would not calculate a sample size? What happens if I test too many people? What would be the most immediate consequences? Cost. So one is cost. Cost in what? Measured in what? How do you measure cost?

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: Well, the first one is, of course, money. I mean, that's the one that hurts the most.

UNKNOWN SPEAKER: If one of the (inaudible), we're assuming that you're here testing an (inaudible) better than the others that you actually are using more people, subjecting more people to address outcome than you need to. Before you learn, you need to know to learn what the best practice is.

JUAN ACUNA: Absolutely. So another one could be the cost of exposing to interventions people that were not needed to be exposed if the intervention is damaging. But if it's helpful, it would be the opposite if the amount of people would be less than the one indicated. What would happen if my sample size is too short, and I would do a study, and the intervention would be useful? I might not be able to even prove it. So there is a big problem. So, here are some of them. Address planning adequacy. So, I need to plan for these research studies and we take that for granted in public health, because we use many--especially

in maternal and child health—we use many pre-collected data. So because we do that, we seldom think on sample size issue. But what if each record to do a birth weight study would cost one dollar? How many of you have analyzed nation-wide data sets for birth weight? Many. How many records?

UNKNOWN SPEAKER: Hundreds of thousands.

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: Hundreds of thousand, maybe millions? One--only one year for birth weight is four million records plus take it—give it a few. Four million. So just one year would cost four million dollars. Would we be thinking in sample sizes if we would need to pay for those records? Of course, we would be thinking much better on how many records do we really, really need to get the conclusions that we really want to get. The (inaudible) potential of harmful exposures within research settings, so that's what Beth was talking about, address ethical concerns, which is, of course, closely related to the previous one but expands even further, and assure the adequacy of the conclusion from the perspective of the hypothesis.

Okay. What are the approaches that we can follow to for sample size calculation? The first one is to specify the desired measure to be addressed by the study. For example, the width of a (inaudible) with measure, a mean, et

cetera, et cetera, et cetera, proportion, and that is the first approach. By (inaudible) approach, which is based on approach which is conditional, that that is what rules are. Conditional thinking such as precision of the study that I need and cost and given that I need those how many people would I need to study. I'm defining the power to test and hypothesis, which is pretty much one of the most common approaches used. So, as it is the most common one, I am going to use that to introduce and go over some of the concepts. In the power approach, the first thing that we do is specify the hypothesis test. On a parameter of importance, such as—whatever. I mean, tell me one parameter of importance that you would be interested on. What are you interested on in your home state right now? What?

UNKNOWN SPEAKER: Maternal age.

JUAN ACUNA: Maternal age. From what perspective?

UNKNOWN SPEAKER: What do you mean? I don't understand the question (inaudible). Oh, I thought you're talking about low-birth weight. (Inaudible).

JUAN ACUNA: Well, it could be low-birth weight or not. I mean, the point being that when we need to calculate a sample size, we need to start from a perspective of something that we need that is the focus, the pinpoint focus of our research, which is what we really need to draw the conclusion upon. I mean, we

really need to go through that measure.

UNKNOWN SPEAKER: (Inaudible) if you wanted to test the risk if there's a difference in cardiomyopathy between African-American and Caucasians in their (inaudible) for testing?

JUAN ACUNA: For instance, we are very interested in the risk of cardiomyopathy in two group of women. So what would be the parameter that we're going to choose to calculate that sample size? At least what type?

UNKNOWN SPEAKER: The prevalence (inaudible).

JUAN ACUNA: The prevalence is a proportion? So it would be a different of proportions for both groups and then we will start from there. But what are the issues regarding those proportions? Because if that would be it, probably this whole course would not be two days but maybe like half an hour or so? So it has to be much more complex. And we are going to try to address some of those. Specify the level of alpha and you all know this. Specify an effect size or an alternative value for the parameter in step one. For instance, if we want to try to test what is the difference between two group of women or the same group of women in two different circumstances, and we have a baseline on the prevalence of cardiomyopathy, what is the next level of that prevalence that we want to be sure that we're going to be able to capture with our study? And that is the effect

size. But we are going to talk more about this probably over and over and over and again.

Obtain evidence of certainty or guidance, at least, that those parameters are actually the ones that are important. For instance, if you try to determine in a study that something that is a risk factor or addressing a risk factor reuses or increases the prevalence of cardiomyopathy, how many points of increase would you be able to tolerate in order for your conclusions to be correct? One point? Two points? Three points? I mean, we're talking a point percent on something that kills women. So the importance of being correct about a one or two or three points is crucial, but we seldom think in those terms. Why? Because we have pre-collected data. If the data is not enough, we don't have more. And if it's enough, sometimes the conclusions might be drawn by large samples which have problems that we're going to talk as well. And specify a target value of the power when the parameter to be tested, which is number one, equals the final parameter which is number three. So, how much power would you like if your prevalence initial one would be equal to the prevalence final one? And that, of course, is a theoretical approach to do and calculate this.

We have problems. And sometimes, sample size, it's just not so important. Why? Because we have the data. So for instance, if we're talking about the economic consequences of preterm birth, and we're talking about hospital charges for premature infants total 18.1 billion in 2003, how sure do we need to be that it was

18.1 billion and not 17.9? Who cares it's a lot of money?

So sample size goes beyond reality and the power intrinsic, power of having drawn conclusions from a population-based sample or a population. Premature infants accounted for half of the hospital charges for infants. I mean, do we really need to put some statistical power in those statements? Probably not. That's why they don't have one, because they are drawn from humongous data sets and humongous populations, et cetera, et cetera. So, do we really need to determine that there was a sample size issue or what was the power on an estimate that in the whole U.S., the difference between a normal birth at $1,700 and a preterm birth at $77,000 is really different? Have you seen a P-value for that estimate? I mean it's totally ridiculous. So remember that one thing that we cannot be is more popus than the Pope, okay? So we need to work under the level of certainty that assures that our conclusions are correct and that we are not misusing data as we have it.

Good. A lot of data that we analyze comes from the U.S. birth certificate. And you will know that and probably you are very savvy in analyzing these data. So we're not going to talk about large sample sizes or large samples, because it would not be relevant. Nevertheless, a good example of large samples and useful information comes from, for instance, this is a longitudinal data set in Massachusetts that is called (inaudible), and that has a large partnership of people working on it, and usually you can draw pretty good conclusions. But just

pretend that you would be trying to test hospital discharge data and birth fetal death certificate regarding the presence or absence of the babies, and you see how many of these squares don't match. So, what would be your question if you would not have population-based data to try to determine if these relationships are true? That's the topic of the rest of today and tomorrow's workshop when we do not have these.

This is one study that I like to present a lot, because it really split the way physicians, at least, do health care, and it was published in the early '90s, 1991, and it determined that the level of knowledge of physicians decreases with the number of years after graduation if they do not engage in specific activities in order to keep that knowledge. And those activities are not CMEs. It has been proven that CMEs does squat. It does. So, this was the birth. And what you're seeing is the birth of the concept of evidence-based medicine, which has changed the way that we practice clinical medicine. I would say that almost throughout the world, because the places that have not joined are very concerned on how to join really fast into the way of living that is evidence-based clinical practices. And you have seen that now we have jumped and we have stolen some of those concepts for good on how to practice evidence-based public health. And that is very important, because accountability has become a major word in public health. So, what is outstanding about this study is that it changed the world. At least, an important piece of the world of health that now goes into public health. And what was the sample size? Just count the red dots.

Forty-six subjects. So, when you have the right conclusion together with the right design, you do not need to measure everybody. I mean that's a paradigm that we have in our minds just because we can do so in MCH, but we do not need to do that. That's what we're going to learn. So the elements in sample sizing are several. We need to start with a good hypothesis, because if we do not have a good hypothesis, we might end up into a type 3 error, which I am going to explain later. The hypothesis testing, of course, addresses something that I will not talk about in-depth, which is the type 1 error and the type 2 errors, because you are going to be tired by the end of tomorrow talking about these type of errors. The level of alpha, which is usually set at five percent but shouldn't, because you need to think really what is it that you'd really need to know in order to adjust these level of confidence to levels different than five percent. Five percent is a good starting point. But when you ask why five percent and not three percent or two percent or 10 percent and people say, "Well, I don't know what--," what do you mean? That you know that the process of sample sizing was not really a thoughtful one, and we will learn that it needs to be a very thoughtful one. Same for power level, which is usually set at 0.8 or 80 percent. And the effect size, which is the point estimator, boundaries and variance of those estimators, and how to deal with them in formulas of sample size for different context.

So the hypothesis is born out of a research question, is set up—and we all know this in the form or format of a null hypothesis, or a hypothesis with no difference or no association. And it needs to be tested, because if you do not test the

hypothesis, you would be doing squat with your own research process or project. So you need to test that hypothesis. And the whole idea is to reject the hypothesis, because proving things is quite easy. I mean, can you prove anything, even things that you do not see? Yes, of course. So actually, not being able to reject things is what makes stronger the fact that you will have to accept or not them. And if it cannot be rejected, the alternative hypothesis will not be true. And that is a pretty simple approach, but it needs to be quite evident in our minds.

So now, I am going to show you just an example of how all these things work together--and I am not going to go into detail, much less about the formulas that I don't even understand myself. So I just put them so you'd go like, "Wow". So anyway, the whole point is how many people would we need if we would need to determine that the average height of a foreign people in the U.S. is 175 centimeters? And that is your statement. Why? Oh, just because you are engaging into a major, major amount of money being invested in clothes for foreigners. And there are many, many, many, many in the U.S. You're going to be either very rich if you do it right, or extremely poor if you fail to prove that this is your most popular size. So what happens if you do that? Well, let's assume that in reality, our sample mean--so, we already sample a couple of—we just found that the sample mean was 180 centimeters and not 175 centimeters. And we somehow know, because we measured the standard deviation of those guys that the standard deviation is five centimeters. So, how sure can we be that that

is a wrongful set of data that led me into thinking that those guys actually measure 175?

So, the point is that we can put that into a test statistic, and these test statistic if we have these many people distributes normal as you can see here with a mean of 0 and a standard deviation of 1, and what we have is that if N, which is the number of subjects in our sample, would be small and small as far as we are not told many, many years ago—it's a bit many, huh—it has to be less than around 30, isn't it? Then you need to go into a study test distribution and we would see what the implications are of that. And you have to determine if it has to be one-sided or two-sided, and one sided means I am fairly sure that my appreciation is that those guys are definitely smaller than what I think or larger than what I think.

The sample size in order to reject the null hypothesis would be outstandingly small, just three. And you can do the calculations. I'm not going to do them for you. I'm lazy for doing that. But if you use the T-test, it's going to be five, which is not bad as a difference. Now, if a sample mean would've been 176 centimeters, then we would have needed more subjects and that is the implication. It would have been 68 people to reject the null hypothesis.

So, if we do power calculations to try to determine how certain are we about our conclusions. This concept is somehow backwards, because we started with a sample which will renew that had an average height and we need to start almost

always from the other approach, which is how many people would we need to really know what we are supposed to know? So, what we do is that we will choose before having our sample the effect size that is important for us. So, is it really different as size in clothes for a person that measures 180 as opposed to 179? Not really. Probably the same size. What about 175 and 176? Probably the same size, not a big difference. What about 175 and 180? Well, that's probably even almost two sizes big. So, it has to be somewhere in between. That is how we estimate effect sizes from a very simple approach.

So, what we have is that we have a very relatively simple formula where the final calculation for our sample sizes depends upon those factors that I already mentioned to you and the size of the overall effect of the size that we need to consider. And the sample size, total sample size is, of course, the sum of the cases that will lead me to reject the null hypothesis and those where the null hypothesis will not be rejected. So, of course, what you need to know is that the standard error is a function of N, which is the whole big deal. So, everything boils down to the same handful of factors that we need to think about before calculating our sample size.

So, what is the real question after all of this? What sample sizes require given that I need a level of certainty that I need statistical power and that my effect size in order to be—in order for my conclusions to be drawn correctly is the one that I really need.

So, we do that and we are going to assume that we have taken a sample of 25 people, let's assume that we took the easy approach, which is—I have already done these several times and I know that there is little people that I need to really measure in order to try to measure differences between these means given that the standard deviation is so narrow. So, I choose 25 people. What would be the consequences? Well, we apply the same formula, and in the case one, we see that the power was really, really, really high. But if the difference would've been one centimeter as in the other example, the power would've been really, really, really small and we would've wasted a lot of time trying to do this. What if each one of these subjects wanted a wardrobe from our clothing company in order to comply with—participate in this study? We would've lost a lot of money. And you can use software in order to plot in different circumstances the level of the power that you're going to have given here is the case one and here is case two, and you see how power increases with sample size given the size of the effect that you need to estimate.

So, in order to do that, of course, you have to start by determining how your hypothesis testing is going to work out. And it's going to work out depending on a setting on a two by two table, and you already know this, so I am not going to talk further about type 1 and type 2 errors, but basically, what you need to be aware of is that depending of the level of certainty and power that you set, the alpha level and the beta level, you're going to have more trouble or less trouble trying

to reject or accept based on your study the null hypothesis. So, you can change the five percent and the 20 percent or the one minus 20, which is 80 percent, but you need to know why. But if you keep them in place, you need to also know why. So, you have to have a reason for every single one of those few parameters that you are going to include in every single one of your sample size calculation formulas. And the bonus is the type 3 error, which is, if you started with the wrong hypothesis, then you're doomed for failure. And it's a very common error to give the right answer to the wrong question. So, if you start with the wrong question, and you get the right answer, what are you going to do? You're going to believe your answer, because we focus more on answers than on questions. So, it is very important is the key part of not only designing the study but in your research process to try to formulate the true questions and the questions that are relevant both for you, for your research, and for what you need to do.

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: Yeah.

UNKNOWN SPEAKER: I thought about that quite a lot and its often—sometimes there's a question you could answer really, really well, but it might not be the right question, and there might be another—the real question is actually really hard to answer and you kind of—the balance between getting a precise answer to the wrong question versus an imprecise answer to the right question to the wrong

(inaudible) and to the right question. So, sometimes you do have to settle (inaudible).

JUAN ACUNA: Well, that's why—when my kids were your age—well, not—your kid's age—

UNKNOWN SPEAKER: Big children (inaudible).

JUAN ACUNA: Yeah.

UNKNOWN SPEAKER: Big children (inaudible).

JUAN ACUNA: Well, I started very early. I used to have more hair, believe it or not. So, I used to tell them, there is right and wrong, and you want to lead them to right and wrong. And now that they're teenagers, I tell them, it's not that easy. There is not really right and wrong. There is just trade offs. So, that is very accurate a statement when you're dealing with questions. In questions in health, there are almost perfect questions but cannot be answered. And sometimes, you have easy answers for very imperfect questions, and you have to really go throughout the process of thinking throughout that range on where does this trade off process actually flips the coin, and you say, "No, I am not willing to really sacrifice more quality on my question because of easiness to get the answer." So, you are totally right, and I think that is a great point because asking

questions in public health has been one of our most important failures and problems.

How many of you think that the current level of training in public health lets you elaborate more questions than give more answers? Our whole system circulates around providing answers. So, we are really good. I mean, just bring it on. Give me—okay. Ask me a question, any question. I'm really good. I am just full of answers. But when some people say, well, "There is this problem of malaria, and these women are dying. What is the most relevant question to address given that you have just $100,000 to do research?" I mean, you have to really sit down and think like for weeks, because you're not sure, because we were not taught into asking questions. So, we have to become better on asking questions rather than giving answers, because answers, we know how to do that. But questions, that's really a problem. And that is one of the major pitfalls of sample size calculation as you will see. I will end up with a statement that my co-presenters might disagree upon, but I would like that to be the initiation of our discussions.

So, when we are going to try to determine statistical difference, of course, we have usually—and I'm going to just pretend that we're working with means because it's just easy for the example—we have one mean and we have another mean and we even try to determine whether there is a statistical difference or they are really different one from each other. And what we end up doing is needing to understand that actually what makes them equal or not is whether the

distance between them actually belongs to a greater picture. So, if they belong to a greater picture, probably they are not different. They might be. But what is the problem? That we are not just seeking statistical differences, we're also seeking clinical or public health differences and meaningful differences. So, if you use large sample sizes, you will find statistical difference in many circumstances when these statistical differences are totally unimportant. Why so? Well, because everything has to do with the boundaries of certainty that we want to work upon given that we have populations and problems and questions regarding that population. So, we will conclude that because there is a broader spectrum in our interest, in our question, there is no difference between our two means that we were working with.

Now, in order to understand these, we need to understand that there are two different samples. The first one is large, humongous samples where statistical difference has to be taken into account extremely carefully, and we need to make sure that the statistical difference doesn't go beyond the clinical utility or clinical meaningfulness of that difference. Now, we have the other circumstance, which is probably the topic of this whole workshop, which is smaller samples, where statistical significance might not be reached even if there is an extremely important interest in actually studying that statistical difference. So, what we need to warrantee by calculating a sample size is that we are going to be able to actually in the presence of clinical interest to establish that difference, to be able to pick up the difference if it is present.

Now, in the previous example, we assume that these two distributions had a distance and that that distance was not enough to warrantee that they were different. Now, where we keep piling information and information piles around the mean, and we keep piling and piling information. What we're going to have is that we're going to overgrow the distributions within pretty much the same range. So, what we're going to do is have edges that are extremely thin and same distances. Distance one and distance two are going to become different just because most of the people just squeezed out the edges and strengthened the middle of the distribution. So this is a very cartoonish approach to one basic concept which is beware of analysis in large samples, because even the hair in the back of an elephant will become significant. And that's what Fran Meider, which is in the other room, taught me many years ago. So the take-home message is if you analyze large samples, have to be careful on clinical significance. If you analyze small samples, you have to be very careful on statistical significance and its meaning.

Effect size, and this is where I might disagree with micro centers, but I think that in my mind, for me as an epidemiologist, as a clinician, the effect size is actually probably the most important single element in sample size calculation, and is an extremely artistic piece of determining a sample size, because it's really hard to understand how much of an effect is important and how much am I willing to strengthen the sample size process by increasing that effect to the point that is

going to shrink my sample size. Why? Because it's part of the denominator. So the biggest difference, the smaller the sample size and vice versa. Is hypothesis dependent? Which means if I do not have a good hypothesis, I will not have a good effect size. And regardless of whether the sample size is calculated, on that effect, if it's not meaningful and related to my hypothesis, then I am going to put in jeopardy my whole project.

Needs a known baseline. So if you say, "Oh, I'm very interested on smoking as a risk factor for birth weight problems," you better have a good baseline to know where you're going to start and to understand what is going to be the level of difference that you're going to tolerate in order to have your conclusions agree with your hypothesis based on a correct sample size determination.

Needs a final estimate. Should you need upper boundary, you need a baseline, and you need an upper boundary, a difference between the two of them is going to be a size effect. And most likely, it needs to be determined by a content expert without statistical epidemiological background, which is really problematic.

Extremely problematic. Because if you are a methodologist and you don't know a lot of other content, and here you have a content expert that knows squat all the implications of what he is saying or she is saying, what you're going to have is people that say, "Oh, no, I need to be really sure," and you have humungous sample sizes that might prove that your differences were accurate and correct,

and they were found, but without—I mean, actually going beyond clinical utility you—so you end up wasting a lot of resources. So you need to understand this part that you guys need not to be working together, but you actually need to overlap. So the content expert needs to understand the consequences of statistical epidemiological background and you need to get a little bit into the content if you're a methodologist.

So, just as an example, and this is another smoking clinical trial from last week. Actually, this is just an ongoing trial. Birth after C-section. The whole point being that people are saying, "Don't worry about C-sections," because with the newer techniques and the safety of the surgical procedures, there is not a consequence of actually delivering most babies by C-section. Not a problem. Period. So, if I am an OB-GYN, I will be extremely happy with that statement. Why? Because I can delivery every single of my patients, whenever I want, within certain limits. So, if these would be the problem and we would be trying to speak on behalf of patient preference, many women today would say, "I would much rather a C-section than a vaginal delivery if there are no consequences for me and for my baby." If it's exactly the same, I much rather have a baby by C-section. I can have it, too, the day I want, same grandma's birthday or not, whatever it takes, whatever it is. So, how important is this in a study? What would be a good question that we would need to ask to—that we would like to answer in such a setting? Important problem in Puerto Rico. In Puerto Rico, 50 percent of births are delivered by C-

section. And we don’t know if that’s bad or not. So what means bad? All of you are MCHers, what would be your interest?

UNKNOWN SPEAKER: The difference in birth outcomes.

JUAN ACUNA: Difference in birth outcomes. Okay, that is a good one because--what about the women?

UNKNOWN SPEAKER: (Inaudible) women’s recovery or women’s complication.

JUAN ACUNA: Women’s complication, that’s another good one. So which one of those two is better?

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: I mean, if I ask the babies question, that’s one thing. If I ask the moms question, that’s another thing.

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: Yeah.

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: I mean, the kid ends up as a very, very jerkish teenager. Let's go to the obstetrician because the C-section was not needed and there might be birth trauma or something like that. Do you know that obstetricians are liable for kids until they're like 16 or something? So what would be an important question? What do you think it is their question? Then again, this is a funded trial, so they already got the money to do this. And these trials are very expensive.

UNKNOWN SPEAKER: Well, this is a (inaudible).

JUAN ACUNA: Well, that's what I'm saying. If it would be a research question, what would be your effect that you would like to measure? Because you need to propose a sample size, don't you? I mean, this is something that you cannot readily answer with vital records or PRAMS, or (inaudible), or--no. I mean, with those we just get some hints on what the risk factors are. But it would really need to prove that this is dangerous or not. If there needs to be a wide-scope policy regarding you cannot do C-sections unless they are strongly justified by what we know already that is harmful both for mom and baby, you should not be doing C-sections. So, let's go with neonatal outcomes, okay? Let's say that you find out that poor outcomes after C-sections or normal low-risk deliveries are into the 1.5 percent level. How much would be too much? Twenty percent? Is that too much, jumping from one 1.5 to 20 percent? Because that would be a tiny sample size, so it would be a very, very, very, very inexpensive clinical trial. How much would

you tolerate? Remember—what is the question? And you can formulate that question in a form of a hypothesis. Null hypothesis. There is no difference in neonatal outcomes in babies born by C-section or by vaginal delivery. No difference.

UNKNOWN SPEAKER: (Inaudible).

JUAN ACUNA: Yeah, among women who have a preference, yeah. By the way, are you aware of why this clinical trial? Because there is a study that was recently published that said that there is no problem by women choosing C-section based on consequences for the women or the baby.

UNKNOWN SPEAKER: And how they measure.

JUAN ACUNA: Huh?

UNKNOWN SPEAKER: What caused it.

JUAN ACUNA: Of course, it was an observational study. Yeah, but I mean, you have to understand who are you telling this to. Are you telling obstetricians that have actually wasted or not, it depends on the perspective, hours of sleep and hard work trying to, you know, be at the bedside--of course, that doesn’t happen anymore—but be at the bedside of the women you know for hours and hours and

hours and hours to deliver a baby vaginally in a safe manner? So that is the standard of care. So what if you could say there's no problem? I mean, these guys are going to be the happiest persons around. So that is the perspective. You give them a hint of anything they will—by the truckloads, there are going to be many that will say, "No, no, no, no, no, I will still do the right thing and the known thing." But there are many, thousands that are going to start doing C-sections, because there is no problem, and that's what happened. So that is why this clinical trial. So I'm telling you right—a true problem. But the point is, for the perspective of this exercise, how many people do you need in order to study this when the baseline level of neonatal complications is around 1.5 percent? So we learn that we need to estimate an upper boundary for that effect. How much do you guys would like to study as a meaningful difference? So do you think that 20 percent complication level makes sense? I would say no. It's way too much. I mean, how many kids in the U.S. would be damaged if there is actually such an association between C-section and neonatal complication, and we determine that with a study that was only able to pick up a difference between 1.5 and 20 percent? I mean, that's way too much. So who can give me a better number? Better number?

UNKNOWN SPEAKER: Five percent.

JUAN ACUNA: Five percent? More or less? Five? Everybody agrees? Five.

UNKNOWN SPEAKER: Three.

JUAN ACUNA: Huh, three?

UNKNOWN SPEAKER: (Inaudible) twice as much.

JUAN ACUNA: So basically, the sample size of this study ended up being 2,314 women. With the standard alpha and betas, those power, standard, which is the only thing that I didn't like, because it's really, really rare that you will come up with these levels if you really think of the consequences. But what is interesting is--and that's why they brought it up--is because the effect size is being able to determine a raise in complications in just two percent, thus the really high—this is a large sample size for a clinical trial. I mean, each one of this one is going to cost several thousands of dollars, so this is a multi-million clinical trial.

UNKNOWN SPEAKER: (Inaudible) multi-side that you're not going to be able (inaudible).

JUAN ACUNA: Absolutely. I mean, this is an expensive investigation. But it's an important one. But look at the effect size. So, what is interesting is that they shot towards an increase of just two percent on neonates because actually two percent neonates from this perspective would mean several tens of thousands of babies with complications in the U.S. per year, U.S. alone.

UNKNOWN SPEAKER: It’s not so more than a doubling, where if you think of it that way, it’s more than a doubling in adverse outcome, which makes it sound big.

JUAN ACUNA: Yeah, well, you have to think the doubling doesn’t work here. So, you have to be very careful in sample size effect when you talk about an increase of 100 percent, because it doesn’t work that way. It works better on the levels and the consequences of the absolute number and its consequences. Okay?

UNKNOWN SPEAKER: But how many babies is that? If you look at—

JUAN ACUNA: Many.

UNKNOWN SPEAKER: --30 percent of babies--

JUAN ACUNA: Yeah.

UNKNOWN SPEAKER: --by C-section (inaudible) 40 percent of those can be avoided if (inaudible).

JUAN ACUNA: Yeah. But if you say, on the other hand, that this year the level of complications was 1.6, our next year is 2.6, what would be your first thought given that you're analyzing birth files?

UNKNOWN SPEAKER: Nah.

JUAN ACUNA: Not a big deal. I mean, we can—even at the population-based level, you can not really estimate that those numbers are different. So, what I guess that these guys did is that they determined that that meaningful difference would be a representation of the real problem when you jump from 1.6 two points to 3.6. Yes?

UNKNOWN SPEAKER: When you mentioned the (inaudible), what you meant is that you needed a higher power for such clinical maintenance?

JUAN ACUNA: Unfortunately, we do not—I don't know if you're going to address that later, but the problem is, remember that the power and the alpha levels and the beta levels means false positive and false negative results regarding testing your hypothesis. So, what you need to think is step back and say, "What are the consequences if I am wrong rejecting such a hypothesis?" So, if you would be testing, for instance, a diagnostic method for HIV and you know that misdiagnosing an HIV case means death, as opposed to over diagnosing, which means, well, maybe more treatment but later somebody's going to retest or

maybe you're going to retest first and then you were going to say, "Oh, this was a false positive." That is much better. So, you want to go with much level higher certainty levels than the standard. You want to have 0.05. You want to have 0.01.

UNKNOWN SPEAKER: And that would affect the effect size?

JUAN ACUNA: Of course, but is much rare than the changes in the effect size, but I--then again, this is going to be a topic of the rest of the day. So, what are the main conclusions? The main conclusions of this first part is or are the good sample size determination does not equal adequate sampling. You could sample size very well and you can screw up—so you can totally miss the problem of sampling, or you can have perfect sampling and you could damage your research because of inadequate sample size determination. So, they are not to be confused. I mean, actually, you have nothing to do except the fact that they both deal with the sample.

Database validation of parameters is very important. Do not use customized sampling numbers. So, never accept 0.05, 80 percent, sample size of five percent—no, no, no. Don't do that. If you do that, then sample sizing is not a big deal, but most of the times you're going to be wrong. And it depends on strong interaction between a teen that has to be composed by statistical support, by epidemiological support, and by content support. And they need to overlap a notch. If they just touch themselves then you're going to have problems because

they need to communicate much better than just providing input for sample size calculation. And despite simple formulas where few elements are involved, sample sizing is a highly technical part of study. So, you have to be very careful on really thinking and understanding each one of those elements that are written down in a formula that is going to help you calculate true sample or an adequate sample size.

And this is my contact information, and I thank you much, and now I will leave you in the hands of the experts. Thank you.