TITLE: SPECIAL ISSUES: P-values and Confidence Intervals; Power and Sample Size

PRESENTER: Julie E. Buring, ScD

Slide 1: In the previous lecture on interpretation of epidemiologic studies we talked about the role of chance as an explanation for our findings. Today, I'd like to talk a little bit more about the special issue of chance, in particular a little more about p-Values versus Confidence Intervals and the issue of power related sample size.

Slide 2: If we review for just a second what we said during the talk on interpretation of epidemiologic studies, what we said is that chance is always an explanation for our data, because we are trying to draw an inference about all people within exposure and/or an outcome based on a limited example of the entire population which is our study.

So chance or sampling variability must be taken into account when we describe our data, as well as if we're going to make comparison between groups. But the overriding principle that we always have to remember is that the size of the sample on which we are basing our conclusions, it's going to play a major role in the likelihood the chances and explanation for our findings.

Slide 3: A common way to measure the effect of chance is by conducting a test of statistical significance where you setup a Null Hypothesis (H0); nothing is going on, no difference between the groups, no association. And you test an Alternative Hypothesis (H1) that something is happening, there is a difference between the groups, there is an association and you perform the appropriate test of statistical significance.

Slide 4: Now there are very different specific tests for specific situations. So if you're going to be comparing whether the two proportions are the same or not, you might use a Chi-Square Test, χ^2 test. If the sample size is very small so you can't assume an underlying distribution, you'll use something like an Exact Test, may be a Fisher's Exact Test.

If the sample size is very large and you can assume normality, you are using a Z score. If it's not quite as big as the Z score situation, you might use a t-test that. But regardless of what test you learn in biostatistics to do, all the tests have the same basic structure.

Each test statistic is a function of the difference between the values that were observed in this study and those that would have been expected under the null hypothesis. If the null hypothesis was true and there were no association between the exposure and the disease.

And we'd have to take into account when we look at the observed to the expected, the amount of variability in the sample which is a function of the sample size.

Slide 5: All tests and statistical significance lead to some measure of the effective chance on the results of the study. And one measure that we talked about before is the resultant p-Value, where the probability in obtaining a result has extreme or more extreme than the actual sample value that was obtained, given that the null hypothesis is true.

And on the basis of this p-Value and based on a priori chosen cutoff which is usually 05 level in the medical literature, we're either going to reject the null hypothesis, conclude that the association is statistically significant if the P equals 0.05 level, if this p-Value is less than 0.05 or we don't reject the null hypothesis if the p-Value is greater than or equal to 0.05, and we say that the association is not statistically significant at the 0.05 level.

Slide 6: Now the problem is that the p-Value reflects both the strength of the association as well as the sample size of the study, so even a small difference between two groups is going to achieve statistical significance. In other words, it will be judged unlikely to be due to chance if the sample size is big enough.

And even a big difference, something that's very important to us from a clinical standpoint or a public health standpoint will not achieve statistical significance; we won't be able to rule out chances and explanation for our findings if the sample size is too small.

The problem is what do we do when we have a small to moderate size difference between the groups and it's not statistically significant? Is it not statistically significant because the sample size wasn't large enough to detect an effect to that size even if one were truly present, or was it not statistically significant because truly nothing is going on, there is no effect, no association between the exposure and the outcome.

Is there any way that we could tease those two explanations apart and try to figure out how much the finding the p-Value that we saw is due to the size of the sample, rather than the size of the effect that we saw.

Slide 7: So to separate out these two components of the p-Value, the Confidence Interval should always be reported. And the Confidence Interval is instead of just giving us one number, the observed relative risk that represents the association between the exposure and the outcome in our data, give us a range of values, not just one, a range of values within which the true relative risk or absolute difference, the true magnitude of effect lies with a certain degree of confidence. So if we get 95% Confidence Interval, then we can say it's the range of values within which the true magnitude of effect will lie with 95% confidence.

Slide 8: The Confidence Interval is actually going to be able to provide all the information of the p-Value in terms of assessing whether the association is statistically significant or not at that level. So if we do

95% Confidence Interval, we can look at that Confidence Interval and know whether the p-Value is significant or not at the 0.05 level.

But far more importantly than that, the width of the Confidence Interval reflects the precision of the estimate. In other words, it reflects what the true value of the association is likely to be. And the interpretation of a Confidence Interval then will depend on the scientific question we are trying to address.

Slide 9: So let's take an example. Let's think, someone comes to you, may be a family member, may be a patient, may be a friend and says, "You know, I'm going through menopause right now, I'm really wondering whether I should take these postmenopausal hormones?

I've read that there are going to be some benefits if I take them on postmenopausal symptoms and they would help the risk of osteoporosis, which I am sort of concerned about. But I've also read that there might be an increased risk of breast cancer and it looks like there might be an increased risk of endometrial cancer. So would you be willing, since you've been taking these -- reading these lectures and thinking about this, would you be willing to look at the literature and tell me what the relationship is between postmenopausal hormones, endometrial cancer and breast cancer."

So you go ahead and do a literature review and this is what you find.

Slide 10: First you look at postmenopausal hormones and endometrial cancer and you pull two studies from the literature, both of which have a relative risk of 7.5, meaning both studies, if you calculate the observed value of the association, it says that women who use postmenopausal hormones have 7.5 times the risk of developing endometrial cancer, compared to women who did not use those hormone.

And in both studies the P is less than 0.05, which means that less than 1 out of 20 times I would see this value by chance alone, given the sample size of this study; therefore I am going to reject the null hypothesis that there is no association between hormones and endometrial cancer. I'm going to say chance is an unlikely explanation from my findings, there is a statistically significant association at the 0.05 level between postmenopausal hormones and endometrial cancer.

Slide 11: But what if I want a little bit more than that, what if I want to know whether 7.5 is the right number, does that really represent the relationship between hormones and endometrial cancer? You could say or someone else could say, that number is awfully big, we don't usually have relative risk quite that big, are you sure it's 7.5? So you would answer, well, let me look at the Confidence Intervals and let me tell you what the range of values are that are compatible with our data.

So Study number 1, you look it up, relative risk of 7.5, but 95% Confidence Interval between 1.1 and 32.1. What that means is that with 95% confidence the true relative risk, the true measure of the association between postmenopausal hormones and endometrial cancer lies between something as low as 1.1, so the women who use the hormones only have 10% increased risk of endometrial cancer.

And as high as 32.1 that women using hormones have a 32 fold increased risk of endometrial cancer. And all of those values are compatible with the data.

Now just looking at the width of that Confidence Interval, I know for sure that that sample size was small. That they did not have many people in the study and thus the data that they calculate must reflect the fact that they are making calculations based on a small number of participants.

And the smaller the sample size, the more variability there is. The more variability there is, the wider the Confidence Interval has to be to include all the alternative values that could be comparable with the study. So a small sample size will result in a wide confidence interval and a large sample size with less variability, will mean that we can give a narrower range of values that are compatible with the data.

So when I look at this first study I go, well, 7.5 might be what we just observed in the study, but I wouldn't count on 7.5 being the true magnitude of the effect, given the alternative number of values that there could be possible compatible with our data as shown in the Confidence Interval.

Then I go to Study number 2, I pull that and I look at it, remember it had an observed relative risk of 7.5.

Slide 12: And remember I told you that you could get the information in the Confidence Interval that says whether the association was significant or not at the 0.05 level. Well, you do that in the following way:

If the null value, relative risk equals 1, risk difference equals 0, if that null value is not contained within the 95% Confidence Interval, then the data are telling us that with 95% confidence, the null value having no association whatsoever is not compatible with our data.

So if the data are not compatible with the null hypothesis, then the corresponding p-Value will be less than 0.05, it will be statistically significant. So if the null value is not contained within the Confidence Interval, the 95% Confidence Interval, then the finding is statistically significant at the T equals 0.05 level.

If the null value, the relative risk equals 1, risk different equals 0 is contained within the 95% Confidence Interval that means, the value representing no association, no difference, nothing going on is compatible with our data 95% of the time. Therefore, the data are compatible with the null hypothesis, the corresponding p-Value is greater than or equal to 0.05; there is no statistically significant association between the exposure and the outcome in our data.

Slide 13: So now let's look at the Study number 2. Again, it had 7.5 as an observed relative risk, but the 95% Confidence Interval is between 7.2 and 8.3. The fact that that Confidence Interval is so narrow reflects the fact that Study number 2 is a much larger study than Study number 1, it has less variability.

And when I look at these numbers, I think very much about playing roulette, going to a casino and playing roulette. You have to pick a number to put your money on. And if we wanted to pick one number to represent the association between postmenopausal hormones and endometrial cancer, we would pick our observed relative risk, 7.5, that’s the number we would play.

But in addition you have to decide how much money you are going to play on that number. And in the first case where the relative risk is compatible between 1.1 and 32.1, I would still put my money on 7.5, but I sure wouldn’t bet very much money, because there are so many other numbers that were compatible with the data.

On the other hand in Study number 2, I would go ahead and play the money on 7.5, but the data are compatible with alternatives that are only between 7.2 and 8.3. Therefore, I would put more money on that bet, because I would be more confident that 7.5 is a good representation of the true relationship between hormones and endometrial cancer.

Slide 14: Now there is another time that Confidence Intervals are very important. In our first example we did it in terms of precision. In both cases in the first example, we showed that there was a relationship between hormones and endometrial cancer that was unlikely to be due to chance P less than 0.05.

So our question at that point was just is 7.5 a good estimate, a valid estimate of the relationship between the exposure and the outcome? The second example is what is going to be when we do not have a statistically significant relationship. So let's look now at postmenopausal hormones and breast cancer with a relative risk in Study 1 and Study 2 were 1.13 that women who use postmenopausal hormones had a 13% increased risk of developing breast cancer.

And in neither case, Study 1 or Study 2 was this finding statistically significant. P was greater than or equal to 0.5, chance could not be ruled out as an explanation of the findings, there was no statistically significant association between the exposure and the disease.

But does this mean that there is no association between postmenopausal hormones and breast cancer? We can see that it was not statistically significant. That chance could not be ruled out as an explanation of the findings, but does that mean there is no association or just the sample size was not big enough to detect an association statistically, even if one were actually there. The Confidence Interval will help us distinguish between these two alternative explanations.

Slide 15: In Study 1, the relative risk is 1.13, P is greater than or equal to 0.05 and the 95% Confidence Interval was between 0.2 and 13. Look at the width of that Confidence Interval. The true relative risk is compatible with a benefit on breast cancer. 0.2, women who use hormones have only 2/10 the risk or 80% less risk of developing breast cancer. It's also compatible with the relative risk of 1.0, because 1.0 is in that Confidence Interval, compatible with no association between the exposure and the outcome and compatible with a huge increased risk of breast cancer, 13 fold associated with the use of postmenopausal hormones.

So basically, if you think of a roulette wheel again, our observed relative risk, we put our money on 1.13, but I have no idea whether 1.13 is accurate or not, because the data are compatible with the benefit, no association and an increased risk, those are what are called a null result. No statistically significant finding that is uninformative; it doesn’t help us sort out what's actually going on in the relationship between the exposure and the outcome.

And remember we said, if the H0 relative risk equals 1 and null hypothesis is contained in a Confidence Interval and 1.0 is contained in the Confidence Interval, then by definition the association is not statistically significant at the 05 level and the P will be greater than 0.05, which it is.

Slide 16: Now how about the second study? Relative risk of 1.13, again P greater than or equal to 0.05, but look at the Confidence Interval and see how narrow it is between 0.96 and 1.2. Now again, we can't

completely figure out whether there is no association; 1.0 or a 20% increased 1.2, but we have narrowed down the option.

We know that at most it's only a 20% increased risk and that really narrows us down and focuses on the magnitude of the effect that we might be having. And then that will allow us to sit down and talk to a woman, look at her breast cancer risk profile, decide whether a 20% at most increased risk would be acceptable to her, we know where we are in terms of the magnitude of the effect. So look at the difference between those two Confidence Intervals.

The first one null result, but really uninformative in terms of advancing our knowledge or telling us how to guide people. The second one again a null result, not statistically significant, but one in which it is informative in terms of narrowing down the magnitude of the association for us.

Slide 17: So when does the Confidence Interval narrow enough? Well, that very much will depend on the question that's being asked.

Slide 18: So for endometrial cancer, Study 1 and Study 2, with the two relative risks of 7.5, both p-Values less than 0.05, and with the Confidence Interval that we saw, if all we wanted to know in our question is, is there something going on? Is there an association? Is the observed association unlikely to be due to chance? Both Study 1 and Study 2 would have answered that question and said yes. Both are p-Value less than 0.05, chances unlikely explanation for the findings, both of them say there is a statistically significant association between the exposure and the outcomes, both would be informative to us.

But if we go one step further and say, how sure we are about the precision of the observed magnitude of the association, Study 1; we are not very sure about it, 7.5 it's uninformative in terms of telling us the precision of the magnitude of the effect. But Study 2 with that type Confidence Interval gives us a much more idea that 7.5 is a precise estimate, it is much more informative about reassuring us that 7.5 is a good estimate of the data.

Slide 19: And for the second study on postmenopausal hormones and breast cancer, again, we have our two studies and if the question was only is there is something going on, is the observed association unlikely to be the chance, then both studies would actually tell you no, it doesn't look like anything is going on, both are null studies, not statistically significant at the 05 level, chance cannot be ruled out as an explanation for the findings, but then the more important question is does that mean there is truly no association or was this due to an inadequate sample size? And Study 1, again with that wide Confidence Interval isn't completely uninformative, in terms of separating out what caused the non-significant finding, sample size or no association.

But Study 2 is much more informative. You can actually tell the magnitude of effect and you know that the fact that it was not statistically significant, means the sample size wasn't big enough, but at most we are talking about a small magnitude of effect, so people might not even choose to repeat the study in that particular example, because they can move forward and act on the data.

Slide 20: So the evaluation of the role of chance really involves three steps epidemiologically. An estimation of the magnitude of the effect or the association, such as relative risk or risk difference, doing hypothesis testing to see whether the association is due to chance, is this a reasonable alternative explanation chance, calculating the p-Value, look at the probability that the observed association or one

more extreme is due to chance alone, given that there is truly no association between the exposure and the disease.

In other words with the null hypothesis it is true, but then finally at this third step, an estimation of the precision of the effect measure, a calculation of the Confidence Interval, or the range of values within which the true relative risk lies, with a specified degree of confidence.

Slide 21: So now I think we have a very, very good idea, how bigger role sample size is going to play in terms of the interpretation of our findings. So let's talk about that just a little more in terms of power and sample size.

Slide 22: For just a second let's go back and think about now where we are in terms of hypothesis testing and the truth. We have used both of those words in the first part of this lecture. So if we do a 2x2 table and we put the tested significance on one axis and the truth on the other axis, the tested significance, there are two things we could have decided, we could have not rejected the null hypothesis that there was not a statistically significant association or we could have rejected the null hypothesis and said there is a statistically significant association.

And the truth is, not what we observe at all, but the truth in the world is that; either the null hypothesis is true and there is nothing going on, no association between the exposure and the disease or the alternative hypothesis is true, there is an association, there is something going on, there is a difference.

Now twice, we got is exactly right. So if we concluded that we do not reject the null hypothesis, we say it's not statistically significant our test, and in fact, the null hypothesis is true, there really is no association between the exposure and the disease, then we get it completely correct. The null hypothesis is true and we did not reject the null hypothesis.

In the other corner of the table, down at the bottom, we're also correct. If we rejected the null hypothesis and said there was a statistically significant finding and we should have rejected the null hypothesis, because alternative hypothesis is true.

So in that case where the alternative hypothesis is true and we rejected the null hypothesis, again, we're correct, we did it exactly the way we should have. But there are two times that we actually can make an error and one of the errors is in the bottom row of the table.

It is when we rejected the null hypothesis and said we have statistically significant finding, and we were wrong. There is nothing going on. The null hypothesis is true, nothing is different between the group, but we rejected the null hypothesis and said that there was a difference.

That is the Type 1 or Alpha Error, but more relevant to us, it is the p-Value. So it is the 1 out of 20 times that we accept that we are going to reject the null hypothesis, even when in fact we shouldn't have, when there is no association between the exposure and the outcome. And that is an error 1 out of 20 times, up to 1 out of 20 times, we are willing to make that kind of a mistake, say there is something going on when there really isn't.

But there is another kind of error that in fact in many ways is worse, and that's in the first row of the table where we do not reject the null hypothesis. We say, not statistically significant, no association, and there is an association and we have missed it.

So it is the error that comes from knowing that the alternative hypothesis is true, but we did not reject the null hypothesis, we missed it. And that's called the Type 2 or Beta error. And related to that is something called Power.

Slide 23: And Power is one minus (1-) the type two error. The power of a study is the statistical ability to detect the difference, find an association when one is truly there.

It is the probability of rejecting the null hypothesis, when in fact the alternative hypothesis is true, when we should have rejected the null hypothesis. And just as we conventionally test at the 05 level for the Type 1 error, the minimum acceptable power is conventionally 80%, meaning that the Type 2 error is accepted at 20% minimum. We'd like to do better than a power of 80%, but we certainly have to have at least a power of 80% when we're starting out our study.

So we can actually calculate the sample size that would achieve 80% power to detect a postulated effect, or the other way around, we can calculate the power that could be achieved to detect and affect, given a fixed sample size. So there are two different ways to look at it. We can sit down and say, you'd like to show a difference of the certain size between two groups.

How big a sample size would you need to detect that size difference with 80% power testing at the 05 level? On the other hand you could come back to a statistician or epidemiologist and say, you know in my clinic or in my environment; I can really only come up with probably a couple hundred people, what can I do for a couple hundred people? And then we would calculate the power that could be achieved with the sample size that you can actually get.

Slide 24: So, I am showing a formula now in this slide not so that you will learn it or we're not going to go through it in detail, but I want to point out to you what components go into that, that if you were sitting down with a biostatistician or an epidemiologist, what questions would we ask you for us to be able to calculate the sample size in your study?

So let's say this was going to be a case control study looking at oral contraceptive use and myocardial infarction. And the first thing I would say to you is, all right, among all women in your population, what percentage of those who do not have a heart attack or do not have cancer, whatever it is that you're looking at, what percentage of them use the oral contraceptive pill for birth control?

And you'd say well, I think about 10% of them do. And I'd say great! So I now know the proportion of the exposure among the control, the people who do not have the outcome in your study.

Now I just need one more thing from you. Tell me, what kind of magnitude of difference you expect to see between those who do have a heart attack and those who do have cancer and took the pill, as opposed to those who didn't? And you'd say one of two things. You'd say, first, I really have no idea what that magnitude of effect is. If I knew all that I wouldn't be doing the study.

And I go, I understand, but you do have to come up with some amount of a difference that you are going to look for. So, did you read the literature, has anybody ever done this before and you'd say, yes, usually they're looking for about 50% difference between the group.

And I'd say okay, well, if that's what they found, we can power the study to look for that. But you also might answer; you know I have no idea. I am the first person to ever look at this question before. And

then we'd say, all right, then from a public health or a clinical standpoint, what magnitude of difference would convince your colleagues that this is an important problem, what is clinically meaningful in the field for us to detect?

So we can get it one of two ways, but we do have to estimate it in some way, before we've ever done the study, before we have any idea what we are going to show, we have to be able to estimate what that difference is to be able to come up with a sample size.

So now we have the proportion of the exposure in a general population, the proportion of exposure we estimate that will be among those who are sick. We are going to test at the 05 level for a p-Value and we're going to do an 80% power.

And what you will get back from your biostatistician is a table like the following:

Slide 25: Where basically for every magnitude of effect that you might be interested in looking at, the sample size will be given to you that you would need to detect that magnitude of effect.

And the first thing that you're going to be able to see from this table is that the smaller the difference that you want to see between your groups, the bigger the sample size that you are going to need. And please note that these sample sizes are the required sample size in each group.

So for example, if you are interested in looking at a 20% difference between the groups, relative risk of 1.2, you would need 3834 people in each group. If on the other hand you are only interested in looking at a twofold difference, relative risk of two, you would only need 196 people in the group. And we would basically go back and forth on this. And may be at the very beginning you would have said, I want to see a relative risk of 1.2 and I would say, then you need 3834 in each group and you said I'll never get that sample size.

So when we go back and forth and you would say all right, let's just go for a relative risk of two. And I would ask you, but is two reasonable, has anybody ever seen a relative risk that big for oral contraceptives in an outcome? And you might say, no they really haven't, there is no way that's what the finding is going to be.

Then even though the sample size might fit your needs, it's not going to scientifically fit your needs to be looking for a relative risk estimate of that size. But everything will be given to you so that you can play off the scientific question you're trying to answer against the logistic issues of trying to come up with a sample of that size.

Slide 26: Often though, you might come back and say, you know this is all very interesting. Now, you could also come back and say, you know this is all very interesting, but when I really look into my clinic or my population, it will not be possible for me to get more than 100 cases of women who have a myocardial infarction or cancer among women of childbearing age and 100 women of that age who did not have that outcome. So we in fact can take that formula where we calculated sample size and we can solve now for power rather than sample size.

So you see on this slide now the same formula expect for it being sample size equals, it's power equals and we will put into that formula that you can do 100 cases and 100 controls and we will figure out what your power would be for different magnitudes of the relative risk that you could anticipate seeing.

Slide 27: And so on this next slide, you can see that the relative risk again go from 1.2 to 3. In the last column it will give you the power and for 100 cases of myocardial infarction among women of childbearing age and 100 controls, you would need to want to detect a relative risk of 2.5 before you reached 80% power.

Slide 28: And so now this is beginning to get it back and forth between sample size and power, between the epidemiologist and the statistician. And this really will be a reality check for you as to whether a study can be realistically achieved as proposed. Because you might have come back at that point and said 2.5 is not a relative risk that is likely at all to be seen in this study.

And we will come back to you and say, all right, you just need to get a bigger sample size. Do you have any other colleagues in your community that you could work with? Maybe they could get cases and controls from their hospital also.

And if you say, no, this is 100 cases, 100 controls, this is all I can do in my community, then we could say are there any colleagues you have in other countries and do a multi-site study, so not just you, but other hospitals also. And you could say, yes; that sounds like a good way of doing it or you could say no, I really have to do it in my own community or not at all. And then we're going to say, then given that you are going to spend years of your life and a lot of money and not have adequate power to show the finding that you want to show, it might be better if you looked at other endpoint, that is in a rare one like myocardial infarction.

So when you are designing a study and writing the grant, you calculate power. At the end when you are interpreting the study, you evaluate Confidence Intervals. Because power is theoretical, power is based on what we think is going to happen in the study and theoretical power is no longer relevant, when the study is over and you've actually seen what your observed values are.

So at that point when you have the observed results, you actually do the Confidence Interval to look at the range within which the true relative risk will lie. And always remember that sample size is based not on a number of people, but a number of endpoints. So you can do a study of breast cancer in a million young women who are under the age of 20 and your sample size is very large, but the number of endpoints of breast cancer among women in that age is going to be too small for you to be able to compare it among women who use the pill or women who are getting used to pill, women who had a one particular lifestyle, versus another kind of lifestyle, number of endpoints.

And also remember that you can calculate a sample size that will be adequate for your main endpoint. Let's say you're going to use a new treatment to see if it prevents a recurrence of the disease, that's going to be your main endpoint and you calculate the sample size and power could do that.

But what if there are other endpoints you also want to look at, and they are rarer endpoints than recurrence is, like a side effect. A rare event will not be able to have adequate power with that sample size to pick it up.

So if you really care about these other rarer endpoints, not just your primary outcome, you're going to need the power of your study, you are going to need to have an adequate sample size to look at them also.

Slide 29: Thank you slide