David Murray:
Hello, my name is David Murray. I'm the NIH Associate Director for Prevention, and I direct the Office of Disease Prevention here at NIH. I want to welcome you to part two of our course on Pragmatic and Group Randomized Trials in Public Health and Medicine. Part two focuses on designing the trial. This is part two of a seven-part course. It's self-paced and provided online, free of charge by NIH. We provide the slides that go with these of each of the modules, a complete set of readings for the course, and guided activities for each of the modules.

The target audience for the course includes faculty and postdoctoral fellows and graduate students who are interested in learning more about the design and analysis of group randomized trials. The target audience also includes program directors, program officers, and scientific review staff here at NIH who are interested in learning more about these designs.

Participants should be familiar with the design and analysis of individually randomized trials. We're not going to dwell on those issues in this course, but we will build on your knowledge of those issues. Participants should be familiar with the concepts of internal and statistical validity, their threats and their defenses. You should also be familiar with linear regression analysis of variants and covariants, and logistic regression.

At the end of the course you should be able to discuss the distinguishing characteristics of group randomized trials, and their close relative, individually randomized group treatment trials, and distinguish them from individually randomized trials. We want you to be able to talk about appropriate uses in public health and medicine. And for group randomized and individually randomized group treatment trials, you should be able to discuss the major threats to internal validity, the major threats to statistical validity, the strengths and weaknesses of design alternatives and analytic alternatives, and to perform sample size calculations, at least for a simple group randomized trial. You'll also be able to discuss the advantages and disadvantages of alternatives that have been offered for the evaluation of multilevel interventions.

The organization of the course is shown here. Today we're going to cover part two, designing the trial, and subsequent sessions focus on analysis, power and sample size, examples, the state of the practice, and alternative designs. We're going to start today by talking about planning the trial, and I always start by focusing the audience on the research question. The driving force in planning any experiment needs to be the research question. That certainly is true for randomized clinical trials. It's true for group randomized trials, and every other kind of trial that I can imagine.

The question identifies the target population, the setting, the endpoints, the intervention. Those factors shape the design and the analytic plan. So the research question is terribly important. The primary criteria for choosing a question, in my mind, are shown here. Is it important to do it? Will the trial address an important public health question? Will the results advance the science, the state of the science, advance the field? The other question that is important to consider is, is this the right time to do it? Even if you have an important question, if you don't have enough preliminary evidence of feasibility and efficacy for the intervention, it may not be time to launch the big trial, or if you don't have good estimates of the parameters that you need to size the study, to conduct power calculations, it may not be the right time to do the study. So these are all important issues.

The investigators need to keep these questions in mind. When I was a professor in academic institutions, I used to tell my graduate students to create a screensaver that was based on their research question so that they would be reminded of it from time to time, and I can share that advice with other investigators, as well.

Let's talk a little bit about the fundamentals of research design. The goal in any comparative trial is to have a valid inference that the intervention, as you delivered it, caused the result as you have measured it, and we need at least these three elements. We need control observations. We need to have as little bias in the estimate of the intervention effect as possible, and we need to have enough precision for that estimate to have adequate power to detect an effect.

The three most important tools to limit bias and improve precision in any trial, including a group randomized trial, is shown here. The most important thing that we can do is randomize. I often say that the three most important things are randomization, randomization, and randomization. So you, I hope, get the sense of how important that one is.

Replication refers to how many groups we include in the trial and how many members we include in the trial. That's also very important, and then we need to take whatever steps we can to reduce variance, so as to improve the precision and the power of the trial, and have a better chance of answering the question.

I'm going to identify four primary threats in a group randomized trial, and I do this for two reasons: one, so that you are aware of them and think about them, as you're designing your trial. The other reason is that every time I wrote a grant application to send to NIH that proposed a group randomized trial I included a section at the end of the application on threats to the validity of the trial and the solutions that we have used in designing the setting, and I talk about these four issues.

The first one is selection. Selection refers to pre-existing differences between the study conditions that are associated with the groups or members that are nested within the conditions. So these are differences that are there after randomization, but it's not something that I caused. It's not something that the intervention caused. It's just differences that exist, and if we don't deal with them adequately, they could fool us into thinking that there is no intervention effect when there really is one or, perhaps, there is an intervention effect, and we're not seeing it because of these kinds of problems.

Another major threat to internal validity is differential history. Differential history is any external influence, other than the intervention, that might affect the outcome, and that affects one condition more than the other. So maybe something happened in some of the groups in the intervention arm, and the outcome is changed in those groups, as a result, and it looks like there's an intervention effect. Well, that might be due to that external influence, rather than the intervention that we introduced. So that's an example of differential history.

Differential maturation is the third major threat to internal validity in a group randomized trial. It reflects growth or development at the group or member level that can affect the outcome, and that affects one condition more than the other. We worry about maturation in randomized clinical trials, especially if we're dealing with young people, children, adolescents who are still growing, and we're looking at outcomes like blood pressure and height and weight because those things change as the children grow.

In group randomized trials, the groups that we're working with, such as hospitals or clinics or communities or schools, may also be maturing, or at least changing -- not always necessarily maturing -- but growing, developing over time. And they may be getting worse on some problem or better on some problem on their own, quite apart from intervention. And if that happens more in one condition than the other, that could either mask an intervention effect or create an intervention effect. So differential maturation is another one that we need to worry about.

Contamination is the fourth major threat, and it exists when important components of the intervention or activities like the intervention find their way into the control condition, either directly or indirectly. I was involved in a study called the Minnesota Heart Health Program at the University of Minnesota many years ago, and we thought that we had isolated our communities well enough geographically that we didn't have to worry too much about contamination. They didn't share media markets. They were hundreds of miles apart. We thought we had things well under control, and then after some time discovered that the person that we had hired in one of our intervention communities to run the intervention program had a brother who lived in one of the control communities, and was the CEO for the hospital in that community. And they, of course, met regularly and interacted regularly, and so all the information about our intervention was finding its way, at least through that vehicle, into one of our control sites. So contamination can happen in randomized clinical trials, certainly in group randomized trials, and we need to be concerned about it.

These are the major strategies to limit these threats. Randomization is the most important one. It tends to balance things evenly. But in group randomized trials we often don't have large numbers of units or groups to randomize, and so the second bullet is terribly important; matching stratification or constrained randomization a priori, in advance of randomization. These are techniques that can be used to help ensure that the intervention condition is similar, at baseline, to the control condition, in terms of the groups that have been allocated to each of those conditions or those two arms. So we're matching or stratifying or constraining the randomization on groups, in group randomized trials, on members in the digitally randomized group treatment trials.

Another strategy that we can use is to use objective measures, and not rely on self-report data. This is something that I recommend strongly. We rely on self-report data only when we have to and try as much as possible to use objective measures. We also try to separate the evaluation staff from the intervention staff, and only collect data from staff who are blind to the study conditions that each group has been assigned to. These studies are often done in the real world setting in communities in hospitals and clinics. It's very difficult to ensure blinding of all of the staff that are involved or the participants that are involved, but we certainly try to blind the evaluation personnel.

Another strategy to limit threats is to use analytic techniques. Where we think that there are imbalances on factors that might affect the outcome, we can adjust for those as covariants, using regression techniques in the analysis and, thereby, reduce the threat of selection bias. We can certainly avoid the pitfalls that invite specific threats to internal validity, and I've listed a few of them here. We avoid regression of the mean, for example, by using very reliable measures, and not allocating only values at the end of the distribution to treatment arms, but rather randomizing people from the entire distribution. We can avoid testing and differential testing by making sure that both conditions have the same schedule of testing, and that we're not doing more in one arm than another.

There are several threats, too, that I'll talk about in particular for the validity of the analysis. The first of them is mis-specifying the analytic model. The easiest way to do this, and it's quite common in group randomized trials and individually randomized group treatment trials, is to ignore a measurable source of random variation. In group randomized trials, we always have at least two members in groups, and the most common mistake is to ignore the group as a measurable source of random variation. Aand I promise you that you'll have an inflated type one error rate if you do that. So that's something that you want to avoid.

A second way of mis-specifying the analytic model is to misrepresent a measurable source of random variation. So if you're outcome is a 01 variable or a binary variable, you don't want to use methods that assume linear -- that fit a linear model and assume Gaussian data. You want to use something that accommodates logistic regression, in some form, or some other technique that will work. So you need to pay attention to the measurement scale that's used for the outcome variable in the study.

You don't want to misrepresent the pattern of overtime correlation in the data either. If you only have pretest/posttest design, there can be only one correlation. You can't get it wrong. But if you have pretest/posttest one-year follow-up, two-year follow up and so forth, you can have multiple correlations across those different intervals. They're probably not all the same, but some of the standard techniques that we use in analysis assume that they're the same, and that's a way of misrepresenting the pattern of overtime correlation in the data.

The other major threat to the validity of the analysis is low-power. So this can come from having weak interventions that just aren't very effective. It can come from not having enough groups or enough time intervals in the design, so that I don't have enough power. It can come from having a large intraclass correlation for the dependent variable, or a very noisy dependent variable, which gives me a high variance. These are also things that we want to avoid. And, finally, we can have poor power if the intervention is not delivered reliably across the groups that are randomized to the intervention condition.

Strategies to protect the validity of the analysis are shown here. We avoid model misspecification by thinking about the analysis at the same time that we're planning the study. You don't want to get to the time to do the data analysis and discover that there isn't a good analysis available for you. So you need to plan that with the design. You need to plan the analysis around the primary endpoint. If it's a 01 variable, be thinking about logistic methods; not linear methods.

You want to anticipate all sources of random variation. Easy to say, but this is probably the hardest part to do in group randomized trials, and you're simply going to have to work with people that have more experience at it, and read some of the literature on the methods for design and analysis of group randomized trials to understand how this works. I've given you one big clue. If you've got any nested factors, they are sources of random variation. Any nested factor should be modeled as a random effect, is a source of random variation, and should be accounted for in your analysis plan.

It's also important, as I said, to anticipate patterns of overtime correlation, and set your analysis plan up to allow for that, if you think it's not going to be constant. If the study is not very long, the correlations may very well be very similar, across the intervals. But if you've got a study that spans five years or eight years or 10 years, I can promise you that the correlations won't be stable over time, and so you need to anticipate what those patterns are.

You might consider alternate models for time. So if you have multiple follow-up measures, you can model time continuously. It doesn't have to be modeled at as a linear function. It could be a non-linear function. You could use spline functions. There are different ways of doing it, so that's something to think about. You always want to think about potential confounding and potential effect modification when you're looking at these studies. Not just group randomized trials, but randomized clinical trials as well.

Things to do to avoid low-power. Well, first, I turned to my colleagues who create the intervention and I say, "Give me a strong intervention. Give me one that works. Give me one that gets to the participants and has an impact on them; affects their behavior or their risk factor or their disease outcome." So that's terribly important. Make sure that the intervention is being delivered reliably. Make sure that you've got observations being collected on the reliability of the intervention implementation, and intervene as necessary to make sure that it is being delivered consistently.

The most important thing that you can do, as a design or analysis person, is to make sure that you've got enough groups randomized to study arms to have adequate power. Having a lot of groups, even if they're small, is much better than having just a few groups, even if they're very large. So the Minnesota Heart Health Program that I mentioned a few minutes ago made the classic mistake of having six very large groups. Of course, we designed that studied in the late 70s, and began it in the early 80s, and we didn't know very much about group randomized trials in those days. So, hopefully, we can be forgiven for that mistake. It would have been far smarter to have had a lot more communities that we were randomizing, with collecting much less data from each of the communities. In parallel, you can employ more and smaller surveys or continuous surveillance instead of just a few large surveys, and that lets you model time in greater detail. That's particularly important if you think you're going to have an intervention effect that's going to develop gradually or fade over time, after you pull out of the study.

And then the last point here is employ regression adjustment. I mention that under protecting the validity of the design because it's a way to correct for potential compounding. It's also a great way to reduce variance and intraclass correlation. We can also reduce intraclass correlations by 50, 60, 70 percent by adjusting for the right covariates. So that's an important technique that can very much help power.

I introduce a little notation, and this is a notation for my book published in 1998. I use "Y" to refer to the dependent variable. That's pretty standard. I use "condition" to refer to treatment in control or to the study arms. So some people use "study arms." Some people use "group" to refer to treatment and control. I use "condition" because I reserve "group" for the unit of assignment. I use "time" to refer to the measurement occasions; pretest, posttest, follow-up, and so forth. I use "group" to identify the unit of assignment, and "member" to identify the unit of observation. I refer to covariates as variables that are usually nuisance variables, but they might be confounders, and I want to adjust for them in my analysis. So I refer to them generally as covariates. I distinguish between random effects using bold type and fixed effects using plain type. In group randomized trials we always had at least two random effects groups and members there, indicated in bold.

Factors that can reduce the precision of the analysis of intervention effects in a group randomized trial are identified here, and I do that by showing you the formula for the variance of a condition mean. So sigma squared Y bar sub C refers to the variance of the condition mean. If it weren't a group randomized trial and I just had M is G observations contributing to the mean, it would be sigma squared Y over MG and the part in parentheses wouldn't be there. But I have a group randomized trial, and so I do have the parenthetical component, the design effect or variance inflation factor, and that's one plus and minus one times the ICC.

This equation can be used very directly in sample size calculations for simple group randomized trials. It needs to be adapted for more complicated analysis techniques, and more complicated designs, but it still can convey the most important factors in determining power in a group randomized trial. And if the goal is to minimize the variants of the condition mean, then you're looking at the right side of the equation and thinking, "Okay, what can I do to make the left side of the equation as small as possible?"

We do that by increasing the number of groups, increasing the replication of groups, since G appears only in the denominator. We can do that by increasing the number of members in each group because M is also in the denominator, but we don't get as much out of adding members as we do adding groups because M is also an enumerator, and that is in the one plus M minus one times ICC parenthetical component. So we get a certain amount out of adding groups up to a point. We get a lot out of adding members up to a point. We get a lot of you get a lot out of adding groups. Sigma squared Y is an enumerator so if we make that smaller, that will help us. The intraclass correlation is also an enumerator. If we make that smaller, that will help us.

The next three slides give you a sense of the power in group randomized trials as a function of the intraclass correlation, the number of groups per condition, and the number of members per group. The Y axis is the detectable difference and in standard deviation units, so standardized units, and the lines that are shown in the figure reflect the number of groups per condition. So the top line in red is for two groups per condition. This is as small a study as possible, and have any variation within a condition. You'll notice that that line is very high on the Y axis and never goes very low.

In other words, if you have and intraclass correlation that's 0.1, which is rather large, and you only have two groups per condition, you better have a very big intervention effect; 1.75 standard deviation units. When I'm giving this presentation live, I usually ask the audience how many of you have ever had an intervention effect that large? No one ever raises their hand because it doesn't happen.

In public health and medicine, we're usually looking at intervention affects that are 0.25 to 0.5 standard deviation units. So that's down toward the bottom of the Y axis, and in this slide, you'd get there with 32 or 16 groups per condition, but not really with the other level. So you need a pretty good size study if the intraclass correlation is as high as 0.1.

Notice also that the lines don't drop much after you get about 50, 75, 100 members per group. They're pretty flat. That's more true with large intraclass correlations than small intraclass correlations, but it is generally true in group randomized trials. We get a lot more out of adding groups than we do out of adding members.

If the intraclass correlation shrinks by an order of magnitude to 0.01, all the lines drop. That's good news. Now we can find intervention effects in the range that we're interested in, even if we have eight groups per arm, maybe even four groups per arm. So it's -- a lot more is possible if the intraclass correlation is 0.1. Two per arm is still not very good. So generally don't recommend that.

If we shrink the intraclass correlation by another order of magnitude, the lines drop a little bit more; not as much, and you're starting to see the law of diminishing returns at work. If we dropped it another order of magnitude, the lines would hardly move. So once you get the intraclass correlation down to this level, .001 or so, there's not a great deal of benefit out of getting it to be much smaller. In this case, adding members has a little bit more of an impact than it did when the ICC was larger, so we're still gaining precision going from 100 to 150; even going from 150 to 200, where previously we got very little out of that.

I want to talk about different kinds of designs that are used in group randomized trials, and these are the topics that we're going to cover in the next series of slides. We'll start by talking about single factor and factorial designs. Most group randomized trials involve only one treatment factor. I refer to that as condition. Most have just two levels: intervention versus control or treatment versus control. Most group randomized trials cross condition with time. So the most common design is a pretest/posttest, to group design or to condition design where I have intervention and control, and I have pretest data in both conditions. I have posttest data in both conditions.

We can have that kind of design with either a cohort design or a cross-sectional design, and we have referred to these as nested cohort designs or nested cross-sectional designs. In the cross-sectional version, you're measuring different people at each time point in each group. In the cohort version, you're measuring the same people at each time point of each group. Some group randomized trials include stratification factors. For example, both multi-center trials generally cross condition with field center. The tag trial, trial activity in adolescent girls funded by NHLBI that I was involved in had six field centers. Each of the field centers created six schools, and within each field center, three of those schools received treatment. Three were control schools. So that would be an example of multi-center trials with stratification on field center. Single center trials often stratify on factors that are related to the outcome, or to the ease of implementation of the intervention.

Time is often a factor. Posttest designs are uncommon, but we do see them. Pretest/posttest designs are the most common. Extended designs occur when we have additional pretest rounds of measurement or additional posttest rounds of measurement. Continuous surveillance would be the extreme case where we're just measuring continuously with some kind of a surveillance operation.

As I mentioned, the nested cohort design measures the same sample of participants at each time point. This is the design that's preferred if your research question involves change in specific individuals, or if you're interested in mediation analysis. The cross-sectional design involves different participants that are measured at each of the time points, and this design is preferred if your research question involves change in an entire population.

Cross-sectional and cohort designs are often compared to one another, and they have various strengths and weaknesses. In cross-sectional designs, I have to worry about people moving into the schools if I've randomized schools, and students moving out of the schools, over the course of the two or three years that I might be doing the study. In a cohort design, I have to worry about students dropping out. I don't particularly worry about students moving in so much, but I do worry about students dropping out. In other kinds of cohort studies, I worry about people actually dying, so mortality taken quite literally.

Cross-sectional designs are good for looking at group change or population change. Cohort designs are good for looking at individual change or mediation. With cross-sectional designs, I have to recruit new members at each time point, so I may spend more on recruitment. In cohort studies, I have to follow the same people, so I have to track them. And I may spend money on tracking and follow-up, and that can get to be very expensive, especially with long-term follow-up.

People argue about whether cross-sectional designs are more powerful or cohort designs are more powerful. I think most people in the audience probably assume that cohort designs are more powerful. That is going to be true if the outcome variable has a high level of reliability over time -- something like smoking status or blood pressure or weight -- but it's not going to be true if my dependent variable is not very reliable over time, that is, changes over time. So something like dietary intake or physical activity, which varies dramatically from day to day, within a person, has a very low overtime correlation within persons, and so the cohort design may actually be more powerful for those kinds of outcomes than -- sorry, the cross-sectional design may be more powerful for those outcomes than a cohort design.

There's also a debate about whether you can -- whether you have to have a cohort design in order to look at the effect of a full dose. You can actually do that as well in a cross-sectional design if you select the sample to include people who have been there during the full-time that the intervention was being delivered. So you can do both of those.

A priori matching and stratification: I want to talk about because it's very important in group randomized trials. You can use either if you want to ensure balance on important, potential confounders or potential sources of bias. A priori stratification is preferred if you're interested in differential effects across the levels of the stratification factor. So if you think that the intervention may be more effective in large schools than small schools, and you want to make that comparison, you want to make sure that you have the same number of large schools and small schools in the two conditions, and so you could stratify on school size, as an example. A priori matching is useful if the matching factors are very well correlated with the primary endpoint, and can improve power more than a priori stratification can.

The choice, then, often depends on the number of groups that you've got available on the expected correlation. If you don't have very many groups that you're going to be randomizing, you lose a lot of degrees of freedom with matching. You don't lose as many with stratification, and so stratification may be more attractive. If you have lots of groups or the correlation is -- the matching correlation is expected to be high, you may get a better power benefit from matching than stratification. I draw your attention to a paper by Allan Donner published in 2007, in which he identified problems with match designs with large, intraclass correlations, and small group sizes. That doesn't apply to all group randomized trials, but it certainly applies to some, and if you're in that situation, stratification is recommended instead of matching.

I want to talk a little bit about constrained randomization because sometimes stratification and matching are difficult to implement. That's particularly true if you have a lot of factors that you want to match or stratify on, and you only have a limited number of groups. So a paper published 15 years ago by Robin Butcher on constrained randomization offered another approach that could solve the problem. "Generate all possible allocations of groups to study conditions. Then identify those that are sufficiently well-balanced on the factors that you're concerned about between the two conditions or across the several conditions, if you have more than two. And then, from within that constrained set of allocations schemes, choose one allocation scheme that you could actually use in the study, and then apply that one." That would be an example of constrained randomization.

And some recent work has focused on finding a best balanced metric that can be used in this context, and I refer you to a paper by De Hoop et al. 2012. "Constrained randomization can be very useful if you want to balance on more than just a couple of things," and recent work has also shown, as we'll talk about in module three, but I'll give you a preview here. "Recent work has shown that constrained randomization can improve power in group randomized trials."

Just a few comments about post hoc stratification: with a priori stratification, we're defining the strata in advance, and then we randomized from within those strata. With post hoc stratification the strata are defined post hoc, and the stratification factor is added to the analysis. That necessarily limits the kinds of things that we can apply post hoc stratification to because these need to be things that are general in measure at an individual or member level; so things like gender, age, race, ethnic group, things like that. If it's a group factor, we may not be able to stratify post hoc at all, and analytically there are big differences between a priori and post hoc stratification, particularly in group randomized trials. We'll talk about those issues in the next module.

To summarize what we've covered today, all of the design features that are common to randomized clinical trials are available in group randomized trials and individually randomized group treatment trials, with the extra complication that you've got another level of nesting. You've got nested cohort, nested cross-sectional designs. So cohort and cross-sectional -- you can do both. You can have pretest or posttest only designs. You can have pre/post designs. You can have extended designs. So time can be included as a factor, with lots of levels. You can have a single factor. That's the most common. We sometimes see factorial designs; less common. I strongly recommend a priori matching or stratification, or constrained randomization as a way to balance, especially if you've got a limited number of groups that you're going to randomize. And we've identified the primary threats to internal and statistical validity, and some of the defenses that you can use to limit the impact of those threats on your study. I encourage you to plan the study to reflect the nested design, to have sufficient power for a valid analysis, and to avoid the threats to internal validity.

Again, I want to thank you for joining us for part two of the course on pragmatic and group randomized trials in public health and medicine. I draw your attention to our website where you can watch the other modules in the series. You can also give us feedback, any comments that you want to share on this component or the other modules. You can download the slides that you've seen. You can download references for the entire course. You can download suggested activities, and pursue those on your own. You can certainly watch this module again, and you can view any of the other modules as you wish.

If you have any questions about group randomized trials or individually randomized group treatment trials or other issues that we've raised in this course, please send an e-mail to grt@mail@NIH.gov, and we'll get back to you and try to answer your question. Thanks very much, and I hope to see you soon for part three.

[end of transcript]