Public Workshop on Advancing the Use of Complex Innovative Designs in Clinical Trials: From Pilot to Practice

Docket No. FDA-2023-N-4965

Moderated by Dr. John Scott

Tuesday, March 5, 2024

9:00 a.m.

FDA Great Room, Building 31

10903 New Hampshire Avenue

Silver Spring, MD 20993-0002

Reported by: Richard Livengood

JOB NO.: 6161834

A P P E A R A N C E S

List of Attendees:

Dr. John Scott, U.S. Food and Drug Administration

Dr. Roger J. Lewis, MD, PhD, University of California, Los Angeles and Berry Consultants, LLC

Dr. Karen Lynn Price, PhD, Eli Lilly and Company

Dr. Herbert (Herb) Pang, PhD, Genentech/Roche

Dr. Stephen Ruberg, PhD, Analytix Thinking

Dr. J. Jack Lee, MD, MS, DDS, University of Texas MD Anderson Cancer Center

Dr. Rebecca Hubbard, PhD, University of Pennsylvania

Dr. Frank E. Harrell, PhD, Vanderbilt University and U.S. Food and Drug Administration

Dr. Dean Follmann, PhD, National Institute of Allergy and Infectious Diseases

Dr. Frank Bretz, PhD, Novartis

C O N T E N T S

	PAGE
Dr. Roger J. Lewis	14-41
Dr. Karen Lynn Price	50-77
Dr. Herbert (Herb) Pang	87-109

P R O C E E D I N G S

DR. SCOTT: Good morning, everybody. Thanks very much for joining us today for our Public Workshop on Advancing the Use of Complex Innovative Designs in Clinical Trials: From Pilot to Practice. I particularly appreciate those of you who came out in person on a kind of dreary Tuesday. I'm sure more will be arriving later.

This meeting is being recorded, and I know we're going to have a really exciting discussion today, and I'm looking forward to hearing from our speakers and panelists and learning a lot from them.

So my name is John Scott. I am the director of the Division of Biostatistics and FDA Center for Biologics, Evaluation, and Research. And I'm going to take just a few minutes to sort of give some background and introduce what's going on today, but really the speakers and the panelists are the stars of the show today.

So, you know, starting with sort of the obvious, we rely so much on clinical trials to support

our critical regulatory and public health decisions. They form the backbone of evidence of safety and effectiveness needed for drug and biologic approval. I think most people know that the cost and complexity of trials have ballooned in recent decades.

Some of the numbers people say are somewhat eye watering in terms of cost and complexity. The questions we're trying to answer just get more and more complicated as science progresses. So consequently, there's really been a need for innovative approaches to answer complex questions and to improve trial efficiency.

And some of those approaches include various adaptive designs, phasing approaches, and potentially incorporating external data in trial analysis and design. And one result of this need was FDA's Complex Innovative Trial Design, which for reasons I won't belabor, we don't have a T in the abbreviation for it. We call it CID Review Program.

A bit of the history of the CID Review Program: Under PDUFA VI, the sixth authorization of

the Prescription Drug User Fee Act, which ran from 2017 to 2022, Congress gave FDA a mandate to facilitate the advancement and use of complex innovative designs in regulatory decision making.

And there were several sort of subparts of that commitment, which included developing staff capacity, convening a public workshop -- we had a workshop much like this one five years ago -- publishing draft guidance, and developing sort of review procedures and templates as appropriate.

But importantly, the kind of main thing was the requirement for FDA to conduct a pilot program for the review of CID proposals. So what that pilot program was, it's a joint effort between FDA Center for Drugs and Center for Biologics Evaluation, and Research.

What happens is sponsors submit designs to be considered by FDA and if the designs are accepted into the program, they have the opportunity to engage with the regulatory review team on those designs via two additional meetings, typically focused

specifically on the more technical aspects of the designs.

FDA will select up to two of these submissions per quarter, and then one of the unique features of this program is that we form an agreement with the applicant on elements on the proposal that can be publicly disclosed so that we can use the designs as case studies for others for outreach and education.

Those CID meetings are led by the biostatistics groups in FDA, so in CDER, it's the Office of Biostatistics; in CBER, the Division of Biostatistics, but there's of course participation from clinical teams and from all relevant disciplines. So again, this ran from 2017 to 2022.

Over that time, we accepted six submissions across several therapeutic areas, including neurology, analgesia, rheumatology, oncology, and several of them were in adult or pediatric rare diseases. And the methodologies or designs that were used included Bayesian hierarchical

models, the use of informative priors to bring in structure and information and master protocol designs.

I mentioned that PDUFA VI committed us to publish guidance on CID. We broke that up into two different guidances, one guidance on adaptive designs for clinical trials for drugs and biologics, and the other a guidance on interacting with the FDA on CID proposals.

So that brings us to PDUFA VII, the seventh iteration of the Prescription Drug User Fee Act, which we're in now. It runs from 2022 to 2027. The objective is similar to PDUFA VI, and says "To facilitate the advancement and use of complex adaptive, Bayesian, and other novel clinical trial designs."

One interesting difference is that the objective now explicitly mentions Bayesian. And we have our sort of goals under this, which include continuing to develop staff capacity for CID review, to continue the paired meeting program (which we no longer call a pilot, we just call it the paired

meeting program), to convene the public workshop that we're having today, and to publish a draft guidance on the use of Bayesian methods in drug and biologic trials.

So in today's workshop, our goals are to discuss aspects of complex adaptive, Bayesian, and other novel clinic trial designs. And specific topics we were asked to consider include considerations for external data sources, Bayesian statistical methods, simulations, that is, trial simulations, for operating characteristics and issues with clinic trial implementation.

And the way we're structuring today, I hope everybody has the agenda for specifics, but we're going to start in the morning with three case studies of successful CID proposals, followed by a panel discussion, including the speakers and additional panelists.

I'm not going to go through people's biographies in detail, but we have a panel today of extremely distinguished speakers and panelists. The

detailed biographies are posted on our event webpage, so I'm just going to introduce people quite briefly now. Let's see. Is it in -- I think it is in order from my immediate left.

Frank Bretz is a distinguished quantitative research scientist at Novartis. Dean Follmann is chief of the Biostatistics Research Branch at the National Institute of Allergy and Infectious Disease at NIH. Frank Harrell, who is joining us remotely, is professor of biostatistics at Vanderbilt University School of Medicine and also an expert biostatistics advisor to FDA Center for Drugs, Evaluation, and Research.

Rebecca Hubbard is professor of biostatistics at the University of Pennsylvania. Jack Lee is professor of biostatistics and Kennedy Foundation Chair in cancer research at MD Anderson Cancer Center. Roger Lewis is senior physician in the Los Angeles County Department of Health Services, professor of emergency medicine at the David Geffen School of Medicine at UCLA, and a senior medical

scientist at Barry Consultants.

Herb Pang is expert statistical scientist in PD Data Sciences at Genentech/Roche. And Karen Price is vice president Statistical Innovation Center Advanced Analytics at Eli Lilly. Oh, and I somehow missed Steve Ruberg. Steve Ruberg is a former employee of Eli Lilly, and is currently the principal of Analytix Consulting; is that correct?

DR. RUBERG: Analytix Thinking.

DR. SCOTT: Analytix Thinking. I'm so sorry for the neglect, Steve. That doesn't reflect your importance to the panel. So before we get started, I wanted to thank -- there were a lot of people involved in making this workshop come together, but in particular, and especially I wanted to thank Tuan Pham, who is the CID project coordinator for CDER.

Tuan was instrumental in every aspect of this, ranging from speaking contact to agenda formation to the Federal Register notice to coordinating catering, and we wouldn't be here without

him today. Christopher Egelebo is the CID project coordinator for CBER and also participated quite a lot in organizing today's workshop.

We have a lot of internal folks at FDA who work on CID and who contributed in some way to this effort. There's a steering committee, a proposal selection committee, and an education subcommittee. And of course, when these proposals come in, there are CDER and CBER reviewers who review them.

And then finally, the White Oak AV team, who I'm grateful that they turned down my mic and are doing many other things behind the scenes. And then finally before we begin, I really wanted to give a very special thanks and recognition to my colleague Dr. Dionne Price, who tragically passed away two weeks ago, but otherwise would be the one here giving this introduction.

Many of you in the audience knew Dionne or were touched by her leadership in some way, but for those who weren't fortunate enough to know her, Dionne was deputy director of the Office of Biostatistics in

FDA CDER and past president of the American Statistical Association. Absolutely nobody had a more significant impact on FDA's CID Review Program and how we treat complex designs than Dionne.

She led the program for CDER, starting with PDUFA VI and through the PDUFA VII reauthorization, and she in particular was instrumental in the PDUFA VII negotiation process that led to the continuation of the CID program. Her influence on every major decision for the program and most minor decisions for the program couldn't be overstated, ranging from process implementation, proposal screening and selection and external communication and outreach.

And CID was just one of many things she did. Dionne brought outstanding judgment, strong leadership, and really truly unparalleled diplomacy to everything she did, and kindness, and we miss her a lot. But we will -- she would want us to have the workshop, I'm sure, so we'll proceed.

And our first speaker is Dr. Roger

Lewis, who's going to be speaking about the CHIPS trial, which was an adaptive storage duration finding trial for platelets. And this trial is -- of the three, this is one that wasn't a formal proposal to the CID meeting program. It just came through the ordinary IND review path, which illustrates that CIDs aren't restricted only to the program.

I wanted to mention that if people have questions about the talks specifically, if you're on Zoom, you can put them in the Q&A, and we may be able to get to them later, but we will also have a specific designated time later for audience questions, both in person and by Zoom. Okay.

Roger?

DR. LEWIS: Thank you. All right. Thank you, John. It's a pleasure to be here. I'd like to thank the organizers for this opportunity. And as John mentioned, this is not a trial that was designed through the CID program, but I think that the positive interactions with the agency throughout the development of the trial illustrate the influence of

the CID program on the agency's thinking.

These are my disclosures. Most importantly, relevant to this, is I'm the senior medical scientist at Barry Consultants and also an inactive special government employee. So the CHIPS trial, first of all, I need to point out that I'm not the person doing the hard work of the trial.

It's led by multiple principal investigators. There's been a very large statistical design team involved in both the design and implementation, a data coordinating center at the University of Utah, and it is funded by the U.S. Army Medical Research and Development Command through the U.S. Army Medical Research Acquisition Activity.

So the background is that in the United States, platelets, a critically important component of our clotting mechanism, are collected and then stored at room temperature. And because of the potential for a small amount of bacterial contamination, the storage at room temperature means they only can be stored for a short period of time, typically five days before

they must either be used or discarded.

There are some provisions for extending that storage time up to seven days if you test the units individually for contamination. But because of this short period of storage, many hospitals with relatively low volumes of blood product usage are unable to maintain platelets. In fact, 10 percent of the hospitals that acquire red blood cells through the American Red Cross actually don't even attempt to acquire platelets.

That lack of availability of platelets put patients with bleeding, either due to platelet inadequacy or, for example, due to a major trauma at tremendous risk for adverse outcomes. If we were able to store platelets for a longer period, that would substantially mitigate the challenges associated with maintaining platelet availability in a variety of areas, including austere settings.

So the objective of the CHIPS trial is to demonstrate that platelets that are stored at 4 degrees centigrade, what we'll call "cold stored

platelets," are non-inferior or potentially even superior in terms of their ability to treat active bleeding, so-called hemostatic efficacy when compared to standard room temperature platelets stored at 22 degrees. And I may refer to that as either "room temperature platelets" or "warm platelets."

And we're going to evaluate that in adult and pediatric patients who are requiring complex surgery and who are actively bleeding as a result of that surgery. This is not because that patient population has any particular bleeding mechanism, but this is a well-characterized setting in which to evaluate hemostatic efficacy.

The secondary objective is to determine the maximum storage time up to a potential of 21 days, over which, the product maintains non-inferiority when compared to room-temperature-stored platelets. So I want to just take a second and circle back to why we use adaptive approaches in many settings in which we have sparse information related to key components of a clinic trial design.

So I think many of us have had the experience that when we are designing a clinic trial, we almost never have enough information in order to know the optimal design. And specifically in this case, we don't know how long you can store platelets over a relatively wide range of uncertainty, say, 7 to 21 days, so we don't know what is the storage duration over which we ought to be exploring the hemostatic efficacy.

But once patients are enrolled in a clinical trial, data start to accumulate that reduce that uncertainty that existed at the time of the design of the trial. An adaptive trial is designed to take advantage of that stream of initially sparse but increasing information in order to make changes according to pre-specified rules and to mitigate some of the risks that were associated with the initial uncertainty.

In certain cases, that can increase the probability of getting the right answer at the end of the trial or improve the trial efficiency. So

specifically in this case, we're going to be using incoming data to help us know where we should be focusing our attention on the continuum of the storage duration of platelets.

So an adaptive trial can be put into this generic framework, where in the upper left, we begin with initial set of sampling rules. So, for example, we might start with a particular storage duration or a particular randomization ratio. We take a first look at the data and analyze the data, ask if there's a good reason to stop the trial; and if not, we may revise randomization sampling or other rules in response to the partial accumulated information at the time of that interim analysis.

We then continue with additional data collection according to those new rules. That process can continue in a circular fashion until we reach a reason for stopping the trial; for example, being able to draw a firm conclusion regarding efficacy or inferiority or reaching the maximum of sample size planned for the trial.

So we're going to place the CHIPS trial design into this framework. Now, the CHIPS design itself is a fixed randomization trial with two to one randomization of cold-stored platelets to room temperature platelets in order to increase the both experience and safety database associated with cold-stored platelets and because we need to explore the storage duration relationship to hemostatic efficacy.

The primary endpoint is a fixed-point bleeding score, a hemostatic efficacy score, and importantly, lower scores are better: 1 is good, 5 is bad. The two arms are treated differently because we consider room temperature platelets to be a homogeneous treatment.

Even though platelets can be stored from zero to five days typically, the processing time means that most room temperature platelets are three to five days old at the time of transfusion, so we consider those to be a single treatment. In contrast, when you receive cold-stored platelets as a patient,

you typically receive a set of platelets that will have a storage duration, and you may have another number of units of platelets that have a different storage duration.

So we're going to characterize the treatment of a patient by the weighted mean storage duration of the administered cold-stored platelets. So in the room temperature arm, platelets don't have an age, but in the cold-stored arm, they do have an age that's defined by this average of the age of the platelets administered.

It's a non-inferiority trial because the advantages in terms of the ability to store the platelets mean that this would be an important part of our options for treating these patients, even if they were not quite as effective as room temperature platelets, and the non-inferiority margin is one unit on the bleeding score.

And we're going to demonstrate type 1 error control through simulation. We're going to have some adaptive rules for changing the maximum storage

duration of cold-stored platelets as the trial progresses, and the trial is designed with a fixed maximum sample size with 1,000 patients and interim analyses after every 200 patients. That's the overall structure.

The underlying inferential model assumes that the hemostatic efficacy score, the mean hemostatic efficacy score for cold-stored platelets is a function of the storage duration X, where, as I said earlier, a smaller score is better. We'll let mu sub one be the true mean hemostatic efficacy score for room temperature platelets, which is a single number. It's not dependent on the age of the warm-stored platelets.

The efficacy of the cold-stored platelets is modeled as a monotonic piecewise linear regression model. The monotonic there I call the "this is not wine" assumption: We assume that platelets do not get better with age. So it assumes that as the platelets are stored longer, their hemostatic efficacy will remain the same, or it will

increase.

The null hypothesis is that there is no storage duration for which the hemostatic effect of cold-stored platelets is within one unit of that of warm-stored platelets, and the alternative hypothesis is that there is some storage duration of seven days or greater for which the hemostatic efficacy of the cold-stored platelets is non-inferior to warm-stored platelets.

At the end of the trial, if it does not stop for futility -- so after 1,000 patients -- we look for the longest duration of storge for which the model-based prediction is that there is a 97.5 percent posterior probability that the cold-stored platelets are non-inferior to warm.

If that is met, then there is a gated superiority hypothesis that has a more stringent criteria of 98.3 posterior probability. And it also has a requirement for super superiority. That's denoted by the little delta sub X, which is required to maintain type 1 error control for the superiority

hypothesis because of the monotonic assumption that's built into the model. Again, that parameter is determined through simulation.

At each interim analysis, we want to have an opportunity to alter the maximum cold-stored duration. So at each interim analysis, we asked based on the current model for the relationship between storage duration and hemostatic efficacy, what is the longest duration for which we predict there is a least 33 percent or one-third chance that that storage duration is truly non-inferior, and we consider that a candidate for a new maximum duration of storage.

If that candidate is less than seven days and the probability of non-inferiority at seven days is less than 10 percent, the trial stops for futility. So that says, if we can't store platelets for at least seven days in the cold and have any reasonable chance they are non-inferior, then we stop the trial.

But if the candidate duration of maximum storage is greater or equal to seven days,

then we take the minimum of three possibilities for the new storage duration for the next 200 patients: either the candidate time of storage itself, the maximum duration plus 5 days, or 21 days.

What these rules mean is that for whatever the current maximum storage duration is, we can only increase by up to five days. We cannot go over 21 days, and we cannot go past a storage duration for which there's isn't a least a one-third probability of non-inferiority. There is no early stopping for success in this trial design.

So I want to take a second to show you what some simulated data might look like because this will be important for understanding how the trial plays out. So on the left side of the graph, you see a pink dot around 2, that is the mean observed hemostatic effect of the warm-stored platelets, and there's some uncertainty around that estimate, and you can see the faints dots for the number of participants whose hemostatic efficacy has been the integral of values 1, 2, 3, 4, and 5.

If you take the observed hemostatic efficacy of the warm-stored platelets and you add one, you get the non-inferiority margin, which is showed by the horizontal yellowish line. And then for the cold-stored platelets in the simulated data, you see that there's data out to approximately 17, 18 days of storage.

And there are mean hemostatic efficacy scores that have been observed from the data shown by the orange dots of various size related to the number of platelets at those various time points in a fitted line based on this monotonically increasing model. The goal of the design is to identify where that fitted line crosses the non-inferiority margin, which will reflect the maximum storage duration for which the cold-stored platelets maintain non-inferiority.

So going back to the overall structure of an adaptive trial, we're going to start with a maximum cold-storage duration of seven days, the same duration that's allowed for warm-stored platelets if they are tested for their sterility during their

storage period.

We're going to start with a first interim analysis after 200 participants. We're going to fit this model for the hemostatic efficacy of the cold-stored platelets as a function of the duration of storage. We're going to make sure that the seven-day storage duration has at least a 10 percent probability of non-inferiority. That's the futility rule.

And then we're going to apply these rules at each interim analysis to revise the maximum storage duration with the hope that, over time, it will gradually increase. As in a duration-finding experiment, we find the maximum storage duration that maintains non-inferiority.

Once we get to 1,000, then we will find the longest storage duration in which the probability of non-inferiority is greater than 97.5, and that will be the primary result of the trial. And if that is positive, we will also evaluate for super superiority against the more stringent posterior probability cutoff.

So here's an example trial based on simulated data to show how this plays out. I want to take a second to orient you to this graph. So in the upper left panel, the structure is the same as was in the simulated data I showed earlier, but at the start of the trial -- this is after the first 200 subjects -- we've only been allowing the cold-stored platelets to be stored up to seven days, so we only have data to support the model fit out to seven days.

Because of that, you can see there's tremendous uncertainty if we try to extrapolate those sparse data out to longer storage duration times. In the lower left, you see the number of subjects that have been enrolled at each interim -- I'm sorry, the range of cold-storage duration that has been -- that is allowed up to that point.

In the middle of the bottom of the slide, you see the fitted probability of non-inferiority based on the model. And if you look at the right-hand column in the table in the middle bottom of the slide, you can see that for all of the

durations, the model says that the predicted -- excuse me -- that the posterior probability of non-inferiority is greater than a third.

So the model would say that it would be acceptable to have your X candidate up to 21 days, but our rule for increasing the storage duration is that at no point can the number of days of storage increase by more than five days.

So the result of this interim analysis would be to, for the next 200 patients, have the maximum length of duration be 12 days, which is the 7 days that we currently have data for, plus the 5 days, which is the maximum step forward that we are allowed to take.

The bottom right-hand panel shows you the storage durations that have a probability of non-inferiority of greater than the 33 percent shown by the horizontal red line. And then the upper right panel shows you the sample size in the various categories, be either warm-storage duration or cold-storage duration in various bins.

So if we let this trial go onto the next 400 subjects -- excuse me, the two 400 hundred subjects, you see that the current max cold-storage duration was 12. We now have data in the interval from 7 days to 12 days, as well as additional data up to 7 days of cold storage; that decreases the uncertainty.

Again, the model identifies all storage durations as potentially having -- being non-inferior with a probability greater than 0.3, but because we're at 12 days, the maximum we're allowed to move to as the storage duration is 12 plus 5, or 17. We enroll an additional 200 patients for a total of 600.

And now, if you look at the middle of the bottom of the slide in the table, you can see that at 17 days of storage, there's a 45 percent probability of non-inferiority, but by 18 days, it drops below a third, so we were studying platelets out to 17 days.

The model says you cannot increase the maximum storage duration because 18 days has less than

a one-third probability of being non-inferior, so we're going to continue with a maximum of 17 days of storage, collect more data.

Here's after an additional 200 patients. The model-chosen duration at this point is 20 days because that's where the predicted probability of non-inferiority falls less than 0.33. And then at the end of the trial, these would be the final results.

Now, it's important to note that the precision in the estimate of the efficacy at the longest storage durations is highly dependent on the support for the model out at the right-hand side, and that's going to be related to a comment that we received from the FDA during the IND review of the process.

Okay. So again, the primary trial analysis after 1,000 patients is that either it's a negative trial because there's no cold-storage duration of 7 days or longer that is non-inferior, or we identify a period of storage between 7 and 21 days

inclusive that is non-inferior, and then we can also evaluate for superiority.

So what are the operating characteristics of this design? In order to evaluate the design, we take the design, and we make lots of different assumptions regarding the true underlying efficacy of cold-stored platelets as a function of storage duration, run thousands of trials, and simply count up the trials that get an answer that is consistent with the underlying assumed truth or inconsistent.

And I just want to point out that this trial can get wrong answers in a number of different ways. It can fail to identify a storage duration that exists. It can identify a maximum storage duration that is incorrect, meaning, the platelets would actually be inferior, and it can have varying degrees of accuracy in identifying the maximum storage duration that maintains non-inferiority.

So in order to evaluate type 1 error control, one has to come up with a variety of

scenarios regarding the possible true relationship against which one wants to evaluate the performance of the trial. So here, you see seven different assumed relationships.

The horizontal dashed line is the assumed hemostatic efficacy of warm platelets. That's at 2. The non-inferiority margin, therefore, is at 3. The broad gray line are the different assumed relationships between maximum cold-storage duration and the hemostatic efficacy of the cold-stored platelets.

Note that every one of the gray lines goes through the point at seven days and three; that's the definition of the null hypothesis. And you can picture that the more gradual the slope, the more difficulty the trial is going to have identifying the correct storage duration because there's many days that are close to the non-inferiority.

Okay. So this graph captures both the rate at which type 1 errors are made and the accuracy and identifying the maximum non-inferior storage

duration. So the vertical axis is an expansion of the distance between the efficacy of the warm-stored platelets and the non-inferiority margin, which is a value of one greater.

The solid red lines are the assumed relationship between storage duration and hemostatic efficacy for cold-stored platelets. The vertical dashed line is the maximum storage duration that maintain non-inferiority. The histograms, the blueish histograms, show the relative frequency with which simulated trials identify a maximum-storage duration of the different numbers of days.

And as long as those fall below seven, not including seven, those are negative trial results. The type 1 errors are any case in which a blue histogram falls on seven or greater, and as I'll show at the end, it's controlled at a 0.025 level. The futility rule is based on this calculation of the posterior probability of non-inferiority at seven days falling below 10 percent.

And under the particular null

hypothesis that's shown in the lower left, the futility rule stops the trial about half of the time. To evaluate power, one also needs a set of alternative hypotheses. So in each one of these case, you may note that the gray broad line that represents the assumed efficacy of the cold-stored platelets falls below; therefore, better hemostatic efficacy than the non-inferiority margin at seven days.

The most difficult in identifying the correct storage duration will be the more gradually sloping curves, and we'll be able to see that. This is the same presentation of the performance of the trial for those alternative hypotheses. So starting at the upper left, that's the situation in which warm-stored and cold-stored platelets have exactly the same efficacy and there's, in fact, no decay in that efficacy for cold-stored platelets over time.

And you can see that the trial is very efficient in correctly identifying that the maximum storage duration of 21 days is the correct storage duration. For the case in which there is a linear

relationship between storage duration and efficacy, and it crosses at 16 days, you can see that the trial identifies multiple different possible maximum storage durations from the different simulated trials, ranging from about 9 days up to 15 days and does not overestimate with any appreciable frequency the maximum storage duration.

The model is specifically structured to avoid what we call "overcrossing," meaning, identifying a storage duration that is too long, and it would therefore put patients at risk for receiving an inferior product from a hemostatic efficacy point of view.

You can see for each of the subsequent different alternative hypotheses, there's similar behavior where the trial systematically is conservative in estimating the maximum storage duration, and there are broader distributions for those curves in which there's more gradual crossing of the non-inferiority margin in a more tightly clustered set of findings across simulated trials when there is

a steeper relationship between storage duration and hemostatic efficacy.

So if one simulates the trial thousands of times and averages this, the behavior, this is the operating characteristics table. And I just want to point out that it's not just a question of type 1 error or power.

So in the first column, you have a traditional power, and you can see that at the bottom across those different potential shapes of the null relationship, the type 1 error is controlled at less than 0.025. For the alternative hypotheses, there is excellent power for detecting the fact that there is a storage duration greater than seven days for which the hemostatic efficacy is non-inferior to warm-stored platelets.

The second column, "Inferior patients," tells you the number of patients out of 1,000 who received platelets that in fact were inferior because they had been stored longer than the place at which the two curves cross, and obviously, we want to keep

that number to a minimum because that reflects potential risk to participants in the study.

The third column is the number of times within -- excuse me -- the fraction of simulated trials that gets the storage duration within three days, which is the correct number or one or two less. And overcrossing is the frequency with which the final result for the trial includes a day of storage, even one, for which the platelets would be inferior to the warm-stored platelets. That's obviously something we wanted to avoid.

So I just want to point out that the original letter on the IND in which we removed the clinical hold had a specific recommendation that we try very hard to maximize the number of subjects exposed to platelets at the longest cold-storage duration in order to maintain the data support for the model out at that location in storage duration.

And I just want to point out that the teams works incredibly hard to manage the platelet inventory in order to maximize the exposure of the

longer cold-stored platelets, and that is quite a challenging thing. The trial has enrolled incredibly well.

I think a lot of people would be jealous of the relationship between the actual enrollment and the planned enrollment, and there's a good distribution of ages in the patient population, which should help in interpretability of the results. The trial has just recently crossed 600 patients and therefore will be conducting its third planned interim analysis.

And I'd like to finish by making a comment about an external event led by the regulatory agency that caused us to change the way the trial was being conducted. So in June of 2023, the FDA released a guidance that essentially made it acceptable to store platelets for up to 14 days when conventional platelets are not available or their use is not practical. And the language is very specific, but it essentially says that up to 14 days is okay.

At that time, the trial, if it had been

increasing its storage duration as designed, would've had a maximum storage duration of 7 to 12 days; it would've been somewhere in that range. So the FDA guidance immediately set an upper limit on storage duration greater than the storage duration that the trial could have gotten to.

I did appreciate the guidance specifically mentions this trial, and I took this as an endorsement of the importance of completing the trial as planned. So in response to that, the DSNB for the trial reviewed a request from the investigator team to immediately increase the maximum storage duration up to 14 days, even though the design could only have gotten to 12 days at that point if everything had been accelerating or progressing as rapidly as possible. And the DSNB approved that request.

This did not leak efficacy information because that decision was based on external information not related to the trial, whereas, if the investigator team had been told the maximum storage

duration based on the algorithm within the trial, that would have leaked efficacy information.

So I'd like to stop there. There's a number of references that give the original design and talk about the trial. And, again, I'd like to thank the agency for their support in getting this trial going, and I look forward to its results.

Thank you, John.

DR. SCOTT: Thanks, Dr. Lewis. Please don't leave yet.

DR. LEWIS: Don't go.

DR. SCOTT: So I think it's a very interesting design. I wanted to turn to the panelists and see if anybody had any questions or comments before turning to others.

Dr. Follmann?

DR. FOLLMANN: Yeah, thanks.

I thought it was a really nice design. It reminded me a little bit of some studies infectious diseases where you might randomize people when to start ART. And so there's this duration question and

also studies that look at the duration of TB therapy where you want to find the sweet spot, and you're sort of pushing the envelope on what the duration is.

So I had a couple points. One is, you know, a lot of times when we're doing non-inferiority trials, we struggle with the margin, and I'd be curious about how you got the margin of one. And the other comment is that you don't randomize to the duration, of course, and so there's, I guess, a potential for a bias to creep in where if, as time goes by, you get slow bleeders later in the study compared to early.

And so the slow bleeders get the longer durations, and then you're comparing them to the lumped control group, which you have slow and fast bleeders. So I just wondered if, you know, there was thought about the concern of a secular trend and sort of, you know, confounded with longer durations in the duration arm in the cold-storage arm. But overall, I was just really impressed with it. I thought it was just soup-to-nuts really great.

DR. LEWIS: Great. Thank you very much. So the model in which the cold platelets are being evaluated of cardiothoracic surgery has a variety of different bleeding endpoints that you can be used. Chest tube output is a standard one.

This is a clinical bleeding score that reflects both quantitative measures such as chest tube output and also the surgeon's experience with a surgical bed during the operation and bleeding during the next 24 hours. I think that over the duration of time that this trial is being conducted, the likelihood of significant secular trends affecting sort of the clinical impression of the bleeding score is unlikely, certainly not impossible.

To my knowledge, there have not been any changes in clinical care practice or the use of ancillary treatments of bleeding during the trial. With respect to the non-inferiority margin, because the goal here is to improve the availability of platelets in both rural hospitals and potentially other austere settings, the non-inferiority margin is

really a value adjustment related to what is the minimum efficacy of a product that would make it a useful adjunct to controlling bleeding in these setting in which no platelets are currently available. And so it's a value judgment of the investigator and blood banking team.

DR. LEE: Yeah, Roger, yeah. Yeah, I think this is a excellent example illustrate the complexity and adaptivity needed, you know, for a trial like this, like in the CID setting. And thank you for the very comprehensive and thoughtful design.

My general question is that, we can think about CID as a very complex machine with many knobs you can adjust, right. So you like to have the -- and the output is also complex. It's not just one-dimensional output, right. So, for example, you like to, say, if a noodle machine make noodles, right, and then there are many sent input and many knobs to adjust. And the output, you know, there are also many measurement of the output.

So my question is that, during

this -- so of course you have certain essential design characteristic you like to achieve, like to control type 1 error rate and reach a certain power, but beyond that, you know, there's some criteria that you would look for "optimize" the trial.

And, again, you know, in a very simple setting, there are many optimal design available, but in a complex setting like this, maybe there's no one single criteria that you try to optimize, okay, and -- but there are many knob you can turn, right.

For example, how many interim analyses when you do the interim analysis, right. And then once a confidence -- put a really confidence level, you know, for futility and for the clear efficacy, right, and randomization ratio, right, so there are many, many things you can adjust.

So my general question for the CID design is that this is a setting -- will illustrate the importance of adaptation, okay. But then at the end, you know, other than meeting the minimum kind of criteria, what do you need to look for to "optimize"

the trial design?

DR. LEWIS: So first, and I think this was implied by your question, but I want to state it explicitly. All of the knob turning and adjustment has to occur before you finalize the design. And once you start the trial, it needs to be a pre-specified design. And when I -- I call it a rigid design, but what's really rigid is the adaptation rules.

The question of -- excuse me -- of optimization begs the question of, optimization for who? Whose utility function are we optimizing? And one of the things that, to me, makes this kind of design activity most rewarding is the fact that it's inherently multidisciplinary in nature.

So when you're trying to decide how many interims, what the randomization ratio ought to be or what your futility cutoff is, it really needs to be not a statistical question in isolation, but it needs to be a collaborative discussion among all of the stakeholders and ideally that involves investigators, clinicians, people, for example, in the

blood banking community who would be using the results of this trial, maybe patient representatives and others.

And it's really a consensus-based process when you have come up with a set of compromised settings for each of the knobs that balances the performance, statistical and otherwise, against the resource limitations, the complexity of the design and complexity, both from a statistical point of view and from an implementation point of view.

So I think the -- one of the things that I really enjoy about adaptive design is the fact that it naturally brings these collaborative groups together, and you actually get greater insights in what optimal looks like than you do if the statistician is working by themselves in designing the trial.

DR. LEE: And also just quickly follow up. You know, seems like it's a overpowered design, right, and why do you do that?

DR. LEWIS: So the reason we do that is because power is not the question, right. This is a duration-finding trial. The value of this trial is not primarily dictated by whether or not there is a storage duration for which cold platelets are non-inferior.

I think people who work in this area believe that there is and, in fact, the FDA issuing a guidance during the trial that says up to 14 days is okay would suggest the agency believes they are a reasonable option for the treatment of bleeding patients.

And the guidance specifically separates, just to be clear, the use of platelets for bleeding versus the use of platelets for prophylaxis, which are -- that's a different clinical question. So the investigators and others clearly believe that the right answer for the question of the overall outcome of the trial is that we should be able to demonstrate non-inferiority because it's likely to exist.

The question is, what is the maximum

storage duration? And if the maximum storage duration is 21 days as opposed to 14, that has huge implications both in civilian and military settings for our ability to make this treatment available to patients who need them. That's where the -- so the accuracy and duration finding was a primary design consideration.

DR. SCOTT: Thanks.

Steve, I saw your hand up. We have one minute, though. Is it very quick?

DR. RUBERG: It's a quick comment. So great study. I loved the graphics and tabulations of the simulations, made it crystal clear about controlling type 1 error, getting close to the true estimate, so good job there. And then the only other comment I make is, this is a duration-finding study.

And I'm thinking about how does this maybe apply to dose-finding studies and maybe more typical drug development where maybe there's a safety issue with a drug, and you just don't quite know how far to go.

So let's do a low dose, collect some data, and then step it up and at some point, we'll bump into the adverse event that people are worried about or whatnot, but I do think what you've presented is very good and perhaps could be generalized to, I'll say, typical drug development dose-finding studies, so that's all.

DR. LEWIS: Yeah. I don't think the mathematics knows whether X is time and days or milligrams. Thank you.

DR. SCOTT: Thanks again, Dr. Lewis.

So our next speaker is Dr. Karen Price from Eli Lilly, and she's going to be talking to us about a very interesting master protocol design for chronic pain indications.

DR. PRICE: All right. Thank you so much. It is great to be here. And just to echo what Roger was saying, thank you for the invitation; really excited and honored to represent this master protocol. I'm going to start first with some acknowledgments.

This has been -- these are several of

my Lilly colleagues. Again, I'm honored to represent this on behalf of this team. It's been quite the journey. We submitted this and so as I'm going to be talking today, all the emphasizing what it was like to go through the CID pilot program and the conversations that we had, but we were one of the first coming through.

And so it's been several years' journey, and we've seen a lot of fruit out of this protocol, and so I'm excited to share that today. So what I'm going to do is start with an overview of the chronic pain master protocol, sort of set the stage and describe the design and the value of it.

I will give some statistical details. In particular, as Dr. Scott had noted, we have an emphasis in the CID program around borrowing as well as simulations, so emphasizing some of the things that we considered during the discussions, I wanted to show an interactive tool that we utilized as part of the CID pilot conversations.

I think that because simulations are so

complex, as we've been talking about and can slow reviews, to the extent that we can make them more interactive and help put into the hands of FDA reviewers the ability to explore some of these operating characteristics more efficiently will see a lot of gains.

And so I just wanted to show that and then talk a little bit about moving forward some things that we need to consider to further advance CID. Okay. So in terms of CPMP, just wanted to start the importance of this case example. So chronic pain is a public health crisis and is one of the main reasons that people seek care.

It's estimated that in the U.S., over 20 percent of adults live with some form of chronic pain. However, the probability of approval of novel analgesics that have completed phase 1 is significantly lower than for other novel drugs across other diseases. And furthermore, the current treatments are things such as opioids and nonsteroidal anti-inflammatory medications, which lack efficacy or

have some safety concerns.

So one of the things, I think, wanted to highlight here is that a lot of times when we think about CID, people do think about rare diseases, pediatrics, which was mentioned, and is certainly an important place for considering these types of designs.

But obviously in common disease states as well, we see a lot of value and, furthermore, is where we have a lot of data that we could really use and enhance these designs, so great setting for us to be as creative as possible to really help meet this unmet need for patients.

There are a number of innovations that can be brought to bear when steady and chronic pain as you've heard what I'm going to focus on. And one of the things we decided to do was this master protocol. It served a number of -- it helped us in a number of ways. One issue is oftentimes, phase 2 studies will focus on one type of pain.

Depending on the sponsor and how that

progresses, if the pain type selected is maybe not efficacious, the drug may be abandoned. And so where it might have been successful in a different pain type because it was only able to be tested in one -- again, we see a lot of abandoned molecules.

So what we wanted to be able to do is within the same protocol to test these three different pain types, and they were selected to be diabetic neuropathic pain, chronic lower back pain, and osteoarthritis pain. And so we were able to look at a variety of novel analgesics, and this resulted in reducing the size and cost of the studies versus had we done these independently and/or may not have even been able to test some of these drugs in these different pain types.

So this slide summarizes the framework. You can see the three pain types coming down the rows there: osteoarthritis, chronic lower back pain, and diabetic peripheral neuropathic pain, really representing three different types of pain, so they were selected very purposefully, and a lot of

discussion went into selecting those pain types.

Across the columns, then, you can see the different assets coming in. The protocol is sufficiently flexible that a given asset does not have to study all three pain types, most have. I believe all have so far. It's not a requirement, but it is allowed, and we have that flexibility to look at the three pain types being enrolled simultaneously.

Again, hard to depict here, but these are coming in. They can come in concurrently, or they may be coming in completely separately. There can be lag time and so forth, so it's an open protocol allowing those molecules to come through.

I just wanted to share also kind of how we set up this protocol. So there were three tiers in this protocol. So we had the high-level master protocol; that established the entry criteria for this master protocol, the randomization scheme, what the common -- the hypotheses that we were testing, the advanced statistical modeling, and various operational features were included in this master protocol.

It's really where we wanted to have the standardization that would be required. The disease state addenda, then that brought in the three different pain types. So anything that was unique where maybe we had an additional measure to be evaluated for a certain pain type, that is represented in the second tier there.

And then the intervention-specific appendices, so that's where the drug information came in. As I will talk about, we really wanted it to be as standard as possible, but if there was some unique feature for the drug, perhaps a drug-drug interaction or some tox consideration that at this moment in time needed to be different, that would be allowed.

But we did have a governing body overseeing this protocol that would make those decisions about whether or not that modification for a specific intervention is necessary. And then this just shows the flow of how we sort of named things and had again this chronic pain master protocol, the three disease state addenda, and then we can add pain types,

so it is flexible to allow for additional pain types to come in if we decide to do that.

And then again, the intervention-specific elements. Okay. So next I wanted to talk about, then, was what were some of the strategic considerations that we were thinking about and highlighting the quote from the Woodcock-LaVange paper from New England Journal. The common denominator here was we wanted to answer more questions more efficiently and in less time.

So what we decided to do from a strategic perspective is we knew that we had several molecules that were going to be coming forward. We had a lot of opportunity in our portfolio and so we -- but that they were going to be coming in at different times, and we wanted to be able to make the best decisions about which ones to move forward and in what pain types.

This is a phase 2 proof-of-concept study only, and so we are very specific that it is about, you know, hitting it hard, understanding is

there a signal or not and proving that concept. Following this would then be more robust dose finding, ultimately moving onto phase 3.

So this particular design was not constrained by registration requirements, although, of course, we wanted to maximize transferability to phase 3. And as I mentioned, we would have phase 2B to follow. We did limit sites to North America. And then, as I mentioned, the master protocol structure is established in allowing for that flexibility within the ISAs as needed.

Some key features of this master protocol. All of the ISAs have the same scale. It's the numerical rating scale. This is the primary endpoint across all three pain types and across all molecules, which is really important. Again, pain is a very subjective indication, and so we have the same scale.

We had the same sites. We had the same -- there was not a washout period, and everything was consistent, which really helped to remove some of

the confounding factors that often enter when we're looking at independent pain trials. There were other scales: physical functioning, emotional functioning and so forth.

We had standard data collection, similar visit schedules, and then, as I had mentioned, there was a master protocol team established to analyze the efficacy data to make decisions about, as I mentioned, any difference that an intervention might need. Is that really necessary to make the difference because we wanted to keep things as standard as we could, while allowing proper flexibility where it was absolutely necessary.

The primary efficacy analysis is a Bayesian mixed model repeated measures. This was the primary efficacy analysis. Again, it was using the NRS. And what I wanted to emphasize here is the primary critical success factor is in the framework as you can see in that sort of second sub-bullet, the probability that the treatment difference is less than an effective interest, and it's less than because

negative is good here, so more negative.

The probability of that difference exceeds a threshold that's again established before the trial starts. And so we had this framework in place at the master protocol level, but we recognized that as an intervention comes in, the effective interest might change over time.

In particular, if there are molecules that are successful, then we need something more even higher, perhaps, but then also that probability threshold may evolve over time. So we allowed that flexibility and just insist that it's specified prior to that specific asset enters into or starts in their ISAs.

All right. And some of this I've talked about, but just maybe to restate and touch on a couple of additional things, you know, anytime working on a master protocol, it really is this tension between what is going to be standard and what is allowed to be flexible. And we were -- it was very important given, as I mentioned, the subjective nature

of pain that we have as much standard as is possible.

So again, same primary endpoint. We had the same probability of getting placebo, so 33 percent randomized to placebo across the ISAs. There is a double-blind period of eight weeks, but the -- on the other hand, if we have a molecule that's coming in that is not able to be studied for eight weeks, we could do -- we had the flexibility to allow it to be studied for four weeks, and then it would be double-blind placebo for the final four weeks just maintaining that same treatment duration.

Again, common visit schedule and inclusion criteria, but from a flexibility -- the ISA can specify a sample size. The amount and type of borrowing will obviously evolve over the course of the study because as we're gaining more information, we'll now have new information that we can utilize.

And then, as I mentioned, there can be inclusion-exclusion changes or scales, visits added, but governed by one group to make sure that it is necessary. There's been a number of statistical

benefits from this. I think probably this top one is one of the most important ones where we're able to directly compare within and between pain types.

We had an advisory board before this, the master protocol formally started, and one of the members was commenting: How often do we wish a drug was in the same protocol, and we didn't have to rely on the debt analysis?

And that's exactly what we were looking to achieve, that we can do indirect meta-analysis or other forms of meta-analysis, but especially in the context of pain, which is subjective, we wanted to reduce a lot of the confounding factors. And it was a lot of enthusiasm from our FDA counterparts during the CID meetings about our ability to do this across pain types.

Again, a lot of times will utilize different endpoints, so maybe it's a VAS. How does that compare to the NRS? It's therefore hard from a meta-analysis standpoint to do that and/or the consistent -- it's lacking consistency and collection

of safety and viral marker. And ultimately, the master protocol affords the opportunity to reduce sample size in both the active and placebo arms.

And so thus far, we have seen as mentioned a great deal of impact from this trial, including reduction and cost, reducing time from, say, protocol approval to when a first patient is dosed, time to data lock, time to decisions, and enrollment time. Completed 12 proof-of-concept studies in 38 months, validated three novel targets.

And this is much shorter than what we would see independently, but is also hard to fully compare because, again, often, we may not get all three pain types because if the first one isn't successful, it may be abandoned, so -- okay.

So let me take a little bit of time, then, and go into some of the statistical details. As I mentioned, I'll focus on the borrowing conversations we had and the simulation. For this master protocol, there were three main sources of borrowing that we considered.

The first source is not unique to the master protocol, can always be considered. So we spent more time talking about borrowing from these second two, the ability to borrow from placebo information, from another ISA within a pain type, and then the opportunity to borrow treatment effect information between pain types.

I'm not going to into a lot of detail on the borrowing approaches. You'll hear more, I believe, from the next speaker on how some of these approaches work. What I will mention is that we typically are thinking about two main buckets of borrowing approaches, one being static, and that includes things like pooling or power priors, dynamic borrowing.

On the other hand, things like hierarchical modeling, mixture priors or commensurate priors. And there's an appeal for dynamic borrowing, where what happens is that if the incoming data is consistent with historical data, it will borrow more. If it looks different, it will borrow less. And

toward the end of my presentation, I'll show an example where you get sort of pooling versus the hierarchical modeling. You can see that when the new data comes in and is different.

So what we ended up doing, and as you'll see later with the tool, is we were -- we spent a lot of time talking through with FDA and comparing things like pooling or power priors more static. On the other extreme, separate, so not borrowing at all, and then something "in-between," and so say that the hierarchical modeling approach.

So that was where we spent a lot of time understanding what were the -- how decisions differ based on incoming data. When the rubber hit the road or, you know, when actually, the reality of conducting this trial and working with teams and data coming in, there were a number of challenges that were encountered, so certainly things we thought about prior to the design.

But then, again, this is an ongoing trial, and assets are coming in and going out and

things change. And so some of this we've sort of figured out, others we haven't. So I think some of these are still outstanding questions.

As I mentioned, you know, when inclusion-exclusion criteria change necessarily, how does that impact the ability to borrow? Do we use pooled placebo or ISA and a safety review? We did allow for patients to repeat enroll into later ISAs after a proper washout period. What do you do with that patient's information from earlier? Is it borrowable? And what does that even mean?

So there was also hesitancy to borrow from some team members, so there's a mixed feeling on whether or not to borrow. And I think similarly at FDA, right. When we encounter, there are different viewpoints on the -- what should be borrowed and whether or not to borrow. Again, what is the best approach to borrow?

We spent a lot of time talking about should we borrow across pain types and then also thinking about placebo expectation bias and how does

that change over time and affect your ability to borrow, so those were the borrowing approaches.

Again, we spent a lot of time simulating these trials -- or this master protocol and then looking at different ways the ISAs would come in and the type of data that would be coming in. Some of the key factors that -- so we did a lot of simulation as part of CID before the trial started and then as an asset comes in, additional simulations are conducted given that now we have new data that's been completed from earlier ISAs.

But before even the master protocol started, we looked at different scenarios thinking about how much placebo data could be available from either completed or ongoing ISAs. What would the different treatment effects between pain types that we could observe, and how does that affect how it's being borrowed?

We looked at scenarios where there could be placebo drift that could occur. What we would do, and how would we handle that? And we knew

that there was the possibility of having a different route of administration. And so if something were, say, injected or oral, or given more frequently, those are things that can affect the subjective endpoint of pain, so how would we handle that?

So we spent time looking at different scenarios and working through those with the FDA as part of our conversations, and we did think about fixed versus longitudinal timepoint settings. This slide just summarizes the types of things that we were looking at in the CID conversations.

Very similar, I think, to quite a bit you heard in the earlier talk: power false positive, bias, standard error, understanding the operating characteristics, as I mentioned, when the underlying true placebo response shifts over time. And again, that could be due to some new drug that is approved, you know, what could happen to affect that true placebo response. Benefits of power increase or sample size reduction.

We explore the various ISA initiation

and lag times in enrollment, dropout rates, as well as the quantity of data that would be available. All right. And then the final thing I wanted to touch on with the -- on the simulation front before concluding here is I wanted to touch on this R Shiny application that was a feature of our CID pilot program conversations.

And as I mentioned at the beginning, the simulations are -- there's a lot of simulation output and typically, it's provided in paper. We may not provide a scenario the FDA is interested in, and then there's an iterative nature there where we're providing that.

And so what we really wanted to do was to build something that could be more interactive, both for us, but also for FDA to evaluate this design, reduce the number of -- amount of paper that's required and provide more interactive visualizations.

And with the ultimate goal that I think -- and maybe part of what we'll talk about more this afternoon is the importance of modernizing this

collaboration to help speed the review of simulations with -- and just to continue to use the best computation we have, the best tools and technology to make this easier on everyone, so we can really get to the heart of the matter and really have those cross-functional conversations that Dr. Lewis was talking about.

That's really where we want to be spending our time, not trying to make sure we're reading a table correctly, but rather really getting into those great cross-functional discussions. And so more interactive simulation output we think could be very helpful. This application you'll see in a moment -- and I just have a short video demo. It's more just to show what we mean.

The specific application has two main parts. One is if you have -- you may want to look at a single realization of a master protocol. So suppose you're really interested in, how would I -- what would be the output, the inference made in terms of the probability of achieving a CSF under a certain given

single scenario.

So it fits a model you can look at different amounts of historical data that would be available, and so that's one part. And then the other part is it simulates multiple trials, again, to evaluate those operating characteristics. And so the user can enter data from completed ISAs, simulate future ISAs, vary the model prior, and so forth. This is just an example.

Again, the -- actually what it would look like can vary. And the intent here is not to go through this whole thing. I'm about to start it in just a minute, and you'll see the interactive nature of it, but the idea here is, again, to give the user the ability to conduct, look at scenarios, or look at situations they may be interested in.

Again, this first part is the model and the priors. And so you can see there's click down boxes to specify the model, prior distribution for tau, for the hierarchical model that's on the placebo effect, and there's a treatment effect prior as well.

And then, as I mentioned, the user can enter the example data, and it will run and provide the posterior distributions, and you can zoom in.

Here, you're seeing zooming into and looking at the different borrowing approaches. Further down, we can look at the probability of achieving the critical success factor, maybe the user wants to look at different effects of interest or probability thresholds.

And the green there -- I think what's really nice about this is you can see which borrowing approaches meet the CSF versus those that don't, so you can see where a different borrowing approach is yielding a different outcome.

There are some diagnostics that are included here, so we can pick the parameter you're interested, check the diagnostics, and then you do get the actual parameter estimates table with posterior means, standard deviations, and so forth.

Now, we're going to look at simulation. So here we have simulation scenarios that have been

included, and this is quite -- the user can play around with this and then run the simulations. Kermit the Frog is going to help us run those in the background, and then provide the output to the user. Again, these scenarios on the upper left-hand side, those are what the -- you can enter different scenarios that you may be interested in.

Here, we're just selecting a parameter and, again, you can determine which parameters you're most interested in and then look at the operating characteristics. And then finally, the bottom portion here, and this is where I'll touch on something I mentioned earlier.

We're looking at -- and you have different borrowing approaches, and we're looking at the probability of achieving the CSF. In this case, we did use a scenario where the historical placebo information was very different from the incoming placebo information.

And so as you click through the different borrowing approaches, you'll see the

differences. And then as we scroll down, here we're looking across all of the borrowing approaches that were considered. You can see in that middle part that that pooling is resulting in a very different outcome versus the hierarchical, which is on the left; or separate, which is on the right, showing us that the hierarchical model is in fact borrowing less in this case because that new data is different.

So just, again, an interesting scenario and the type of things that can be evaluated through an app like this. And then couple things finally, and as we to conclude here, generally, our experience with the CID pilot program, there were very positive interactions between Lilly and the FDA.

Absolutely improved the master protocol, and we think advanced how we're thinking about borrowing of information, master protocols, and simulations. So again, very collaborative, progressed how we could communicate these methods, simulation plans and results.

And because they were so positive, we

do need to continue to have these pathways and avenues to have these types of collaborative discussions. Really to the extent that we can have some informal pathways as well, which this was -- while it was formal, it had a more informal feeling because we were talking about the different methods, talking about what data sources might be useful, these sorts of things.

Again, I touched on the R Shiny collaboration. In terms of opportunities for improvement, I think our biggest recommendations would be the timeline is long. So it was actually -- because of the situation going on at Lilly with the assets that were coming in, we did not delay anything by doing this, but if in normal cases, we would not have been able to probably go this route because it is a timely process.

So we knew well in advance what was coming and therefore sort of had the luxury, if you will, of going through this. We do think that it would be great to have also opportunities especially

for something like a master protocol, but any design -- as data comes in, what are we learning?

And I realize this forum is part of that, but having some more conversations with sponsor and FDA continuing, we think would be very helpful to continue to share learnings. And then ensuring that we continue to have consistency in those attendings. And so finally, what are some things that we can think about moving forward?

And maybe part of what we'll talk about -- you know, how do we continue to share learnings across divisions? We have taken additional master protocols into other divisions, and there are different types of feedback that's being provided.

And so how can we have more consistent across divisions in terms of the feedback and the expectations. I mentioned improving the infrastructure, continuing to look for cutting-edge technology to enable us to be faster as we evaluate simulation results and interactivity.

I've already touched on the meeting

schedules that can accommodate the speed that's needed, improving the education of statisticians and medical, both again in sponsors as well as at FDA. And then some other things that will continue to play into how do we advance these innovative designs?

What's the role of AI/ML, other new technologies? And then how does the fact that there's so much move towards decentralized trials and digital health technologies. How does that affect how we borrow, if we borrow, these types of designs, so things to think about. So with that, I will close. Thank you.

DR. SCOTT: Thanks so much, Dr. Price. Would anyone on the panel like to comment or have a question?

Dr. Ruberg?

DR. RUBERG: Well, that's complex and innovative. So I'm just wondering. You had different probability thresholds and effect sizes in your decision criteria. I'm just wondering if you can talk a little bit more about so how are decisions made.

You know, one asset completes osteoarthritis before it completes lower back pain and -- or they all complete at the same time. In one disease state, you meet the critical success factor, one you fail miserably, one's right on the borderline. Can you talk about the complexities of the decision-making about --

DR. PRICE: Sure.

DR. RUBERG: -- what goes forward or what gets held back, or of these three assets, this one has the biggest probability of having a clinically meaningful effect? And I was just wondering, it's not only complex in its setup and the analysis options, but also in the decision-making framework.

DR. PRICE: Absolutely.

DR. RUBERG: So a little -- maybe a little --

DR. PRICE: Sure.

DR. RUBERG: -- bit about how that worked out, or how it is working out, I should say.

DR. PRICE: Yeah, sure. So yes, it is

complicated. I think this is where having that centralized group has been vitally important. So we have one group who's overseeing this master protocol and of course, there is some rotation, but you have some members who've consistently seen this protocol, know it inside and out, so that has helped.

It typically, in terms of your question on the timing, there is some lag that can happen where, like you said, maybe OA pain finishes and then a little bit later chronic lower back pain, but we know well in advance when these things will happen, and so there are decisions made before we start seeing data.

Are we going to pull the trigger on if it's highly positive? Are we going to go pull the trigger on the next phase 2B, or are we going to wait for the data? So that is -- the importance of making these decisions before we see data, I cannot say that enough. You know, once you start seeing data, then it gets very confusing.

So we have basically a preplanned,

here's how it's going to be communicated; here's how we're going to make the decision. What's the level of efficacy required in order to do that phase 2B versus -- so there could even be different thresholds for different decisions.

So I think I said the main thing would be preplanning that, understanding the timing. If it's a few weeks, it's probably worth waiting. If it's going to be a while, then maybe we need to do something different.

So really, like I said, the centralized group, preplanning, and understanding how does it impact -- the last thing I'll say is sometimes that maybe we say, okay, we think we're going to move this one forward, so we start sort of planning that trial, but we're not going to do too much until the next thing comes out. Lot of things to think about.

DR. SCOTT: Dr. Lee?

DR. LEE: Yes. Hi, Karen. I really like your Shiny app.

DR. PRICE: Thank you.

DR. LEE: Okay. And I think it really is a very good tool to enhance the collaboration and even communication with FDA. So the question I have is that -- I'm sure this is not probably available yet; right? But a certain point, it'll be great that if you can turn it to a more general-purpose kind of a tool, as an education tool and maybe publish it and so that people can learn how to do this. So any plan for that?

DR. PRICE: Yeah. Well, so I think there are -- first of all, I should acknowledgment people like Eric Nantz and Michael Sonksen for working on that. Eric Nantz is very active in the R community and is working with counterparts at FDA to really understand what technology is required.

And so it's not just Lilly; right? It would be -- these are more scientific-working-group-type opportunities where they're looking to see what technology does the FDA have? What technology would the FDA need? Because that could certainly be part of some future PDUFA

negotiation or whatever the case may be as we need to advance this to allow for both FDA and sponsors to have it. So yes, a lot going on. And there are things that could be available to do this, but appreciate the feedback, and we should continue to try to push that.

DR. SCOTT: Dr. Lewis?

DR. LEWIS: So on the same theme, I think one of the more challenging discussions people often have in discussing designs that use borrowing has to do with a choice of the hyperprior that's used for higher -- for dynamic borrowing with a hierarchical model.

And I really like that you were able to show with the app the performance in a specific situation where the historical data did not well match the data that was coming in currently. But there's the challenge in picking the hyperprior is that you have to account both for that situation where you have discordant information and the opposite situation which might occur where the information is highly

concordant.

I'm wondering whether you've thought of a display that shows both of those simultaneously so a person can just glance sort of at a prior picking dashboard and sort of see where it does well and where it doesn't because we've struggled, frankly, in figuring out how to communicate the implication of these choices to these collaborative design teams.

DR. PRICE: That's great.

DR. LEWIS: And make sure that we have a prior that reflects, you know, the best balance between performance when you have concordant data and performance when you have discordant data.

DR. PRICE: So great. We'll think about it, I guess, is the -- yeah. Currently, the scenario is entered. We've evaluated another scenario. And so we look at -- to your point, we look at -- everything looks great; things don't look so great. But to do that all in one, yeah, we'll think about it and get something in there.

DR. FOLLMANN: Yeah. I guess I'm also

a fan of the Shiny app. What I particularly liked was when you could have no borrowing and get the posterior distribution for that to compare it to various levels of borrowing, so it's sort of made visually and precisely how much, you know, the borrowing was contributing.

Had a comment, I guess, inspired by an earlier question. So you have a particular drug and you might be evaluating in back and osteoarthritis, and you might fill up the -- you might stop randomizing in the back, but continue osteoarthritis. Is that how it might play out, or do you just keep randomizing until you get the answer for that drug?

DR. PRICE: Oh, so what happens -- let me -- what I think you're asking. So there's a preplanned sample size for each pain type. And so the -- what you were seeing there, those were all simulations or made-up examples. And then the trial, let's say that we're going to have a couple hundred patients in a given ISAs randomized two to one, we enroll until that is done for that ISA.

But now, if osteoarthritis is ahead of chronic or back pain, we don't keep enrolling osteoarthritis. Once it's done, it would stop, so that's why they can finish differently.

DR. FOLLMANN: Right. Well, then just a minor comment. So osteoarthritis might stop and you continue enrolling, you could in theory sort of borrow into the future where you get placebo group for osteoarthritis --

DR. PRICE: Yes.

DR. FOLLMANN: -- that was concurrently enrolled, so --

DR. PRICE: That is correct.

DR. FOLLMANN: Yeah.

DR. PRICE: Yes, yep, that's exactly right.

DR. SCOTT: So we are almost at time. Dr. Price, just one quick question from me. I hope it's quick. You mentioned some internal hesitancy to borrowing. What was the nature of that hesitancy? Was it scientific or strategic or regulatory?

DR. PRICE: I believe it's probably a -- in the case because it was -- this is a proof-of-concept, I don't think it was so much regulatory concern. This would be more if the borrowing changes the decision; I'm concerned about that. So it's more like a personal, maybe not fully trusting the borrowing.

DR. SCOTT: Thanks, yeah. It's not that dissimilar to some people internally -- that way, yeah.

DR. PRICE: Yes, yes.

DR. SCOTT: Okay, all right. Thanks, everybody, for two great talks and for your attention and for the questions. We now have a 15-minute break, and we will resume with Dr. Pang's talk at 10:50. Thanks.

So thanks, everybody, for coming back after the break. We'll give people a moment to get settled. And in the meantime, I'll introduce our third and final case study for the morning, and then it will be lunch. Herb Pang from Genentech/Roche is

going to be talking about their CID case study involving the use of a hybrid control in diffuse B-cell lymphoma.

Dr. Pang?

DR. PANG: Thank you, John. And thanks, everyone, for coming. And also just like Karen and Roger, I would like to thank the organizers for the invitation. So today, as mentioned by John, we'll talk about a case study of hybrid control design in diffuse B-cell lymphoma.

First, I will briefly go over some introduction about the study designs with external controls and then follow up very shortly talking about the CID pilot program as well as the timeline, and we'll go into details about the Genentech/Roche CID pilot.

After that, we will also cover hybrid control ongoing research, and I will also conclude with a summary. As you may know, for study designs using external controls, there could be different ways you can go about doing it. On the left-hand side,

there's threshold crossing benchmarking approach, where in the external control arm, you will have applied randomized control trial inclusion and exclusion criteria to have a restricted external control arm, and then it would read out the outcome.

In the middle one, you have a single-arm external control study where you have the external experimental arm in combination with their external control arm and in forming the cohort, but then you would do covariate balancing or adjustments so that you can make the comparison between the two.

And finally on the right-hand side, you have a hybrid external control design, which is the one that we will talk about today. You have 200, for example, randomized control trial subjects, and then you spread it into, for example, three to one randomization and then having 150 in the experimental arm, and then 50 in the control arm.

But unlike the typical randomized control trial setting, you also incorporate and augment the control arm with external controls. In

this case, for example, you have roughly a hundred. And then you would run this trial and read the outcome. Okay.

So in considering the choice of when to do a hybrid design, you may think of two extremes: on the left-hand side, you have a randomized clinical control trial; and on the right-hand side, you have your fully external control trial.

One factor is whether there's a medical need. And for the randomized control trial setting, you can think of it as if you have an effective control available, it could be a good choice. On the other hand, for the fully external control trial, it could be that there's clear-met need; however, there's no effective control available. And it is kind of like a spectrum, and hybrid design would fall in between.

For the target indication on the left-hand side when you think about a randomized control trial setting, you may think of things that are more tied to first label or a new broadline

extension. And then for the fully external control, you may have line extension and similar indication or indication with well documented standard of care. Choice of endpoint is also an important consideration.

So for the randomized control trial setting, you may consider whether there's a specific endpoint that you don't have data readily available from external sources, then you would -- probably it's wise to choose the randomized control design.

On the other hand, if you want to do the fully external control part, you want to have a robust endpoint where you have data available from the external sources. For example, overall survival, et cetera. So again, the hybrid design would fall in between the two.

In terms of anticipated effect size is also an important consideration. For the randomized control setting, you may think about having it as in cases where you have modest effect size anticipated based on some observed prior information.

On the other hand, for the fully

external control study, you may want to have more compelling effect size. And then finally for population size, having a large population and no changes in recruitment is probably an important factor. You want to get the trial completed on time and within a reasonable time frame so you can consider NLCT.

On the other hand, for a fully external control study, you may have potential issues with recruitment or some ethical challenges, and you would choose a fully external control. And again, for the hybrid design, it's somewhere falling in between.

So as we may know, there are different potential biases that can come from the use of external control sources. These include, but not limited to the set, such as selection bias, where patients enrolled in clinical trials are different in some ways compared to patients that are treated in clinical practice.

Other biases could include calendar time bias, where patients treated in the past do

differently than those treated today. Regional bias. If you have enrollment of subjects, not just within the U.S., but in other parts of the world that patients coming from different regions could have variations.

Assessment bias. Knowledge of therapy that can influence assessment. And study bias would be patients in clinical trials, they have different outcomes than in clinical practice, for example, placebo effect or different care. As I mentioned, this is not an exhaustive list, so there are other biases that you also need to consider.

I think the FDA also had a guidance that provides some knowledge about these potential biases, but framed under more the design setting, which is also very important in terms of how to think about these biases. Some important thinking going behind how to mitigate potential biases to understand the Pocock Criteria, which was developed many years ago.

Receiving a precisely defined standard

treatment the same as randomized controls is one important factor. Being part of a recent clinical study which contain the same requirements for patient eligibility is another one. Matters of treatment evaluation should be the same.

Previous study must have been performed in the same organization with large the same investigators, as well as there must be no other indications leading one to expect different results between the randomized and the controls that are historical.

Distribution of important patient characteristics should be comparable in those in the new study. So I think, earlier, Karen also touched upon this about the dynamic borrowing design. So let's take a look at this illustration on the left-hand side.

At the top, or the extreme top, is the no-borrow scenario where you only use the randomized clinical trial. At the other extreme is the bottom, which is borrow more, and that would be the case where

you consider full borrowing, where you just simply pool the two controls together and then estimating the treatment effect. And dynamic borrowing is somewhere that falls in between these two extremes, so it belongs to a spectrum.

And even within dynamic borrowing, you can consider different types of prior that would be more; for example, skeptical prior would be more conservative. And then you can have a more aggressive prior if you have more -- understand it will be more optimistic on the external control.

And later on, we will illustrate in our case how we consider choosing between these two. So Bayesian method presents a natural way to handle combination of data, external trial data can be use to setting up the study prior, and dynamic borrowing framework, as mentioned, could be an important, allowing you to kind of understand the difference between the internal and external control.

So when you have more differences between internal and external controls, you will

likely have the prior which is more flat, so in those instances, you would borrow less. And then on the other hands, if you have more trust, then you have a prior that has more, like, weight, and then that would provide a borrowing scenario of borrowing more, so that's the commensurate prior.

In the publication that was published about ten years ago, they discussed dynamic borrowing methods, were able to achieve similar power gains, which is the color in green and blue. But then in the full borrowing scenario, which is more extreme, that we described earlier, which is in red, it has more type 1 error inflation.

So the dynamic borrowing method can achieve similar power while having better type 1 error control than full borrowing. So I won't go into detail about the CID program here because John already described it very well. Just want to mention that it has been a great opportunity and also a wonderful experience for us to join the program.

So this is the typical timeline of the

CID program and the process how things happen, think. And like the scenario that Karen had, I think in our case, we actually needed more time in the end to do the simulation, so it was actually us asking the FDA to allow us to provide more comprehensive understanding of the method for the second meeting, and the FDA were very collaborative and flexible in allowing us to meet later.

Why innovative design was needed for our case? Unmet medical need, as described earlier, in certain subgroup of DLBCL was the case, DLBCL is more common non-Hodgkin's lymphoma worldwide with 25,000 newly diagnosed patients in the United States annually.

And standard of care for first-line DLBCL has been established many years ago and is well characterized and well understood. Patients in certain subgroup of DLBCL have a poor prognosis and consequently, a high unmet medical need.

So borrowing patients from control of another study can help us enroll fewer patients in the

control regiment, allow us to shorten the study time, and also conduct more efficient trial by sharing control between trials. Here was the timeline of the brief phase 3 of development for the first-line DLBCL and pathway to the CID pilot.

And initially, we receive encouraging data about the phase 2 study compared to historical RCHOP control, especially in the biomarker-positive patient group. How the control can potentially limit the number of new patients exposed to a well-established standard of care?

FDA Type C meeting on proposed phase 3 in the biomarker-positive of experimental, plus RCHOP versus RCHOP, three to one randomization, plus externally borrowed control were selected from the internal study. And the agency recommended, actually, the primary and asset population and assets planned to be on the randomized trial were found in the external control.

And an analysis population can be used for support of analyses. The focus of the updated

design with the external control was for the secondary endpoint overall survival because overall survival is a clinically meaningful endpoint with minimal ambiguity in its assessment.

So then we joined the FDA CID Program, which really provided a great opportunity for us to build on utilizing these external controls and discussion and also within a very collaborative framework. So eventually, we came up with this design where we would randomize subjects two to one, with the novel combo with 280 subjects and the RCHOP group about 140.

And then we supplemented the external controls with about -- internal controls with the external control of about 100 subjects. As I mentioned before, the primary endpoint for the study is PFS, but that was assessed only with the randomized subjects and a key endpoint is where we actually combined randomized subjects and also augmented the internal control arm with the matched external controls.

So the external control patients were selected from a contemporary ongoing internal clinical trial in an intent to support early OS analysis at a time of primary PFS analysis, so that's the shortening the time to get the readout. Randomized study with external control arm used matched external controls through, as we mentioned before, dynamic borrowing.

So the rationale for sources of external control arm is for prospective plan to select external controls from an ongoing contemporary interval randomized control clinical trials is consistent with the eligibility criteria planned, aims to targets similar and investigators of sites.

Overall survival, as mentioned before, also very clear and clinically meaningful. So five out of six proposed criteria with the Pocock Criteria that was mentioned earlier were met. So what was our kind of analysis flow diagram look like? So the first one is to look at control comparability evaluation by applying inclusion-exclusion criteria. As far as flagging based on factors that have significant

differences between internal and external trials.

The next step was to utilize propensity score matching to match patients between internal and external control trials using propensity score matching, which enhances the covariate balance between the two groups by filtering out unmatched patients.

Finally, we applied Bayesian dynamic borrowing method, which automatically downweighs external control based on the agreement between internal and external controls. And again, as we mentioned, we provide inference for the overall survival for the borrowing scenario, and sensitivity analysis would follow the main analysis.

I won't go into all the details about the simulations, but briefly we'll cover kind of at the high level what we looked at. And the main goal of the simulation scenario is to evaluate the proposed statistical method, which is a combination of propensity score matching with Bayesian commensurate prior approach.

We examined a few things to understand

the operating characteristics, including a varying magnitude of differences in baseline characteristics, which we will see in the coming slide, as well as the different choices of prior, which may influence the degree of borrowing. We will at high level cover how we looked at violation of various assumptions.

So on the left-hand side, you see a plot of type 1 error ad at the bottom, you have different types of scenarios. So there's a scenario on the bottom left, which is the no difference, followed by the moderate difference between the internal and external controls and also the scenario where you have a large difference, plus the benchmark, which is the no borrowing case.

So as you can see, the no -- full borrowing case would have the highest type 1 error inflation, while the dynamic borrowing with half- Cauchy is doing better in the type 1 error control. And in terms of the no-borrowing reference, it's not too far from that.

As for the power gain, they're also

shown on the right-hand side with the different scenarios: no difference, moderate difference, and large difference. And of course, given that the borrowing case demonstrated an increase in power gain, but we also need to take into account the type 1 error inflation and as we can see, the dynamic borrowing with half-Cauchy does have a decent power gain.

In terms of violation of various assumptions, we won't cover the results in detail here, but what we looked at include understanding and simulating to see anywhere else where we have, for example, observed unmeasured confounder, as well as understanding if the survival curve distribution is different from the assumptions that was made.

And we also looked at if there's a nonlinear or non-additive effect model, how does the operating characteristics performed. So in conclusion, we found that in general that dynamic borrowing with the conservative half-Cauchy prior was able to have a good average error rate, weighted type 1 error rate, and a slightly inflated maximum type 1

error rate, but is the most conservative one that we observed much better than the aggressive as well as the full borrowing scenario.

So what was some feedback on this, the potential using OS with external controls? There are several aspects we learned, including model assumptions assessment, which is standard analysis typically requires fewer assumptions, so these borrowing scenarios have more assumptions, can be less standard, so we need to understand and assess them just like as we demonstrated with the evaluation of assumptions simulations.

The need for pre-specification, I think earlier speakers also alluded to that. And another consideration unique to this case would be what could hamper inclusion of overall survival in the label. For example, whether the model assumptions appear to be met and the outlying subgroup effects, is the endpoint credibly captured or not, overall conduct of study, missing data, as well as, for example, whether the baseline characteristics are the same.

In addition to statistical consideration is there are other considerations that you need to think about, which include, is the analysis of the summary -- summary of analysis clear? Can it be interpreted by clinicians as well as, would it provide available information?

So with these more novel designs, it's unlike the typical case where you can decide on parameters and have fixed scenarios, so you need to do more extensive simulations. And for the case that we did with the CID program, we actually had many scenarios per FDA meetings. So we have a couple of meetings, and each of them has more than 20 scenarios that we investigated.

So the implications would be that we have to plan earlier, allocate sufficient time and resource, as well as having, for example, software being available. So right now, we have this open-source software. I think Jack asked a question earlier about having open-source resources so that other people in the industry can also benefit from it,

so psborrow2 is one such example.

And of course, within internally, we need help from other Roche statistical software engineering team and methods experts. So learning from the CID program certainly helped us a lot and a real initiative that we have that we will cover very briefly in the coming few slides is another FDA initiative, which is the U01 grant.

So in 2020, FDA awarded full grants for the U01 mechanism for the exploring the use of real-world data to generate real-world events in regulatory decision-making and think they actually have a new batch that came out last year as well. And if you are interested in learning more and you're a member of the ASA BIOP Section, there's actually a session this Friday that would actually have some speakers that have been awarded these grants.

So this is an ongoing research and which is entitled Applied Novel Statistical Approaches, develop decision framework for hybrid randomized clinical trial design, and combining

internal controls with patients from real-world data sources.

So this is in collaboration with University of North Carolina. One of our research work has been recently published, utilizing something different than what we presented in the CID program, which is a case weigh specific power prior method.

So as we saw in the dynamic borrowing case that we have, we actually have this maybe assessing the agreement between internal and external controls and then putting a weight, but the weight is the same across all subjects. So this particular approach would actually have a case-specific weight and is soon to be available online, but there's also an archived version of this paper.

I think this point was also touched upon by Jack when he asked a question about -- or talking about turning the different knobs, different ways, different parameters that can go into it. So one of the questions that came up during our discussion of the grant that we have a monthly meeting

with the FDA on is the randomization ratio, so how do you choose and understand that?

So in this paper, I won't go into all the detail about it, but essentially for the scenario, very similar to the CID case that we have. We try to understand what's the impact of the various parameters on, for example, the internal and external control ratio.

So we discovered that the internal-external control ratios is one important aspect, but also the randomization ratio can also affect how the operating characteristics can behave in terms of type 1 error and power. So this work is published in the ASA BIOP Report.

This work is another piece that actually something that came out from our CID collaboration. And when we were doing the step about the -- before the dynamic borrowing, FDA suggest that we should use propensity score matching. One of the reasons for that is that there's not much literature on how does it work with other ways of handling.

For example, propensity score weighting or even covariate adjustment. So in this work, we actually investigated also their operating characteristics how different approaches in handling covariates can have an impact on the design.

So there's ongoing research on this topic. And what we covered mostly today are tied to more time-to-event outcome, but there are other outcomes that, I think, in Karen's talk and also in Roger's talk that it was covered that there are different outcomes that, for example, rare diseases, pediatric outcomes, could have different endpoints and different characteristics.

So oncology was the application today, but we also have, for example, other studies as well. So in summary, the CID pilot program, we facilitate a very collaborative scientific discussion with the FDA and we had alignment on critical concepts, design proposals to boost confidence in future designs and outcomes. Agency also demonstrated openness to the proposed design with external controls while providing

key feedback.

As we learned, early sponsor and health authority engagement is paramount when we want to do these novel trial designs. And successful adoption of novel innovative designs also requires collaboration effort between health authorities, academics, and industry as we see from the panel that we have today.

So we highlighted one example of how the MATHIS work has been used to fill the research gap. So this work takes a team effort and for the preparation of the CID event and also the CID program, as well as the grant, these are the colleagues that have helped us. And thank you for your kind attention, yeah.

DR. SCOTT: Thanks so much, Dr. Pang.

I'll turn to the panel to see if anybody has any questions or comments.

Dr. Hubbard?

DR. HUBBARD: Thanks for that really nice example of a hybrid-controlled trial and all the sort of many considerations that go into how you put

it together. I had a question about a specific element of it, which was the choice of the number of external controls.

You know, sometimes we're in settings where you have a very large pool available from a registry or other external source. I'm not sure that that was the case in your example, but in general, how would you go about thinking about considerations there, the practical constraints, the effects on the operating characteristics?

And then also I think there's sort of a gut feeling that you just sort of don't want to have too many external controls, and how do you weigh those things?

DR. PANG: Thank you, Rebecca, for this important question. I think this is one thing that probably in our talk we didn't cover as much in detail. One of the reasons is that in the CID program and what we have for the case, we actually have a concurrent ongoing trial, which makes it much easier in terms of having high-quality data that really

follows the Pocock Criteria.

Think in the case of if you have a scenario where you can borrow a lot more, but you may be more skeptical about the quality, and data is so important, right, the quality of data. So having the screening of inclusion and exclusion criteria to make sure the two groups are aligned, the internal control and what you use to augment it is extremely important.

So our filtering step is certainly important, but how you go about doing the filtering, there are many different approaches to do that. So one such would be maybe propensity score based approaches, but there could be other ones, like setting up ways to make sure that you just exclude certain subjects, right, because there are some differences in terms of calendar time biases. Like, you want to ensure that things align in terms of when you enroll the subjects.

So just like you have to consider all the biases that may be involved and read through, for example, the paper we discussed, but also the FDA

guidance on how to evaluate, right, the subjects in the end. Because in the end, if there are more uncertainty, there's also risk, right, to reading out something that things are less aligned.

But I think we are very fortunate in our case that we can borrow from something that's concurrent and ongoing, and which is sort of also blinded, right. So I think that scenario is very ideal, so I think that's also why we were selected because this is one of the better scenarios that we can have, yeah. So thank you for the question.

DR. SCOTT: Dr. Follmann?

DR. FOLLMAN: Yeah, just to pick up on that comment. So I guess you were underpowered for overall survival and so, like, in theory you could at the end of the study look at how many deaths you have in the randomized trial and then do a power analysis to decide on how big the external control could be.

It could be that you have enough deaths that you don't even need it. That's probably unlikely, but anyway, you could calibrate it to

answer, you know, to formulate it with a power, based on a power analysis.

DR. PANG: Yeah, thank you, Dean, for a great comment. So yeah, I think in some scenarios, definitely that could be a good choice. So, for example, you can just understand whether is already enough. I think in our case with the DLBCL population, the overall survival actually takes a long time to mature. So even with the borrowing, I think we are not really at 80 percent, so it's more like 60 to 70 percent.

So that's one of the reasons why in this particular case, the borrowing could be a very good approach and plus, it's a secondary endpoint less sensitive to -- if it were to be a primary one. So -- but I think the idea of maybe having something pre-specified that you can adapt, whether you borrow or not, could be a good approach, but in addition to the simulations that needs to be run, but also the other scenario, how to go about thinking about that, right. So -- but that's a good idea, yeah, thank you.

DR. SCOTT: Dr. Lewis, then Dr. Ruberg.

DR. LEWIS: So I think one of the challenges in dynamic borrowing of control data has to do with how you quantify the benefit of the approach and communicate that benefit to decision makers and others involved in deciding what the approach is you're going to take. So I'm curious in this setting. How did you quantify what you gained through this approach since obviously a lot of work had to go into it, and how effective was that communication?

DR. PANG: Yeah. So that's a great question. And of course, communicating the results was, like, even the simulation settings and the results is also very important. So in terms of understanding whether -- I think one of the keys things we look at is really the type 1 error control.

I think FDA, as you know, is also very keen on making sure that that's under control. So, for example, when we decided between the aggressive prior and also the conservative one, we can see that the aggressive one tends to have a lot higher type 1

error inflation. So we want to calibrate against a strict control of type 1 error, considering all the scenarios that we investigated.

Definitely, that has the disadvantage, so it's -- in our case, this was quite clear that the conservative one would be the ideal choice in our scenario. I think there are many other things. For example, thinking about the different assumptions, the variations, or the sensitivity analysis that goes about understanding violations, right.

Like, unmeasured confounder and what's the impact on that, and whether you actually know enough information about the different prognostic factors in your studies. That's also important. I think the advantage of what we had was we have also some older studies that we can learn from, understanding participating population one of the important confounders, and then what kind of information you have.

And we got to that, and we control for that. Assuming that, for example, I think in our

setting, which I didn't cover today, we actually simulated taking out some of the, like, known cofounders and then see what was the impact on the type 1 error control, and also power, and we actually see not so much because I think we have very good confounders all measured, yeah.

DR. LEWIS: Thank you.

DR. RUBERG: Yeah, thanks for that presentation. You've got me thinking. There's many phase 3 programs in drug development where a company will two identical phase 3 trials pretty much contemporaneous in time.

And those trials, let's just say each trial has 1,000 patients. You put 500 on drug and 500 on placebo, and you've got two of those. Why can't we put 500 on drug and 250 on placebo in study A and 500 on drug and 250 on placebo in study B and borrow the placebo information across.

Should be very few questions about exchangeability, or you could apply dynamic borrowing, but I would suspect you could borrow quite a bit of

information, given it's an identical protocol done contemporaneously in time. And as long as your investigators were scattered kind of evenly between Europe, U.S. or whatever.

You could have even geographic balance, if you would, between the two studies, and we could cut down on the exposure of 500 patients, placebo patients and reduce time and cost. So is that an extension of what you're presenting here, or is that something that we should be thinking about routinely in drug development?

And, John, what would the FDA think about that idea?

DR. SCOTT: So obviously, I can't answer for the FDA, but I will say --

DR. RUBERG: It's a review question; right?

DR. SCOTT: Exactly. It's totality of the evidence, but it does seem like low hanging fruit, and it's also the kind of thing that is not dissimilar to a master protocol. And so we do have sort of

processes for thinking about those and reviewing them.

DR. PANG: Yeah, so I think I agree with what John mentioned and especially for a non-oncology setting like running two phase 3 is very common. And so actually, as far as I know, there are some examples of that thinking behind the scenes. I'm not directly involved, but the methodology could be very closely related to the master protocol and similar.

So certainly, I think related methodology can be used, but also, I'm sure that you need to make sure that whether there's any issues that may come up, right. With such a design, yeah.

DR. RUBERG: Yeah. I would just say one of the criticisms of some Bayesian and borrowing and all that as well, you still got to have enough patients explored to your drug for safety and all that other kind of stuff. In the scenario that I mentioned, or we've discussed here, you still are exposing the same number of patients to your experimental treatment, and so you're still

accumulating sufficient safety data.

And anyway, John, I like your phrase about "some low hanging fruit" or consider it as a master protocol for your phase 3 program. Seems like it's imminently doable without huge leaps of assumptions or social or cultural or scientific barriers to overcome.

DR. SCOTT: Dr. Lewis?

DR. LEWIS: I just wanted to point out something which you already know, which is when you share that control data, you're introducing a correlation in the results between the two trials, so you lose the independence of the two trials. So one has to think very carefully about whether you care about that for the particular development program.

DR. RUBERG: I mean, I was even thinking, it doesn't even have to be Bayesian; right? It's like a multiple comparisons problem where you're comparing back to the same control group, and you could adjust for those correlations. Of course there's the Bayesian.

I don't know which one would work out to be better in terms of operating characteristics, but I suspect maybe somewhat dynamic borrowing might be a reasonable approach taking into account the correlation between the placebo groups. Anyway, just a thought.

DR. SCOTT: I mean, it also raises the question of, why not one larger trial? And I know there's an answer to that question, but thinking about what we get specifically from independent or quasi-independent replication versus more precise estimates from one trial. It's, you know, it's an interesting conversation, I think.

Dr. Lee?

DR. LEE: Yeah, following the discussion, I think we all like borrowing, right, partially if it's your money, right. And how to borrow it properly and really efficiently and to get really the more accurate decision. Yeah, it's why we are interested.

So in this case, as, Herb, as you

mentioned that R-CHOP has been used in the large B-cell lymphoma for over 20 years. There are ample of data, right, and the treatment's very standard. So you talk about hybrid control, which is makes sense.

So question one is that, again, how much -- you have cancer data; right? So how much to borrow in this case, the hybrid control? I see many of the designs say that, oh, we want to borrow up to the number of experimental arm; right? But -- them, or maybe, you know, you can think about a factor of that, right?

But in this case, there are so many data, I would argue that, you know, nowadays with the very good electronic medical record and real evidence, you know, this kind of getting -- the quality getting better and better. So one may think about a synthetic control, right. You don't even need a control group, right. So again, that -- I'd like to know your thought.

And also, John, the FDA's thought on that.

DR. PANG: Yeah. So I think certainly the real-world data setting can also contribute to the external control trials, so I think there are, like, two instances if you can borrow from a concurrent ongoing one and that already kind of is sufficient for the purpose of shutting time to read out for the overall survival when you do the primary PFS analysis.

I think that's an ideal scenario you have because I think even with very high-quality data coming from a real-world data setting, you may have other additional biases because subjects can be enrolled differently in trials versus more of the observational database.

So there will be a higher instance of, like, other biases that could be involved, but just need to make sure that, like, the whether the -- I think going back to, I think, one slide that we have is whether you have compelling effect size. I think if you have more compelling effect size, then you have more wiggle room, right, to have some of the risk, right, from the biases.

But if it's more modest, I think it's safer to have a cleaner and higher quality data. So I think it really depends on the scenario. It's really case by case. I think in this particular setting, you have very ideal situation where you have a concurrent one. But in the case, maybe you don't have one.

It's because the concurrent one also is blinded, right, so I think that's the advantage too the FDA also likes. But in scenarios where you don't, then I think you have to think about other options, but with higher quality real-world data, this may not be a big issue, but you have to simulate right. Also understand the scenarios where you have any other issues that may come up, yeah.

DR. SCOTT: Yeah. I won't answer it in detail, first of all because we're mostly here to listen to you all and draw our conclusions afterward, but I agree strongly with what Herb said, especially in terms of when you have a large treatment effect size, a lot of things are possible that may be less acceptable in a more marginal case. We see that in a

lot of areas.

And we are five minutes over for lunch, so we'll break for lunch. For folks who are here, there's lunch available for purchase right outside to the right, and there's also some drinks behind you on the tables. Feel free to help yourselves. And we will resume at 12:30.

DR. PANG: Thank you.

DR. SCOTT: Hello again, everybody. We are back from lunch and going to proceed with the panel discussion part of the day. This will be the next two hours, and then we'll have another break, and then we'll have audience Q&A.

So starting with the panel questions, I wanted to start before we get into the sort of preset discussion questions, with a couple questions we've received online that I thought might be informative about the talks we heard.

So there was a question for Dr. Lewis: How are futility analyses built into the simulations for type 1 error and power? Are the thresholds of

0.975 and 0.983 changed in the presence of futility analysis?

DR. LEWIS: It's a great question. So the general principle, which I think many people are well aware of is that the futility rule decreases the observed type 1 error if and only if the futility rule is followed. And so as a general rule, when we're simulating trials, we control type 1 error without the futility rule, and then we add the futility rule, which results in better control of type 1 error and sometimes a very small loss, typically a few percent, in power under the alternative scenarios.

Of course, the gain is that if the trial is trending towards a negative result, you can get out of it quicker, save resources and save participants from avoidable risk. I think the type 1 error control, as I recall, is controlled without the futility rule.

And if one was relying on your futility rule to maintain type 1 error control, then it must be absolutely clear that it is a binding futility rule,

and that there is no option for the trial continuing and sort of what we sometimes informally call "blowing through the futility rule" because then you've lost your type 1 error control.

I'm well aware that this is an area of some controversy, and I think the important thing is that those designing the trial and those who will ultimately be involved in reviewing the results and designing it -- deciding its clinical impact have a clear understanding of the precise assumptions that underly the calculation of the type 1 errors control, and then are able to verify that the trial was conducted in a way that was consistent with those simulations and therefore has the desired operating characteristics.

DR. SCOTT: Thanks, Roger.

And we also had an online question for Herb: Does the hybrid approach require access to patient-level data from external controls, or could it be done with summary-level data?

DR. PANG: Yeah, thank you for the

question. So I guess in our case, for the analysis, we definitely need individual patient-level data, so because we are looking at the similarity between internal and external control and then using it as a way to understand whether we should borrow more or less, but I think there are other approaches probably that can use summary-level approaches, but this is not the case for our CID, yeah.

DR. SCOTT: Thanks.

Okay. So having done that, we'll move on to the sort of preset discussions. Just a reminder, these are our panelists for today. The speakers are joining the rest of the panelists as panelists.

And, Steve, I'm so sorry for excluding you earlier, but I should've put you first to make it up. I apologize, yeah.

Okay. So question one: Each of the case studies this morning use a Bayesian statistical framework in one way or another. Did these studies need to be Bayesian, or could similar study designs

have been implemented using frequentist approaches? And what advantages, if any, did Bayesian methods provide in these examples? So it's sort of open field to anyone who wants to chime in.

Roger?

DR. LEWIS: So I think it's commonly stated that many forms of Bayesian adaptive design could have been created using a frequentist approach and moreover with many of these Bayesian designs were very interested in frequentist operating characteristics. So from a certain perspective, that of evaluating the performance of the designs, we actually have a mixed approach and actually care about frequentist characteristics, so they're frequentist in some sense.

I think it's interesting that as complex and innovative designs have progressed in their sophistication and the degree with which the designs have been carefully customized to the clinical setting, how much more common Bayesian approaches have become.

And I think that indicates a practical as opposed to a theoretical consideration that it is just easier to use the Bayesian machinery than a frequentist one to address multiple competing priorities, understand what the design is doing, implement things like hierarchical modeling, or understanding the way external or historical data is being used.

So I think there's a pattern that we're seeing, which is that the Bayesian approaches facilitate the kind of interdisciplinary decision-making design activity evaluation of alternatives that is necessary to realize an optimized design, and frequentist approaches just don't seem to be as practical in accomplishing the same things, even if in principle they could accomplish the same outcome.

DR. SCOTT: Thanks.

Jack?

DR. LEE: Yeah, I just want to add a little bit. Indeed, many of the design is specifical

can be accomplished using either approach, right. But conceptually Bayesian inference is a more coherent way of thinking about a problem, right. So just quick reminders that a frequentist approach will look at the probability of data condition on the parameter.

And Bayesian approach is a probability of parameter condition on the data, right. So these are kind of complementary to each other, and it depends on how you look at it. But, well, after all, what we are interested is in the parameter, okay. And whenever we don't know the parameter, it has a distribution, okay.

So Bayesian is very natural to quantify uncertainty, but on the other hand, the frequentist, you know, look at how likely you observe the data given the parameter. So it's kind of taking the inverse approach; right? And things can be done in either way, and there's a intersection between the two; right?

And we all know that if we use non-informative prior in many setting, the Bayesian

approach give the same answer as a frequentist approach, right. But that being said, still, I feel like Bayesian approach provide a more flexible and the natural way to adapt, right.

And also, it has a formal way to quantify uncertainty, and you can add the prior information there very specifically, you know, kind of way it spell out, rather than like a frequentist approach, the priors is in the head, right. Kind of, you know, you -- oh, everybody's tried to do a reasonable thing, right.

And one last thing is that we've touched this a little bit in the morning is the Bayesian approach you can easily use a -- construct a utility to synthesize the information and put a multidimensional kind of problem or thinkings and then, you know, construct the utility to make decision, and this will be harder to do this in a frequentist approach.

DR. SCOTT: Steve?

DR. RUBERG: Yeah, I'll pile on here a

bit. But the thing that I think is most about the Bayesian -- well, most important -- one of the important things is that it's directly interpretable; right? If you have decision rules or results, let's say the p-value is 0.02.

Okay. We declare that as significant, but I don't know exactly what that means in terms of effect size and everything else, whereas in a Bayesian approach, we all know we can say the probability that the effects, treatment effect, is greater than 1 is 90 percent. Okay. That statement is a very clear and interpretable in and of itself. You don't need a lot of other information or context, so I think that's good.

And I think, Roger, your example in the CHIP study, the non-inferiority margin of, okay, the probability of this, this, and this; that's how we built it into our decision rules and futility and et cetera. And it's very clear and straightforward as to -- as opposed to perhaps, well, we're going to do these interim analyses, and if the p-value's less than

something, then we'll do this or, you know, that's just harder to get a feel for what that really means. Yeah, you can do frequentist approaches.

And I'll just -- Karen, I don't know if you have an answer to this or not, but if you did the frequentist approach, and you just say, we're going to use, do nine clinical trials instead of this platform, you mentioned that Lilly has a lot better throughput through this platform trial with discovering or advancing drugs, et cetera.

And I'm just wondering, is it a matter of it was the platform and the standardization, et cetera, or how much of it was the Bayesian analysis of borrowing and integrating data? Again, I don't know if there's an easy answer to that, but I guess I'd like to think by borrowing information, you can do more efficient studies, less time, less cost, better decision-making, et cetera.

But in your concrete example, I'm just wondering if you might speak to the -- how much of it was the platform, and how much of it was the Bayesian

approach to the synthesis and analysis of the data?

DR. PRICE: Thanks, Steve. It's a great question. I think I'll answer it by sort of doubling down on something that Dr. Lewis mentioned, which is a part of it was that it forced us to have the conversations ahead of time, the cross-functional conversations around what is meaningful, so we really had a lot of discussion about what is that effective interest.

Also, how did the pain types relate? Would we borrow across pain types or not? Because the fact that we could and we had the Bayesian analysis there, it really facilitated a lot of the internal conversations that maybe don't always happen when you're doing independent trials and so probably a combination equally really the -- having that platform, but facilitating the conversations and then being able to borrow the information.

One other thing if I could add while I have the floor, is that another thing a Bayesian approach allows is that it keeps functions of

parameters maintain coherency. So once you have that joint posterior, not only can you make inferences about the parameter of interest, but functions of that parameter and everything stays, probability stays between zero and one and things maintain that coherency, which isn't always the case.

And so I think just one additional thing. All of the points I was going to echo additionally just to put that one in there as well: No relying on large sample theory.

DR. SCOTT: Following up on something Steve said on the interpretability of posterior probabilities, are they still interpretable if you've chosen a prior for pragmatic reasons to optimize some operating characteristic, rather than to capture your prior state of belief?

DR. RUBERG: I mean, I guess I'd say at face value, they're still interpretable. You probably have to be transparent, and I think the Bayesian approach leads to that transparency about declaring what exactly is your prior information. I think in

many situations, people intuitively make decisions about effect sizes and are they clinically meaningful in some more abstract, intuitive, internal mental process or whatever, so all of that's going on, even in a frequentist world.

And I guess I would say the Bayesian approach is not only more interpretable, but you have to declare, you know, up front quite explicitly what assumption, what prior is going in and feeding into that probability statement.

DR. SCOTT: Thanks.

Frank Bretz?

DR. BRETZ: Yes. Maybe I'm just reacting a little bit to this he's a Bayesian or frequentist. So in my view, I think the studies today, they used actually both in the senses that, you know, that the decision criteria they calibrated, so essentially, they were doing a frequentist analysis, I guess, by calibrating to decision criteria.

And so I think there's a marriage of those two methods that I saw today for large part. Of

course, I can well imagine about fully Bayesian approaches in the sense of using Bayesian influence just based on posterior probabilities without calibrating decision criteria.

I'm not sure, but I have seen that earlier today. So it's just either/or, I'm not sure. I'm so comfortable with, and I think it's more important that you have the right design. It's fit for purposes and be addressing the right questions. And which methods we use, I think it's to some extent secondary.

But since the question is also there, maybe it's also good to remind us that there are some purely frequentist methods based on, say, meta-analysis where you can incorporate historical information if you wanted to, or frequentist propensity score methods. So there a variety of methods out there, but again, I don't think it's an either/or, so it's just, you know, what is fit for purpose.

DR. SCOTT: Thank you.

Yes, Herb?

DR. PANG: Yeah. So I agree with what Frank said and in the -- I want to speak to the scenario of the hybrid randomized trial design, which is our CID study. In that case, actually, even though the dynamic borrowing is certainly a Bayesian approach, there's also the aspect of the propensity score, right, that went before that. So that's kind of a mixture on top of the Bayesian.

And Karen was asking me, it was a little bit after the talk, which is, we had a paper about a covariate handling approaches on top of Bayesian approaches. So there's also, like, other considerations, right, that needs to be in place.

So as Frank said, like, it's not just purely, like, not the way to do it, but there could be different ways to do the same thing, but we just need to make sure that we're doing it right. And then I think another point is for the hybrid control trial setting, there's a recent paper by one of our colleagues working on using adaptive lasso methods for

hybrid control design.

So think in their scenario and also what they investigated was that it can be computationally more efficient, like, from the computing time standpoint, but then I think the -- in the settings they investigated, there still the issue of type 1 error inflation in some scenarios. So it doesn't get around, right.

So potential issues that may come up to for Bayesian, so I think, yeah, goes back to Frank's point about the -- not the particular type of method, but how to do it well, yeah. So thank you.

DR. SCOTT: Thank you.

Anyone else?

Okay. We have lots of questions. Question two: For late-stage studies with a frequentist design, the maximum type 1 error rate is typically controlled at 0.025 one sided. Is there a direct analog for Bayesian designs? What are the specific design characteristics that you see as most critical to support regulatory decision-making for

Bayesian trials, especially trials that use informative priors to incorporate external data in the study analyses? Anyone? Steve's making eye contact.

Please.

DR. RUBERG: All right. I'll give it a shot. So first of all, I guess we have to realize that controlling the type 1 error is not the same as controlling the false positive rate or the probability of a false positive finding. It's conditional; right? You have the probability of A given B. The probability to reject H nought, given H nought is true.

What we really want to know is, what's the probability of a false positive finding in some sense the joint probability, the probability of A given B times the probability of B; right? The probability of B, the probability that H nought is true is, like, our prior or the probability that it's false; right? So it's really two different concepts, and I'd much prefer thinking about controlling about a false positive finding.

So, you know, the real question is now that I have a p-value that's 0.05 or 0.03 or 0.01, if I decide to reject the null hypothesis as might traditionally be done, and I'm talking two-sided p-values here, what's the probability that that's a false decision; right?

And so it's kind of what's the probability that the null hypothesis is true given that I've observed a p-value of 0.03 or 0.01, two sided. Well, that's decidedly a Bayesian formulation of the problem, and it would be much more interesting to me through simulations or whatever that say I am using an informative prior.

Okay. That inflates the type 1 error, but it does not inflate the probability of a false positive finding because I'm going in with a notion, perhaps, that I'm having some slight favorable prior toward my drug works. I've got phase 2 data, you know, another phase 3 trial that I did in another area, et cetera.

So you're going in with a notion that

the probability of the null hypothesis is true is probably quite low; right? And so the chances of making a false positive finding are, you know, commensurately decreased.

And so I'm, you know, going to use an informative prior with all the right considerations about the prior, but I'm going to show that my false positive finding rate is sufficiently low, even if my type 1 error rate appears to be inflated by using that informative prior.

So that's kind of my perspective on designing trials and taking the Bayesian approach, and I do bristle a little bit. I understand, I think, but I do bristle a little bit about the frequentist characteristics of a Bayesian procedure. I don't know. Those kind of things, it grates at me a little bit because it's kind of like trying to mix two different philosophies.

So if we're going to do Bayesian, then let's control the false positive finding rate, et cetera. That'll maybe start the conversation here.

DR. SCOTT: Roger?

DR. LEWIS: So I think I'd like to address the second half of the question, which has to do with this question of, how do you think about type 1 error control when you're using informative priors safe from external or, you know, previously existing data. And to me, there's a fundamental question about what we mean by an error rate in this setting.

So I want to consider two possible scenarios. In the first scenario, we're doing a single trial. We take a look after a quarter of the data have accrued. We take the distribution for the unknown parameter from that quarter of the data, and then we update it with the last three-quarters of the data.

So we've done one complete trial. We've started from a non-informative prior, and we have a final estimate versus a situation in which that first quarter of the data came from historical, had the exact same information in it, so it resulted in the same now external prior, and we updated it with

that last so-called three-quarters of the patients.

Mathematically, there is -- if we don't discount the prior, they're exactly the same. My point is, when we're conducting a single trial, we never stop at an interim analysis and say, oh, by the way, we need to make sure our type 1 error rate is now again controlled from this interim going forward.

So there's something inherently inconsistent about the phrase "controlling type 1 error risk and using informative prior." If you believe that the prior information is informative in a way that is likely to be valuable enough and relevant enough so you want to use that information, in my view, there is no such thing as type 1 error control.

Apologies to my neighbor. And so I think the question reflects a logical inconsistency based on our habit of thinking of type 1 error control as a characteristic of a trial. And I think therefore the answer to this question is that, if we make the decision to use informative priors, it no longer makes sense to have the same criteria for error control.

What we really care about is the sensitivity of our final decision, getting the right answer or the wrong answer as a function of the degree with which the informative prior was drawn from a different set of data fundamentally different estimate of the treatment effect or reflected a different underlying treatment effect because we were wrong that it was drawn from a similar-enough patient population outcome measure, whatever it is.

So to me, the last question has two parts: First, realizing that when you're using informative data, if that makes sense, type 1 error control no longer makes sense as a criteria; and number two, all of our focus should be on deciding what evidence informs our assessment of the likelihood that the treatment effect reflected in that informative prior is valid as a predictor of the treatment effect in the subsequent data.

DR. SCOTT: Thanks, Roger. I tend to agree with you about type 1 error in these settings in terms of evaluating the degree to which our

informative prior matches the population from which we're drawing the new data. Is that something -- how much of that can be planned at the designed stage versus how much of it is based on observed heterogeneity after the data are collected?

DR. LEWIS: So just responding to that directly. I think at the design stage, depending on the context, one has to make a decision about whether your design is specifically structured to mitigate the risk of associated with mismatch of the historical data versus the current data.

So we can picture settings, and I think Dr. Pang's setting was a good example where there was tremendous similarity with the historical -- excuse me -- the external data and the concurrent data because they were similar protocols concurrently administered, et cetera. In that setting, I think there's a very good argument for not -- for having a relatively fixed approach to strong use of that prior information.

On the other hand, if I'm drawing

historical data, I'm using historical data for a prior, and it's from a different setting or a different time or different centers or different practitioners, now I'm much more likely at the design stage to want to use a dynamic borrowing approach so that I can anticipate and mitigate the risk associated with their turning out to be a mismatch in the treatment effects between the prior and the subsequent data.

And my dynamic borrowing will naturally borrow less aggressively and still give me a valid and interpretable estimate of the overall treatment effect.

DR. SCOTT: I'm going to turn now to our online panelist, Frank Harrell, to weigh in on this question.

DR. HARRELL: Thank you very much. Can I share my screen, John?

DR. SCOTT: I think so, yes. I'm getting a nod.

DR. HARRELL: Okay. Let me try

clicking here. I want to -- it says I can't share screen while the other participant is sharing.

DR. SCOTT: Let me stop -- oh, I can't.

DR. HARRELL: Think you're still sharing something.

DR. SCOTT: Someone will take care of that.

DR. HARRELL: Okay. I wanted to elaborate on what the last two speakers said so beautifully and to give a simple example. And while I'm waiting on the screen share, I think it's just so ironic that Bayesians are asked to study frequentist operating characteristics of Bayesian procedures, but frequentists are never asked to demonstrate good Bayesian operating characteristics of their frequentist procedures.

It's just very weird to me because, as Steve said so well, the Bayesian procedure has to do with decision-making. And what you care about is not what you planned before a study began and what might happen, which is related to alpha and type 1

probability, but what you care about is the accuracy of the decision after everything has finished.

So the Bayesian operating characteristics are so different from frequentist ones. And if I could show you a simple slide right now -- it's still not letting me do it.

DR. SCOTT: Yeah, still working on it.

DR. HARRELL: I've laid out what -- okay. I've laid out the main Bayesian operating characteristics. Number one far and away is the correctness of the decision that you make with Bayes. That's all important. The other things are minor. The second is the Bayesian power.

Do you have the sensitivity to detect, and in fact that's at the minimally clinically interesting level, and then what is your expected stopping time? That's a Bayesian operating characteristic that's about efficiency and cost.

And then what is your precision of estimating efficacy if you have evidence for efficacy. And so those are important. And I just would like to

be able to show a very simple simulation.

DR. SCOTT: Okay. You should be good now.

DR. HARRELL: Good, good. I think it's going to come up. This is a very simple simulation. It's actually a very dangerous simulation because it's simulation under a radical situation where you would expect Bayes to run into trouble.

And I say that for two reasons: It's because it was simulated with unlimited data looks, and it was simulated under a universe of treatment effects that does not match the prior that is assumed during the analysis. In other words, the universe of treatment effects uses a much more skeptical treatment effect than the prior that's used in the analysis.

So even under those two situations, the Bayesian performance is pretty amazing. So the first thing to understand is, how do you know you're doing a Bayesian simulation? Well, the number one clue is that you never get the same treatment effect twice.

So if you're simulating, as I did here,

10,000 clinical trials, no two of those trials have the same treatment effects, so you're recognizing the Bayesian goal is to uncover the treatment effect that generated the data, whatever that is. So I chose the universe of treatment effects to be disadvantageous to what I'm showing.

Do 10,000 trials with sequential assessments and unlimited looks, except I'm restricting the first look for efficacy to be the first moment where you have sufficient precision for the treatment effect if you were to stop for efficacy at that moment.

So what happens when you have unlimited looks at the data essentially and you want to judge the Bayesian operating characteristics? Well, of those 10,000 trials, which allowed stopping at any time for inefficacy, and if you have an inefficacy as a formal stopping rule, you don't need a futility assessment anymore.

So what happened in this is over half the trials were stopped earlier with its conclusion of

inefficacy. The average sample size at which that happened was 62. The frequentist sample size for this study was about 234. And then the question is, are you accurate? Did you get the right answer?

This was what Steve was getting at. This is just putting numbers on that. So of those 5,184 trials that are stopped early for inefficacy, 5,020 reached the correct conclusion. In other words, 5,020 out of 5,184, the true treatment effect was lifting the threshold for trivial treatment effect.

I took gamma to be one-third of the MCID for this particular simulation. So that means that when you stop for inefficacy, you are correct 97 percent of the time. What if you stopped for similarity? Well, that's actually hard to demonstrate, but 634 trials stopped early for similarity at an average sample size of 423. 607 of those, that was the correct decision.

So it was 607 out of 634 times, which is 96 percent of the time, the underlying truth that generated the trial that you stopped early for was

similarity of treatment effect. How often did you never stop? I set a maximum sample size of 750. There were 172 out of 10,000 trials that went to the full maximum and without stopping.

So you're really avoiding wasted money with these Bayesian sequential designs. And that what is it that made you unable to reach a conclusion, the median treatment effect that was in play at the end of the study was exactly the threshold for non-trivial treatment effect.

But here's the most important part: Stop at any time that you're greater than 60 sample size for minimum precision for efficacy; otherwise, you stop at any time. 4,010 stopped early; the average size at which it stopped with evidence for non-trivial efficacy, average sample size 102.

So how often were we correct? 3,643 out of 4,010, which is 91 percent of the time, the decision to stop early for non-trivial efficacy was correct. In other words, the true efficacy that generated the data was greater than gamma. Even more

impressive is how often were you correct in saying that there was any efficacy if you stopped early for more than trivial efficacy? You were right 98 percent of the time.

So I think in terms of operating characteristics, I can't think of anything more important than showing you that you get the right answer after the data are in, and I have just two quotes to try to get your attention about this:

"Asking one to compute type 1 assertion probability alpha for a Bayesian design is like asking a poker player who wins more than $10 million a year to justify his ranking by how often he places bets in games he didn't win," or "Do you want the probability of a positive result when there is nothing, which is alpha, or do you want the probability that a positive result turns out to be nothing?"

This is exactly what Steve talked about earlier. So thanks for letting me share that.

DR. SCOTT: Sure. Thank you, Frank.

Dean?

DR. FOLLMANN: Yeah. So getting back to this question, you know, I agree what was said earlier. You know, this is sort of cross-purposes. If you're a Bayesian, you believe in your informative prior. You have a different way or describing evidence, and it's not really your cup of tea to talk about the type 1 error rate, but I think, you know, it's relevant to do that.

And I appreciate, like, when Bayesians will evaluate the performance of their method under -- evaluate the frequentist performance of their methods. Informative priors, I, you know, I think they have their place, probably for rare diseases and probably for sort of evaluating sort of a series or streams of trials, and I think that's what Frank was doing.

He's saying you adopt this approach, and in the long run, you know, you'll have certain performance characteristics, which I think is a certain calculus. It's a bit different than saying, I

really want to get the answer right for this particular trial for this community of people with this disease, and I want it to be based on evidence and not belief.

And I think these are two different perspectives and it's sort of, I would say, a decision for the FDA or someone else, which one do you want to adopt. You know, always getting it right, really controlling the type 1 error rate or have a stream of sort of better decisions over a long horizon.

So that's one kind of general comment, and the other thing about informative priors is, sometimes it's hard to figure out how much information they're really taking or how much providing or how much they're adding to the data. And so if they're being used, I like them to be interrogated and to understand how much they're deriving the evidence.

So, for example, you could compare the posterior probability for your -- the posterior distribution for your informative prior analysis with the posterior distribution where you have a

non-informative prior and see to what extent the informative prior is driving it. Is it worth, like, 100 percent of the sample size or 50 percent of the sample size, something like that, so you're transparent.

And I think, you know, that everyone should be in favor, transparency, and I think there's different ways to try and interrogate those, so it's clear what you're doing and how much the informative prior's driving the data. Thanks.

DR. SCOTT: Thanks.

Jack?

DR. LEE: Yeah. I try to answer the question in a different way. You know, should we look at the hypothesis testing using p-value less than 0.025 to make a decision or not, right. We all know that the p-value is heavily influenced by the sample size. Even the magnitude of treatment effect's the same, right. When you have a huge sample size, anything can be significant.

So in making a regulatory decision, I

think really we probably need to look more about estimation rather than just the hypothesis testing, right. And, again, your estimation, then you worry about the precision. Of course, you worry about accuracy of estimating that treatment effect, but also the precision, and that can be kind of measured against the what's a clinical meaningful difference, right.

Much better than, you know, just based on the p-value alone, so I think I'd like to kind of expand the problem a little bit. Regarding the regulatory decision-making, then I think the estimation's very important.

DR. SCOTT: Thank you, agreed.

Anyone else. Steve has a follow-up.

And, Dean, your mic is still live.

DR. FOLLMANN: Oh.

DR. RUBERG: Yeah. This is an example that I've used before is, so study A is done in pancreatic cancer and the agent that's being tested is an extract from a leaf or plant in the jungles of

Brazil, and it was noted that this tribe that lived in that area and used that never developed pancreatic cancer.

And so somebody extracted an active agent out of it and a study was done in two centers -- one in Argentina, one in Peru -- and looked at 200 pancreatic cancer patients and did a randomized trial and got a p-value of 0.02; right? Trial B is results from research done at the Max Planck Institute in Germany where they've looked at the biochemical mechanism of action and pathways and found some biomarker or whatever, developed -- and somebody developed a drug or a biologic that goes in and interferes with that pathway and stops cell proliferation.

And you go off and do a study of 200 patients with pancreatic cancer with that drug, and it comes up with a p-value of 0.02 on survival, whatever. If you're like me, I'm more likely to think study A is a false positive finding than study B; right? I think most people would kind of align with that kind of

thinking, and so it just goes in my mind to demonstrate that the p-value is really not related to the probability of a false positive finding.

You got to take the whole context into play. Call it your prior; call it whatever you want. And in fact, people who would evaluate, I think most scientists, who would evaluate study A would in the back of their mind be skeptical whether this leaf extract or whatever from South America is really something that I want to invest in a major phase 3 trial, et cetera, et cetera because they know -- you know all those things.

Somehow in your brain, you're taking all that prior information into account in your gut. And, I guess, I would just say it points out -- at least, I've used that example, whether you like it or not, I use that example to say 0.02 is not the probability of a false positive finding, and oh, people intuitively are using context, intuition, whatever.

And as I said earlier, and I think many

people in the Bayesian would say, at least in the Bayesian, somebody's going to tell you, well, write down what that prior is for that leaf extract versus that biomolecular mechanism of action, antibody that binds to the right receptor in the right place that stops cell proliferation, et cetera, et cetera.

Write it down, what's your prior? And then do the study and the analysis and take that data in the context of that prior. So anyway, just a couple other thoughts about emphasizing, again, I think, if you're going to do Bayesian, then I think you got to start talking about what's the probability of a false positive finding and not what's the alpha level for a frequentist-like approach that you might have taken in this context.

DR. SCOTT: Thanks, Steve. And that also raises the question of, what information is in scope when you're forming your informative prior? Which I think is at least adjacent to some of the upcoming questions. So question three: Regarding the use of external data in trials, how should external

data sources be chosen? How would you advise us at FDA to evaluate a proposed external data sources, and what are some approaches to identifying and mitigating bias in the use of external data?

Rebecca?

DR. HUBBARD: So there are obviously a lot of considerations that go into the comparability and the appropriateness of different data sources, and I feel like today so far, we've been talking a lot about the comparability of the patient populations, drift in treatment effect or in placebo effect over time, things about the contextual effects, et cetera. But I think at least as important as all of those things is considerations about the data quality, the assessment, the timing of assessment, the methods of assessment and so on.

So when I think about pulling in data from external sources, a point that Herb made in his presentation that I think is really important is, do we feel confident that in these different data sources, the outcome measure, key inclusion-exclusion

criteria, et cetera have been assessed in ideally the same way, at least similar, but ideally the same way.

So I think the examples we saw this morning were excellent because they provided kind of the best-case scenario for having a set of external data where you could feel fairly confident that things were being done in a similar manner and hence, you know, we're comparing apples to apples when we compare across those trials or when we pool together those data sources.

I think as we try to get more aggressive and more innovative in using more modern, novel, real-world data sources, it's where these issues really become challenging because with the exception of just a few really hard endpoints like overall survival, comparability of almost anything else is really, really challenging.

So when we think about comparability of the patients, to go back to where I started off, we might be able to access that explicitly, empirically just using the data that we have in hand, but when you

think about comparability of the data source itself and assessing data quality, that requires all sorts of information on the metadata.

So we really need to know, you know, where it came from, how it was collected. What did it look like? What did the assessment look like? What factors affect which patients get assessed according to which timing? et cetera.

So from my perspective, that's a really, really challenging task if we're using anything outside of the context of a controlled trial, you know, even at the point of pulling in registry data, I think it becomes very, very challenging.

And I think I'll stop there because the idea of identifying and mitigating the bias, I think, becomes enormously challenging when you don't even have a good handle on the quality of the data.

DR. SCOTT: Thanks.

Frank Bretz.

DR. BRETZ: Yeah. So I think that's a great question, and I think Herb already mentioned

this this morning, the paper based with Pocock, but I think there are several frameworks out there that could help us in putting such a framework together for our needs within pharmaceutical development.

Certainly, the FDA recent guideline on externally controlled trials, I think there are some considerations on the appropriateness of external data. I think that's a great first step, but of course, there are other frameworks like the Cochrane Collaboration.

I think they have for a very long time thinking hard about systematic meta-analysis, and they have thought about, you know, important considerations of historical data and similarity of data. And the target trial framework is yet another one recently put together by Miguel Hernán and others more on the estimate framework or the cause of influence framework, which I think allows us to disentangle biases more into external biases versus internal biases.

You know, what would have been the

ideal trial that if you could have wanted and how does the data that you're connecting and how observation study fit to that. I guess, there must also be other frameworks in the real-world evidence, the real-world data community.

I remember there was a whitepaper by Duke-Margolis a few years ago, which, you know, distinguished a bit between data reliability and data relevancy. So I think there's lots of frameworks out there, and I think it would be helpful to have one framework within our context in pharmaceutical drug development, but I don't think we need to reinvent the wheel, so to speak.

DR. SCOTT: Thank you.

Roger.

DR. LEWIS: So I think that one has to worry both about what data sources were chosen and which ones were not chosen. And I think that -- I think we're all worried about selectivity and choosing data sources, but this is really a challenge to be objective when frequently the treatment effects that

are suggested by different data sources are known before the decision is made whether or not they're going to be included.

And obviously, what we would like to do is be able to write down a priori the criteria for selection of data to be included, for, you know, external data sources to be included without any knowledge whatsoever of the direction of the treatment effect that would be reflected by those data sources, but that's rarely possible.

And I worry that it is relatively straightforward in some settings to choose data sources that are likely to be supportive of the treatment effect that the sponsor is hoping to demonstrate and then to retrospectively write criteria for the selection of data sources that exclude contrary data that the sponsor also knows exists.

And so for the second part of the question, what would I advise FDA in evaluating them? I would say the first is to make sure that the FDA is aware of all alternative data sources that could

plausibly be related to the treatment estimate of interest that have actually not been brought forward, and that puts a tremendous burden on the agency in order to find those.

And then building on the point that Frank made, there certainly are systems for evaluating similarity of data or data quality, and I think there is work that could be done to look at the sensitivity of analyses based on the variable inclusion of various amounts of external data, with the inclusion gated by different thresholds for similarity or quality, to look at the sensitivity of the result based on those decisions, but I worry much more about the data that you don't see than the data that is brought in.

DR. SCOTT: Thanks, Roger. Yeah, in terms of the selectivity, a thought experiment I often try to pose to people is: Let's say you're developing an Alzheimer's drug. You had a positive phase 2. You want to borrow into phase 3. Should you also be borrowing from the information you have that Alzheimer's is a very difficult target; that there

have been many failed late-phase studies?

Would the answer to that depend on whether the drug had the same biological target? I think these are really important but hard questions.

DR. LEWIS: Or should you borrow from all of the negative phase 2 studies of similar compounds and downweigh your estimated treatment effect and then not proceed at all?

DR. SCOTT: Right, exactly. We don't want to be filled with despair, you know. I think there are pathways to developing effective medicines even for difficult targets, but knowing how best to form our informative prior stance I think is very difficult.

Herb.

DR. PANG: Yeah. I think Rebecca gave a very comprehensive answer to one, so I won't add anything more to that. And Frank and Dr. Lewis gave a very good answer to the second one. One thing I want to add is, think maybe one thing that FDA should also evaluate is -- think the sponsor would also do

that -- is to also look at the alternative, which is to not borrow at all; right?

So also consider that scenario given that would be maybe the least biased in terms of the data. And then for the third question about the medication and kind of identifying strategies, I think the CID program with the dynamic borrowing is a good illustration of what can be used, but in addition to that, like, understand the covariant handling approaches on top of these approaches, propensity score based or a covariate adjustment is as important.

I think also Dr. Lewis pointed out earlier about with these approaches, sensitivity analysis, thinking about the plan how to do that appropriately is also very key to success, yeah. Thank you.

DR. SCOTT: Thank you, sir.

Steve.

DR. RUBERG: Yeah. I'm going to focus my comments about external data on what I think is maybe could be most impactful in drug development and

that is the typical drug development program, perhaps it moves from phase 2 and into phase 3. So I'm going to talk about external data, but internal to the company; right?

So there you have access to the raw data and everything else; right? And I think that's where there could be the biggest impact is to say, how much can we borrow? Here's my drug for psoriasis, now psoriatic arthritis, now something else, you know, related to skin lesions or whatever.

You know, can I borrow from the phase 2 programs and the other phase 3 programs all around them? I will say without talking, I don't think, out of school, with my Lilly colleagues here. When I was there, we often talked about the probability of success when we had a positive phase 2 result.

Now, what do we think the probability of success is in phase 3. And I think you can -- we had kind of a checklist from my recollection that we used, and I think the same kind of checklist can be considered when thinking, well, then do you want to

borrow any data from phase 2 to phase 3?

And the checklist included things like dose, route, formulation, and even batch processes; that there were times when the manufacturing process changed from the formulation for phase 2 to phase 3. Geographic sites and countries involved, investigative site types, research hospitals versus community hospitals, et cetera, can have an impact.

Of course, inclusion-exclusion criteria, any changes there. Disease state severity or duration of disease or subgroups or biomarkers that might be modified or focused on in the phase 3 trial, study duration. The outcome variable, is it the same or different? How different? What's the time, course, and trajectory? Is it measured in the same way, right, in short-term, long-term trials?

And then, you know, we would think about this regression to the mean from phase 2 to phase 3 or, hey, by the way, all these -- you know, we got this great drug for stroke. Well, so did the 38 other companies that came before us had a biological

mechanism and a positive phase 2 study, but none of them had a positive phase 3 study.

So somehow we got to take that into account. And I think maybe an interesting one that didn't think about, maybe six, eight, ten years ago when I was working in the industry but has emerged recently is the whole estimand and analysis approach. I think that was mentioned a little bit here early, or touched on.

But you're borrowing data from some study that did a true intent to treat analysis versus some study where the treatment effect was estimated using MMRN and a different set of assumptions about missing at random. Okay. Now, can I combine those treatment effects, or how do I bind those treatment effects, or somebody else used a composite -- I don't know.

All the stuff that's emerged with estimand and what is the treatment effect you're estimating and how it's estimated now has heightened consideration in my mind about, well, I just

can't -- when I don't have access to the raw data, I can't just borrow that straight away without knowing some of that.

If you're in the scenario that I mentioned, which I think could have a big impact, phase 2 to phase 3 within a company, you would have access to the data. So you can analyze it however you like to do that, but I guess I think it would be great to focus on or have at least greater focus on this typical drug development scenario, which could really have a big impact on drug development broadly as opposed to well, can I borrow some external oncology trial that was done at MD Anderson? Now, somehow I want to use that at Lilly. I don't know.

That's a harder problem, more controversial, but since many companies do the phase 2 to phase 3 or multiple phase 3, I think it'd be very interesting to focus on that and say low hanging fruit or whatever. How can we make that work in a Bayesian paradigm and be acceptable to all stakeholders that are involved, so anyway.

DR. SCOTT: Thanks, Steve.

And Frank Harrell has something to add.

DR. HARRELL: Yes, thanks, John. If I could share the screen briefly too, that would be great.

DR. SCOTT: All yours.

DR. HARRELL: Thank you. So I just wanted to put up an alternate viewpoint on how data should be borrowed and suggest that we do it more with raw data than using a summary of previous data, where you summarize the posterior distribution from the previous data and turn that into a prior for the new study, and that is to do joint models of multiple data sources.

And I think there's many advantages to doing this and this relates to a comment I put in the Q&A online, which is, I don't really trust meta-analyses based on summary data. I really want to see meta-analysis based on raw data.

So this is a related concept. So what happens when you do joint modeling instead of using

priors for discounting in the way that we're talking now is you can use standard simple priors in the discounting. You don't need anything strange, power priors or anything like that. You make the assumptions a lot more explicit.

You're explicitly modeling the bias in, say, historical data, and you can get more accurate analysis by not assuming normality and such. And then this is the most important thing: covariate adjustment. I don't really trust the use of any historical data unless there's careful covariate adjustment to account for covariate drift.

And this is especially tricky when you have non-linear covariate effects, things that even propensity scores may miss can be very important to adjust for covariate differences using the raw data. So I just wanted to show very briefly what is it I'm talking about here with simple joint data models.

You have a model for the randomized controlled data, trial control patients. They have, let's say, a mean mu sub C for the control arm normal

distributed response variable and then for the active arm, you have some unknown Mu sub A for the active arm. And then for the historical control data, you just model it very explicitly that the historical control data have a mean of Mu C plus B, where B is a bias term.

So what you're doing here is just being very obvious about the fact that we don't assume that the historical control data are estimating the same thing as what you're estimating from the RCT. So the bigger this is, the more different it's estimating something.

And so you have a prior on the bias, and that prior is, like, a simple normal prior, which will control the amount of borrowing, and you have a lot of flexibility in how much borrowing, but just to put limits on it.

If you have an infinite variance on the bias, that means this historical data are completely irrelevant, and they're not used at all. And if the variance were zero, that means you're trusting the

historical data just as much as within study control data because they're estimating exactly the same quantity.

So I just want to make a suggestion that sometimes we kind of rush into things and assume that the way to harvest the powers of Bayes is by having discounting priors, but I think the idea of doing joint modeling, which Bayes allows this to be done very flexibly and powerfully, including multiple data sources, not just one extra data source as this example indicated.

And then how would that extend one more level to extrapolation? This is really dissimilar sort of thing, but let's suppose that you're extrapolating on a continuous variable such as age. So you're talking about using adult data to inform kids.

So at the root of extrapolation is, what assumptions are you willing to make about treatment by the extrapolating factor interaction. So if you write down a joint model for all of this,

you're going to have the interaction effect beta 3, traditional interaction effect, no restriction of beta 3 means the new study stands on its own.

It doesn't get driven by the, say, the adult data. A skeptical prior in beta 3 means you're borrowing information and so you're assuming there's commonalty and similarity between the data sources. And you could have another level of extrapolation that allows for additional complexity, such as nonlinear effects of the interaction.

And so just think about whether using joint Bayesian models gives an attractive alternative to the sometimes-difficult decision we have to make about the family of priors that we choose for discounting. Thanks.

DR. SCOTT: Thanks, Frank.

Karen.

DR. PRICE: Going to just add a couple of quick points. So thank you, Steve, for sharing some of the things that we're thinking about internal at Lilly, and I think you're exactly right. I think

you're exactly right. A lot of things you would want to think about in terms of whether or not to borrow externally.

A couple of things that we do is, in order to allow us to make the most informed decisions about whether or not to move forward to phase 3 and, if so, how, is we do systematic literature searching. And that will get at the comments we've had of ensuring we have a robust understanding of available data.

And, you know, obviously, the FDA also has access to more data at an individual patient level, tying back to what Frank was just talking about. It puts a burden on FDA, but maybe there's something there for how are we able to then utilize some of the internal data that you may have access to that others wouldn't, or the individual sponsors wouldn't?

The other thing -- or another thing that comes up in addition to the points, Steve, that you brought up, are now that we're moving to, and I

mentioned this in the talk as well, decentralized trials or different ways of measuring, that's of course going to add a component of what can we borrow if the measure was done in person versus in a more decentralized way.

And then the final point I wanted to mention I think is very useful is to consider where can we use structured prior elicitation conversations to help inform whether or not external data sources could be utilized.

I don't mean to formally borrow that elicited prior, but rather to facilitate structured conversations about, hey, if you knew this information, and even, let's say in the pain type, if you knew this information about a drug's performance in pain type A, what do you think that means for the pain type B?

And if they can't answer that question, then that's very insightful, or if it's very variable across experts, that's also very insightful. So again, just a few additional thoughts on how to think

about whether or not to use data.

DR. SCOTT: Thank you, Karen. Those are good thoughts. One small thing you mentioned, FDA's access to blind data from sponsors, this has come up in various settings in the past, and it turns out there are significant legal barriers to us even internally using other sponsors' data in the review of an application, but it is a good thought.

Discussion question four: So I think we actually covered some of this in the previous discussion, but we'll see if there's anything to add. So consider a phase 3 trial conducted after a very similarly designed phase 3 or 2 trial of the same treatment in the same population.

What are the advantages or disadvantages of analyzing these trials independently versus borrowing versus doing a meta-analysis? Does anyone have thoughts?

Roger.

DR. LEWIS: Sure. And I don't want to be repetitive with my earlier comment, but I think the

key consideration here is what I'll sort of a "unit of evidence," which is starting with the sort of example of two completely independent phase 3 trials there in which there's no overlap in the patients enrolled, and you get two independent estimates of what you assume to be the same or very similar treatment estimand that gives you particular statistical characteristics regarding the strength of evidence under the situation in which both trials give a positive result.

And that's sort of two units of evidence. As soon as you borrow information, or you make your primary analysis based on the combining of information in any way, you no longer have two independent units of evidence.

That may be a very, very good thing to do when there are practical financial time-based issues that make conducting two separate independent trials either unnecessary or infeasible or suboptimal for a patient population, for example, that doesn't have access to an effective therapy.

But to me, it's not just a statistical

question. It's matching your approach to the challenges of the area in which you're trying to develop a treatment. What I would -- the second point I'd make has to do with us trying very hard not to fool ourselves.

And what I mean by that is if, for example, you run a phase 2 trial or a phase 3 trial and they are positive and therefore you decide to do a second confirmatory trial, and you borrow information from the second, you're just doing -- the only reason you did that other trial is because of the first one.

And, Steve, you already mentioned regression to the mean. I tend to think of it as the fact that you failed to borrow from all the negative phase 2 trials that you didn't carry forward. It's the same concept.

But we tend to -- as human beings with our inherent limitations, we tend to sometimes fool ourselves in our enthusiasm and our hope to develop effective therapies where we aren't really honest with ourselves when we're double counting information or

discounting information that was negative and therefore excluding it from our interpretation of the future data.

And I just think we have to be very careful to not do that so that we have as accurate information as possible regarding the strength of the evidence that we're generating.

DR. SCOTT: Thanks, Roger.

Steve.

DR. RUBERG: Yeah. What are the advantages and disadvantages? I just jotted down some notes here. Independent trials. Okay. So there's a value to independent replication of results no doubt. And in fact, that might be the strongest evidence one can possible generate. In some sense, it's conservative and safe.

It's solidly on scientific ground, et cetera, but it may also be the most expensive, time consuming and unnecessarily conservative. It often includes more placebo patients and an experimental drug trial, et cetera. So that's advantages and

disadvantages, perhaps, in a nutshell.

Borrowing phase 2 from phase 2 -- for phase 3 trials, I don't know. It does make a lot of sense to a lot of people, statisticians or non-statisticians to build knowledge sequentially. Science is kind of we build on each other, and we stand on each other's shoulders for a totality evidence, but assessing that totality evidence, I emphasize, in a quantitative way, because I think it's obvious that FDA, others inside companies, whatever, you're always evaluating the totality of evidence.

It's just how quantitative or qualitative are you at doing it? And of course, as we've mentioned here, the Bayesian approach assumptions and priors and weights are explicit and clear. There's some benefits potentially for using less time, fewer patients, and more direct probability statements.

I'll just note that there's an American Statistician article that I co-authored with Jack and Karen and Frank Harrell and a few others, Lisa LaVange

and a few others. And in there, we have an example of a lupus drug that two phase 3 trials didn't meet the p-value less than 0.05. In fact, one of them had a p-value of 0.051. You know, it was close about as you can get.

And yet if you took a Bayesian perspective -- now, this again, a retrospective analysis, very, very modest borrowing from phase 2, but then the first phase 3 trial was done; borrow from that to make the second phase 3. And the probability of a drug effect was -- I can't remember exactly -- it was 0.99 or greater.

All right. So when you looked at phase 2, these two phase 3 trials that didn't quite make it, it's clear the drug works. Now, you can have all sorts of debates about safety profile, and what's the magnitude, clinically meaningful. But if your first question is answering does this drug work, there's a really, really strong case to be made that the trials that were done in the development program.

And yet, the drug was never ever even

submitted because we had two trials with p-values that were above 0.05. So that's the -- is it a type 2 error; right? Those are the kinds of at least pros and cons of borrowing or using independent trials.

DR. SCOTT: Frank Bretz.

DR. BRETZ: Okay. So maybe it depends also a little bit on the context where these questions could appear. So I'm thinking about if you have a very difficult endpoint where you need a lot of sample sizes, then, like, number of successive patient is your -- the trial.

Maybe then a standalone trial or two standalone trials would be difficult by itself, to reach conclusive statements, and maybe that's a possibility where we could combine information from both trials, maybe later down, then -- give to primary endpoints, FEV1 or some lung function parameter is significant first.

So maybe that's one part of the answer or one particular type of context. The other one I was thinking about is, do we need -- or do we want to

differentiate between approval state decisions versus labeling decisions. So maybe for approval decisions, I think this area of application, that also what Steve mentioned, I think it's hugely variable, but when it comes to prescribing information or, like, just information for investigators and physicians later on, would it be then be helpful on -- you know.

If you feel comfortable of pooling the data and have more precise treatment effect estimates that we can provide to different stakeholders. So maybe the answer depends a little bit of the context or various context.

DR. SCOTT: Thanks. I think that's a good point. And it reminds me of, you know, labeling for subgroups and possibly using things like shrinkage models for that.

I think we'll move onto the next question. How should exchangeability be accessed in late-stage trials that borrow external information? Are some methodologies more robust than others to violations of exchangeability, and what should be done

in cases where there is strong evidence of heterogeneity between prior data sources and trial data.

And I'm going to ask Roger to weigh in first to make sure we're starting from a common understanding of exchangeability.

DR. LEWIS: Thanks, John. So often when conversations of exchangeability come up, especially with folks who didn't study and suffer through this in graduate school, there is confusion between the concept of exchangeability and similarity of the patient populations.

All the criteria, for example, we've talked about when making a decision whether external data is similar to the data in the current trial. But exchangeability here means that based on what's known about the sources of data, one cannot make an informed decision about the direction of the inequality of the treatment effect based on the different data sources.

So if you have trial A and -- or data source A and data source B, and you estimate the

treatment effect from each of those, if you cannot know, based on what you know about the sources of those data, which one of those is going to show the larger treatment effect, then those are exchangeable. It does not mean the treatment effects are equal. It means you can't tell based on what you know which direction the inequality will be.

So it's a very specific criteria that is a necessary -- it's a foundational piece for the validity of many of the kinds of models we've talked about, like hierarchical models. The reason I make that point is that there's many situations in which we know the data sources are different.

One is from one geographic region; one is from another geographic region. One, something is being used in one kind of outpatient center, whereas another one, the data source is a different type of outpatient clinic, but we actually have no idea what that means.

From a modeling perspective, for the treatment effect, those are exchangeable data sources.

And therefore, in that setting, it is both reasonable and appropriate to form hierarchical models that require exchangeability. And it's important to avoid falling into a common -- or in a conversation or sometimes a heated conversation, about the equivalence of the data sources.

DR. SCOTT: Thanks.

Jack.

DR. LEE: Yeah. Actually, more technically, like, you know, the definition of exchangeability by divinity, you know, kind of fun letters, you know. So I think that one can -- another way to think about this is this: It's a weaker assumption than IID, right, identical independent distribution, right.

So, for example, we can draw the response rate of the different cohort from a common distribution; right? And then you draw the sample from that -- after you draw that -- the parameter from that distribution. So you know that without knowing which is which, then it's exchangeable. That's why

it's called exchangeability, right.

So we know that the Bayesian hierarchical model is built under the exchangeability assumption. So when exchangeability is not met, then one can do many different things, right. Like, for example, the cluster Bayesian hierarchical or more recent -- actually, not that recently.

It's called multisource exchangeability model, right, so it can identify which subgroups, you know, are exchangeable, which subgroups are not, and they model accordingly. And also a related thing is that, again, when we talk about external data, then we worry about the measured cofounders and the unmeasured confounders, right.

So for the measured confounders, we typically use, say, propensity score matching, or regression method, try to adjust for that. But for unmeasured confounders, then we are stuck, right. I mean, so people use, like, a "robustified" version of the hierarchical model, try to address the unmeasured confounders. And these things are all kind of

intertwined together, okay.

And there are more recent method that, for example, like, the SAME approach, right, self-adapting meta-analytical approach or some kind of elastic hierarchical model. You know, these are all different methods, try to address the, you know, "exchangeability" or measured, unmeasured confounders and try to get the good estimate, efficient and accurate estimate, of whatever estimate we are interested in.

But that being said -- actually, I was going to make a comment early on that is, no matter how good the statistical method is, we all know it cannot substitute good data, okay. So that's still, you know, very important that, you know, there are a lot of good statistical methods and really advanced statistical methods that's been developed, and we should know about this and use it appropriately, but no statistical method can rescue bad data.

DR. SCOTT: Yeah. At the end of the day, it's really important that a drug works.

DR. LEE: Right.

DR. SCOTT: Anyone else?

Karen.

DR. PRICE: Maybe just a couple comments, especially on this last piece. And maybe I'm not fully tracking, except that I think what that means is, if we have prior data, we observe our current trial and they look very different, what do we do?

And I think that's -- one thing that's important is if that happens, it shouldn't be a surprise what you would do. And what I mean is, when you're designing the trial, you should look at those observed cases and understand how does the borrowing of this information with potential outcomes of my trial -- when did I meet the CSF or the critical success factor, the probability threshold when borrowing versus when I don't, and oftentimes, that's in that borderline region.

They're usually not -- I mean, I guess maybe if you're pooling or there could be cases, but

it's usually in those borderline regions where the borrowing gets you over the threshold, or you may miss because it doesn't always go the direction of achieving the threshold, but we should know that.

And so I think mitigating against that in the design phase is important in understanding what's going to happen when the outcomes come; that you've looked at that and really understand that. So that would be my suggestion there is to make sure that you understand what that would look like, what we would do, and could even have a decision tree in place for here's what we would do in these various cases.

DR. SCOTT: Thanks.

Dean.

DR. FOLLMANN: Yeah. I guess, you know, this is sort of taking as a precondition that exchangeability is what you want to -- do you want to borrow information? I think a lot of settings you can do -- if you can do a randomized trial at two to one, why not just make the control group a bit bigger, and then you have no issues about exchangeability or, you

know, what the interpretation of the study would be.

So I think, you know, I think -- I feel like I'm different from most people on the panel, that I think external data borrowing, I worry that they degrade the importance of randomization, which I think is really precious and that we want to elevate that and keep it, you know, very solid.

And I think of borrowing data as sort of degrading randomization, and it should be reserved for very special cases where there are no alternatives. And we can do a two to one randomization. You can do a one to one randomization, properly power it. Maybe it takes longer, more money or whatever, but I'm just very weary of degrading a randomization.

Randomized trials, something that we've, you know, come to appreciate, has been responsible for a lot of great drugs over the decades and so on, and I think it's at our own peril we sort of just go and ignore that. And just I would prefer to reserve it for very special situations.

DR. SCOTT: Sure. Can I ask what kind of special situations?

DR. FOLLMANN: Well, like, earlier today, I thought we had a nice example where the primary endpoint wasn't subject to borrowing, but a secondary endpoint was. And I can sort of see it for that, I think.

For therapies that have very great large effects, maybe you can get away with seeing a very high cure rate in the experimental arm and then augmenting it with maybe some historical controls, or possibly rare diseases, but, you know, it's sort of, what is the universe's special situations where I think it would be?

I don't really have a good catalog of that, but I think it should be, you know, reserved for special situations. And not just sort of, let's build it into all the trials that we're going to do going forward, and the tone of this meeting to me is really that, like it's a given.

We're going to go this way, and let's

just, you know, think of randomization and the importance of randomized trials in the rearview mirror, and I don't think that way.

DR. SCOTT: Yeah. I think that's kind of the nature of the discussion topic, but the majority of trials, we certainly still do not do this, but --

DR. FOLLMANN: Well, I mean, this is meant to be a little provocative, I guess.

DR. SCOTT: No, no, that's a good point.

DR. FOLLMANN: You want to get the discussion going.

DR. SCOTT: Yeah. I think, Roger.

DR. LEWIS: Yeah. I just want to clarify my earlier comment. You know, when I was thinking about exchangeability, I was actually thinking about exchangeability, for example, of subgroups, where all subgroups have data that's randomized within the trial.

I was thinking, for example, of

Dr. Price's set of three different pain types and the exchangeability of the treatment effects between those same types. So I actually didn't -- even though the end of the question is specifically talking about heterogeneity between prior data sources and trial data, I didn't assume any of that would be necessarily not randomized.

The second comment I'd make, and maybe this is also an attempt to be provocative, is when we talk about the value of randomization in terms of improving the likely balance of unmeasured covariates or prognostic factors, I think it's a very strong argument.

When we talk about it as a cornerstone of our ability to successfully develop effective products, I think we tend to forget about all the products that haven't been developed because the barriers posed by randomization, so it's very easy to be aware of the successes.

What we don't know is what would've happened had we allowed there to be more flexibility

in this regard over the last few decades, I don't know if we'd be in a better place or not, but we don't -- you know, one of the things about time is you don't get to retry the last few decades again to see how it would be, so it's important not to just look at one approach and assume that it was the best path forward.

DR. SCOTT: This is provocative because lots of people want to talk now.

But, Dean, go ahead.

DR. FOLLMANN: No. I don't agree with that. It's sort of a big perspective on how important is type 1 error rate versus the type 2 error rate if you want to put it that way, and I think FDA traditionally has a certain view. And so when you say something about a drug, it means yeah, yeah, for sure it works.

And, you know, we're not talking about all the drugs that you didn't get approved. And it's a fair question, I guess. Suppose if you want to change that paradigm, it's that sort of calculus about

the importance of those two errors and, you know, if you move in a different direction, it's sort of a different kind of FDA, the way you speak, the authority you have will be different. And maybe it's better and maybe it's not, but it's -- it would be a change.

DR. SCOTT: We definitely have very few ways of estimating type 2 errors in the population. I find different people have different intuitions. My intuition is, we don't miss a lot of great drugs, but beyond that, I'm not so sure.

Steve.

DR. RUBERG: Yeah. I don't want to diminish in any way the importance of randomization, incredible development in science and what we do. I do think -- and I have to remind myself -- it's a tool. It's a means to an end. It's a very powerful tool, and it's a means to an end, but the end is estimating what is the treatment effect.

Did this treatment cause that outcome? How big is that effect? And then is it statistically

credible? Is it biologically, medically, whatever, meaningful, et cetera, et cetera. And it's worthwhile, or it's okay to bring other kinds of ways of evaluating or quantifying evidence to help answer that question.

And even in randomized controlled trials that produce p-values or whatever, as I've pointed out and given some examples earlier, everybody wraps their head around the context of how to interpret that p-value intuitively, experientially, whatever.

The trial from South America has a lot less credibility than the trial, you know, from a biological basis. So I absolutely insist on doing randomized controlled trials. I love randomization. It's a tool.

It's a means to an end, and there's other means that -- other tools in the toolbox that help us -- that can help us answer quite credibly the question, did this treatment cause that outcome, that outcome being an efficacy outcome or a safety outcome,

an adverse event, or whatever it is.

DR. SCOTT: Thanks.

Herb.

DR. PANG: Yeah. So I think for the last question here, thinking about our case for the CID, which is using it for overall survival, because we don't anticipate very compelling treatment effect, so when we do the investigation, we decide on using a more conservative prior, even though, right, the external control is a very, almost like a very ideal situation.

So I just want to bring out that point that, like, I think depends on the treatment effect scenario. Like, like, that's really also an important aspect, to decide on where, how you should borrow, and what should borrow from or not, right. Thanks.

DR. SCOTT: Thanks, Herb.

Oh, Karen, yeah.

DR. PRICE: I'll just make one more quick comment kind of on this topic of how we think about these various errors and the impact of borrowing

or not borrowing. And so we do have an example where we wanted to borrow phase 2 information into a phase 3 trial. The primary endpoint was death. We weren't able to do that, so the trial ran longer.

It was bigger, and we ultimately exposed more patients to placebo. And what we had wanted to do was more quickly move on to better understand the dose and have it be more of an active comparator trial. So we ultimately didn't even generate that evidence of comparing the -- having an active comparator trial because the placebo portion ran longer.

So I think there's also an element of evidence that just is not able to be understood better when there was a clear effect. And it was a very hard endpoint, probably a really important case where we could've done it, but because it wasn't understood, and we didn't get the alignment. We didn't.

And I think those are also missed opportunities to think about as well as drugs that aren't approved, what evidence are we not generating.

DR. SCOTT: Thanks, Karen.

Okay. The next question is a little specific, but important in the context of Bayesian trials that borrow external information in particular. So the question is, if you are borrowing information, and your prior distribution in some way governs the amount of borrowing, either sort of explicitly as in a power prior, or somewhat indirectly in a variance parameter in a hierarchical model, how do you choose those parameters?

Should you do it by quantifying the amount of data or the effective sample size that's borrowed, and if so, how do you do that, especially for the dynamic borrowing cases? And if you were in my shoes and somebody was saying, we're going to borrow X percent, how would you evaluate that proposal?

Roger.

DR. LEWIS: So I'll take the first crack at this and therefore be able to choose the simplest part of it. From my point of view, the key

is to evaluate it from multiple different ways. So the work we do, we tend to be using dynamic borrowing, so we're in that situation where we're putting a prior on the variance parameter for the borrowing, and so the amount of borrowing is not a fixed thing. It's a data-dependent thing, depending on the alignment of the different data sources, whether it's historical and current or in different subgroups.

So in order to understand the effect of the choice of, for example, that variance parameter, you have to look at many different case examples that have different underlying treatment effects so you can see how that choice affects the behavior both when the data are concurrent from the different -- I'm sorry -- concordant from the different sources or discordant.

In my view, the calculation of an effective borrowed sample size, especially in that setting where it's a variable number is only one small piece. I think it's useful when it can be calculated in a way that's transparent. I think it's a way of

communicating the amount of borrowing, but it does not replace just looking at lots and lots of different examples and seeing how your choice actually performs. And then in the criteria for what is good performance, I think it is strongly influenced by the scientific and clinical context of the development program.

There are settings in which we have good reasons to believe based on our understanding of the underlying biology, the behavior of different therapies, for example, within class, that a larger amount of borrowing seems to make sense because we think we understand how things work, and then there's other settings in which the underlying mechanisms may be much less well characterized.

There's less experience in developing therapies in the area where a smaller amount of borrowing or a requirement of greater evidence of concordance between the data sources ought to be required before we allow the dynamic borrowing to add much effective sample size, so I think there's not a single right size.

I think it's informed by the science, and I think to communicate the effect of borrowing, whether it's to your investigator team and collaborators, or to regulatory agencies, requires a multidimensional presentation of the actual performance of the borrowing strategy.

DR. SCOTT: Thanks, Roger.

Jack.

DR. LEE: Yeah. I'd first like to talk a little bit about how to quantify the amount of borrowing, yeah. You know, of course, effective sample size is a very intuitive way and also natural way to quantify it. But there are also other ways to do that by constructing different kind of borrowing index.

So, you know, we and others have trying -- you know, working on this area, like, using the overlap index and try to come up with some index. It's sort of like correlation coefficient, okay. You like to -- that correlation coefficient, we know it's between minus one and one, right.

And the borrowing index, you know, we'd like to make it between zero and one, right, either no borrowing or full borrowing, right. But how to do that, still there's no clear way or no one way of doing it, and I think still that's a active research area in terms of how to quantify.

You know, what's a various way to quantify the amount of borrowing and particular, under the Bayesian hierarchical model with a cluster, with clustering. You know, how do you quantify this amount of borrowing at the cluster level and at within the cluster?

So we have some work in this and it's still -- you know, our work's still under review. But I think it would be nice to have a more objective way to measure, you know, the amount of borrowing. And of course, you know, we all do sensitivity analysis, right.

I mean, once we -- to measure and quantify the influence of prior and -- but again, it would be nice to have some more development in terms

of how to quantify the amount of borrowing.

DR. SCOTT: Thanks, Jack. As difficult as the methodological and quantification problem is, I honestly get more stuck on the next step, which is a sponsor comes in and says, you know, we propose to borrow 40 percent of the phase 2 data. And some people can look at that and say, "Oh, that's too much." And I have no idea how you do that.

Frank Bretz, I think you had something to add.

DR. BRETZ: Certainly not an answer to your question.

DR. SCOTT: I wish you would.

DR. BRETZ: But it's more pragmatically speaking. I think it's, you know, maybe three things I wanted to say. First, I think it's probably helpful to go through different hypothetical data scenarios, so you do understand a little bit by, you know, borrowing that much, of that much information, you get that sort of positive trial result in the end.

I think you get a better understanding

also of the influence of the prior, so to speak. So that would be one answer. Another one, and maybe Herb said that before. I would certainly also look into what's the outcome if I would not borrow any information just a way to -- as a benchmark. I think that would be of interest to me.

What I also think is it would be good to understand, should we put a hard limit on the maximum amount of information that we can borrow from. You know, think about a situation where you have, say, a handful of trials between 2000, 2015, and today's 2024, and those handful of trials among themselves don't have much of heterogeneity, so they look pretty similar.

Doesn't mean that we can just do fully borrowing, so to speak, without any discounting. Well, nine years later, maybe not. Doesn't sound very reasonable, but then the question is, well, if that is not reasonable, what do we do? And should we limit the effective sample size of the historical controls or limit influence of the prior on the posterior.

For example, insisting that a certain minimum percentage of total information should come from trial information. And so I think understanding this would be important considerations, I guess. So I don't have any answers, just more questions and maybe some pragmatic solutions.

DR. SCOTT: Thank you.

I think Frank Harrell has something to add.

DR. HARRELL: Yeah. I think mixture priors have an advantage here because the mixing proportion is the probability of applicability of the other data you're borrowing from. And I would encourage everyone to look at the work of James Travis and others in the Office of Biostatistics CEDR, who have some really nice examples in pediatric studies borrowing from adults, where I think James and his colleagues also did a study of eliciting the applicability probability from experts.

And I think it's important to do this kind of elicitation exercise to get the amount of

borrowing from people that don't have any vested interest, and maybe you're not even informed about the previous results or the current results.

DR. SCOTT: Thanks, Frank.

Steve.

DR. RUBERG: Yeah. We talked a little bit last night at dinner and one of my favorite statements to make is evidence is continuous; decision-making is dichotomous. You're in a situation where you've got historical data, external data, whatever you want to say.

You have evidence that exists on a continuum, and you want to know how to map that down to a here's what I should borrow decision. And I guess how I would advise regulators, I would say, don't look for that objective answer because it won't exist.

There's no mathematical formula that says, here's how you take any and all kinds of evidence, and therefore it maps to 33 percent borrowing of this data. There's always going to be

subjective elements that people are going to look at the same evidence and weigh it different ways. People are going to look at, as Frank was saying, the temporal lapse.

And some people are going to say, "Oh, but therapy hasn't changed that much over the last decade," or "This is a standard therapy." Other people are going to go: "The medical world's a lot different than it was ten years ago. You can't use any of it."

So I think you're going to be stuck in just having cross-functional collaborative thoughtful conversations, and I guess I would say in the end, as a regulator, you're probably going to have to think about, what's the maximum that I'd be allowed to borrow without kind of being excessive or whatnot. What is that upper threshold? And there's no magic answer.

DR. SCOTT: So, Steve, I agree, and I respect your answer, but we have -- there are two problems. One is that a decision has to be made at

the design state, but the other problem is, it's not always clear how to structure the scientific discussions, even to distinguish between 10 and 90 percent of the borrowing. It's like picking a number out of a hat. And the question to me is, how do you structure the scientific discussion about that? And it's a hard one. I'm not faulting you for not answering.

DR. RUBERG: Yeah, yeah. And the only thing I can say is, you know, the list of things that I mentioned earlier about dose and population and duration and outcome measure and trying to understand what those similarities or dissimilarities and at least getting some judgments about that's a big jump, a leap in faith, or that's not such a leap in faith or whatever.

So yeah, I mean, I recognize it's hard to do, but I guess you're in the regulatory chair and ultimately you can say, here's our best understanding internally, and that's what we're willing to agree to and so be it.

DR. SCOTT: Roger.

DR. LEWIS: Just very quickly. I think that one of the challenges is that intuition on a percent borrowing scale doesn't work very well for most people. And one of the things that, if I understood correctly, Frank Harrell's suggestion about the reparameterization of the down weighting, where you explicitly have a prior on the degree of discordance between the treatment effects estimated from the historical and the current data, is that may, for some, provide a more intuitive way of thinking about how different are different degrees of discordance.

How plausible are different degrees of discordance between the data sets, and it's possible to take -- I believe that it's possible to take that formulation where people may be able to give you an informed opinion and actually recalculate it as a percent, basically variance inflation or effective sample size discounting.

So it may allow you to map it to the

parameter that you want, which is this percent discounting, but based on opinions on a parameter people actually may have some intuition about.

DR. SCOTT: Thanks, Roger.

And, Dean.

DR. FOLLMANN: Yeah. So this might be kind of a long-winded answer but if you had, like, a platform trial that was doing borrowing and did maybe 20 trials, you could look at the treatment effect estimates for those 20 trials based on borrowing along with their uncertainty. You could also make the 20 trials with zero borrowing.

You get 20 estimates, and they would have more uncertainty. And I'm thinking, like, harkening to the Efron and Morris paper where you looked at batting averages in April and some were -- you know, you could look at the actual batting averages or shrink them, and the shrinkage estimates were better when you looked at the batting averages in September.

So now we have, like, two competing

sets of 20 estimates. How can we see which one is better? Maybe you could do, like, a meta-analysis of the unshrunk estimates, which are going to be unbiased, and then with that meta-analysis, you might get more complex and say, yeah, some treatments are going to be null; some are going to be non-null.

So you do a two point mixture meta-analysis and say, this is my truth. These are the two means for the successful ones and the unsuccessful ones. Which line up better: the shrunken estimates that are 20 of those, or the 20 unshrunken estimates.

So it's a thing in principle that one could do under certain circumstances, and I think it would be interesting to do. It would be something that would maybe validate or support borrowing in an empirical sense, but I don't think it would help you tomorrow necessarily. But anyway, that was just a comment.

DR. SCOTT: Great, thank you.

And, Steve.

DR. RUBERG: Yeah. One last quick comment. It just comes back to me now from my days working at Lilly, but there was a publication in the early to mid-teens, 2012, '13, '14. I remember the lead author was a guy named Hay and some others published in Nature Drug Reviews, I believe.

And I looked at 5,200 drug development programs over a 20-year period and look at the transition probabilities from phase 1 to 2 to 3 to approval by therapeutic area: autoimmune, oncology, you know, et cetera, et cetera, et cetera, by biologic, by new molecular entity.

And I remember using that article, and people would say: "Okay, so what do you think the chances are we'd have some positive phase 2 result. What are these things' chances they're going to be a success in phase 3?" In some sense, what's the prior -- what's your prior for doesn't really work?

And I'd say, "Well, I'm going to start with this article, and it says 17 percent." You know, of all the autoimmune drugs or oncology drugs or

breast cancer drugs, you know, 17 percent that were successful in phase 2 actually went on and were successful in phase 3 and commercialization.

So in some sense -- you know, and people would come, teams would come: "Oh, we got the right mechanism action. We got the right this," you know, blah, blah, blah, blah, blah. I'd say, "Okay, well, let's see." In this study, the other 179 molecules that had that same notion, only 17 of them -- percent went on.

So I used to tell people like, "That's where I'm starting from, and maybe I'll give you a little leeway up or whatever, but you better have some really compelling arguments because history shows that." So anyway, don't know if you could use that, but I do remember having some conversations with teams inside the company and saying, "Okay, I just want you to be clear. Here's my starting point is the historical data in the pharmaceutical industry for this disease state or this whatever."

I think that article by Hay and all has

been updated with some more recent kind of trends or patterns and approvals and things like that as well.

DR. SCOTT: That's good. Yeah, I remember that one. I actually remember wondering, what if you shrink all of those therapeutic areas toward each other, how different were they actually? Although I think oncology was an outlier. Okay. So we definitely want to get to this topic before we run out of time. Under what circumstances can clinic trial simulations provide enough confidence in trial operating characteristics to support a confirmatory trial design proposal?

I think, Frank Bretz, you may have some insight on this one.

DR. BRETZ: Yes, some sort. Maybe not insights, but --

DR. SCOTT: Close enough.

DR. BRETZ: Yeah. So maybe first of all clarifying that I do think clinical trial simulations sometimes are needed, and I emphasize this for demonstrating type 1 error control. I emphasize

this because I know many colleagues, at least in Europe, who believe that they would not accept a confirmatory trial if there was not analytical proof of type 1 error control.

And I think it would restating that sometimes clinical trial simulations could help us in not demonstrating type 1 error control, but at least supporting false positive claims, and that they are limited or controlled. With that out of the way, then of course there are some settings where we do have analytical type 1 error controls and sometimes maybe where such a proof is not available, then we should run simulations.

And I think sometimes there's also confusion in the sense that many people believe that -- or if I use frequentist methods, by definition, I have analytical type 1 error control, and if I use Bayesian methods, by definition, I have to run simulations, and I don't think this is necessarily true.

I do believe that there are Bayesian

methods out there if you conjugate prior or so, then you do can -- yeah, you have a closed-form solution and vice-versa. There are frequentist methods for which you don't have analytical type 1 error control.

So -- and then I think you should have a good framework for planning, conducting, and reporting simulation studies. And I think the three case studies today were excellent examples of, you know, how simulation studies could be planned, conducted, and reported efficiently in a good way.

And I think that starts with having the questions upfront, explicitly stated what the simulation study is supposed to answer, understanding what are the candidate trial designs or analysis approaches, including a benchmark design like maybe we just talked about, not borrowing external data as a benchmark, just like a -- you know, a standard RCT.

I think such a benchmark design should always be included. We should understand the key operating characteristics that you would like to simulate. And then we talked about borrowing

information or where you get the external information from if you have some, so document any existing knowledge so that you can also describe the scenarios that you want to run your simulation study for.

So this is all very structured process at the design stage. And then, I mean, a simulation study is almost like a clinical trial study. It's an experimental design, so we should really plan for that. And but then consider implementation, we should also be careful about implementing the simulation study.

We should have details about data generating process, for example. Often, I just see someone just starting simulations and I don't know what they're simulating actually from. So understanding data generating process is important and then how you actually report and summarize those would be very important.

I can easily imagine, you know, people being bombarded with simulation results and then not knowing what to do with all this. So -- and I

saw -- we saw some great examples.

I think, Karen, you showed this. The Shiny app, which was very excited to see, where you go through the different possibilities, and you see some results in an interactive way. So anyway, it's a long answer, with some opinions.

DR. SCOTT: Thank you.

Jack.

DR. LEE: Yes. Following what Frank just said, again, you know, what evidence or what kind of simulation is sufficient, provide enough confidence. And, you know, the simple answer is that you need to cover all bases, right, and do it in a kind of fair way, right.

And but, you know, I think that I want to make a comment or have some discussion about the software availability, okay. So if I were a regulator, I'm sitting, you know, at FDA, okay, at a desk, and then I receive this elaborate simulation scheme and, you know, I think that first I want to understand it, but second, I want to reproduce it.

And third, I may want to try something that's not specified, you know, in that packet, right. So without a easy, accessible, this is not possible, right. And oftentimes, the CID can get very complex very quickly, right. So if I have to depend on whatever the sponsor submit, I feel I'm a little bit uncomfortable. You know, I want to be able to reproduce it, and I want to be able to run it using different parameter settings, right.

So we know that there are some commercial available software like BaCIS, you know, Cytel software, but then not too many -- but this can be expensive. So, you know, I would like to see more, like, open source, freely available software. And in this regards -- well, a little bit self-promoting is that at Anderson, we have -- it's trialdesign.org website, which is completely free.

And, you know, has many available software, including many of the hierarchical-based models, you know, basket trial, platform trial, you know, things like that, but it's far from really a

complete suite, you know, that allow, like, a sponsor or whoever interested in running.

It has many good element, but I feel nowadays that we need to have more this type of software, and I just want to mention a few more. For example, like Octopus, you know, like it's available, it's open source by Kyle Wathen.

And Herb talk about psborrow2, right, and that's open source. And there are quite a few R packages like basket, you know, the multisource exchangeability model, and there's a recent one called Simple, okay, or NCC, you know, and classical Bayesian hierarchical models CBHM and the BaCIS.

There are a bunch of them that available as R packages, but so far, I still think that we need to have a more -- this kind of freely available software and empower the, you know, stakeholders to really learn and study this. Okay.

Lastly, again, as I said, I'm very impressed with the Shiny app that Karen present, and this is good, and there are more and more such kind of

thing available. But, you know, Shiny app is at the point-and-click phase, right. So it's very easy to use, and it can provide beautiful, you know, graphics and table, graph, et cetera.

But in the CID, I think many time -- I also look -- I like to have kind of batch program because many of these can take a long time to run, right. And then, you know, I think what's lacking is if you -- no, if there are some way to run it as a batch and it can come up with a reproducible result, and that would be great, so we need both.

We need to have a point-and-click type of software, and we also need to have a batch job, you know, and so that it can all be reproducible. You know, something like our markdown kind of thing, you know, a steroid version of that, and you can actually run it, and then you can get exactly the same output of the report, right, and that would be wonderful. So look forward to have -- to see more people develop in this area.

DR. SCOTT: Thanks, Jack.

Roger.

DR. LEWIS: So I was rereading this question, which was blissfully short, but I think it covers -- I think there's a couple of different sort of pre-questions and post-questions. So I guess the first pre-question is, is not how do we determine operating characteristics, whether it's through simulation or through analytic methods, but what should our trial designs look like?

So if we can design a trial that's fit for purpose, for which we have good analytical understanding of its operating characteristics, why not do that. And so the decision to use simulation should be based on the need to do that because you want the trial design that requires simulation to understand operating characteristics has other objective advantages that are necessary for the development program that you're participating in.

And I think some of us work in areas where there are often analytical solutions. The group I work with, I don't know if we would recognize an

analytical solution if we ran into it because we're just not familiar with them for the problems that we try to solve.

The second point is both analytic analysis of type 1 error control and the use of simulation to understand operating characteristics are just two different tools. And like almost all tools, each of them can be done well, and they can be done badly. And we've certainly all seen, I'm sure in the review work we've done, absolutely standard frequentist approaches done badly and wrong.

And so the question is, do you recognize when it's done well and when it's done poorly? And I'm sure that many people have more experience making those distinctions for analytic-based approaches than for simulation based.

The point that was made earlier that if someone presents a clinic trial simulation to justify the operating characteristics they're claiming for their design, they better tell you enough about those simulations so that you understand what they are

doing.

The second piece is that they must provide simulations over a broad enough range of hypothetical situations so that you believe that the hypothetical situations that are plausible as actual things that might happen in nature are covered.

So we once had -- we once received feedback on a simulation analysis of type 1 error in which an unnamed regulatory agency very geographically close to where I'm sitting suggested that we evaluate the operating characteristics in a parameter, but if you looked at the parameter, they were suggesting we consider the situation in which all the patients became immortal.

And it's just -- it turns out that's not a pressing problem in some areas of oncology. And so that area of the space wasn't an area of space for which one had to explore the operating characteristics because it just wasn't going to happen, or if it did, it would be a good problem to have.

So the point I'm making here in a

long-winded way, and I apologize, is that your clinical trial simulations need to cover the plausible parameter space, but we really don't need to worry too much about the implausible parameter space.

If we find ourselves at the end of the trial finding out that we were in an area of space that we didn't adequately explore, there is a role for posttrial simulations to fill in the gaps. We should try very hard not to be in that position when at all possible, but I do find it difficult to justify an argument that we need to protect ourselves from errors that only occur in situations that can't plausibly occur in nature.

DR. SCOTT: Thanks, Roger.

One, second, Herb. Let me follow up on Roger's, and then you, and then I -- we're going to have to bring the session to a close.

So on the first point you brought up, Roger, I think it was very well framed. You know, you turn to simulations when you've chosen a design for which simulations are appropriate. And I guess -- and

clearly there's no simple answer to this question, but the question is, when is the benefit of the complex design versus whatever difficulty and lack of confidence there is in the simulations due to, e.g., multiple endpoints, or very, you know, geometrically complicated parameter spaces.

How do we evaluate that tradeoff versus telling people just to go to a simpler alternative, which is question eight, which we don't have time for, but I would invite people to think about and take home with them.

Herb, what was your comment?

DR. PANG: Yeah, just very -- two quick comments, is you really need to have sufficient time to think about the scenarios. If you don't have time, I think it's risky, right, to run into trial. And then another point is -- actually, it's very related to this is, I think the opportunity at the CID really gave us a good understanding of the sponsor and also the regulator side, how to do these things appropriately. So I think that's actually a very good

platform to do so, so I hope this kind of endeavor will continue, yeah.

DR. SCOTT: Thanks, Herb, appreciate it. And then the final question was, do you have any suggestions for ways FDA can support the appropriate use of complex designs in addition to the CID Paired Meeting Program.

We are about to break for the afternoon and then come back for Q&A. But I would like to note that there's a public docket open for this meeting and if panelists or anybody else has an answer to this question, we'd love to hear it. The docket is open until April 4th, if I remember correctly. So anyway, we'll break now and return at 2:45 for public Q&A. And thanks again to the panel. That was really helpful.

DR. SCOTT: Hi, everyone. We're going to be transitioning into the Q&A session. Okay. So to close the meeting, we have an opportunity for public comment or Q&A. We'll start with folks in the room if anybody has comments or questions, but we're

also taking questions on Zoom. If possible, if you're comfortable, it would be helpful give your name and your affiliation when you give your comment in the interest of transparency, but it's not a requirement. I know we have one in the room.

MS. BUTTS: Thank you so much. My name is Cherie Butts. I'm at Biogen. And I've really appreciated all of this discussion. It made it worth coming down here, although I used to work at FDA, so that was really good. My conundrum is that for all of the questions that were raised in the presentations, I added "and rare diseases" because that's what I focus on.

So I would love to get all of the panelists' comments on two things: number one, we are encouraged to borrow, but you know that there's a lot of heterogeneity in rare diseases. So perhaps comment given the populations that are small numbers, what do we borrow? Like, what's reasonable? And then number two also relates to heterogeneity, usually -- oh, good. It didn't record my name.

Also related to heterogeneity, when it comes to rare diseases, there are very, very few established efficacy endpoints. And so as it relates to, I think it was question number seven about confidence, so we will evaluate a series in our first study and then we hope that that will increase our confidence in the second, but we might have all of the wrong participants.

So I also wanted to get your comments, all of the panelists comments on this idea of how do we have confidence when we expect our population in our trial will probably be heterogeneous. So borrowing and confidence: Those are the two things I wanted each of you to comment on.

DR. SCOTT: Thanks so much for the question.

Would anyone like to start?

DR. PRICE: Oh, sorry, go ahead.

DR. SCOTT: We'll take Karen, then Dean.

DR. PRICE: Okay. One thing just to

mention is there was a workshop on the use of Bayesian methods in rear diseases a few years ago. It's Duke-Margolis. Just FYI, it's online and might be worth checking out there because I think some of these things were discussed very much focusing in the rare disease setting, so just some thoughts there.

And some of the things that I recall are the importance of the caregiver insights and as well as patient advocates in the context of rare diseases because these are individuals who really -- they're dedicated to these rare diseases.

They understand what's going on. They understand the impact to patients whether their own self or to the people that they're taking care of, their family members. And so that is, I think, part of what help in this scientific conversation around what can we borrow, what is useful and how do we do that?

We also talked about -- and we've talked in this session, but I think is relevant in rare diseases on the role of structured prior

elicitation to help understand what do these relationships -- how do people think about if you knew certain information, how does that inform you about other future trials or other information.

So there's a lot of lessons, I think, in the prior elicitation literature as well that you might explore to help -- you understand what is the level of confidence and then how to best borrow, so just a few thoughts there.

DR. SCOTT: Thanks, Karen.

Dean.

DR. FOLLMANN: Yeah. So I think for rare diseases, like, everyone's different. They're heterogeneous and so on and I think really the focus would be on specific designs like randomized withdrawal or trying to characterize, you know, where you have both groups on drug, and then you randomly pick a time to withdraw the drug from one of the groups.

And you can also do a randomized trial where after the placebo versus treatment period is

over, everyone on the placebo arm gets the drug, which gives you some additional information. So I think I would try and bind, you know, specific designs that are relevant to that disease that can try and answer the question.

In terms of borrowing, you know, that sounds very generic and so on, and I think I would try and get, like, history -- so basically, enroll people in a protocol so you see and under the auspices of that protocol, you're measuring endpoints and categorizing things, and then either introduce the drug or take it away.

So you're borrowing, like, the historical data from that, perhaps, but you're not, like, borrowing a different dataset from different people in a different part of the world. So people are sort of acting as their own control in a way.

And then confidence, gosh, that's kind of general too, but I think for rare diseases, you know, the calculus is a little different and maybe their people are more willing to, you know, have less

confidence in their result, perhaps, changing -- you know, accepting a higher false positive rate.

DR. SCOTT: Thanks.

Herb.

DR. PANG: Yeah. So I can speak to the hybrid control designs because we work in that area and then we also have a grant. I actually studied this in particular in addition to, as we mentioned, oncology setting. We do have some scenarios where we planned how to do the methodology for designing these hybrid controls in a rare disease setting.

So we have some examples from the spinal muscular atrophy setting, and then fortunately, similar to the case for the oncology setting in which we actually have a trial that we can borrow from that we are simulating. So the data quality's a lot better than, like, other sources of real-world data, but in that scenario, we are developing some new methodology that can help doing the hybrid control setting.

So as we mentioned, augmenting the internal controls with external control arm, and we

are studying, like, also the properties. So we actually have that, and the work has already been resubmitted to the journal, so it should be released sometime later this year.

In addition to looking at it from the primary endpoint perspective, there's an interesting thing about the rare disease scenario in which sometimes you don't want to keep the subject for too long after the primary endpoint readout, right. We want to enroll them to trial the new drug.

So in the SMA case that had from already-approved drug, we actually can learn from it and utilize it and utilize some approaches to actually infer subjects who actually didn't get the treatment or didn't continue it as a control, but how to estimate the treatment effect, even though in the open label extension phase, you can still estimate.

So that's really important for the rare disease setting, and we do have some methods probably would be published in the next few months, yeah, related to that. So it also fits in the high unmet

need, right, which is a good thing. That's a consideration for these kind of designs. So yeah, so we are covering that. I think not just us, but many others also working in this area, so thank you.

DR. SCOTT: Frank Bretz.

DR. BRETZ: Yeah. Another example of borrowing information is in pediatric drug development, where you would like to borrow information from adult trials. And some of my colleagues actually went to the CID program. It's a trial in multiple sclerosis, so if you're interested, I'm happy to share information on papers that they have published on that.

DR. SCOTT: Thanks, Frank.

Anyone else?

Okay. Thanks again for your question.

Dr. Irony.

MS. IRONY: Hi. I have a question. I'm Telba Irony from J&J Innovative Medicine, and I wanted for the panel to comment on one point that I think could help answer many of the nine questions

that you presented. For instance, the amount of information or the amount of sample you borrow from previous or for external sources and that's first, the rarity of the disease; that wasn't mentioned.

So that could be a factor in how much you should borrow. And also on the unmet need of the treatment. For instance, if you're talking about the survival and you have to wait for long time to recruit a lot of patients to get enough evidence and patients are dying, even if it's not a rare disease, but it's an unmet need, doesn't that justify borrowing more information, more external information?

So I didn't see that to be commented among the speakers, and I wanted you to talk about that. Isn't it important to have the same amount of internal or clinical trial evidence when you look at the benefit risk of waiting until you get enough evidence?

DR. SCOTT: Would anyone like a crack at it?

Karen.

DR. PRICE: Sure. And I touched on this briefly with the example I gave earlier where we did have a trial where the endpoint was death, and our intent was to -- or what we wanted to be able to do is borrow some of the events from phase 2 and combine into an estimate of the treatment effect, placebo versus the drug.

And then we wanted to move on to answer other questions, including looking at our drug relative to active comparators. And so that was an instance where I think the unmet need was high. The endpoint was quite objective. So, you know, maybe there could be some discussion along with what you're talking about around the role of borrowing information when something subjective versus objective.

Obviously, the pain master protocol is a highly subjective endpoint, but we've put that into a master protocol to enable that more. But maybe for regulatory approval setting, something more objective I could see would be more likely to be accepted, to have the borrowing.

However, again, we were unsuccessful in getting to that point and it did -- it is important to think about that impacted the patients. There were more patients on placebo, and we missed out on answering some, I think, really useful questions, and so that is part of the calculus of thinking through the benefits and risks here, and it can't only be about some level of type 1 error, for example.

DR. SCOTT: Jack.

DR. LEE: Yeah. We talked a lot about dynamic borrowing today and generally speaking, you know, this related to raw data conflict, right. So the less conflict, the more borrowing, and the more conflict, the less borrowing, right. And what Telba just mentioned is another dimension of the things, right, like, you know, what type of disease, how severe, the severity, you know.

We can even include in some other dimension like the toxic -- no, efficacy-toxicity tradeoff, the cost, and whatnot, right. So I haven't seen this being done, but, again, we mentioned earlier

that we can construct some utility function and then when you try to turn the knob, right, and you can try to maximize that relevant utility function. I think it can be done.

DR. SCOTT: Herb.

DR. PANG: Yeah, just to add to the point about the overall survival endpoint. I think Telba's point is actually very important, which is something that happened kind of after the CID. I think more recently, the FDA also see that overall survival is very important endpoint, like, to emphasize and also to study.

So having the borrowing and allowing you to look at it earlier with better power I think from the responses perspective and maybe with some good control of type 1 error is advantageous, right, which is something that I actually didn't bring up.

But definitely a good point that more recently FDA emphasize on the important of OS and there's some discussion in other forums about that topic as well, yeah, so thank you.

DR. SCOTT: Rebecca.

DR. HUBBARD: I've been thinking about the tension between continuous measures and needing to make a dichotomous decision, for instance, whether or not to move forward with a particular trial design.

And it seems to me that we can accrue a lot of information on a continuous scale, things like effective sample size, measures of robustness of the parameter estimate to different prior choices, et cetera, but at the end of the day, there's a dichotomous decision that has to be made about those.

And I think the point that you're making gives us the additional contextual information to decide where to set that threshold, which I think goes to Jack's point about utility functions. So I think that's how we sort of harmonize those two sort of seemingly incompatible things.

We have this continuous information about what we have learned or what the value of borrowing, or what was the amount of information that was borrowed. And now, we need to make a decision,

you know, are we confident enough? Is it good enough? Is it robust enough?

And I think the only way that we can make that decision formally quantitatively is by bringing that information about the strength of the endpoint, unmet need, et cetera, incorporating that into a utility function and then deciding.

DR. SCOTT: Thanks.

Steve.

DR. RUBERG: Yeah. As much as we've been talking about Bayesian approaches and borrowing, and I'm generally very favorable for that. Telba, it kind of relates to your question around, particularly survival outcomes, and I'll focus on oncology, for example.

Being careful about what you're borrowing because if you're looking at overall survival, usually people don't get to that terminal endpoint of death without having disease progression. And typically, when you have disease progression, you switch to another line of therapy or additional

therapies.

All right. Now, that overall survival outcome depends not only on what you initially randomize to, but what second-line therapy. And then when you look at the overall survival estimate, the proportion of patients that went on second-line therapy or third-line therapy.

And I know that cancer treatment is not uniformly done across the United States, let along across the world. So Harvard Medical School may use this second-line therapy. Somebody else may use some other second-line therapies, et cetera.

So then you got to start asking yourself, well, what am I borrowing here? What's going to happen to my trial? What proportion are going to get to progression-free survival? What second-line or third-line therapies might there be, you know, and all that kind of -- so I don't know.

At least in the context of oncology, that overall survival outcome is usually a mixture of many treatments along the way. Best supportive care,

following progression, et cetera, et cetera. And just difficult to think about, can I borrow that? Should I borrow that? If I'm borrowing it, what am I actually borrowing?

Do I think that the scenarios that played out in those trials are similar to the scenarios that might play out in my trial, especially with the rapidly changing environment in oncology, where a study that was done three years ago, first-line, second-line therapies are changing considerably from one year to the next.

So anyway, while I like the idea of borrowing in general and Bayesian approaches, in the oncology world for overall survival, I don't know, it makes me a bit nervous as to what I'm borrowing. For progression-free survival, okay, now I can look at, here's the initial randomized treatment, et cetera. Overall survival is a much more complex thing, I think, so --

DR. SCOTT: Thanks.

Anyone else?

Okay. I think we have another question.

MS. MO: Hi, May Mo from Amgen. I do have a question. Not loud enough?

DR. SCOTT: That's good.

MS. Mo: So, Steve, I hear you talk about evidence is continuous, decisions dichotomous that give a lot of people like high blood pressure, right, because we're afraid of making mistake.

So the question is, I know in diagnostic, sensitivity specificity basically is false positive, false negative. Depend on different prevalence rate, right, your decision rule actually adjusts to that.

So the bottom line is there's no fixed number. It's the context and the risk. What really is the risk of a false decision? And how we can, in our work, view that so basically, we are not talking about a fixed p-value or a fixed posterior probability, let's say, but really thinking in that scenario when we make a mistaken, what that means, and

how big is the risk.

I think that's a area potentially we can all work together and think together. Like, what is the prevalence in our innovative trial setting, like, something we leverage?

DR. RUBERG: Yeah. You make a good point. If you think about clinical trials or drug development as a diagnostic process, you're trying to answer the question, does this treatment work or not? And I think there's a direct and almost near-perfect analogy.

The prevalence is kind of like I mentioned that article from Hay et al. or whatever updates from that, that kind of gives you the background prevalence of drugs that are successful in phase 2 and what actually goes on. And in the past, and even in the present time, I look at this sometimes and say, okay, if that's the prevalence and you design a trial with alpha 0.05 and power of 0.08 or 0.09, I draw my little two-by-two table.

And I say, well, the positive

predictive value, if the study's positive, the positive predictive value might only be, you know, 55 percent, right, or 60 percent, and the negative predictive value kind of perspective as well. I mean, PPV and NPV from a diagnostic test are Bayes formula. I mean, they are one and the same. They're identical.

So yeah, I think I've tried to as I've learned more Bayesian statistics and drug development over the last 15 years from people like Karen and others at Lilly, I take that diagnostic view quite often and say, you know, what's the positive and negative predictive value for this phase 2 result, or this thing that I'm looking at is probably a more accurate representation.

And then, again, you can start adding values to false positive and false negative decisions, true positive, true negative decisions, et cetera, et cetera. So anyway, with a little bit of self-aggrandizing, some colleagues from Pfizer and I published a paper in Harvard Data Science Review last September with relationship to machine learning and

artificial intelligence algorithms to do clinical predictive kind of diagnostics, et cetera.

And we talk about these kind of things and values, so there might be some analogies to the clinical trial work that I think are very useful.

DR. SCOTT: Anyone else?

I would just add I agree that missing from the picture of the threshold for approval is the chance of -- real chance of an error and also the cost of an error, which is not always the same. A type 1 error for a drug that's unsafe or a drug that is going to be a barrier to more effective drugs coming down the pike is worse than a type 1 error in other situations.

Does anybody else in the room have any questions? Okay. In that case, we have several questions from Zoom. Going back, I think this first one was for Karen. "Did your placebo arm run throughout the study? To the non-statistician audience, would you mind explaining how you account for any changes in placebo response over time?"

DR. PRICE: Okay. So the way that it worked is that patients -- so we hadn't given intervention that came in. Patients identified what was the biggest complaint as it pertains to pain, and that indicated which of the disease state addenda they would end up in.

So some patients had multiple types. It's the one that was the biggest complaint. And then patients were randomized to placebo or that intervention. And if there multiple going on, then they would be randomized to the intervention and then to drug or placebo.

So the placebo was concurrent with the LY, the Lilly drug during the duration of that while the Lilly drug was being studied. So I think the question is, was there a continued placebo arm, and the answer would be no. It's according to the -- when the intervention is in.

What was the second half, then? Sorry.

DR. SCOTT: No, no problem.

DR. RUBERG: Changing placebos --

DR. PRICE: Changing --

DR. SCOTT: Yeah.

DR. RUBERG: -- time. So if you're pulling any from --

DR. SCOTT: Are there any -- do you have any way of dealing with trends in placebo response?

DR. PRICE: Sure. So a lot of that would come through the modeling. I don't think that we've seen it, but I think if there was an instance where the placebo, the true placebo response, underlying placebo response was believed to be different, then we probably would not borrow the earlier data.

So then that case, things we talked about around exchangeability, those are clearly violated. I guess the other place would be, as I mentioned, with route of administration in cases where it is known that the route of administration can truly influence the placebo response. Again, then we would borrow more from those that had a similar route of

administration.

DR. SCOTT: Okay, thank you.

And we also had a question for Herb. "The propensity score adjustment seems very conservative to me. The 'external controls' are from a contemporaneous internal study. Was it more about handling the covariates?"

DR. PANG: Yeah. So thank you for the question. And so after the first CID meeting -- the initial CID meeting, we actually didn't propose propensity score adjustments. And then after the first meeting, FDA actually asked us to consider propensity score-based adjustments, and then we actually, in fact, were thinking of either doing weighting or covariate adjustments.

So but in the end, it was decided that it's better to go with the propensity score matching. For the propensity score matching, it's essentially to try to just filter as a way to filter out the external controls that are quite different from the randomized subjects.

So it's just an additional step to make the subjects more comfortable, so -- and there were not that many subjects that were removed after that step, so they are quite comparable. But from a simulation, yeah, so think to essentially just make things more similar to the randomized subjects, so that's the goal, yeah, thanks.

DR. SCOTT: Thank you. A couple of the questions that came in are sort of -- I guess they're for me. One of them was -- sorry, I lost it. There was a question about whether all Bayesian or adaptive proposed studies should be discussed with the agency through the CID program, or can they be handled under IND Type C or Type D meetings. The answer is no.

You do not have to submit Bayesian adaptive or CIDs through the CID Paired Meeting Program. The program is useful for sponsors who can benefit from the extra interaction within the timelines afforded by the program, but all proposals, no matter how complex, will be considered under IND like any other protocol.

Let's see. I think there were multiple questions about whether we're posting speaker slides after the workshop. We will be asking the speakers' permission to do that formally. I'm optimistic that they will say "yes," but no pressure. All materials will eventually be posted to the event website and will also be linked from the CID program website. Let's see.

Yes. Oh, hi, please.

MR. COLLIGNON: Hi, thank you. Olivier Collignon from GSK. So I just wanted to come back to the question you raised, Scott, before the break, is that how the FDA can help, in particular, the industry, you know, use CIDs. And we saw through all the examples today the amount of resources that it takes for us to show the operating characteristics that we present to FDA or EMA or PMDA for that matter.

How do we build up the prior? And really my question is, could we think about the process that is a little bit more iterative, you know, rather than engaging straightaway into a full-on, you

know, fully fledged package. All the examples we saw today, that was really a context where the experimental context was way more challenging, you know, how the -- we are faced to with -- Frank talked about pediatrics.

I mean, another example we didn't cite was the Pfizer COVID vaccine, where you had an informative prior for the primary analysis is clearly a space where regulators are more willing to take more risk, right. So where I'm going with this is that I'm nervous engaging resources in a setting I know I have a 99 percent chance to have a no because there's no space for more risk taking.

You know, there are case in immunology where, you know, there's loads of drugs on the market already standard CDPR3 phase 3 trials. So could we think about the process that is a little more iterative. You know, first up, are we willing as regulators to take a little bit more risk than usual, certain stuff.

Let's have a discussion around the

prior, yeah. We think that the prior makes sense, but you forgot a few sources of external information that we know about. Let's factor that in. Your prior is too informative. Let's increase the variance a little bit and maybe we can start playing together, and then let's look at your operating characteristics.

So that's really my question. I'm a little bit directed to you, Scott, but I would like also to hear the point of view of the speakers. And I thank you very much.

DR. SCOTT: Thanks. I won't answer it in detail, but it's an excellent question. I think one of the goals of the CID meeting program was to be able to bypass a little of that iteration by having people bring their, you know, their fit-for-purpose complex designs to us in a way that we could then share publicly and could be used, if not as a template, at least as inspiration for what's possible with other proposals.

Frank.

DR. BRETZ: Yeah, no. Thanks for the

good question. I think what we saw today is that there's no one-size-fits-all. I think any CID will be highly specific to the setting, to the clinical setting, and better fit for purpose and better it addresses the question that it's supposed to address in a specific setting.

Now, I think the question is really, you know, how do we ensure beyond the paired meeting program, how to ensure a more sustainable use of CID. And yeah, I think it's -- we need to get an understanding as a community what are the objective advantages of running a CID versus a non-CID as a benchmark, so to speak.

Obviously, we need to understand still from these various case studies, some of them we have heard today, you know, what was specific that made these designs to be fit for purpose. I think it would also be good to get some consistency in decision-making all the time across therapeutic areas and sponsors.

So -- because you mentioned, why

does -- as a sponsor, you would like to have some critic ability if I do engage in a certain complex design, I invest a lot of resources, what are the criteria for success so to speak on the other stakeholders' side. So I think getting somehow a consistency in place. I think that would be helpful to have.

And finally, I guess, there's also something like precedent setting, right. So if you have the one CID is now applied and successfully applied in a certain setting, what does it mean for similar settings? Can we build upon the successes in similar settings later on?

So I think this will be to me questions to help having a more sustainable use or sustained use of CIDs in a more regular way, rather than always seeing this as a one-off solution, and we can never do it again, so --

DR. SCOTT: Thanks, Frank.

Karen.

DR. PRICE: Thanks so much for the

question. And I think mostly, I will echo what you've said. And I had mentioned a little bit in the presentation earlier, but I do think the identifying pathways that are more interactive, maybe a little bit more informal where could we submit a set of slides that summarize the really key questions and rather than a full briefing document with a protocol and an SAP or what -- I mean, I don't remember all of the things that were included, but that type of arrangement.

So again, CID Paired Meeting Program was very helpful. What allowed us to do it, though, was that we were not going to delay the start of something internally, and that was because the molecules that were coming in were doing other -- they were in tox studies and things like that.

So we were able to go through the meeting program. But in normal, fast-paced development, we probably couldn't always go through it. And so, like I said, we would love to understand maybe alternative way -- I'm not pointing only at you,

Dr. Scott, but I think that would be usually beneficial to have some more informal iterative approach for that, those types of conversations.

I also think, then, a lot that we've talked about with the open source, our Shiny apps, speeding iterative simulations is important that we don't have back and forth of paper. So the cloud computing, improving the infrastructure, those sorts of things all vitally important.

The final thing I might just throw out since we're talking about this is, and Steve and I were talking about this a little bit is, do we need the word "complex" in this whole conversation? And could we just remove it? Direct development is complex, period.

And so, you know, we're talking about maybe opportunities for additional conversation really to understand the scientific elements, operating characteristics. We know how to do these things. They're not necessarily "complex," per se. May require additional conversations, may require some

additional learning, but I think it dissuades people because, generally speaking, sponsors do not want to do things that are more complicated than it already is. And so just something to think about.

DR. SCOTT: Fair point. I think there's sort of a legislative history behind the word, Karen.

DR. PRICE: Fair enough, okay.

DR. SCOTT: Which makes it a little difficult for us to strike, but I hear you.

DR. RUBERG: Don't drag Congress into this, please.

DR. SCOTT: They dragged me into this, Steve.

Frank, yes.

DR. BRETZ: Yeah. I hear you, Karen. I just want to make the little comment that in Europe, we don't use the term "complex innovative design." Unfortunately, we kept the term "complex," but we have struck out the term "innovative," so we call it "complex clinical trial," CCT. That's the European

version.

DR. SCOTT: Yes, Olivier.

MR. COLLIGNON: So just one last addition about templates, et cetera. One type of information that would be helpful is some form of statement around what are the camera tricks that we always have to present in order to have an informed discussion.

I mean, clearly, we are all going to come up with power and type 1 error. I mean, the example we saw from Roger around average type 1 error. There are several people working on that in Europe at the moment. I think that's a very important metric to be presented when we are engaged in this type of design.

So some form of positioning from -- yeah, this is something we working on looking at or we'll never look at that, I think that would be extremely helpful. Thanks again.

DR. SCOTT: Good comment. I would put in a plug for these two guidances. They might be

helpful in terms of here's things that need to be submitted. And this brings -- actually, segues nicely into closing remarks.

Okay. So we're in the last few minutes of the workshop. And just to summarize what we heard today, we heard three case studies of innovative designs that may or may not have been complex. The CHIPS study of cold-store platelets, Eli Lilly's chronic pain master protocol, and the Genentech hybrid control in diffuse large B-cell lymphoma.

And we followed that with, I would say, a quite robust panel discussion, covering multiple topics, including the use of external data sources, Bayesian methodologies and trial simulations. I wanted to thank again all of our panelists.

We really value the input we've heard today. Once there's a transcript, I will personally be reading it and taking notes, and taking that home as we move on our next steps of policy development. And I'd also like to thank the public participants, both people who asked questions and also people who

spent their time with us today.

In terms of next steps, moving on from this workshop, I mentioned earlier there's a docket open for public comments. It's open until April 4, 2024. The link to that is available as part of the Federal Register notice for this meeting, and you can get there from the event website through a series of clicks.

And what we're going to do is take the feedback we've received today, the comments to the docket, review them, digest them. And among other things, what was discussed today will really help us in terms of our movement toward publishing a draft guidance on the use of Bayesian methodology in clinical trials for drugs and biologics, which was a PDUFA VII commitment.

That guidance is supposed to published in draft form by the end of September next year. Eventually, a transcript and link to the video of today's workshop will be posted on the event website and we'll also put a link on FDA's CID website when

available. My understanding is the video itself will be hosted on YouTube. I don't know the time delay, but I don't think it will be long. I think it'll be posted not too far from now.

And I think that was it. For more information, this is our CID website. You can also find it just by googling FDA CID, which is what I do every time I need to find it. But thanks again, everybody. Thanks to the panelists, and I hope everyone has a safe trip home.

(Whereupon, the meeting concluded at 3:26 p.m.)

CERTIFICATE

I, RICHARD LIVENGOOD, the officer before whom the foregoing proceedings were taken, do hereby

certify that any witness(es) in the foregoing proceedings, prior to testifying, were duly sworn; that the proceedings were recorded by me and thereafter reduced to typewriting by a qualified transcriptionist; that said digital audio recording of said proceedings are a true and accurate record to the best of my knowledge, skills, and ability; that I am neither counsel for, related to, nor employed by any of the parties to the action in which this was taken; and, further, that I am not a relative or employee of any counsel or attorney employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

RICHARD LIVENGOOD
Notary Public in and for the
State of Maryland

CERTIFICATE OF TRANSCRIBER

I, JUSTIN BISHOP, do hereby certify that this transcript was prepared from the digital audio

recording of the foregoing proceeding, that said transcript is a true and accurate record of the proceedings to the best of my knowledge, skills, and ability; that I am neither counsel for, related to, nor employed by any of the parties to the action in which this was taken; and, further, that I am not a relative or employee of any counsel or attorney employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

JUSTIN BISHOP