U.S. FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

AND

JOHNS HOPKINS UNIVERSITY CERSI WORKSHOP

Addressing Challenges in the Design and Analysis of Rare Disease Clinical Trials: Considerations and Tools

Day 2

Wednesday, May 3, 2023

9:00 a.m. to 11:58 a.m.

Meeting Roster

Frank Harrell, PhD

Expert Biostatistics Advisor to CDER, FDA and Professor of Biostatistics, Vanderbilt University

Rima Izem, PhD

Director of Statistical Methodology Group, Novartis Pharma

Kelley Kidwell, PhD

Associate Professor & Associate Chair for Academic Affairs in Biostatistics, School of Public Health, University of Michigan

J. Jack Lee, PhD, DDS

Professor, Department of Biostatistics, University of Texas MD Anderson Cancer Center

Gregory Levin, PhD,

Associate Director for Statistical Science and Policy, Office of Biostatistics, Center for Drug Evaluation and Research (CDER), FDA

Dionne Price, PhD

Deputy Director, Office of Biostatistics, Office of Translational Sciences, CDER, FDA

Karen Price, PhD

Associate Vice President and Statistical Officer, Statistical Innovation Center, Eli Lily & Company

Michael Rosenblum, PhD

Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Noah Simon, PhD

Associate Professor, Department of Biostatistics, University of Washington

Nigel Stallard, MSc, PhD

Professor of Medical Statistics and Deputy Director of Clinical Trials Unit, Warwick Medical School, United Kingdom

C O N T E N T S

C O N T E N T S (continued)

P R O C E E D I N G S

(9:00 a.m.)

Welcome

DR. D. PRICE: Good morning, everyone, and welcome to day 2 of our workshop on Addressing Challenges in the Design and Analysis of Rare Disease Clinical Trials. Our focus yesterday was on quality and fit-for-purpose data for use in rare disease trials and how that data collected from patients could be used to inform drug development. Today, we will focus on design and analysis methodologies that may be useful in settings with small numbers of patients.

The first session will focus on adaptive designs and the second session will focus on analysis methods in small populations. We've assembled an experienced group of speakers and panelists who will share their insights and knowledge with us today.

Both sessions will be moderated by Dr. Michael Rosenblum. Dr. Rosenblum is Professor of Biostatistics at the Johns Hopkins Bloomberg

School of Public Health. His research is in causal inference, the design and analysis of clinical trials, and enhancing capacity in low- and middle-income countries in statistical methods for clinical trial design and analysis, and he is a fellow of the American Statistical Association.

So, without further delay, I will turn it over to Dr. Michael Rosenblum. Thank you.

Session 1 - Michael Rosenblum

DR. ROSENBLUM: Thank you, Dionne, and welcome, everyone. Thank you for joining today. I want to remind everyone, if you have questions, please put them in the Q&A. We won't be recognizing the raise-hand feature, so any questions, please put them in the Q&A, and we'll try to answer them.

The setup for this first session, as well as the second, is 20-minute talks, three of them, followed by 20 minutes of Q&A. And I want to cut right to the chase and introduce the first of our three outstanding speakers in this first session.

Kelley Kidwell is an Associate Professor and

Associate Chair for Academic Affairs and Biostatistics at the University of Michigan School of Public Health. She has multiple areas of expertise, including sequential multiple assignment, randomized trials, also called SMART trials, and other novel trial designs. She's collaborated in a wide variety of fields, including oncology, mental health, and rare and chronic diseases. We're very lucky to have Kelley Kidwell speaking today, and I will hand it over to you.

Presentation - Kelley Kidwell

DR. KIDWELL: Thank you so much.

OPEN SLIDE SET. So yes, I'm so grateful to be here. Thank you. I'm grateful to start the day off. I'm going to talk about SMART Design and Bayesian Methods for Rare Disease Trials. Next slide please. As many of you already know the challenges in rare disease research -- I won't spend too long on this -- quite obviously, the small number of patients make this very difficult to come up with robust conclusions, and sometimes when we're interested in especially binary

endpoints -- success, failure, response, non-response -- there's even a smaller number of those endpoint events.

Something else that is really notable is that because of the small number of patients, it's quite challenging to run separate dose-finding trials, along with additional separate confirmatory trials. So trying to think of designs in which those can be combined is really important in this area. Then it's also very difficult to meet those standard frequentist benchmarks, like 80 percent power and 5 percent two-sided type 1 error that we generally think of being somewhat almost necessary or expected in clinical trials. Next slide please.

There's a clear need for clinical trial innovation, both on the design side and also in the analytic aspect. In the design side, we want a trial that is of minimal size but will still provide robust evidence. We also would like the design that we know it's not necessary that the patients are going to do very well in the trial, and it's more for the future benefit; however, in

order to get patients interested, recruited, and retained in the trial, we'd like to have some sort of benefit for participants. So maximize the chance of receiving therapy and minimize the number of individuals receiving placebo or standard of care. It'd also be great for design to consider more than one dose of treatment and confirm its efficacy.

On the analytic side, we'd like to provide estimates of the treatment effect with clinical interpretability, and if possible, incorporate external data such as natural history studies, registries, previous trials, expert opinion, and formalize the content of that.

The Complex Innovative Design program, particularly thinking about Bayesian design and analysis, has really kind of allowed us to think more innovatively about clinical trial design, especially in small samples in rare diseases. Next slide please.

A lot of my work started about seven years ago when we were discussing this trial in isolated

skin vasculitis, a rare disease. This is the first snSMART. We call it the small sample SMART that's in the field. The clinicaltrials.gov number is here. It's called a randomized, multicenter trial in isolated skin vasculitis, or ARAMIS for short, and it looks like this.

The physicians were using multiple treatments; however, they didn't know which one was really better. So this is a comparative effectiveness study where we have treatments A versus B, versus C, and patients are followed for some time in ARAMIS at 6 months, where we expect the treatment to be effective. And if it's not effective, then participants are re-randomized. So those who didn't respond are re-randomized to the other two treatments that they didn't first receive. So if they got A, then they'd be randomized to B or C. If they did respond, they continue on that treatment.

The goal of this study, while there are two stages, is actually to estimate the first stage treatment response, so to say is it best to get A,

or B, or C, but we want to use data across those two stages, so we've developed methods that are appropriate for whether the outcome is either binary like a response yes/no or continuous like a score. Next slide please.

I want to describe this design a little bit more. SMART comes from the term "sequential, multiple assignment, randomized trial," and these designs really came together in the last two or three decades, two decades or so, in the early 2000s, really thinking about more in the phase 2 space for large samples, and to think about these treatment sequences. We're kind of forgetting about the original intent of SMART, but thinking about using that design, a multistage design, where the second stage of randomization depends upon response to the initial treatment, and using it in a small sample to get more information from a small number of individuals.

This is a multistage design. Patients can be re-randomized based on response to initial treatments. You might want to think about it as

like a restricted crossover design. Crossover designs are used quite often in small samples. First, a patient gets A, and then they get B, or they first get B, and then they get A. Here, we're going to say, "Well, I don't want to give everybody that second treatment because if they did well, then they probably just want to stay on that treatment, and that will help with retention. Instead, if they didn't do well, then we can re-randomize them." So the SMART design in this aspect is more of this restricted crossover, but we're going to be able to use both stages of data in our analytic method to get a more efficient treatment estimate.

Now, this design isn't appropriate for all rare diseases. It's really only appropriate for ones which are chronic and relatively stable over the trial period, so the two trial stages. Next slide please.

After we did some work on comparative effectiveness, the three active treatments, snSMART, we thought, you know what, this design

would actually be maybe more useful, considering how we could register a drug for a small sample. So the fact that there isn't often dose finding in rare diseases, because we don't have enough patients to go from a dose finding trial to an additional confirmatory trial, we thought we could actually do this in one trial.

We've been developing these snSMART designs considering two dose levels and placebo. Here's a design where all participants are randomized between placebo low and high dose. They're followed for some time in which we expect to see the treatment effect, and then both placebo and low-dose groups are re-randomized between low and high dose. So you can see here that while there are some patients that received placebo, at the end of the two stages of the trial, all participants received some dose of the trial. So this, again, may help with recruitment and retention in the trial.

For those on high dose, if the individuals didn't respond, as long as that dose is safe, they

could stay on high dose, perhaps they need it longer; whereas if they did respond, perhaps low dose could be just as effective but maybe have less side effects, so we could re-randomize individuals between low and high dose. Next slide please.

We've modified the design slightly, which is perhaps maybe more patient-centered. So all those on placebo would be re-randomized as before, but for those on low dose, if they didn't respond, they'd get high dose, assuming it's safe, and if they did respond, they could stay on low dose; perhaps they don't want to change.

For those who don't respond on high dose, perhaps it's not as accepted to continue to give high dose; instead, they'd sort of be off trial and given physician's choice, whereas those who did respond, similarly to the last design, could be re-randomized.

Again, the goal here is to use all of the data from stages 1 and phase 2 to decide is a low dose effective? Is it more effective than placebo? Is high dose more effective than placebo? And

could you go forward with low dose as opposed to high dose? Next slide please.

Actually, as we're developing these methods, we noticed that some trials in the rare disease world were using something very similar. This is an actual trial. It's the SPITFIRE trial. The clinicaltrials.gov number is here. It was a two-phase, placebo-controlled study of two dose levels of treatment in ambulatory boys with Duchenne muscular dystrophy or DMD. You can see that here, only the placebo group was re-randomized in stage 2 to low or high dose. Again, this is something that is useful for recruitment. Those who got low or high dose, they continued their low or high dose. Here, it was likely to see long-term effects.

Now, in their primary analysis, they only were planning to use that stage 1 data. Their primary outcome was the change from baseline to week 48 in the 6-minute walk distance, and that secondary outcome was the NSAA score or ambulatory score. You can see that this design is quite

similar to the snSMART design that I just showed you; however, we use this in comparison to what if they did an snSMART and actually use that second-stage data. We could see more efficiency, and I'm going to come back to that in a few slides. Next slide please.

There are many advantages of this snSMART with the dose levels; the fact that many participants will receive a higher dose level in stage 2, or at least some dose level, so this can really help with engagement and retention, and also recruitment. The design allows for both dose finding and dose confirmation, which allows us to find the dose effect and also register the dose within one trial as opposed to requiring two. This is really helpful as one of the most common reasons for not being able to register a drug, being the fact that the correct dose wasn't found.

The analysis can formally incorporate expert opinion or external control data, external co-data, from previous trials or registries, and this can really help with the efficiency of the treatment

effect estimates but also decrease the number of individuals on the placebo arm. Next slide please. We've worked the analytic methods in the Bayesian framework, and I'm not going to show you any equations here, but I want to give you some intuition for what's happening, and if you're interested in the actual equations, I'm happy to share those papers with you, so feel free to follow up with me.

Our goal is to estimate the first stage response rate, or the mean outcomes, of each treatment by pooling together the data across both stages. We want to provide credible intervals of the effect, or the differences between the dose levels and placebo, or the differences between the treatments, and those are going to contain the true effect with some particular probability. This is actually the definition of a credible interval that most people think is what a confidence interval is.

With this Bayesian framework, we must shift our focus away from significance and p-values and really focus on the estimate itself, and we can

formally incorporate expert opinion, historical data, or external control data, which is going to help increase our precision, and we can formally test the effects of that as well. I know people are very worried about Bayesian framework in terms of the priors set, but you can rigorously test the effect of these priors, the effect of that prior information. Next slide.

The Bayesian framework is really set up such that, actually, we think we don't know the true population parameters, so we're trying to figure out what they are. They're random; they can change. So we take our best guess at these response rates based on our current knowledge, which might be expert opinion, registry of prior trials, and that's this red curve. This is our start of the trial idea of what we think the estimates are going to be.

Then we collect data to observe the response rates in the trial, so that's going to be the data that we see, or that distribution of the effect that we see in this blue curve, the likelihood.

Then we combine the prior information with our likelihood, and we get this purple curve, the purple distribution, which is now our updated estimate of the response rates or our posterior distribution. Next slide.

These prior distributions can be informed by clinical investigators or historical data, and we do want them to be slightly informative; however, we don't want them to overtake the data in the trial, so we only want them to be a few people's worth of information. For example, say we plan on having 60 individuals in our trial, our prior distributions might be worth the strength of two or three individuals, so not remotely going to take over the amount of information coming in from the trial.

We also can use what's called a mixture approach, which says, well, we can have some informative prior, but we can combine that with a non-informative prior if we're worried about the effect of that prior being overly informative; then let's combine it with a non-informative prior and

see how that affects our results. And like I said, we can really test the sensitivity of our results, given different prior distributions, to make sure that we're not being overly informative.

Then for our Bayesian analysis approach -- we call it the Bayesian joint stage model -- we model the first stage simply, then we model the second stage data conditional on the outcome, whether they responded or not. Basically what we try to do is link the first stage response rate to the second stage response rate via what we call linkage parameters, and this induces some within-patient correlation and says, well, we know the second stage outcome isn't as good as the first stage outcome, if we're estimating the first stage outcome, because they already responded or didn't respond. So we want to basically kind of discount that data in some way, but we also want to be able to use it because it's still very helpful in getting these treatment effect estimates. Next slide.

Obviously, if we're going to incorporate

external data, it's very important to have a careful choice of this control data. There's some some criteria, the Pocock criteria, to assess the similarity between external controls and trial data that include trying to make things similar across inclusion/exclusion criteria, endpoint definition control treatment, distribution of demographic data, et cetera. And something important, as I've mentioned already, or as I mentioned for the prior, is that we also want the number of external control patients, or really the effect of sample size of this prior information or this external control data, not to exceed the number of controls on our trial. But what's really advantageous about incorporating this data is that we can then lower the number of participants on the placebo arm, which makes the recruitment perhaps a bit easier for an snSMART or trial that includes a placebo arm in rare diseases. Next slide, thank you.

Our models do have some assumptions. Right now, we don't incorporate patient or disease characteristics or covariates into the model. We

often make the simplifying assumptions about the parameters that link the first stage and second stage outcomes together. We do assume that there's some sort of washout period or no carryover effect between the first and the second stage, so much like a crossover trial.

Right now, we're only working with one endpoint of interest, and we also assume no to low missing data; however, we're working on extensions to all of these assumptions. Also, once you use our model, you can test many of these assumptions and how appropriate they are in your settings in various ways, and we provide some guidelines to that. Next, please.

What we found is that our Bayesian joint stage models provide estimates with very low to no bias and are much more efficient or have lower variance than other estimates. I have "double whammy" on the page because, really, with the two-stage design plus the Bayesian model, we have this double-whammy effect in which we've got these more efficient and robust estimates of the treatment

effect, and we can test our sensitivity to our various assumptions. Next slide.

So going back to that SPITFIRE trial, we simulated some of the data and reanalyzed the data, incorporating some external control data from one of the largest natural history studies in DMD, CINRG. What you can see is the first row if we just use the first stage of data, like a traditional analytic approach, which they used for the SPITFIRE analysis. The next two rows are using methods which we've developed, the Bayesian joint stage model and a robust meta-analytic combined model, which I didn't go into, but it really formally incorporates that external control data.

What you can see is that these estimates are quite the same, 1.8 and 1.6, but the credible intervals are smaller, or we have more efficient estimates with these Bayesian joint stage models, or the Robust MAC, because we've used both stages worth of data, and we formally incorporated external control data, so we can see the savings. Next slide.

So you might be thinking, "Okay. If I'm interested in running this, well, how do I size it?" We have two sample size calculators out there in Rshiny applet ready to go if you're interested in three active comparators and you have a binary outcome. We can provide the sample size such that you have some amount of probability, say 80 percent probability, for the 90 percent credible interval of the difference between the best and second best treatment to exclude zero. Here are some results, for example, where you'd need 28 participants per arm or just under 90 patients, depending upon the treatment difference that you're interested in seeing. That Rshiny applet is available online. Next slide please.

We also have an Rshiny applet, which is soon to be online, for the placebo high and low dose trial, where you have continuous outcomes, and here, we can really quantify the sample size savings. If we look at the first row, scenario 1, you can see that if you had a one-stage design and a frequentist analysis, say like comparison of the

means, like a t-test or so, you would have 50 participants per arm.

Now, because we've formally incorporated the external control data or have some prior information, we save four patients with a Bayesian analysis, so we need 46 per arm. However, if we have a two-stage design with the snSMART and use our Bayesian joint stage models, we only require 31 participants per arm. So you can see there's this massive savings, 62 to 67 percent of savings in our sample size by using the two-stage data and the Bayesian joint stage model. So again, that double-whammy effect, where we're seeing that we can reduce the sample size from a one-stage design by 15 to 60 percent. Next slide please.

All of our current methods, Bayesian, and then we have similar frequentist methods, are available in an R package called snSMART, which you can download and use. We have a number of papers already in the field, and there are two snSMARTs in the field that I know of right now. There may be one or more additional ones, but ARAMIS was sort of

our initial motivating. Then there is a more recent one called MISTIC, which has just been written about, the protocol paper, and is in the field. Next slide please.

In summary, the snSMART design and Bayesian joint stage models fit under this complex innovative design program for comparative effectiveness, but also for confirmatory dose-finding and confirmatory drug comparisons. But really, I'm not saying this is for all rare diseases. This is for chronic stable rare diseases, but this design has potential to aid in recruitment and retention. The design and analysis can both dose-find and dose-confirm that best dose level. When we use a two-stage design and the Bayesian framework, we're allowing for more efficient, unbiased treatment effect estimates. We have developed software to disseminate these methods, and we hope that this design will aid in identifying more effective treatments for rare diseases. Next slide.

I just want to acknowledge some of the

people that helped with this work, some of those pictured in this picture here, and also through contracts with PCORI and also the FDA. I think that's the end of it, so thank you so much, and I'd be happy to take questions later or offline. My email is kidwell@umich.edu.

DR. ROSENBLUM: Thank you, Dr. Kidwell. That was a fantastic presentation. There are multiple questions in the Q&A, and we'll come back to them after all the speakers have gone, but you kept your talk perfectly on time as well. Thank you for that.

We'll turn next to Noah Simon. Noah is an Associate Professor in the Department of Biostatistics at the University of Washington. He is an investigator for the Therapeutics Development Network at Seattle's Children's Hospital. He has a variety of research expertise areas, including biomarker development; clinical trial design, including adaptive clinical trial designs; and machine learning. He primarily engages with trials in oncology and cystic fibrosis. He's, as all the

speakers and panelists today, an outstanding researcher, and I turn it over to you, Noah.

Presentation - Noah Simon

DR. SIMON: Perfect. Thank you so much, Michael. It's a pleasure to be here.

I'm going to be talking about adaptive enrichment designs, and we'll get into that in a little bit. I think there are some real challenges in employing those in rare disease settings, but I also think some real opportunities. I like to work on machine learning. We're not going to talk about machine learning here. We're not going to be talking about anything too fancy, I think, because we're going to be very limited in the number of people we can engage within these settings. Next slide.

I wanted to give a shout out to a certain Richard Simon, who I've chatted with a lot about these ideas, and that's a fairly recent picture of him. Next slide. I guess I want to say, normally, I love to engage with questions during the talk, but I'm probably not going to be able to do that

this time because we're a little time limited, so I'm looking forward to engaging with questions during the Q&A.

In many diseases, and in rare diseases a lot as well, rather than engaging with general purpose treatments that target maybe every person with the disease, we often are engaging with very targeted treatments that may only target a subset of people with the disease. I guess something that we might call one disease could actually be a heterogeneous collection with a similar phenotype but different mechanistic causes, and maybe slightly different phenotypes that, again, are very similar.

So again, our new treatment might only target one of those mechanistic causes, one type of dysregulation, and often we have some idea of which patients suffer from that dysfunction versus have a disease caused by another, but we don't have a perfect characterization of essentially which subset of patients who have the disease will benefit from our new treatment versus which subset maybe have a slightly different dysregulation

before we start our phase 3 trial, before we start our pivotal trial. The idea is when we have that uncertainty, what do we do moving forward, and we'll talk about how adaptive enrichment is one class of designs we might use. Next slide.

So what might you do? I guess classically you might say, "Well, we don't really know who's going to benefit from new treatment, so let's just enroll everyone with the disease," or you might say, "Oh, here's our best guess, so rather than just enrolling everyone, let's enroll our best guess of the subset of people who we think will benefit from the disease."

So the first I'm going to call an all-comers design, and the second is what's called an enrichment design. Generally, you need some rationale for why you're not going to enroll a subset of patients with the disease, but it's often very hard, as you enter a phase 3 trial, to identify exactly who should be enrolled. Next slide.

An adaptive enrichment design ideally

provides a happy medium here. You start by enrolling everyone and randomizing them to your new treatment, maybe in standard of care in the case of a randomized trial, and as people progress through the trial, assuming you can measure at least some short-term endpoint, you can start to identify -- based on one or more covariates you have measured, generally, at baseline -- who you think will or is benefiting from treatment and who is not, then you can modify your enrollment criteria as you go, and maybe drop a strata of patients for future enrollment based on what you're learning.

As you can imagine, it's hard to learn a lot during a trial, but maybe you can learn a small amount that could still be really effective. So these modifications will use outcomes and treatment assignments from earlier patients, so you obviously have to be careful about blinding, and bias, and such, but there are ways to engage with that. Next slide.

Before I jump into that, again, we said you

can have an all-comers design, you can have an enrichment design, or maybe you can do this adaptive enrichment design, where, on the fly as the trial goes, you say, "Oh, this trial targets people whose" -- in the oncology example -- "tumors express this protein too much." But what does too much mean?

Okay, we're starting to learn as the trial goes along what a reasonable cutpoint for too much versus too little is, and you sort of stop enrolling patients who you think have too little, and allow them to engage with other treatments, again, as the trial progresses. You obviously have to be very careful about statistical conclusions there, but there are a number of frameworks for engaging with that, which we will touch on as we move forward.

In oncology -- and this is not a rare disease example, but I want to give this because I think it's a good archetypal example -- there are exciting new immuno-oncology therapies that target basically expression of this PD-1 or its ligand,

PD-L1. And again, we basically know that this should work in tumor or tumor microenvironments that have high expression of either PD-1 or its ligand, and we have observed that. We've observed the efficacy of treatment increases as that expression increases, but it may be effective even in relatively low expressors.

So the question is, what do you do when you run a trial? It's become increasingly common to engage with trials where, sort of on the fly, we try to identify that cutpoint and initially allow fairly low expressors to join the trial, and slowly cut that off. Next slide please.

Another example where it wasn't used, but it could have been used, and I would argue maybe should have, is another cancer treatment. Cetuximab is a treatment widely available and used. It targets the epidermal growth factor receptor protein on cancers, and the original pivotal trial in colorectal cancer was an all-comers design and didn't initially find significance for this drug being effective at treating that cancer. But when

they looked back, did a retrospective analysis, they actually identified it wasn't the amount of expression of eGFR in the tumor that was most important, which they hypothesized initially; it was whether or not that was the pathway by which the cancer was acting or not. So there was actually this mutation, this KRAS mutation, and what you needed was for this mutation not to happen in order for cetuximab to work.

So again, we were able to look retrospectively, reanalyzing very strong evidence that treatment was effective in KRAS wild-type patients. But maybe things could have been done more cleanly. There was some chance we were going to miss this, so maybe initially we could have said, "Okay. Let's look at eGFR expression and KRAS mutation status when we jump into this trial, and think about modifying enrollment as the trial moves on," based on who appeared to be benefiting from new treatment over standard of care. Next slide.

In all of these cases, there was a clear

molecular target, but before the trial was run, it was hard to identify exactly who the right subgroup of patients was, either because we weren't sure what the cutpoint was or because we had maybe two covariates that we thought could be involved in who benefited from treatment.

So I would say these are prime choices for adaptive enrichment. We don't want a zillion features. We don't want someone to say, "Well, we have expression of 20,000 genes measured on the tumor, and we think any of them could be informative," because you can't really learn that during one trial and evaluate the efficacy of treatment at the same time. You're asking for trouble there, so you want a handful of features that you think have a priori strong scientific relevance, maybe just one, and you need to find the cutpoint, which is pretty common. Next slide please.

So we talked about some examples that are not, maybe, so appropriate for a rare disease forum. I also engage with cystic fibrosis, and

here I'm going to talk about a sort of example in cystic fibrosis, where one could think about applying these ideas that I think can map to a lot of other problems.

Cystic fibrosis, as many of you probably know, is a genetic disease that results from dysfunction of the CFTR gene/protein, and there are a lot of different mutations in the gene that can cause various types of dysfunction that are all termed "cystic fibrosis."

There are some of these mutations that are very common, and there are some that are much more rare. Sometimes they mechanistically might have very similar effects, although maybe not identical effects, and in terms of developing treatments, often treatments are developed for those classes of dysfunction, and for common mutations, we can also identify, here's the dysfunction, and we can actually evaluate the effective treatment in people who have that mutation, that dysfunction. For rare variants, it can be much harder. We maybe don't have enough patients to actually just enroll

patients with that mutation.

So again, there's this new class of treatments, these modulator therapies that provide essentially replacement for those mechanistic dysfunctions that we see, and these modulators have been extremely successful. I mean, this is like a real success story in medicine, in general, and in the rare disease setting. Next slide please.

There's growing evidence that these modulators -- and in particular this triple combination therapy called "Trikafta," which is essentially the most recently approved modulator -- in addition to working for particular common variants that were tested in the main pivotal trial for approval, they also work in some of these more rare mutations.

Given a rare mutation, you can actually run an in vitro screen to look for how well, say, Trikafta, that we believe it will work. You can basically take cells that shouldn't express CFTR protein. You can treat them so that they actually do express a mutated version of CFTR protein, and

then you can treat that with Trikafta, and leave it untreated, and then look at something like conductance -- and conductance gets messed up generally when you have that mutation -- and see if conductance is fixed.

This is a screen by which we can say, okay, does treatment work, not in a human, but sort of in this in vitro very sick setting? You've got a continuous measure of activity. There is strong reason to believe that an increased value of that measure should result in an increased clinical outcome. So one question is, how much is enough activity? What should a good cutpoint be?

I want to note I'm going to talk about running trials here. There has been some label expansion of Trikafta just based on this, where additional trials didn't need to be run. It was expanded to different mutations beyond the original one, but one could imagine that in a scenario like this, say before a pivotal trial had even been run, one might want to employ an adaptive enrichment design. Next slide please.

Again, is it a good fit for adaptive enrichment? It's a potentially very effective therapy but only for a subset of people. Excellent. Are there good alternatives for people to try? Maybe not great, so maybe that's a strike against it. Adaptive enrichment is better when there are good alternatives. Do we have a larger potential pool of patients than we can likely enroll? Are we in the case where we have five patients with the disease in the enrollment window, and we basically need to enroll all of them, and cystic fibrosis is a relatively common rare disease, so we're not.

Again, I would argue that it's a pretty good candidate. I generated simulations roughly based on values observed in the pivotal trial of Trikafta. This is in people who had at least one allele of this quite common variant, F508del, where we hypothesized Trikafta should be effective. Next slide please.

Here, again in my simulation, I generated a Biomarker X. You can think of this as that

conductance measure as uniform between 0 and 1, and I had a treatment effect -- in cystic fibrosis, the outcome is this continuous measure of lung function, which we measure at the beginning of the trial and at the end. I'm going to evaluate average improvement; that's commonly what's evaluated, between the beginning and the end.

Here, for standard of care, regardless of biomarker value, I'm imagining you have no average improvement, where for Trikafta, in this case for biomarker values here, I'm saying about 0.5, but I vary this in the simulation setting. I'm imagining you have an average improvement of something like five points on this FEV predicted scale. I'm going to vary, again, where the cutpoint is and the jump just to show what gains we might get from using the adaptive enrichment versus something like an all-comers design here. Next slide please.

In these simulations, we have 60 patients who we've randomized 30-30, the new treatment and control, and I've run an adaptive enrichment design where I have two blocks of 30 patients randomized

15-15. In the first block, I include everyone. In the second block, I try to identify what that cutpoint is and then restrict enrollment to only those with the biomarker value above that cutpoint, and I use a hypothesis test that combines the p-values from block 1 and block 2, and I give a citation for that at the end. Next slide.

What we see here is, to explain this, is we have three panels. On the left, I'm imagining that 30 percent of people benefit from new treatment over standard of care, so that cutpoint is actually at 0.7. In the center, 50 percent of people should benefit from new treatment over standard of care, and on the right, 70 percent of people should benefit from new treatment over standard of care.

Here, I have the actual effective treatment in those who benefit, between 5 and 10, what the average change is for those patients, and on the Y-axis, I have the power ranging from 0 to 100, and in the dotted line, I have the power of a simple, non-adaptive design, where I just enroll everyone. And here, I have the power of an adaptive

enrichment design, where I identify who to enroll after that first block. Maybe I could have also had an enrichment design where I just, a priori, choose a cutpoint, and that probably would have been a reasonable comparator as well.

Here, we basically just see, as one could sort of expect here, that we can get a large improvement, especially if only a small subset of patients truly benefit from treatment. We can identify those, and then in that second block of the trial really only enroll those people. So again, in all of these cases, it's a little bit of a straw person scenario. It's set up such that adaptive enrichment will do well, and in fact it does great, but I think this is a pretty realistic scenario we might see in practice. Next slide.

My takeaways here, we can have a large improvement in power, and we're likely not perfectly identifying that cutpoint, that threshold, and we're not perfectly saying nobody below this will benefit, but it gives us statistical evidence to justify the use of

treatment in people with a large biomarker value. A rare disease setting is a very hard setting to identify whether a treatment is effective or not, so for these very effective treatments, I think identifying that the treatment does work in a subset of patients is a good goal, although, again, there's discussion to be had with the FDA. Next slide please.

I'm going to jump into some discussion points, which I will likely run out of time for in a minute or two. But some points; I think we don't want to let the perfect to be the enemy of the good, so I've used an adaptive enrichment design formalism that doesn't prespecify subgroups and then testing them. I think that's great if you have a thousand patients. I think in the case that you have 60 patients, you can't really do that.

So again, there is a hypothesis test that we're running. We got a p-value, but there's a little bit of subtlety there, and I'm happy to answer questions about that during Q&A, online, or it's in one of the papers I cite. And I added a

Yogi Berra quote that I think is definitely not a real quote, but I kind of like it. Next slide please. Again, I think statistics here is meant to support decision making, but it's never going to give us guarantees. Here, we're in the absence of perfect info, so we can't make perfect decisions, but hopefully we can make good ones. Next slide please.

In this stereotyping example, it's possible that no clinical trial is needed at all because Trikafta is extremely effective, in general. We have strong evidence that it will be effective in these patients in vitro. In fact, again, there was label expansion based on that without running a trial.

That said, you can imagine a slightly different world where we didn't originally have that positive trial. We didn't originally have that common variant that Trikafta was so effective with, and instead, everything was a rare variant; then maybe when we run that first trial, we have to figure out where to threshold that assay. So we

have that problem from the beginning, and then adaptive enrichment would be useful. Next slide please.

I throw out the number 60 people that may be a pipe dream in some settings. It may not be totally appropriate. You can do this with fewer, and maybe you're combining early- and later-phase data. There is something called a seamless phase 2/phase 3 that one could engage with there.

I'm going to cut it there. I have a little bit more on the slides that mention maybe use of registry data or historical data for a control arm and some thoughts on specifics on what null hypothesis is being tested. But again, those are technical details. [Go through next several slides and finish on last slide.] On my last slide, I think I have citations, so there's some discussion here and papers that I did with Richard Simon. So anyhow, I look forward to answering questions. Thank you all for your time.

DR. ROSENBLUM: Thank you very much, Noah, for that outstanding presentation. I liked

especially the Yogi Berra pretend quote, but very good. We will address questions at the end, and we'll have at least one FDA panelist respond to questions.

But at this point, we're honored to have Dr. Nigel Stallard. He's a Professor of Medical Statistics and Deputy Director of the Clinical Trials Unit at Warwick Medical School in the UK. He's an Editor in Chief of the Journal Statistics in Medicine. He has a wide range of research expertise, including statistical design and analysis of clinical trials. In particular, he's worked on optimal design for clinical trials and rare diseases in small populations, and on methodology for trials with interim analyses and adaptations, such as treatment for subgroup selection.

We're lucky to have Dr. Stallard here presenting, and I will turn it over to you.

Presentation - Nigel Stallard

DR. STALLARD: Thank you, Michael, for the introduction, and thank you for the opportunity to

speak. I was going to say this afternoon, but it's the afternoon here but probably the morning for most of you.

This talk does fit with the general theme for today of design and analysis methods for clinical trials for rare diseases, but although I am interested in adaptive designs, I'm not particularly going to be talking about adaptive designs today. In fact, this will be less of a talk about the methodological details or about a particular approach, and more of an overview, both an overview of current practice, and also thinking about what we might do differently in terms of methodology for rare disease clinical trials. I guess I mean different in two ways, different to what we're doing currently and also different to what we might do in non-rare disease settings. Next slide please.

Let's just start with some acknowledgements. Most of the work that I'm talking about today has really risen from two projects on clinical trials in rare diseases or small populations. The first

is a project that was funded by the EU that I led, called InSPiRe. It stands for innovative methodology for small populations research, and the second was a task force on small population clinical trials, organized by IRDiRC, the International Rare Disease Research Consortium. In addition to acknowledging the EU funding, I should acknowledge the input from a number of colleagues on these two projects. I won't list them all, but there were numerous comments, including Simon Day and Tim Friede, whose names appear at the bottom there. Nevertheless, I should say, of course, the views that I'm expressing are my own and not necessarily theirs or anyone else's. Next slide please.

I want to step back from some of the details of a particular trial design and really think about what it is that we're trying to do when we do clinical trials in rare diseases, so I'm going to start off with these three quotes from our regulatory guidance that many of you may have seen.

The first two come from the EMA guidance,

which, of course, applies in the European community and also is now adopted by the UK. There are two quotes here. The first says, "Patients with rare conditions deserve the same quality, safety, and efficacy in medicinal products as other patients," and then they go on from that to say, "Orphan products should therefore be submitted to the normal evaluation process." There's a similar statement from the FDA and the U.S. saying "The Orphan Drug Act does not create a statutory standard different from common conditions."

The title of my talk asks should we be doing clinical trials in rare diseases differently, and that means differently to what we do in non-rare disease settings, then these statements at face value would suggest that the answer to that is no. Nevertheless, of course, things are not quite as simple as that, or we wouldn't be needing this workshop today. So I want to start by looking at what's actually being done in the world of rare diseases. Next slide please.

This slide presents some work that was done

by a group in Liverpool, and what they did was they went through the clinicaltrials.gov database, looking at trials in rare diseases and non-rare diseases, and comparing the recorded sample sizes for those trials. So this slide is just a summary of what they found. The first column gives the ranges of trial sizes, naught to 50 patients, 51 to 100 patients, and so on; then the second and third columns give the percentages of trials in each of those size ranges, the first one for non-rare diseases, and the second one for rare diseases.

So looking first at the non-rare diseases, you can see many of the trials, even in non-rare diseases, are quite small, with fewer than 50 patients, but there's quite a sizeable chunk, around 30 percent of trials, with between 100 and 500 patients. I think that's slightly bimodal, if you like distribution, and probably reflects a mix of trial types. What they looked at was all of the trials on clinicaltrials.gov. That's represented by early phase and confirmatory late-phase trials, so perhaps that's why we have a bunch of small

trials, and then a bunch of larger trials.

If we look at the rare diseases, then perhaps, not surprisingly, we see something rather different, and here it seems, as we probably would have expected, the larger trials are much less common. So here, the large majority of trials have 50 patients or fewer, and really, there's no evidence of that second mode in the distribution, if you like, of trials with more than 100 patients. Next slide please.

Here's a summary of similar work, which focuses particularly on rare diseases and splits up for the phase 2 and phase 3 trials so we can look at those separately, and also splits up diseases by the disease prevalence because, of course, a rare disease, which affects between 1 and 5 in 10,000 people, while still rare, is obviously very different from one which affects perhaps one or two in a million.

The left-hand panel shows phase 2 trials, and the right-hand panel here shows phase 3 trials. The sample sizes are on the Y-axis, and you can see

that this is on a logarithmic scale, and it does go right down to 1 at the bottom there. In each of the disease prevalence groups, we've got two sets of bars. The first one is for actual trial sizes, where the trial's been completed, and the second one is for target trial sizes, where the trial is still ongoing. Of course, quite often, the actual trial sample size is rather smaller than the target.

Looking first at phase 2 on the left here, apart from the very, very rarest of the diseases, the trial sample size doesn't seem to change that much across the different prevalence groups, and here, the typical sample sizes are just under 50, so there may be kind of 30 or 50 patients in each trial. So clearly that appears to be achievable even in these rare diseases.

If we move across and look at the picture on the right, here the story is rather different, obviously, because these are phase 3 trials, these are confirmatory trials. Trials typically in this setting are rather larger, and it certainly looks

as we consider more rare diseases, the trials are smaller than for the less rare of the rare diseases, and we can see it's something of a trend with larger trials on the right and smaller trials on left within that panel. In the very rarest of diseases, those affecting fewer than 1 in a million, really, phase 3 just appears to be completely absent in this picture. There are very, very few trials which describe themselves as phase 3 that were done in these very rare diseases.

Now, obviously the designation as phase 2 or phase 3 was taken directly from entries on the clinicaltrials.gov database, but it really appears that, perhaps unsurprisingly, whatever the regulations might say, trials are being done differently in rare diseases to more common ones. Trials in rare diseases are smaller, and for rarer the disease, definitely [indiscernible] the trials. And in particular, it looks like in phase 3, in the very rarest of diseases, these might be skipped altogether. Next slide please. But clearly we are doing trials differently in rare to non-rare ones.

This is the well-known problem, and trial sample sizes are small, often necessarily so, yet we want as much high-quality data as possible for clinical and regulatory decision making.

So how might we do them differently? How might we do them better? I'm going to talk briefly about three possible general approaches. The first is to get more data. Assuming it's impossible to increase the sample size, we nevertheless obtain more data for our analyses and our decision making somehow from the patients that we've got in our trial.

Secondly, can we get more information for our decision making from the same data? So use our data as efficiently as possible. Then thirdly -- and this isn't really talking about doing trials differently, but more about the way in which we use the information, and using that differently -- can we consider changing the level of evidence required for our decision making? Next slide please.

So first of all, let's think about how we

might try to get more data, and first from within the trial itself. So without increasing the sample size, this means collecting more data from the same patients, and sometimes we can do this by thinking about the data that we collect. So we can ensure, for example, that we collect continuous endpoints where we can, rather than binary endpoints, and certainly not dichotomize a continuous one into a binary response. We can make sure we collect relevant baseline covariate data, and we can collect longitudinal data where we can and have a long-term follow-up, if possible, and we can collect secondary endpoints that can be used to aid our decision making.

Now sometimes there are good clinical reasons why we shouldn't do these things, but, generally, if we can, then if we can collect and use more data, then that can only be good and can only help us in our decision making. Next slide please.

So those are pretty much no-brainers, I think, but slightly more controversial and

certainly requiring some more careful thought is the use of additional data from outside of the trials, and this may be from historical control data, so taking the control data from previous trials in the same population and using that either to completely replace the control group in the current trial or perhaps more likely to augment it so we can use a smaller control group in our trial.

There's been quite a lot of recent statistical work or methods which enable us to do this. In particular, there's been a lot of work on what are called dynamic borrowing methods, where external control data were weighted according to their concordance or discordance with the observed data; so the idea being that if the historical controls look similar to the current controls, then we should include them, and if they look rather different, then they could be ignored.

I think there's lots of nice work there, and it's a really nice idea. The challenge is, if you think about the decision as essentially being a hypothesis test and about the type 1 error rate for

that hypothesis testing, the probability of a false positive result, taken over all the possible sets of data from the trial, with the historical control data considered fixed, then by including that additional control data, you can inflate the conventional type 1 error rate. Now, there are ways around that. You can make the tests more stringent to bring the type 1 error rate down again, but generally that removes the very advantage of including the external data at all.

Now, obviously, hypothesis testing is only one way to think about the decision making process, so how much you worry about type 1 error rates really depends on what it is that you think you're doing in the trial. But if you are going to use historical control data in this way, then I think it's important to at least consider this. Maybe it's something that could be discussed in the discussion session later on.

Extending the use of external data further, there was quite a lot of talk in the session yesterday, of course, about using real-world

evidence with a whole range of methods possible, including the generation of synthetic controls or even conducted so-called in-silico trials, using models based on real-world data. Again, I think this probably comes with many of the challenges of historical controls, and perhaps more. It does have real potential, perhaps particularly for exploratory work, or for planning trials, or using alongside trials to interpret the results when the samples are small, but I think it's something, again, that needs to be very carefully thought about. Next slide please.

So having collected that data, we obviously want to make the best use of it possible, which really makes sure that our analyses are as efficient as possible. We're not going to say very much about this because I know it's the theme of some of the later talks today.

We also want to make sure, of course, that our designs are as efficient as possible, which we've heard in the last couple of talks; so using methods like group-sequential approaches to stop as

soon as we have sufficient evidence to reach a conclusion; or adapting our design to remain efficient as we learn about treatment effects, or about use of parameters, or about patient heterogeneity; or going back to something more similar to what we heard in the first talk, designs which allow us to have patients receiving multiple treatments, so they act as their own controls and, obviously, there are a number of methods enabling us to do things like this. Next slide please.

The final thing I want to talk about is really thinking about what level of evidence is required for decision making. Of course, when we conventionally have a confirmatory trial in a non-rare disease setting, then the focus is primarily on type 1 error rates.

Now, I know the p-value isn't everything, but in practice, much of our demonstration of efficacy comes down to whether or not we have a conventionally statistically significant result. So when we plan our study, we typically fix the type 1 error rate, and then find the sample size to

give us a certain power to detect a specified effect. And when we analyze our study, we typically control the type 1 error rate, and if the trial is smaller than we'd hoped, then we accept it can be underpowered.

But why are we really thinking about error rates so much at all? Of course, the answer is because we're worried about the consequences of an incorrect result. In particular, our focus specifically on the type 1 error rate is because we're usually more concerned about protecting ourselves and the patient population against false positives and against false negatives. So maybe it makes sense to focus explicitly on those consequences and to think about the decisions that we might make either in terms of patient outcomes or in terms of health economic benefits to make those decisions as well as we can. Next slide please.

The health economic approach of value of information analysis allows us to do this. Essentially, in this setting we have a trade-off

between larger and smaller trials. If we do a larger trial, then that will obviously give us more information. It will allow us to better control our error rates, and we can maintain our alpha as a conventional level and ensure good power. But the consequence of that is that the trial will take longer, of course, than a smaller one, and that's a particular challenge in a rare disease because it might be a lot longer. And the longer the trial goes on for, the longer it will be before the patients benefit from the results of that trial.

So we're trading off getting better evidence to make a good decision against having more people more quickly benefiting from that decision, and that balance changes as the number of potential patients in the population changes. So it's different for a rare disease than what it will be for a more common one. And as the population gets smaller, so it becomes optimal to use a smaller size trial, and to achieve this, to use larger type error rates than conventionally used.

As we see, this might be happening anyway.

Trials are smaller, and some of the methods used probably don't control the type 1 error rates at conventional levels. So maybe we should consider this more explicitly to enable us to formalize these things rather than making these rather ad hoc decisions. Next slide please.

The final slide, I'll just end with another quote from the EMA guideline here that says, "No methods exist relevant to small studies that are not also applicable to larger studies." This is kind of true, of course. In many respects, the methodology for clinical trials in rare diseases is not that different to that for other diseases, but in rare diseases in small populations, the challenge is often greater as we seek to base decisions on small samples.

So we somehow need to be more efficient, faster, and smarter, so we need to make sure that we do consider all information as much as we can, and that our designs are as efficient as possible, and that our decision making reflects the patient populations to try to ensure that they benefit from

the clinical trials as much as they can. Thank you.

DR. ROSENBLUM: Thank you, Dr. Stallard, for the excellent talk, and I want to thank all three speakers, Kelley Kidwell, Noah Simon, and Nigel Stallard, for being crystal clear in presentations of somewhat complicated topics. You've done very well on that, and I appreciate it.

We'll turn next to panelist, Greg Levin. Greg will give some feedback, based on the talks, and then we'll go to answer questions with the remaining time.

Dr. Levin is the Associate Director for Statistical Science and Policy in the Office of Biostatistics in the FDA's Center for Drug Evaluation and Research, CDER. He has experience supporting drug review across a wide range of therapeutic areas and has represented CDER on several policy and guidance working groups, including efforts related to adaptive design, master protocols, benefit-risk, and the evaluation of effectiveness, and it's my pleasure to turn it

over to Greg.

Summary of Presentations - Gregory Levin

DR. LEVIN: Thanks, Michael, for the introduction, and I really appreciate the opportunity to participate today. I really enjoyed the three talks, and thought I would share a few brief thoughts on each one, and then make a few additional remarks related to adaptive designs in rare disease settings.

First, Dr. Kidwell gave an interesting talk about SMART designs. I like the ability to incorporate dose finding, which is often inappropriately skipped in rare disease settings. I also think there can be value when incorporating a second stage with re-randomizations, and this will often be better than simply rolling all subjects over into an open-label and controlled extension phase, which is something we often see. The re-randomization can provide data on relevant secondary clinical questions, things like she discussed, for example, to evaluate whether there's benefit in uptitrating to a high dose in subjects

you don't approve on a low dose.

That being said, I also think there are some limitations to consider. Separate analyses of the second stage to answer secondary questions, that's always a good thing, but if the primary efficacy analysis seeks to formally incorporate the second-stage data, then we need to rely on fairly strong assumptions about the stability of the disease, and the lack of confounding, and the lack of carryover effects.

The potential borrowing of external control data that was discussed also adds additional risks and complexities. These may be able to be adequately addressed with careful approaches and in certain settings, but there are challenges there. I think the recent FDA guidance on externally-controlled trials gets into a lot of that.

In the second talk, Dr. Simon gave a nice summary of adaptive enrichment designs, and I think such an approach can definitely have value in settings where there is a targeted treatment with a

fairly well understood mechanism of action that's thought to be more likely to benefit a certain subpopulation of the disease. I agree with the recommendation that there should ideally be one feature with strong a priori evidence considered for adaptive enrichment.

Unfortunately, such evidence can be hard to come by in the rare disease setting, and if it's not available, it may be better to enroll and obtain data on the full population, and then just explore the relationship between the biomarker and the treatment effect.

One additional comment I would make is that the approach may not be as advantageous if the true relationship between the biomarker and the treatment effect is, say, more continuous in nature rather than a step function, and this might be a bit more realistic. And finally, I'd also be interested in learning a bit more about approaches to not just control type 1 error but to provide reliable estimates of treatment effects in the targeted subpopulation.

In terms of the third talk, I very much agree with Dr. Stallard's recommendations to maximize efficiency with simple approaches that are often underutilized, things like analyzing continuous endpoints and using all of the information rather than dichotomizing; adjusting for prognostic baseline covariates; and collecting longitudinal data.

With respect to longitudinal data, I've seen multiple trials in rare disease settings where a short-term control period was carried out. They didn't see evidence of a treatment effect, and then there were post hoc hypotheses that the controlled period was not long enough for the treatment effect to manifest. I know this is challenging, but if there's uncertainty at the design stage about the duration of treatment that may be necessary to provide benefit, having a longer controlled period is really critical to increase the chances of establishing benefit and making the effect of a treatment available to patients.

I also think something like the information

approach to sample size determination that Dr. Stallard discussed can be valuable, and I do want to note that the 2019 FDA Effectiveness Draft Guidance states that in certain settings where flexibility is warranted, like in ultra rare disease where sample sizes are limited, a higher significance level than the typical 0.05 standard may be acceptable if it's prespecified, justified, and agreed upon with the agency. But we definitely have more work to do on a framework, and factors to consider, and trade-offs to consider in making such determinations.

Finally, I'll just add one comment of my own. The type of adaptive design in the rare disease setting that I think is often overlooked and should be considered nearly all the time, is actually the simplest one, and that's the use of a group-sequential design with multiple interim analyses to potentially stop the trial early for efficacy. Trials in rare diseases are often planned with tremendous uncertainty about the potential effect size, and the choice of the

alternative hypothesis that determines the sample size is often pretty arbitrary and overly optimistic, and the trial with a single analysis on a fixed number of subjects may be substantially underpowered to detect smaller effects, but effects that would still be plausible and still be clinically meaningful.

So to address this, a trial can be designed that has interim analyses with prespecified stopping rules, and this can lead to a high probability of stopping for efficacy with smaller sample sizes if the treatment truly has a large effect, but it can also allow the trial to proceed to larger sample sizes and ensure reasonable power if the treatment truly has a more moderate effect size. Such an approach can also incorporate futility stopping so that more resources are invested only if the interim results are promising. So that's something I think should definitely be considered more often. But I'll stop there, and thanks again for the opportunity to share some thoughts.

Q&A

DR. ROSENBLUM: Thank you, Dr. Levin. That answered many of the questions that I had and some of the participants had put into the Q&A and have submitted beforehand.

First, I'd like to thank all the speakers and panelist, Dr. Levin. We have 10 minutes, if I'm not mistaken, to answer questions that have been submitted, so if all the speakers open your video. Great. Fantastic. I'll throw out questions, but feel free to jump in, anyone, with answers.

One of the questions that came up most often -- and this is in more than 100 submitted questions prior to the workshop -- is when are single-arm studies appropriate for regulatory decision making, specifically approval in rare diseases? Actually, one of the live attendees submitted this as well.

It's a tough question. That points towards Greg, and it's, in some ways, not a fair question because it's hard to say anything with generality.

But nonetheless, I still wanted to toss that your way, but feel free to decline, of course, but that does seem to be on many people's minds, at least based on the questions that were submitted.

DR. LEVIN: Thanks, Michael. Yes, it's a difficult question. I guess I would point people towards the recent FDA draft guidance on externally-controlled trials, which typically have a single investigational arm, and then there is some sort of comparison, whether it's to patient-level external data or whether you're in a setting, where there's understanding of the disease process such that there's knowledge that the outcome would not improve in the absence of an effective treatment, and whether that's an oncology situation where it's known that the tumor would not shrink, and tumor shrinkage by a certain magnitude would be considered meaningful.

I think it's a very complex set of factors, that I would point to that guidance and some of the considerations there about both the settings, maybe more fit for purpose, and also some of the

approaches that would be necessary to try to mitigate, to the extent possible, the biases that can be introduced with non-randomized comparisons.

DR. ROSENBLUM: Thank you. That's a very clear response. I haven't read that draft guidance, so I'm going to read it myself after this.

Any thoughts from the three speakers? I mean, it's hard to comment on regulatory --

DR. SIMON: I can't comment on the regulatory.

DR. ROSENBLUM: Just on your thoughts, in general, on the evidence produced by them.

DR. SIMON: From what I've seen, one needs extremely good historical data, and you need to be very, very careful to make sure that things are appropriately matched or that statistical methods are in place to account for the fact that there may not be a perfect match. I love statistical methods, and I am very cautious when they say, "Oh, just do an analytic correction for the fact that you don't have matching populations." That rings

alarm bells.

I also want to say the perfect is the enemy of the good here, I think, and it's an incredibly challenging problem. I think sometimes we need to be a little bit more aggressive, and noting that there are type 1 and type 2 errors that we have to be concerned with.

DR. ROSENBLUM: Those are great points.

Dr. Stallard or Dr. Kidwell, any thoughts? You don't need to.

DR. STALLARD: Not really. I don't think anything about what Noah just said. I think it is something which you need to be very sure that you really understand what's going on from your historical or external data and have very good confidence that that really is comparable.

DR. ROSENBLUM: Yes. That's quite reasonable as well.

Good. Thank you. We'll move on to the next question. The second most common question that came up was about sample size, choosing one sample size, given limited information. I liked what

Dr. Levin proposed about doing group-sequential design to try to handle the large uncertainty about effect size, and also I liked Dr. Stallard's approach for potentially inflating alpha and looking at decision theoretic criteria to decide on the study design and sample size.

Dr. Kidwell, do you have software for sample size selection? Do you want to talk a little bit about that?

DR. KIDWELL: Sure.

DR. ROSENBLUM: It's an opportunity, and any others who want to talk about the sample size question, that also I think is on many people's minds.

DR. KIDWELL: Sure. Yes, sample size, we all have that question when we think about clinical trials, and especially in rare diseases when we know that the numbers are so limited. We have Rshiny applets that are available via if you search my name, I think. I guess one is in the draft stage. It should be out very soon. They ask for information that we think statisticians, along with

clinicians, could easily, or somewhat putting their heads together, come up with and put in there, any try to provide some guidelines; mainly, expected treatment effects.

We ask some information about linking the data from stage 1 to stage 2, and then what's the probability, sort of asking about power and type 1 error, but since we're in a Bayesian framework, we phrase it slightly differently. But it's very similar in terms of what power do you want, and how much are you willing to allow for an incorrect conclusion.

So we hope that that helps facilitate the use of these snSMART designs, as it can be intimidating, if you're not used to the Bayesian framework, to think about sample size and/or how to use these. So we've created both, on the front end, design software, and then also an R package on the backend to actually analyze the data to help with the whole process.

DR. ROSENBLUM: That's great. Thank you.

Any other thoughts on the sample size

question?

DR. STALLARD: First of all, really a great report in what Gregory said about the use of group-sequential designs here and allowing stopping as early as possible. I think there's also a question in the Q&A specifically about using Bayesian methods in group-sequential design whenever that's possible. And the answer is yes; there are quite well-established Bayesian sequential analysis methods only if they're stopping in terms of the amount of posterior information or controlling frequentist type error rates using a Bayesian stopping rule. So yes, if you want to be Bayesian, you can also still have rules which allow you to stop early.

DR. ROSENBLUM: Great. Thank you.

DR. SIMON: I'm a naïve guy, so I'm going to make a kind of naïve comment here. I think if you have a blockbuster, you don't need a big sample size, and if you have something that's pretty incremental, at least my feeling -- from engaging with some of this stuff, and other people may feel

totally differently -- you may never have the sample size in a very rare disease if it's a really small improvement to identify, where in oncology, you could make those incremental advances because you can enroll a lot of people in your trial.

Again, I definitely agree with Greg's statement. If you have a blockbuster, you can run a group-sequential trial and maybe identify it very quickly, and/or maybe say, "Ooh. We thought it was a blockbuster, but it's not." I know, of course, everyone wants a treatment that's a blockbuster, and we don't always have it.

I like adaptive enrichment because maybe it's a blockbuster in some subset of people. Now, you don't want to fish for that subset; you want to maybe be able to identify it beforehand, but I think there's obviously value in very carefully identifying what the sample size should be, but it seems like you do need a lot of flexibility there, and maybe to go in noting that you're likely going to only find blockbusters in some of these cases.

DR. LEVIN: This is Greg. I just want to

follow up. I agree with Noah. Unfortunately, drugs that have very small incremental improvements, it's probably going to be challenging to identify treatment effects in rare disease settings.

I do think there's kind of a middle category of non-blockbusters that have moderate effects. A lot of times I see hypotheses going into trials, but it is a blockbuster. There's always the optimism that it's a blockbuster, and that leads to a fixed sample size that will be large enough for power for a blockbuster effect size, but would be substantially underpowered for a more moderate but very meaningful and still plausible effect size. And that's where I think allowing the trial to go to a larger sample size, if results are promising but not persuasive, that an interim analysis can be valuable, perhaps in combination with something like an enrichment approach, if there's a targeted subpopulation that's realistic and there's evidence to support.

DR. ROSENBLUM: Fantastic.

We'll stop here, and I want to thank, again, all of the speakers and panelists. I learned a lot. As a participant, if you enjoyed these presentations, which I'm sure you did, please come back in 10 minutes. We have our second session of the day, where the focus is on analysis methods for small populations.

Thank you to all the speakers, and we'll return in 10 minutes.

(Whereupon, at 10:26 a.m., a recess was taken, and workshop resumed at 10:36 a.m.)

Session 2 - Michael Rosenblum

DR. ROSENBLUM: Welcome back, everyone, from the break. We'll now start Session 2 of today's symposium. The focus is on Analysis Methods in Small Populations, and we'll be discussing different analysis methods and also differences in the target of inference, also called the estimand, the assumptions required when you have small populations, and also performance of different analysis methods.

We'll have three panelists and a discussant,

the same structure as earlier this morning. Each panelist will be for 20 minutes, followed by some discussions of feedback by the panelists, and then we'll have Q&A, and please put all your questions in the Q&A. Let's get started.

We're honored to have Dr. Karen Price as our first presenter. Karen Price is Associate Vice President and Statistical Officer at the Statistical Innovation Center, which focuses on innovative design and analysis of clinical trials. This is at Eli Lilly and Company. She has a wide range of research interests and expertise, including Bayesian methods, innovative clinical trial design and analysis, and quantitative decision making, and I'm excited to pass it to you, Dr. Price.

Presentation - Karen Price

DR. K. PRICE: Thank you so much, and it really is an honor to be here to give this presentation. It's been such a great series of sessions in all of the presentations. I've learned a lot, and I'm just honored to be here.

What I'm going to do is give an overview of Bayesian approaches and master protocols, if time on the master protocol front. I think it's great. A lot of what I'm going to talk about, especially up front, was touched on in the earlier session, so I think, really building on one another, there may be points that I will just re-emphasize and/or go into in just a little bit more detail.

On the next slide, before I get going, I wanted to give some acknowledgements to several of my colleagues who have been involved in much of this work over the years and helped provide slides and/or were involved in the design in some of the trials I'll talk about.

So in the next slide, what I will do is give a super quick overview of rare and pediatric diseases just to really touch on and motivate the Bayesian framework, very much in line with how Dr. Kidwell did earlier. Then I want to go into motivating a Bayesian framework, but probably spend the majority of the time on the third item, showcasing some examples of Bayesian applications,

and then if time permits, conclude and touch on master protocols.

From the next slide, again not needed really for this audience, but on these next two slides, obviously the key here and what is motivating the use of a Bayesian approach in this instance is that we are dealing with low prevalence of these diseases. On the next slide, I think it's important, of course, to note that it's also a very vulnerable group of patients. Many of these are children, and of course we have a very low number of approved treatments for many of these diseases.

So we're definitely motivated. It's incumbent on us to use all that we can from this data to really squeeze everything that we can from this really valuable data, whether it's external or internal to an ongoing trial, so that we can make better decisions and get these compounds to these patients as quickly as possible.

So with that motivation, I wanted to talk about, then, the Bayesian framework, so maybe a couple of slides. I like to motivate Bayesian

thinking in a couple of different ways. I think there's been a lot that's been discussed about the fact that humans do struggle with prediction and uncertainty. We do have a tendency to over-index information, and we really need in place quantitative approaches to help us frame the available data, especially when it's diverse and coming from multiple sources.

We can think about a Bayesian approach, as is depicted on the right side, and you saw a similar thing, again, in Dr. Kidwell's discussion earlier, that this is very much about a continual learning process. This is a very natural way of thinking. We do this all of the time in our decisions throughout the day as we utilize our previous experience, but when we move into business and scientific decisions, the information we have is in the form of data, so we need a platform to synthesize that information. Bayesian methods will provide that so we have what we knew, and that's the prior distribution; what we see, and that's our likelihood; and then what we now know is our

posterior distribution, and that's going to continually be updated.

On the next slide, we can get a little bit more motivation on this as we talked about, and you'll hear throughout my discussion today it's very much about iterating upon the science and allowing a continual learning process. Today's posterior is tomorrow's prior, and when the facts change, I change my mind. So this is, again, a very natural way of thinking and updating the science. It provides rigorous integration of what we know already within the analysis of new data so that we can shed light on what we don't know.

I think one of the things I really wanted to emphasize here is the transparent nature of a Bayesian approach. Whenever we look at data and we're analyzing it, we do bring to the table our previous experience, and that is weighing into how we are thinking about that specific decision. But a lot of that then is done without awareness of others who are also viewing that data, so a Bayesian approach can allow for that transparent

understanding of how other data sources are being brought to bear in various decisions, and can improve and allow for more efficient decision making.

So it's a very transparent approach. It does allow for straightforward statements of probability and uncertainty, so I'll talk in a little bit about, again, a Bayesian interpretation is very straightforward and typically what people would, I think, prefer, and wish that they could interpret a p-value that way. A Bayesian design can help reduce sample sizes or study durations. Additionally, there is a tremendous flexibility through hierarchical modeling, and now with computational conveniences, we can fit a wide range of models and synthesize information in ways that we couldn't historically.

So with that, on the next slide, as I've mentioned already, I think bringing a Bayesian approach to bear here is very important in this rare disease setting, where we do have the small sample sizes, limited data, and few treatments,

that a Bayesian approach is going to allow us do more borrowing of information and increase precision, so again, it's providing that mechanism for us.

With that motivation, I wanted to move into the next slide. One more slide. Going into a little more detail. Again, some of this you heard in the earlier session, so let me maybe just talk about it in a little bit more detail, and I wanted to show some examples because I think it's important leaving here understanding that it isn't a scary thing, that it isn't a black box, but you can have an understanding of how does this borrowing impact inferences.

Borrowing approaches, as you heard, there are two main types, a static type of borrowing as well as dynamic borrowing. Some examples of static borrowing would include pooling, single-arm trials, and also power priors. Dynamic borrowing examples include hierarchical modeling, mixture priors, commensurate priors. There is an appeal, as was noted as well in the earlier session, to dynamic

borrowing, at least bringing that to bear, as it can borrow more when the current data are more similar to the historical data and can help protect against over-borrowing and some of the errors that were discussed in the earlier session.

On the next slide, as we're going to look at a couple of borrowing approaches, thinking about what some data sources are, I always like to mention expert and caregiver opinion in the rare disease setting. These are diseases that people devote their lives to understanding and live with day in and day out, so there's just a tremendous wealth of knowledge that can be brought to bear. In a moment, I'll touch on that just a little bit more, but certainly a really valuable use of information that we should bring to bear when we're thinking about trial design.

Natural history studies, summary level data, individual patient data, PK/PD, preclinical, many of these were discussed yesterday, so really what we're talking about today is maybe ways to better utilize this type of information. Of course, any

time we would then be looking to use the Bayesian approach and bring historical data to bear, we need to be thinking about its relevance, similar indications, patient population, the relevance of the endpoints, and so forth.

On the next slide, as I mentioned, I did want to touch on the role of opinion or expert knowledge. Just so you're aware, if you're not yet, there is large literature on this topic in eliciting beliefs about endpoints; there is a lot available. There are formal, well-tested protocols for eliciting distributions about belief, and all of the methods we would talk about then can be applied such that one can downweight that information when thinking about it relative to maybe formal data that would be coming in.

We've used it quite a bit at Lily in a variety of settings. Oftentimes it's to help us in the design of a trial, to be thinking about maybe relationships between endpoints, or doses, or populations. It has also been used to inform about the relevance of historical information. So in a

case where maybe we want to borrow some adult data in a pediatric setting, partnering with a patient has been used and published to help in thinking about how much to borrow.

There are examples available. MYPAN is a rare inflammatory disease in children, where prior elicitation was used. Again, I think there are a lot of unexpected benefits. I've heard others talk about the unexpected benefits on elicitation that may not be fully used and could help with us as we're thinking about designing trials, and really setting things up to be as efficient as possible.

With that, on the next slide, just some general comments about borrowing, and then I'll show an example. Of course, there are some questions about how much to borrow and things to think about, and what data is eligible to be included we need to understand. I won't go into detail about simulating operating characteristics, but that's a really important element. Considering things such as prior effective sample size or the prior probability of success is important, and of

course, understanding prior to posterior sensitivity. We may borrow different amounts for different treatments based on the medical need and so forth.

Just one thing I'd like to mention is sometimes there's a feeling that borrowing is brought to bear to help favor a sponsor, but in fact it's really just getting at what is the most likely effect here, and getting the best estimate of the true event we're interested in, so to dampen it but, again, seeking to get at get what that true effect is.

Next slide. What I wanted to do next is go into an example. This is a hypothetical example. I presented this elsewhere, but just to show what happens on the backend when we analyze data in this way, looking at a static and a dynamic prior. Suppose that we have previous data on a control group -- and it could be a trial, it could be a set of trials, whatever -- and somehow we've synthesized this information. We have 120 subjects with 72 responses, so the historical rate is

60 percent. What we're going to look at is keeping that historical rate constant, and then what we're going to do is say, suppose we have a future trial looking at 70 for the control and 140 in the new treatment?

Next slide. What I want to show here is just looking at what those priors look like for the power prior versus the mixture prior. On the left hand is a power prior. The power prior amount of borrowing is governed by this value a_0. The a_0 can range from 0 to 1, and you can see as you go across and sort of down, as a_0 goes from 0 to 1, you have the non-informative prior up to, really, the whole of the prior at a_0 equals 1 or borrowing the entire prior.

Then on the right-hand side, we have mixture priors. This is just an example of, again, that same prior, where here the p is governing the amount of borrowing from the informative part of it, so we can compare. Just taking the blue, for example, you can compare a_0 equals 0.25, and you can see that, versus on the mixture side with p equals

0.25, you can see the difference in the weighting of the non-informative portion of this mixture prior is represented there in the graph.

All I wanted to highlight in the next couple of slides is how does this affect what we observe on the backend. This is showing for a power prior, so suppose on the left-hand side, we observed 20 out of 70 in the control arm. What will happen is you can see the prior is the blue, which on this slide is the rightmost distribution. The likelihood is the red, so then the purple lands in between, pushing forward, in this case, the inference to be closer to that prior and downweighting that observed data. On the right-hand side, you then see suppose the observed data is higher than the prior, and, of course, the posterior then would end up in between.

Then we can contrast that, on the next slide, with the mixture approach. Here, very similar observed cases, and the purple you can see then on the left-hand side is now shifting more, and it's coming more towards the likelihood, and

you can see on the right-hand side, again, similar. So in this case, where there's mismatch, the posterior is going towards that data. So that's the interpretation of results or what results look like with a couple of different priors.

I just wanted to, then, touch on one more example here on the next slide, again, to say how this works and to emphasize that it's really quite straightforward and, again, intuitive. Here's an example where we have an open-label study. It could include an active reference arm, but what we're really interested in is analyzing and looking at that test treatment.

Suppose a control isn't feasible or it's unethical. What we would use here is maybe this historical information to establish what a meaningful response would be, and then we need to set up our decision rule. In this case, suppose that we have utilized this historical information and we've determined that 57 percent is the effective interest and that 80 percent is the probability threshold. Then all we're doing is

we're just setting up a decision rule that says the probability that that response rate is greater than 0.57 needs to be greater than 0.8.

On the next slide, we can apply this. Suppose we have a study with 30 patients in juvenile idiopathic arthritis. In this case, we're just showing we aren't using a formerly informative prior on the treatment arm, but to say, okay, we've observed now 20 out of 30 responders at week 24 in this trial, so our observed response rate is 0.67, roughly, and then on the next slide, all that we need to do -- and it may not come over in this form. I'm not sure if it builds.

Do you mind to go to the next to sort of build that out? Okay. It does build.

So basically, all we're trying to do is look at that effective interest, calculate the area under the curve to the right of 0.57, and that's 85 percent. On the next slide then, applying that decision rule, as success criterion has been met, so that's how that simply would work.

The final thing in this part, on the next

slide, I wanted to touch on is predictive probabilities. Again, we talked a lot about this being a continual learning process, and in particular focusing on using information that may be external to the trial. Obviously, we can use information within the trial and make decisions as the trial progresses.

Again, this was touched on in the earlier session, but basically the idea here is just showing an example where suppose we have a final sample size of 100, and we're going to declare success if the probability of a response being greater than 0.5 is greater than 96 percent. Then this is just showing how those predictive probabilities of being successful at N equal 100 progress, so we can see early on that predictive probability is 54, and then as we progress, it becomes very small. We would establish a decision rule early about what predicted probability would be -- if we saw that, then we would terminate the trial to begin allowing for that continual learning.

I know I just have about another minute or two left, so let's go ahead and jump real quick then. I wanted to touch maybe on this next slide, and then I can close here.

Master protocols are a great way to look at multiple indications or multiple drugs, and can be very, very efficient in many settings, but in particular, in the rare disease setting. Bayesian methods will often feature heavily in these and, of course, it also is allowing for more consistent data, getting back to these data are then collected under very similar circumstances, same timing, similar inclusion/exclusion, and so forth.

This is an example trial that we did in oncology at Lily. We had two rare oncology tumor types, and randomized patients to the drug or control in these two indications. So at the end of the day, we were able to bring in real-world information and prior elicitation as well, and use a dynamic model to analyze this data, being able to more efficiently use the information between these two tumor types.

The final on the next slide, this is an example of a platform type of trial. This is in pediatric IBD. Patients are randomized to a drug, and then they follow a similar schedule of activities. Again, in these cases where there are very few patients, we want to have these trials together so we're getting same endpoint, same time points, and able to really use that information together.

With that, I would like to conclude on the next slide just to say Bayesian design analysis can help facilitate rigorous incorporation of the relevant data, especially in settings with potentially limited sample size. Again, as I mentioned, it's prespecified, it's transparent, and needs to be studied via simulation to really understand, before the trial is conducted, how this will play out. It can result in an increase of power. It, in some cases, can maintain low type 1 error, although it will often be inflated, but there are methods to help mitigate against that, and very much around efficient continual learning.

Designs such as master protocols can help enhance this learning. In all of these cases, the collaboration between the sponsor or regulatory is very important to statisticians, others, and bringing that patient experience to bear is critical. We just need to continue to have experience with these designs and analyses to continue advancement.

DR. ROSENBLUM: Great. Thank you, Dr. Price, for the excellent presentation.

Our next speaker is Dr. Jack Lee, Professor in the Department of Biostatistics at the University of Texas, MD Anderson Cancer Center. He's also the Kennedy Foundation Chair in cancer research there. His areas of expertise include design and analysis of clinical trials, Bayesian adaptive designs, statistical computation and graphics, drug combination studies, and biomarker identification and validation.

I just met Dr. Lee at the previous ENAR in person for the first time, and I'll pass it off to Dr. Lee.

Presentation - Jack Lee

DR. LEE: Thank you so much for that kind introduction, and I also thank you for the opportunity to share with you the Bayesian Adaptive Design and Information Borrowing for Efficient and Accurate Statistical Inference in Rare Diseases.

Next slide please. I will talk about the statistical challenges and solutions in drug development for rare diseases. I will give a brief overview of the Bayesian statistical inference, then give some examples of the clinical trial design and analysis, and finally give concluding remarks.

Next. As shown here in this dart-throwing example, in the top-left panel, the result is not accurate nor precise. In the top-right panel, the result is accurate but not precise. In the bottom-left panel, the result is precise but not accurate. What we want to be is in the bottom-right panel, where the result is both accurate and precise. So we need to have methods that can reduce bias and also increase the

efficiency.

Next slide. In drug development, we are facing many challenges, shown on the left, and we propose some solutions on the right. For example, we know that randomized controlled trials are gold standards, but then it requires a large sample size, and hence not feasible in rare diseases. So what's the solution? The solution is to have novel adaptive designs.

For example, we can take all-comers with adaptive randomization to put more patients in better performing arms. We also need to implement more frequent interim analyses. Lastly, we need to make the study enrollment and conduct easier. It's been discussed that single-arm trials are subject to bias because there's no comparators, and it's hard to make robust inference.

So how can we do better? We can borrow information from the concurrent control or historical control. As discussed yesterday, nowadays we have many good registries and EMR data, and these are large sample sizes that often times

comes from heterogeneous groups with mixed data quality. So how can we better use real-world data to turn real-world evidence? We need to have a clever way to form synthetic controls, and we can do propensity score matching and network meta-analysis.

Next. This slide just shows you a simple example of how statistical inference can be made. Let's assume we want to estimate the unknown response rate for a new drug and we conduct a phase 2 trial with 30 patients. At the end, we have 14 responses and 16 no responses, so the point estimate of the response rate is 0.467. The bottom left-hand panel shows how the posterior probability of response rate is updated. The gray line in the background shows how the previous step was done. And at the end, we have the red line showing the posterior probability of the response rate. As it can be seen, as the trial moves along, the distribution picks up. Why? It's because we have more data; therefore, we have a more precise estimate. On the bottom-right panel, it shows the

frequentist 90 percent confidence interval.

Next. With the Bayesian posterior probability, we can do many things. For example, we can calculate what's the probability of a response greater than 0.3, 0.5, or 0.6, as being shown in the figure with the area highlighted in red. In the bottom panel, you can also see that we can calculate the probability that the null hypothesis is true or probability of the alternative hypothesis is true. With that, we can calculate the odds and compare the posterior odds over the prior odds to form the base factor. In this case, after we observe data, we can conclude that the alternative hypothesis is 17 times stronger than the null hypothesis, based on the data.

Next. In the Bayesian inference, all information pertinent to the parameter of interest is contained in the posterior distribution, so in our case, we start with data prior, and then we can model the unknown parameters with statistical distributions, and then we can properly address

various levels of uncertainty. In this talk, I want to emphasize two things; that the Bayesian method allows us to use all available information -- prior, current, and even predict into the future -- to make decisions, and we can use information within and outside of the trial via dynamic borrowing to increase the study efficiency and also allow more frequent monitoring and decision making.

You can incorporate subjective utility in the decision making; for example, like the toxicity and efficacy trade-off, which I'm not going to talk about in my talk. But we need to be aware if the data and models are compatible, then where they are biased due to some data heterogeneity, we also can do the sensitivity analysis to evaluate how robust the inference by varying priors.

Next. This slide shows us the current status and the enhancement of clinical trials. In current status, we conduct a clinical trial with one drug, one study population, and one trial at a time, and we have discrete-phase drug development,

starting from phase 1, then pause to go to phase 2, and then pause to go to phase 3, so it's not efficient. As Dr. Karen Price mentioned, we can use master protocol with a seamless phase design, umbrella basket, or platform trials, and we can also include adaptive randomization to put more patients in more promising arms.

Again, the two important things in data analysis is that we should do more frequent interim analysis if the drug is too toxic, or not promising, or highly promising. Then we can have early stoppings, and then we can borrow information from historical data, or across subgroups, or even across different trials.

Next. I just want to quickly mention about one adaptive design that we developed called the Model-Assisted Design, which is the design that uses a Bayesian framework. We can have precalculated decision rules such that the design is very easy to conduct, that includes the Bayesian Optimal INterval design, or BOIN design, for dose finding, and also the Bayesian Optimal Phase 2

design, or BOP2 design, for complex endpoints. Notice the font in the red color that shows all these designs allow us to incorporate historical data as informative prior.

Next. This slide shows you that at MD Anderson, we not only develop methods, but we also provide software. We have more than 30 freely available shiny apps that allow you to design and run clinical trials. All you need to remember is our URL at the bottom, trialdesign.org.

Next. For example, we have a family of BOIN designs for single agent, for combined agent, and for late-onset toxicity, and we also have design, like BOIN design, BOIN12, to allow us to find the optimal biological dose.

Next. This schema shows us how to choose the right design. On the left, you can see that if you want to find OBD, then you can use only one tool, U-BOIN. If you want to find MTD, you can use iBOIN, TITE-BOIN, or BOIN combination. The one highlighted in purple, I just want to highlight the importance that we use the prior data or historical

data.

Next. This is a simple diagram for the BOIN design. All you need to do is determine the target probability of toxicity ϕ, and calculate the observed probability of toxicity p_j at dose level j, and compare with the prespecified boundaries to determine if either dose should be escalated, retain, or de-escalated. The iBOIN design can incorporate prior data, as you can see on the next slide, but I want to make a really important plea, that it is long overdue to abandon the commonly used 3-plus-3 design because there are many better alternatives.

Next. So as can be seen here at the left panel, we can see that we have different doses and have different expected probability of toxicity, but then we can put the effective sample size that's corresponding to historical data and how strong the historical data is, and then to model that when we do the decision. On the right, you can see the decision diagram, and with low probability of toxicity, you escalate, and high

probability of toxicity, you de-escalate the dose, and if it's the middle, then you retain.

Next. At the end, you plug in the number of patients treated and number of dose-limiting toxicities observed, and can calculate the MTD with its confidence intervals. As you can see now, the dose chosen is the dose level 3 MTD.

Next. This slide shows that we have different types of model-assisted designs for a phase 2 trial as well, and this we call BOP2 Suite.

Next. The BOP2 design allows us to run the phase 2 trial with a unified framework with different endpoints. In particular, depicted on the right, you can see that, for example, in the control arm, the probability of efficacy is assumed to be 0.3, but we have data with the prior efficacy showing that it's corresponding to a sample size of 20, and then for the experimental arm, we expect real efficacy, 0.5, but we don't have much data, so we only put the prior effective sample size of 1.

Next. This shows you that under the BOP2 design, you can have different endpoints; like in

example 4, we have objective response and toxicity as endpoints, and we have a different stopping rule. For example, with 10 patients, if we observe two or less responses or five or more toxicities, then we stop the trial and declare the drug is not working. This shows you the importance and feasibility of monitoring the trial during the interim.

Next. This shows a graphical presentation. On the left, you can see how the trial evolved in terms of responses, and on the right, in terms of toxicity. The green zones are the go zones, and the pink zones are the no-go zones. In this case, the trial stopped after we observed 35 patients because the drug was too toxic.

Next. Just to quickly go over the platform design for adaptive enrichment, since it has been discussed by previous speakers, I'll quickly run through the study schema.

Next. The idea for the adaptive platform design is that we can have the control group as the backbone of the platform and experimental

treatments as modules, which can be plugged in and out of the platform.

Next. For example, if we want to compare experimental 4 with control, we can either use a concurrent control, shown in the light brown, or a historical control, shown in purple. Each one has its own merits and also disadvantages, so we just need to use them carefully.

Next. This shows you the normal design with master protocols on the left. This phase is an exploratory kind of design analysis to try to find the signal. After we find the signal, then we can design the confirmatory trial with more focus on the phase 3 trial in the selected patient population.

Next. My last part of the talk will cover the Bayesian hierarchical model for synthesizing information for the subgroups analysis in basket trials. Because a clinical trial often has different subgroups, how do we model them such that the information can be borrowed? The Bayesian hierarchical model can synthesize multisources of

real-world data.

Next. To increase accuracy and also increase the efficiency if we do the right borrowing.

Next. This is just to depict that we have five subgroups and different age groups. On the left, you can see the prior distribution of the response rate, and on the right, you can see after borrowing, showing in the red, the posterior distribution. The information tends to move to the center and peaks up.

Next. This shows you that if we apply one of the methods we call BaCIS, then you can classify these five subgroups into two clusters, then borrow information within the clusters.

Next. In cluster 1, we have 2 arms or 2 subgroups, and you can see the red curve is higher than the blue. What does it mean? It means we have more information,

Next. Cluster 2 has 3 arms or 3 subgroups, so then, as you can see, the information is closer to the center of the three, and the red curve is

higher than the blue curve.

Next. We can also use another package called BCHM, Bayesian classified hierarchical model, and in this case you will determine the number of clusters automatically, so in this case we formed 3 clusters.

Next. In conclusion, statistics can help us in extracting signals from the noise in the data to avoid bias and increase efficiency. There's no free lunch, but there are some lunch specials if we apply some novel design and analysis. The Bayesian paradigm takes the "we learn as we go" approach, and is particularly useful in rare diseases because it allows flexible, adaptive, and continuous learning, to naturally and easily incorporate and synthesize all relevant information.

Bayesian adaptive designs are efficient and robust in the drug development process, but there is one caveat. All signals found need to be validated in prospective trials, so please work closely with statisticians from beginning to end, and apply rigorous statistical methods to maximize

the success of every project. Thank you very much.

DR. ROSENBLUM: Great. Thank you, Dr. Lee, for an excellent presentation.

I'll turn to our third speaker. Dr. Rima Izem is Director of Statistical Methodology in analytics at Novartis. She has expertise in regulatory statistics using causal inference for comparative safety, signal detection, and survey research. She also has experience and expertise in comparative effectiveness in rare diseases at Children's National Research Institute. She works with real-world data, including claims data, electronic health record data, international registries, and electronic clinical outcome assessments. We're honored to have Dr. Rima Izem, and I'll turn it over to you.

Presentation- Rima Izem

DR. IZEM: Thank you so much. I've really enjoyed the workshop so far. My talk will switch gears a little bit from the previous, although we will continue discussing design and analysis that tries to get as much efficiency from the

participants in the rare disorders. My talk will focus on leveraging longitudinal data, or in other words, leveraging time.

Next slide please. As this is the last talk of the session, I wanted to put the take-home messages right away, and those are the following. First, there are multiple ways to incorporate randomization in your study design beyond a parallel control, and a lot of these designs exploit within-subject comparison, or what I would call sometimes self-control, and that could be used to establish efficacy or safety.

In the same fashion, there are also multiple observational study designs beyond the cohort study, or even single arm with an external control, that tries to leverage longitudinal data or repeated measure. However, anytime you move away from randomization, you have to control for multiple sources of bias, like a confounding and selection bias.

So why include time or why include within-subject comparison? Because it has a lot of

advantages compared to between-subject comparison, in that you're increasing your analysis unit, you're reducing potentially outcome variability, and if you're using an observational study design, you're reducing confounding compared to between-subject comparisons.

Next slide please. With that, this is the outline of my talk. I will spend some time discussing how to implement randomization, maybe slightly differently than the speakers earlier today, but maybe less novel; that's my disclaimer. I use this as a motivation before getting into observational study methods because some of them, to make them more rigorous, will try to emulate a randomized design with an observational study, and then I'll finish with some design and analysis consideration.

Next slide please. This is basically my outline and my take-home messages all in one slide, but also showing the running thread as I move from randomized to observational studies. In the left-hand side, I show randomized study designs

that, again, with randomization, you get a lot of benefits; that is that you control for all confounding, and also you have a really good idea of time zero for a lot of these designs. The parallel arm is probably the most commonly used, where you're randomizing different people to different treatments; however, you can see also that there are other designs that I'll go into with a graphic in the next slide.

On the right-hand side are observational study designs. Again, the top one is probably the most commonly known, is the cohort study, or using an external control arm, but the bottom two may be less known, but they're still useful. Now, the reason I have arrows going from one to the other, it's not really to say that they're equivalent, but rather that if you are going to use observational studies like self-control, case series, or cohort study, or sequential control for confounding, it's helpful to keep in mind what would have been the randomized study equivalent, or what would have been ideally, hypothetically, the randomized study

that would correspond to this observational study. The reason you do that is that, in itself, with already existing data, emulating a randomized study design is a way to control for multiple sources of bias. And again, I'll be following the same outline of the paper that's shown in the bottom.

Next slide please. One more. This is my favorite graphic. It's a little bit dated now because it doesn't include a lot of study designs that you've heard of before this talk, but I really like it because it shows you the diversity of study designs that you can think about, and especially how to incorporate randomization in your design.

In A is the parallel group design, but then in B and D are designs that rely solely on within-subject comparisons. In the crossover design that was mentioned earlier today, each person receives both treatments, but what's randomized is the order in which they receive the treatment. So again, it's a randomized design, and yet each person gets to test both treatments.

In the N of 1, it's kind of taking the

crossover to the limit. You only have one subject, but you have multiple treatment periods. In this case, that's what the N in N of 1 refers to, and then what's randomized in each treatment period is which treatment will be received in that treatment period. So again, that's a randomized design with only one patient.

In the bottom in blue are designs that are not only relying on within-subject comparison, but also between-subject comparison; however, they use kind of both. For example, in the randomized placebo phase, what's randomized is whether the patient is receiving treatment right away, after they enter the study, or after a placebo period they will receive the treatment.

In the randomized placebo, and the stepped wedge, and the randomized withdrawal, all of the patients receive the treatment, but what's randomized is when they receive it and how long they receive it. So in the analysis stage, the analysis unit is a period rather than a patient, so you're using both within-subject and

between-subject comparisons.

Next slide please. Just to show you that these designs are feasible, but also that they can lead to significant finding, or statistically significant finding, I'm sharing here an example from urea cycle disorder. This was an N of 1 design. Here it used 6-week periods, so there were 6 weeks that were split in 3 pairs, and in each pair, the patient was randomized to either receive placebo or active within a given week. This was one particular patient that had a mutation for this rare disorder and had to have a washout period prior to entering the study.

There were three outcomes of interest. One was the patient-reported outcome, a questionnaire score to ask about their symptoms, but also there were two laboratory measures of their well-being. Although there were only 6 weeks worth of observation, at least in this study, it was sufficient to show that the treatment of L-arginine significantly improved symptoms, as evidenced by the questionnaire score, but also their glutamine

level.

Next slide please. If we try to both use within- and between-subject comparisons, there's also another example also in urea cycle disorder that recently used a design that used a hospitalization episode for hyperammonemia as the analysis unit rather than the patient. Every time the patients that were participating in this study were admitted into the hospital, that was considered an analysis unit that was either randomized to receive CARBAGLU or to receive placebo. With only 24 patients, there were 42 analysis units in CARBAGLU and 48 in the placebo arm.

What we're showing on the right is something that was shared in the label for this drug. We see how over time in the horizontal axis, the proportion of events that occurred in both the treatment arm and also the placebo arm, and in the events here, higher is better, so it's a responder rate. And we see that the response rates, or the two curves, kind of differentiated after day 1 off

hospitalization, and it continued to differentiate as time goes on. So it is possible to think of the analysis unit as the treatment episode rather than the patient.

Next slide please. In summary, there were a lot more studies that were cited in some of the examples that I showed earlier, that you can go back to, to show not only a stepped-wedge design, but also early withdrawal or delayed therapy. The main advantages of using the designs is, again, increasing the analysis units, maybe reducing variability, and then also getting information on the natural history of the study. However, these designs are not always feasible, and what's the tricky part is to figure out the timing or the duration of these time periods. They should be long enough to observe a change in the outcome, but they shouldn't be too long. They should be short enough to assume that these time periods are independent or may be including some washout period.

Next slide. I'll move now to observational

longitudinal design, and I apologize, but I'll have to skip some of the slides as we go through this.

Next slide. Next. Next slide please. I skipped the cohort design, and that was an example that you can refer to, and I'll be happy to answer any questions you may have on that example during the Q&A, but I wanted to share a few thoughts on using the single arm with external control.

Whenever you talk about using or leveraging observational studies, that's probably the first idea that comes to mind, at least for my collaborators in academia, but also industry. They think about conducting a single arm, and then using the observational study instead of randomizing their control. That is, frankly, a very challenging approach, especially in the regulatory setting, and we've reviewed, actually, a lot of applications that tried to make the argument for using an external control, and this is the paper that's referred in the bottom.

The difficulty with using external control is that it's very hard to show that two different

data sources have similar information. It's very hard to show that there is a comparability of population, treatment, outcome, and frequency of assessment, and that the start and end of follow-up was similar. It's also hard to control adequately for confounding because sometimes the information on confounders is not even collected, or not collected sufficiently frequently, and it's hard when you have to tune the control for confounding methods to prespecify everything at the beginning. So there are more exceptions than rules in using observational study of an external control.

Next slide. I'll skip this example, but this was an example of actually the exception. The only thing maybe that I will say here before our skip this slide is what's shown in the label on the left is the effect size, and if you see what effect sizes we're talking about here, we're using an external control that's huge. So definitely, when the effect size is really large, maybe it's worthwhile to discuss having an external control.

Next slide please. I've moved very quickly

for a cohort study and use of external control because I wanted to spend some time to discuss self-controlled case series, and also maybe sequential control for confounding. The reason I wanted to spend time on those is because I don't think they're used enough, or they're not thought of enough. Especially early on when you're doing proof of concept, they could be leveraging already existing data in your natural history studies to see whether maybe there's a treatment that's promising. It's helpful for those studies, to think of them as the observational study equivalent of crossover, or N of 1, or a sequential randomization study.

Next slide please. What are some of the questions that you can ask yourself if you already have a natural history study like some of the ones that were discussed in the workshop yesterday? Can the unit of analysis be subject time rather than just subject? Can the duration of the look-back, that is the medical history, instead of it being static, can it be dynamic? Can you actually look

at it longitudinally before diagnosis?

Similarly, do you collect information on potential confounders sequentially over time or do you collect this once and for all at baseline when people enter the study, and then you don't look at it again? All of these can inform whether a self-controlled study, or something that augments self-control, with between-subject comparison is good for you.

Next please. One successful example that will leverage existing data is the alpelisib, and this is a relatively simple self-controlled study where each person was serving as their control, but you only have two periods, a pre-period before the treatment and a post-period after the treatment. The data here that was used was from a compassionate-use program, so this product was already approved, and this was just to expand the indication for this new population.

There were already some findings that supported that maybe this treatment was probably effective, so this study kind of formalized this

comparison of the pre-index period to the post-index period. Although the outcome of the responder was not apparent right away, just by calling it a responder because you're comparing to baseline, it is a pre-post comparison that's determining whether this treatment is working or not. So again, just by using pre-post comparison in this case when there was a lot of information already available on this drug, you were able to see that it would work for this rare disorder.

Next slide please. One way to generalize this and this type of design, this self-controlled case series, is not really common in rare disease, but it could be used. It's used a lot in postmarket safety to try to leverage observational data for rare outcomes, but I think it has promise. And again, for me the self-controlled case series is the observational study equivalent of the N of 1 design, although you could have more than one person, obviously, if you have multiple periods for each person.

The study is anchored at the first exposure

to therapy, and I'm showing here the exposure period to a particular test drug in yellow; however, you can still use information prior to this first exposure therapy, especially if there was standard of care or exposure to placebo. The periods may have to not be exactly contiguous if the test drug can have an effect beyond when the subject was exposed, so you may need to include washout period. But the point here is that if you have a transient exposure, where a person is receiving the treatment, not just once but multiple times over their journey, then that could be exploited in this analysis to see whether the test drug is effective compared to the placebo drug.

Next slide please. A little bit more complex to explain design is what I call sequential cohort entry. Just to give you an idea of how this works, I'm going to be using an example that I worked on in urea cycle disorder. In this example, the interest was to leverage natural history study data from hundreds of patients to see whether liver transplantation was improving multiple outcomes,

including hospitalization for hyperammonemia, or quality of life, and survival. The outcomes vary a lot based on the disease severity.

What was difficult in this particular leveraging observational data is that the time of intervention was very clear for those who received liver transplantation. That was the day they received the transplantation. However, for those who didn't receive transplantation, finding a start of follow-up was really hard. Because they're defined by not receiving an intervention, there is no anchor and, again, this is an observational study, so you don't have a randomization that sets everything to time zero.

Because liver transplantation happened early in age, the time scale here that I'm showing on the red line is age, so you can think about it as age in days. Here, I'm showing 2 days where that intervention had occurred in the database, but of course there were a lot more days. There were around 100 eligible patients that received the liver transplantation.

One way to get around thinking about when to start follow-up is to think about this sequentially or as a sequential cohort entry. For example, at 63 days, we could ask ourselves who would have been eligible to receive transplantation that maybe either received it at 63 days or later, and then match those who actually received liver transplantation at 63 days with those that did not receive transplantation. By doing this, you're allowing someone who received a liver transplantation later in their life to be a control for someone who received a liver transplantation early in their life. You do this sequentially, and that's a way to solve, first making sure that you're not including immortal time bias, but also control for confounding in every strata.

Next slide please. In summary, for a non-randomized comparison, you have similar advantages to randomized comparison. When you're using within-subject comparison, you're augmenting between-subject with within-subject. I'm highlighting here two maybe additional arguments

for using self control in longitudinal design, which is that when you're comparing one person to themselves, you're often controlling also for confounding, and also by looking at time periods, you're solving the probability of time zero, making sure that you're starting at the right time.

Next slide please. As with randomized study, timing is really important, and that's the trickier part. In addition to the duration, making sure that it's just right, not too long and not too short, you have to also ask yourself whether some of the confounding that you're worried about is actually time-varying because you would need to adjust for that. In terms of the analytical considerations, most of the time, whether you're using self-control, you have to correct for a correlation of measurements within the same subject. So if it's only self-controlled, you just need to use paired tests or paired analysis; however, when you're augmenting between-within, then you probably need to use a hierarchical model for the adjustment.

I didn't discuss the case control, but you need to think about anchoring in a lot of these designs.

(Alarm sound.)

DR. IZEM: Sorry. I had put a timer, and I'm out of time, so next slide.

This take-home you've seen before. You can skip a few more slides. I just wanted to put a plug. These are acknowledging all of my collaborators. This is a plug for IDeAL, which is on the same side of the Atlantic, I guess where I am now, which has a lot of very useful work on rare diseases and study designs that is very relevant to the audience today.

Then finally, the last slide is my email address if you have any questions. Thank you. Sorry for going a little over.

DR. ROSENBLUM: No, that was great. Thank you, Dr. Izem, for the excellent presentation.

We'll now turn to our discussion.

Professor of Biostatistics at Vanderbilt University, Dr. Frank Harrell, he is an expert

biostatistics advisor at CDER, Center for Drug Evaluation and Research at FDA, and he has a wide range of expertise, including development of accurate prognostic and diagnostic models; model validation; clinical trials; observational clinical research; cardiovascular research; technology evaluation; pharmaceutical safety; Bayesian methods; quantifying predictive accuracy; missing data imputation; and statistical graphics and reporting.

We're lucky to have Dr. Harrell here today to give his thoughts on the presentations and the topic in general.

Summary of Presentations - Frank Harrell

DR. HARRELL: Thanks very much for the nice introduction, and what a privilege it is to be commenting on the talks by these three amazing speakers. Not only are they leaders in this field, but they presented very complementary work to each other, so I really enjoyed the presentations. I'll start with just a few comments about each one, and then I'll do some general comments.

Karen Price, she gave us a great Bayesian background, and I love that little diagram of knew this, saw this, and now know this, and in feeding that back into a continuous learning cycle. She made really a key point, which is that the Bayesian approach makes integration of pre-study knowledge and new data transparent. There's a lot of ad hoc approaches to do that. They're not very satisfactory, and they're very subject to bias of the observer or of the analyst, so Bayes gives, really, the way forward for that.

She talked about different kinds of borrowing, and data sources, and a nice overview of knowledge and belief elicitation, which is a very important component of Bayes. She mentioned Bayesian decision rule for a single-arm study, which does require a bit of an arbitrary response probability threshold, so I worry about that a little bit.

Just one other comment that I could maybe bother caring about a little, she mentioned Bayesian procedures that have good alpha or type 1

assertion probabilities. That's really kind of at odds with optimum decision-making, and in the chat, I put a link to some more details that has a link to a long discussion about this. So controlling alpha is sort of counter to what Bayes is trying to do, which is to maximize the probability of making a right decision.

She talked about continuous learning, where you have maybe a sample size goal, and you might be dealing with Bayesian predicted probabilities to decide when you've learned enough. That does require sort of treating an ultimate sample size as a magic quantity, and I would like to argue that one of the best things we can do in the future of clinical trials is to do away with sample size calculations altogether because that's really how we'll really respect continuous learning.

Jack Lee had another fascinating talk, and he gave a nice background about bias versus precision, which are always important to keep in your mind as you're looking at any method. He talked about several novel adaptive designs and how

information grows as the sample size grows, advantages of the Bayesian paradigm, and he made a push, which I think is very wise, for moving away from discrete drug development phases but making the process more continuous.

He talked about different ways to adapt and different reasons to stop early, such as toxicity, futility, efficacy, and handling complex endpoints. Jack has a long track record of developing, really, state-of-the-art free software for helping people use these sometimes complicated ideas in designing studies, so you should look into his software.

He talked about minimizing the expected sample size, which is a way to learn faster and maybe save resources also, and the different sorts of environments, platforms, master protocols, and avoiding noisy subgroup analyses. And I loved this quote, "There is no free lunch, but there are lunch specials." I'm writing that one down.

Rima Izem really hit something I was hoping somebody would hit in this session, which is capitalizing on time. I think this is extremely

important because in the simple case of doing a univariate outcome, let's say you measure a patient outcome at 2 months versus having the outcome assessed at more times, the payoff of having the longitudinal data is huge.

She went further than that to talk about the value of within-subject comparisons in general, including a lot of designs that are not used very often and should be used more often. She mentioned how observational longitudinal data, that even though you can have confounding, you probably have a little less confounding when you have time-oriented data. A big plus for the longitudinal way of thinking is increasing the effective sample size, and your unit of analysis might be a patient response measured in one day, or one week, or one month.

She mentioned crossover and N of 1 studies. I learned a lot from her talk. There are other things that she touched on which relate to rare disease and ethics, and the push to not randomize because there are things like delayed treatment,

there are randomized withdrawal studies, and stepped wedge, which is another kind of delayed treatment design that really should be considered to answer those concerns that a lot of rare disease communities worry about. She mentioned database fitness for purpose and several other good things.

I'll just add a few overall comments. I'm not one of these statisticians that's very optimistic about borrowing historical data, so I tend to be very afraid of historical data. I just want to note that when you do use historical data, it takes extreme diligence, and one of the pieces of that is you have to include a lot of historical data that's very unfavorable to what you're trying to show. You cannot be accused of cherry-picking favorable historical data.

The use of historical data actually requires you to use the raw data. There's no way I know to do a really satisfactory analysis using only summary statistics of the raw data. For one thing, the raw data should almost always be covariate adjusted. I see a lot of use of historical data

where they're not even adjusting for age differences between historical data and the new patients.

Jack mentioned the need for validation studies, especially when you're doing things that are complex like adaptation or borrowing data. When you are borrowing data, you really need a validation study, or if you're doing an adaptive study that might actually lower your effective sample size for the favored treatment regime. And if you really need a validation study, it goes against some of what you're trying to gain by having limited sample size now; so how are you going to get a sample size for the validation study? This is an over-generalization, but I think there's less need for a validation study if you're not borrowing historical data.

I want to emphasize what I think is the biggest bang for the buck. The biggest bang for the buck is to have a high-resolution outcome variable that has very high test-retest reliability. An example of that -- this is kind of

an idealized example -- would be you're looking at bone mineral density measured with a DEXA scan, and the patients get a DEXA scan at 6 months, 7 months, 8 months, out to 18 months, so you spend a lot of money per patient but you don't spend a lot of money on a lot of patients.

So if you have a high-resolution, high test-retest measure like bone mineral density, it's amazing what sample size you can get away with and learn a tremendous amount of information. So trying to emulate that with the data we usually have, such as having a 60-level clinical outcome scale, or a 60- level patient-oriented outcome scale, and then you have clinical overrides for events that happen to the patients and you do an ordinal analysis that, say, has 62 levels in the ordinal response, now you're starting to approximate what you can get out of the DEXA scan sort of high-resolution data. Then if you measure that multiple times, you'll have much more information, and that increases statistical power a lot. So the idea also was to not have a sample

size but to mimic what physicists do, which is to study until you have an answer, you have enough evidence, and that's when you stop.

I will go out on a limb and say something nobody will be ready for, which is there's something more valuable to borrow than borrowing data from historical data, and that is to borrow knowledge about how consistently a treatment affects different endpoints, and you could form a Bayesian prior for how the treatment affects mortality versus how it affects functional status and other patient-reported outcomes. Borrowing that kind of information allows you to learn about multiple endpoints when you don't have enough sample size to study any one endpoint, especially a mortality outcome.

Then I'll just close by saying, avoid the worst possible mistake. The worst possible mistake in rare diseases, I saw this actually done, and I'm still kind of in a state of disbelief that a sponsor would do this. They had very great difficulty getting patients because of the rare

disease. They were able to randomize about 80 patients, and they had an established ordinal scale that had maybe seven levels to the scale, but then they dichotomized the outcome and did a responder analysis about whether or not you had like a 2-point improvement from baseline. That was their definition of a binary response.

So what they've done is to say, we've got 80 patients -- that's not enough really -- but we're going to do a responder analysis and pretend we had 30 patients because that's what that particular responder analysis did. It reduced the effect of sample size from 80 to 30. So whatever you do, don't lose information in making your sample size smaller than it already is.

So thanks for listening, and, again, I just tremendously enjoyed these three talks.

DR. ROSENBLUM: Thank you, Professor Harrell, for the excellent feedback on the three talks.

This is a perfect place to stop, and I'll pass it over to Dr. Dionne Price, but I first want

to thank all the speakers and panelists for an excellent second session of this two-day symposium. Thank you all.

Concluding Remarks - Dionne Price

DR. D. PRICE: Thank you, Michael.

I started the day by welcoming you all to our workshop, and I was so eager to jump right in, that I neglected to introduce myself. I am Dionne Price, and the Deputy Director of the Office of Biostatistics in the Center for Drug Evaluation and Research. In this role, I also co-lead the CDER and CBER's, Center for Biologics Evaluation and Research, Complex Innovative Trial Design Paired Meeting program and related efforts, and I'm actively engaged in CDER's Accelerating Rare diseases Cures program, so I encourage you all to explore the FDA website for information on both.

Now, in our first session today, we heard about small sample sequential multiple assignment randomized trials, adaptive enrichment designs that could be mapped to some rare diseases, and the need to get more data and/or get more information from

the same data.

Specifically, we may be able to get maximum information available from limited data via adaptive designs with the goal -- and I quote Dr. Stallard -- "of efficient, smarter, and faster trials." The panel also touched on the potential advantage of group-sequential trials with interim analyses, single-arm trials, and sample size considerations.

In our second session, which nicely complemented the first session, we heard about Bayesian thinking, borrowing approaches, master protocols, including adaptive platform trials and basket trials; Bayesian adaptive designs; frequent interim analyses; and leveraging longitudinal data using various designs. So if it's not obvious, I will say it. There is no one size that fits all, but we certainly have heard options and some thought-provoking ideas for designs and analysis methods that may aid in drug development for rare diseases.

On behalf of the FDA, I would, again, like

to thank all of our speakers, our panelists, and our moderators. I would like to thank you, the participants, for your time, your questions, your engagement, your attention throughout both days of the workshop, and I will conclude with a reminder of our June 7th and 8th public workshop on Novel Endpoints for Rare Disease Drug Development, and that link has been added in the chat. So once again, thank you all.

(Whereupon, at 11:58 a.m., the workshop was adjourned.)