David Murray:
Hello, my name is David Murray. I'm the NIH Associate Director for Prevention and Director of the Office of Disease Prevention. I want to welcome you to part three of our course on pragmatic and group randomized trials in public health and medicine. Part three focuses on analysis approaches. This is part of the seven-part self-paced online course that's free and provided by NIH. We provide the slides for each of the parts, a set of readings for the entire course, and a set of guided activities for each module. The target audience for this course includes faculty, post-doctoral 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 officers at NIH who are interested in learning more about these issues.

Participants should be familiar with the design and analysis of individually randomized trials. They should also be familiar with the concepts of internal and statistical validity, their threats and their defenses. We want participants to be familiar with linear regression analysis of variance and covariance and logistic regression.

At the end of the course, participants will be able to discuss the distinguishing features of group randomized trials and individually randomized group treatment trials and how they differ from individually randomized trials. We want participants to be able to discuss their appropriate uses in public health and medicine, and for group randomized and individually randomized group treatment trials, to discuss the major threats to internal validity and their defenses. Also to statistical validity and their defenses, to discuss the strengths and weaknesses of design alternatives and analytic alternatives, and to perform simple calculations for group randomized trials in terms of sample size calculations. Participants will be able to discuss the advantages and disadvantages of alternative designs for group randomized trials for the evaluation of multi-level interventions.

The organization of the course is shown here; there are seven parts. Today we will discuss part three, analysis approaches. We'll start by considering the classification scheme for statistical models. This is defined in terms of the nature of the outcome -- the primary outcome for the study -- which might be Gaussian or normally distributed, or non-Gaussian, non-normally distributed. These are also classified based on the number of random effects that are allowed in the model. The upper left hand corner of this two by two table is the general linear model, characterized by normally distributed or Gaussian distributed data and one source of random variation or one random effect. The upper right hand corner is the generalized linear model, many of you will be familiar with logistic regression and Poissan regression, those are examples of the generalized linear model. Here we have non-Gaussian outcomes still one source of random variation.

The lower left hand cell in the two by two table is where we have normally or Gaussian distributed data, but now we allow two or more sources of random variation. This would be the general linear mix model. The lower right hand corner is the generalized version, so this would be

a mixed model logistic one, this mixed model Poissan regression. Here we have two or more sources of random variation and the outcome level, and sometimes other levels, are non-Gaussian in terms of their distribution. We distinguish between fixed effects where we're only interested in the levels that are used in the study, and random effects where we're interested in the general population of levels that the levels in our study represent.

In terms of preferred models for designs with one or two time intervals, this could be a posttest only design or a pretest design. I can recommend to you mixed model analysis of covariance and analysis of variance. This is group randomized trials or individually randomized trials where we always have two sources of random variation or more. So we're in the lower row of the table that we were looking at a minute ago. And if we have normally distributed data, we're in the lower left hand corner, non-normally distributed data, lower right hand corner.

The mixed model ANOVA and ANCOVA are extensions of the familiar ANOVA/ANCOVA based on the general linear model. The output looks very similar to what you would get if you've run the analysis based on the general linear model, but it's fit using the general linear mixed version or the generalized linear mixed version. These models accommodate regression adjustments for covariance which is very important. In group randomized trials and individually randomized group treatment trials, there is only one correlation possible in the data, at least over time, because at most we have a pretest/posttest design, so one interval between observations. So we can't misrepresent the overtime correlation with these models.

Using the mixed model ANOVA/ANCOVA can take several forms. We can look at the posttest data only using either ANOVA or ANCOVA we can look at an ANCOVA of posttest adjusting for pretest data. This is a very common analysis, especially where we adjust for the baseline measure of the outcome. Or we can conduct a repeated measures analysis of variance or covariance where we take both the pre and the post into account on the deep inner variable side of the equation. This is the second most common analysis applied to group randomized trials.

If we have more that two -- sorry -- simulations have shown that these methods have the nominal type I error rate across a wide range of conditions common to group randomized trials. Now if we have more than two time intervals, we need to change the model substantially. The mixed model ANOVA, mixed model ANCOVA is no longer appropriate. We need to use models which I often refer to as "random coefficients models." These are sometimes also called "growth curve models." Here the intervention effect is estimated as a difference in trends that are estimated for each condition. In fact, we estimate trends for each person if we have repeated measures for each group, whether we have repeated measures or not at the member level. And then average trends for each condition. And the intervention effect is estimated as the difference in the condition average or mean trends.

The trends can be linear, that's common. They don't have to be, though. They can be non-linear. The mixed model ANOVA or ANCOVA, the simpler

analysis that we talked about for one or two time intervals, assumes homogeneity of group-specific trends. That may not be true for the data that we're working with. And simulations have shown that that simpler mixed model ANOVA or ANCOVA has an inflated type I error rate if those trends are heterogeneous.

The random coefficients analysis allows heterogeneity in those trends, that's the strength of this approach, and that's why we recommend it for analyses where we're trying to model more than two time intervals in the analysis. It's perfectly possible to have three, four, five measurement occasions in a design if we're only including two, however, in the analysis such as the last one adjusting for base line, we can use the simpler mixed model ANOVA or ANCOVA. But if we want to model three or more of those time periods in the analysis, then we need to use the random coefficients approach. Simulations have shown that the random coefficients approach has the nominal type I error rate across a wide range of conditions common in group randomized trials.

What about some other approaches that might be taken for the analysis? I've just talked about model-based approaches and, specifically, the mixed models that are commonly used with group randomized trials. It's also okay to use randomization tests. With a randomization test, the intervention is effect -- or, quite generally -- the intervention effect in a group randomized trial is a function of unadjusted or adjusted group-specific means or slopes or other group level statistics; that's what we're examining with the model-based methods. Under the null hypothesis of no intervention effect, the actual arrangement of the group level statistics in the intervention condition on the control condition is one of many equally likely arrangements; it's the one that we got by chance.

With the randomization test we actually compute every possible arrangement and the intervention effect that we would get given every possible arrangement. The probability of giving a result more extreme than the one that we actually observed, is the proportion of effects that are greater than that observed once we've generated the entire distribution. There are no distributional assumptions required, and in the unadjusted form of this test, there are no other assumptions required. This is the beauty of the randomization test. It really is assumption free. And so many statisticians like it for that reason.

Mitch Gail published a lovely paper in 1996 in "Statistics and Medicine" in which he reported that randomization tests have the nominal type I error rate and good power across a variety of conditions common to group randomized trials. This is what makes this approach quite attractive. This is true even when the member level errors are not normal, even when very few heterogeneous groups are assigned to each condition -- whether the ICC is large or small -- as long as there's balance at the group level. And that means that we have the same number of groups in each condition. So as long as you hold on to that feature, the randomization test is quite strong and robust against other kinds of issues that may arise. Programs for randomization tests are available in print and on the web and so they're much easier to do than they were 20 or 30 years ago.

The major weakness of the randomization tests is that you really just get a point estimate and a p-value. You don't get standard errors; you don't get confidence intervals. The other weakness is that the unadjusted version doesn't provide any more protection against confounding than other unadjusted tests. Now we can certainly make adjustment for covariance in the first stage of what is really a two stage analysis with this randomization test. And if we do that, the regression adjustment requires many of the same assumptions that we have in the model-based tests and brings the randomization tests much closer to those model-based tests in terms of the assumptions that we're making.

The model-based methods, of course, do provide parameter estimate standard errors, confidence intervals, and they hold the nominal type I error rate, even if the member and group level errors are non-normal, unless they're very heavily skewed or heavy tailed. Even when there are a few heterogeneous groups assigned to each condition, whether the ICC is large or small, so long, again, as there has balance at the level of the group.

So the summarizing at least this first part of the presentation today, permutation tests, model-based tests are equally good when we're analyzing group randomized trials, whether we have a large or a small interclass correlation, whether we have mini or a few heterogeneous groups assigned to each condition. And even if the data are decidedly non-normal as long as they're not extremely non-normal, so long as there's balance at the level of the group.

What about a method like GEE, generalized estimating equations? The attraction of GEE is that it's robust against misspecification of the correlation structure or the random effects variance/covariance matrix. Methods-based like this use an empirical sandwich estimator for their standard errors, as I said, that estimator is asymptotically robust. Asymptotically means "given a large sample size." And here that means "given a large number of groups that are randomized to conditions." If we have a smaller study, fewer than 40 groups involved overall, the empirical sandwich estimator has the downward bias, and we can end up with an inflated type I error rate. Recent work provides corrections for this issue, and several have been incorporated in some of the common statistical packages, for example, proc glimmix and SAS beginning with version 9.1.3.

So methods that employ the corrected sandwich estimator may have very broad application in group randomized trials including small trails methods that use the standard empirical sandwich estimator should be reserved for circumstances when we have 40 or more groups in the trial. Recent work published just in the last few years suggests that these corrections do not work adequately when we have very small trials, so fewer than eight or 10 groups per condition. In that case, you're actually better off to use the model-based methods that we talked about earlier.

What about methods developed for analysis of complex survey samples? And there are a variety of those. These methods suffer from the same

problems that the GEE methods suffer from. If we don't have a good number of groups in each condition, we can get into difficulty with an inflated type I error rate. Many of these methods then are useful in large group randomized trials, but not when the number of groups per condition is limited.

What about fixed effects methods in two stages? These methods have been around for a long time. In fact, this was the original way to analyze data from a group randomized trial even before the term "group randomized trial" was created. This is commonly known as a "means analysis." It's very simple to do, very easy to explain. We essentially calculate means for each of the groups, whether adjusted or not, in a first stage. And then in a second stage, we apply a t-test to the means as the raw data points, with degrees of freedom based on the number of groups. If we have data that are balanced, we're going to get exactly the same result that we would get from a mixed-model analysis of variance or covariance so long as both the mixed model is properly implemented and the two stage analysis is properly implemented. If we have unbalanced data, we can still get the same result, we just have to wait the means in the second stage of the two stage analysis to reflect the uneven sample sizes in each of the groups.

This approach can be adapted to perform a random coefficients analysis. It can be adapted to very complex designs where one stage analyses simply aren't possible. And we see this approach used in a variety of very large trials including CATCH, the Minnesota Heart Health Program, REACT, the Community Youth Development Study, and the Trial for Activity in Adolescent Girls. Two stage methods can be very useful in group randomized trials so long as they're implemented correctly.

What about analysis by subgroups? So this is a situation where perhaps we only have a limited number of groups randomized to each arm or each condition and, as a result, we don't have many degrees of freedom at the group level. Some have been tempted to analyze the data at a subgroup level, perhaps classrooms instead of schools, or physicians instead of clinics. This approach rests on the very strong assumption that the subgroup captures all the variation attributable to the group. Analysts may be drawn to this because they're many more degrees of freedom available at a subgroup level. The problem is that this approach will have an inflated type I error rate even if the subgroup captures, say 80 percent of the group variation, in part because of the extra degrees of freedom that are provided, in part because the subgroup is not capturing all of the variation associated with the group. So this approach is something that we don't recommend.

What about deleting the unit of assignment from the model if it's not significant? This is an approach that we see in a number of papers, even these days. The analysts will examine the data using a correct approach, perhaps a model-based approach, test the effect for the group or the interclass correlation which is equivalent, and if it's not significant then go back to an analysis that ignores the group all together. The problem with this approach is the degrees of freedom for the test of the interclass correlation are often limited and as such they're power of those tests is limited. The standard errors for the variance components

are not well estimated if the variance components are close to zero, as they often are estimated to be, especially in small trials. Even with a small interclass correlation, if we ignore it, we can have type I error rate inflation if the number of members per group is moderate to large. And so, the prudent course is to retain all random effects associated with the design and sampling of plan, even is statistical tests suggest that we might be able to delete them. So I do not recommend deleting them based on tests of significance.

What about studies based on only one group per condition? The example that I often cite is a study I saw when I was on study section at NIH some years ago where an intervention was being delivered in Kansas City, San Antonio was begin used as the control site. The first question that many would ask, of course, is, "How are San Antonio and Kansas City similar? Why would we ever pick those two?" But the statistical question is, "How do we separate the variation associated with the group from the variation associated with the condition?" And the answer is we can't when there is only one group per condition. So the city itself is completely confounded with treatment assignment, and we don't know if any affect that we're observing is a result of differences between San Antonio and Kansas City or differences between treatment and control.

Any approach that we apply to this kind of design rests on a strong assumption. There are several approaches that have been tried. A paper by Sherry Varnell published in 2001 examined these and suggested that they have an inflated type I error rate if there is a variation associated with the group, and we can't test those assumptions. So we do not recommend this approach if you're interested in statistical evidence for causal inference. You may still find this approach useful in a preliminary study, but please don't try to analyze the data; don't try to publish the data where you're talking about intervention effects.

What about Kish's effective degrees of freedom? So effective degrees of freedom are the design effect, divide by -- or sorry -- the individual degrees of freedom divide by the design effect or the variance inflation factor. We've looked at this in simulations and found that it doesn't work very well, the type I error rate is still inflated, often badly. And so, we don't recommend this approach in the analysis of data from group randomized trials.

What about end balance designs? So I've talked about the importance of group-level balance. You can get into difficulty if you have a different number of groups in treatment than in control or if you have three arms, a different number across the three arms. So it's very important to maintain the same number of groups in each arm of the study, or in each condition. Member level imbalance though is quite common and can create type I error inflation under certain conditions, and the risk increases as the level of imbalance gets worse. So let's talk about that a little bit.

In general, if the largest group compared to the smallest group -- if we calculate a ratio of those group sizes -- if that ratio is no worse than about two to one, we can usually ignore the imbalance in the data at the member level. But if the imbalance is worse than that, we need to pay

attention to it. A paper by Johnson et al. published in 2015, looked at 10 different model-based approaches to dealing with member imbalance. And they identified two that worked pretty well, but only under certain conditions.

A one-stage approach mixed model approach using Kenward-Rodger degrees of freedom and unconstrained variance components worked well in larger trials, that is studies that had 14 or more groups in each arm. A two-stage model weighted by the inverse of the estimated theoretical variance of the group means also with unconstrained variance components performed well so long as we had at least six groups per condition. None of the models performed well when we had very small studies, and so let me caution you very much against substantial imbalance in your group randomized trial if it's a small trial with a limited number of groups.

What about constrained randomization? How does this impact the analysis? We talked about it in part two in terms of design implications, now let's talk about it in terms of analysis implications. Leigh Addall published a paper in 2015 that evaluated model-based and randomization-based tests in the context of constrained randomization and group randomized trials. The unadjusted version of the randomization tests maintain the nominal type I error rate; the unadjusted model-based test as it turned out, was somewhat conservative.

More commonly though, we used adjusted tests, and both the model-based and the randomization-based test were similar when we used the adjusted version in the context of constrained randomization. Both maintained the nominal type I error rate, both had better power under constrained randomization. So in this instance, we can recommend constrained randomization not only because of the benefits it provides in terms of protecting against confounding, but because it improves power if we take it into account in the analysis. It's important, however, to use the correct specification of the permutation distribution if we use constrained randomization in a permutation test. And details are provided in Leigh's paper, published last year.

What about the non-negativity constraint? This is something that many investigators and analysts aren't even aware of, but should be. As it turns out, most of the software that's based on maximum likelihood routinely constrains variance components to be non-negative. Most of you would think about that for a moment and say, "Duh, that makes sense. Why would we allow a variance estimate to be non-negative?" Variances have to be positive. Well, in the context of variance estimates, yes, it has face validity to limit them to be positive. But when we think about interclass correlation where it could be negative or positive, that's what raises the issue of, "Oh, well, maybe we ought to get negative variance components another look."

If we combine this constraint with traditional methods for calculating degrees of freedom, we introduce a positive bias in the estimation of those variance components. And that actually depresses the type I error rate, often dramatically. And if that happens, we are dramatically depressing power. Swallow and Monahan showed this in a paper published years ago. We sort of re-discovered this issue in a paper that we

published in 1996. So the earlier advice was to avoid the non-negativity constraint. And there are ways to do that in some of the software packages. Recent evidence suggests that use of the Kenward-Rodger method for degrees of freedom addresses this problem. It was shown in a paper published by Rebecca Andridge in 2014. And so if you use Kenward-Rodger degrees of freedom, the non-negativity constraint and the -- it doesn't seem to raise the same problem that it does with regular degrees of freedom.

What about analyzing individually randomized group treatment trials. Many studies randomized participants as individuals, but then deliver those treatments in small groups. And we've talked about this several times in the course of this course. Psychotherapy trials, weight loss, smoking cessation trials are examples of that. Participants are nested within these small groups, the facilitators may be nested within conditions, there may be little or no group level interclass correlation at baseline, but we anticipate that it develops over time and in proportion to the magnitude of the -- an intensity of the interaction among group members in those small groups.

Analyses that ignore that developing interclass correlation risk an inflated type I error rate just as we see in group randomized trials. It's not as severe, but certainly can exceed 15 percent under conditions that are common. And we want to maintain the type I error rate at five percent, so this is a problem that we need to address. The solution, as it turns out, is the same as in a group randomized trial, we need to analyze the data to reflect the variation attributable to small groups, and we need to base the degrees of freedom on the number of small groups in arms where they exist, not just on the number of members.

What about individually randomized group treatment trials where members belong to one or more group at the same time or change groups over the course of the study? The literature in individually randomized group treatment trials assumes that each member belongs to just one group and that the group doesn't change over time. That pattern is not likely to hold in practice. And we found in the same paper published in 2014, that failure to account for multiple group membership can result in an inflated type I error rate, even if you are accounting for the original group membership.

Roberts found a similar pattern and found that multiple membership, multi-level models address this problem. They required data on membership time in each group, which is not routinely collected in individually randomized group treatment trials. But, hopefully, going forward from having been exposed to this material in this course, you'll collect that information and be able to take advantage of it. Let me note, as well, that this general issue also applies to group randomized trials. So if group membership changes over time, or as can happen, for example in a school-based study where we start collecting data in junior high school and continue to collect data after the students have moved on to high school. It's important to take into account the time spent in each group and not just take into account the original groups that were randomized.

In summary, group randomized trials are different animals in terms of their designs. They require different analyses. In particular, they require analyses that reflect the nested design that's inherent in these studies. If we use the general linear model or the generalized linear model and these are the familiar linear regression, logistic regression, analysis of variance, t-tests sorts of methods, in a single stage, we're not going to get a valid analysis, and we're going to have an inflated type I error rate. Instead, if we're going to use models, we need to use models that are based on the general linear mixed model, the generalized version of that. If we have one or two time points that were included in the analysis, we can use the mixed model ANOVA/ANCOVA. If we have two or more -- sorry -- three or more time intervals that we're trying to model, the random coefficients approaches are recommended.

Other methods can be used effectively as long as we're careful and implement them correctly. And those include randomization tests, GEE, and two stage methods. Other approaches are not appropriate including analysis at a subgroup level, deleting the unit of assignment if it or the interclass correlation is tested and shown not to be significant, designs with one group per condition, and Kish's effective degrees of freedom. Unbalanced designs, as I noted earlier, can create special problems and an inflated type I error rate if we don't pay attention. Special methods are required, and I've given you references to those. Constrained randomization can be helpful, both in balancing potential confounders and in improving power so we can recommend that to you. Individually randomized group treatment trials face similar problems, the solutions are similar, it's important to model the small groups or common change agents as nested random effects and, of course, that has implications for degrees of freedom and testing.

So I want to thank you for participating in part four -- sorry -- part three analysis approaches, and I want to draw your attention to the next part, the next module in this series, part four, which addresses power and sample size calculations. You can go to our website shown here, and provide feedback on this module or any of the others, you can download the slides that you've just seen, you can download all of the references, and you can download suggested activities that will help you apply some of what you've learned in this module. Certainly you can view the material again, and you can watch the next module in the series, part four, Power and Sample Size. If you have any questions, please send them to grt@mail.nih.gov and we'll get back to you. Thanks very much.

[end of transcript]