Want to comment? Your input is important. OLAW welcomes questions and comments from viewers of this recording. OLAW will post the comments, questions, and answers on the OLAW website. Please go to the OLAW Webinars and Podcasts page and click on the seminar title for further information.
Note: Text has been edited for clarity.

Foundations for Evaluating Study Design and Statistical Approaches for the IACUC
Speakers:

- Penny Reynolds, PhD, Assistant Professor of Anesthesiology, University of Florida College of Medicine

Broadcast Date: March 9, 2023 View
Recording: https://youtu.be/pzn5OikUFwY

>*Nicolette Petervary:* All right, why don't we get started? We still have some folks coming in. But I'd like to welcome everyone. Good afternoon. I'm Nicolette Petervary, part of the NIH Office of Laboratory Animal Welfare (OLAW). And today is Thursday, March 9, 2023. I'm pleased to welcome you and our speaker to our webinar today titled "Foundations for Evaluating Study Design and Statistical Approaches for the IACUC." We have just a few housekeeping details before we get started. If you have any questions, please enter them in the Q&A box, or you can use the chat. Dr. Reynolds will be taking questions at the end of the webinar. If the question is a little more nuanced or context-specific, or if we don't have enough time to address all of them, we'll forward those questions to her after the webinar, and then we'll append the question-and-answer to the end of the transcript. We'll monitor the chat the best we can, and we encourage you to use it to interact with us and with other participants. The slides, transcript, and webinar recording will be available after the webinar on our website. They do take some time to be processed for 508 compliance compatibility before posting, and this can take a few weeks, so please bear with us.

All right, let's get started with an introduction for Dr. Reynolds. And I'm very excited to have her here today. Dr. Penny Reynolds obtained her undergraduate and master's degree in wildlife biology and zoology from the University of Guelph, Canada, a master's in biometry and PhD in zoology and statistics from the University of Wisconsin, Madison, and has American Statistical Association GStat professional statistician accreditation. At the University of Florida, she is an Assistant Professor of anesthesiology in the College of Medicine, and is a member of the IACUC. Dr. Reynolds has been deeply involved with the development of the ARRIVE guidelines. She was part of the international ARRIVE 2.0 guidelines revision working group, and is a co-author of the ARRIVE 2.0 revised guidelines and ARRIVE 2.0 explanation and elaboration document. She was awarded the 2021 UK Animals in Science Education Trust 3Rs prize, and together with collaborators Maggie Hall and Elizabeth Nunamaker, the 2022 IQ Consortium and AAALAC International Global 3Rs award for significant and innovative contributions to the 3Rs of animal-based research. She's a published advocate for the improvement of animal-based research design through application of statistically-based experimental design principles and quality improvement strategies. Welcome, Dr. Reynolds. I will cede the floor to you. But before we get started, we have a few questions for our audience today. Would you please put up the poll questions? They vanished. Oh here we go.

So if you please answer these questions as best you can, and it'll inform us about how we should be approaching the content of this webinar. Thank you.

One thing while you're filling out the poll questions— unfortunately, Dr. Reynolds' camera is not working. We have gremlins in the system…so please bear with us. But she is excited to be here as you

will hear when she presents. Okay, I'm not sure, are we still getting people answering or can we put up the results?
> *(Technical Support):* It looks like we have gotten everything, so-- Oh, one more person. All right. Poll is ending.

> *Nicolette Petervary:* All right.

> *Penny Reynolds:* Oh, wow!

> *Nicolette Petervary:* Interesting. So the majority of you don't have training on experimental design. And it seems like confidence is somewhere in the middle when it comes to evaluation of experimental design. And looking at how does your institution evaluates the rigor of experimental design; it looks like all reviewers consider it for the majority of you out there. Interesting. Okay, so it looks like some of the attendees can't see the poll results. Can you check on that?

> *(Technical Support):* Looking, unfortunately, I'm not able to see what might be going wrong.

> *Nicolette Petervary*: I do apologize, there have been gremlins in the Zoom system. And we've had our IT folks working on this. If there are serious issues, we will stop the webinar and reschedule it. But so far, right now, I just will read the results. And as I mentioned, most reviewers on the IACUC consider elements of the experimental design as opposed to having a dedicated statistician who just does that. All right, why don't we get started with Dr. Reynolds? Dr. Reynolds, the floor is yours.

Slide 1: Foundations for Evaluating Study Design and Statistical Approaches for the IACUC
> *Penny Reynolds*: Well, thank you very much, Nicolette. I will probably just share my screen here. Can everyone see my screen? Can you see my screen?

> *(Technical Support)*: We're getting yes. Yes, yes.

> *Penny Reynolds*: Okay. Okay. Oh, look all these little raised hands showing up. All right. So today I'm going to talk about the foundations for evaluating statistically-based elements of statistical design. This is not going to be highly technical, or even mathematical.

Slide 2: Outline
Instead, I'm going to spend some time going through a bit of the background and the rationale and why it is now featuring so largely in the NIH radar. And then I will spend most of the time talking about statistically-based design essentials, some of them are "need to know" like bias minimization, but the design skeleton and quick methods for screening sample size definitely should be in the IACUC toolbox.

Slide 3: Learning Objectives
So at the end of this, you should be able to identify the five components of an experimental design which conveniently are also the five major elements which are needed for a proper research question. You will be able to describe three essentials for assessing whether or not an experiment is right-sized, and all the way through, I will provide little red flags, so you can identify where you should start asking questions in the protocol.

Slide 4: Background: Why Does it Matter?
So why does this matter?

Slide 5: Enhancing Rigor, Transparency, and Translatability in Animal Research
Well, about a year and a half ago, in June 2021, the advisory committee to the NIH Director came out with a report: Enhancing Rigor, Transparency and Translation Potential for Animal Research. And as you know, motivations are threefold. The biggest one, of course, is the really dire rates of translational success for most of preclinical research. The industry rule of thumb is less than 50% of preclinical animal-based results even make it to industry, and about less than 5% ever make it to clinical trials. And for some areas of research, like Alzheimer's research, it's more like zero. There are also well-documented problems and replication translatability of animal-based research. These have been in the primary literature and in major publications like Nature and Science for the past 15 years. And also, just as scientific practice generally is evolving, best practice certainly is, and we need to adjust to that and inform researchers of what the new expectations are.

Slide 6: Two Themes
Although they made recommendations for the five major domains, there were two things which were common to the entire report. The first one is that animal-based research is poor overall. And the second one is that the degree of understanding and training in statistics and statistical concepts on the part of investigators and researchers is very, very poor. Now, to be fair, the number one priority identified both by the report and by investigators themselves is that they really, really need better training in statistics and study design. But the big thing that comes out from this is that it doesn't matter how good the science is or how good the animal models are; unless studies are properly designed, properly analyzed and completely, appropriately and thoroughly reported, nothing of that will even matter— their models are useless.

Slide 7: Spoiler Alert! None of This is New
Now, that report is not a wake-up call. In fact, it's another in a long series of reports which say that study design really needs to be considered more rigorously than it is. In 2014, Francis Collins and Lawrence Tabak outlined an entire restructuring initiative that would improve preclinical research, of which the end result was that report to the director. But five years before that, NIH teamed up with the NC3Rs in the UK to do a review of published government-funded animal-based research, and they found what was very disturbing was that there were major deficiencies in reporting and omission of key information, all of which were related to scientific value— meaning that published research on government-funded money was essentially not very informative. Those reports led to the development of the original ARRIVE guidelines in 2010 to encourage better and more transparent reporting of animal research ,and their subsequent revision in 2021.

Almost 40 years ago, Patrick Bateson came up with the Bateson's cube. That is: In order to understand the harms and benefits trade-offs in doing animal research, we also have to consider the quality of the study as the essential third rail. And even before that, before most of you were probably even born, there was the foundation document of the 3Rs, Russell and Burch, where they explicitly identified statistically-based study design as an essential component of animal research.

Slide 8: The 3Rs and Statistically-based Study Design
So the 3Rs are not just a checklist of “replace, refine, reduce.” Instead, it's a central ethical framework for everything we do involving animals, [the] principle being [that] maximal information should be obtained from minimal harms. And Russell and Burch were very clear that that's achieved by statistically-based, appropriate and relevant experimental designs.

Slide 9: Barriers

Now, starting in about 2015, NIH started to recommend that IACUCs could be a bit more proactive in facilitating a good quality research. There has been a considerable amount of pushback on the grounds that, first of all, "I don't want to," either because it's not our lane or because it represents a greater regulatory burden that we don't want to deal with, and also because people just don't know what to do (knowledge barriers).

Slide 10: Perception Barrier I: Not Our Lane?
The first perception barrier: "Not our lane?" The answer is, well, yes, it is. [According to] OLAW Frequently Asked Question D.12, yes, it's true that scientific and technical merit is not in IACUC's lane; it's the purview of the scientific review panels. But the IACUC has to consider the Government Principles, the first one being that the relevance of an animal-based study has to do with whether it contributes to human or animal health, whether or not knowledge is being advanced, or the good of society. Now, those are all fairly vague and ambiguous. But the PHS policy review criteria are far more explicit. There has to be a sound research design, there has to be a scientifically justifiable rationale for using animals in the first place, and the research has to be scientifically valuable. So OLAW's take on this is that a study which can't meet these basic criteria is unnecessary, it's wasteful, and therefore it's not justifiable.

Slide 11: Perception Barrier II: More Regulatory Burden
The second perception barrier: That it's more regulatory burden. Well, that is a real concern, especially when people try to impose these sorts of things as a checkbox. It's not. What this should be is an agreement by the scientific community as a whole that best practices evolve. Best practices, including statistically-based design, are part of a responsible conduct of research to make sure that we're getting the best possible research for the dollars and the animals that are invested in it. So the goal is not to burden people with more things that they have to do, but rather instill a culture that good responsible research is embedded in as part of routine research practice, to enable researchers to develop the necessary skills and resources, and to make sure that there's close alignment between what is proposed, the animal use protocol, and how the protocol is actually implemented.

Slide 12: Better Science – Better Animal Welfare
And of course, in the *Guide*, it says on page 26 that this is also part of our duty of care to minimize those indirect harms caused by wasting animals in research which will ultimately be non-informative. It is true that [as the IACUC] we cannot evaluate specific experimental design, nor is it necessary. What we can do and should do is evaluate the design components. And that's going to be the substance of my talk.

Slide 13: Knowledge and Cognitive Barriers
Now, as far as knowledge goes, many IACUC reviewers consider (and rightly so) that they don't have the statistical expertise to evaluate a study design. Very few IACUCs in the United States include a statistician, but in any case, that really doesn't matter because very few statisticians have actually done an experiment in their life. In the United States, most of them tend to be mathematical statisticians, and so they're not really practically trained in terms of the nuances and the actual demands of a statistically-based design of experiments. And then again, very few IACUC members or researchers even understand what is meant by statistics.

Slide 14: Knowledge Barriers: Attitudes and Myths
Most people say, which makes it really not very fun for me at parties, is that they hate statistics. "Oh, fine, what do you do?" "I'm a statistician." Okay. But there's three basic myths that are underlying that. The first of all is that the reason why most people hate statistics is because it was taught to them as

being something that's very highly mathematical, and really only accessible to the gifted few. And therefore, statistics is seen as something magical. The second myth, which is also promoted by what I consider very poor teaching in this country, is that statistics is only analysis. And so the idea has come about that, "Well, we only need to consult a statistician after the data are collected at the very end of the process." Paradoxically, the third myth is that “It's so easy, anybody can do it, because after all, we have all the software.” None of these are in any way true.

Slide 15: Take Home: “Statistics” is a Process not a Thing
What has been recognized almost from the get-go but was formalized almost 50 years ago is that statistics is a process. It's not a thing. Statistics involves the careful formulation of a research question in the elements that will direct the statistical design. And the design itself encompasses the planning, the actual mechanical conduct of the study, data collection, analysis, and then interpretation. But it has to be built in before the data are collected, or even before animals are ordered, because one of the things that one learns in quality process control is that if you have poor upstream basics, you can't fix anything that occurs downstream. Fancy analysis can't rescue a bad plan. It can't rescue a bad design. And unfortunately, again, as a consulting statistician, at least half of the time when I'm called in, it's after the publication stage when the paper has been rejected. Well, since most of the problems have occurred upfront, there's not very much we can do.

Slide 16: Statisticians Should Be Consulted at the Beginning, not the End
In fact, the man who invented most of this, Sir Ronald Fisher, who invented not only methods of statistically-based study design, but also most of the analytical methods that we use routinely in analyzing our data— he said himself, in 1929: “To call in a statistician after the experiment is done, is no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.”

Slide 17: What is Design?
So part of our ethical mandate, and this is what Russell and Burch were getting at, is that statistically-based, clearly articulated study designs are important, not only for good science, but also for maximizing animal welfare. So what is design?

Slide 18: What is (Statistical) Design of Experiments (DOE)?
It's often confused with descriptions of technical methods and materials, and that seems to be the major problem with a lot of publications. But what it is (design of experiments) is an operationalization of your research question. More technically, it's a formal, logical statistical structuring of all those elements of the research question, your independent variables, your inputs, and your outputs. And what a good statistically-based design will do is primarily discriminate the experimental signal from any extraneous noise, so it has power to detect real treatment differences without necessarily demanding very large sample sizes. A good design can detect interactions (that is, any synergisms or inhibition responses) which is where most of the serendipitous scientific discovery occurs- it's the study of interacting factors. A good design controls noise by controlling unwanted variation by statistical methods such as blocking, subsampling and so on. It minimizes bias and it's right-sized. And I will talk about more of these things in a bit.

Slide 19: DYK? Statistically-based Designs Have Been Available for Over a Century
So this is sort of a family tree of how comparative experiments can be structured in terms of formal statistical designs. Most of the ones in the top half like *completely randomized, randomized block, Latin squares*, they've been around for almost a century, which is always a surprise to people. Some of them

like *optimal designs* and *response surface methods* had to wait until the development of more high-speed computing methods. The second thing, though, that really surprises people is that there's no mention of ANOVAs and t-tests on here. And there's a reason for that. They're not designs. Those are methods of analysis that implicitly assume that one of these formal statistical designs is already in place.

Slide 20: A Good Design Makes a Valid Study
I don't have time to go into the specifics of those designs, but that could be a subject for another seminar later on. But what a good design does is makes the study valid, and validity has been defined by the degree to which results are likely to be true. It assesses the strength of your causality between your inputs and your outputs. And so you'll see sometimes in the literature [that] we talk about internal validity. This is methods-based, it's assessing the truth within the study, and the purpose is to ensure that results are both true and reliable. So the second take home message is that design comes before your statistical inference, and data quality comes before analysis.

Slide 21: Benefits of a Statistically-based Study Design
Why should we bother? I mean, we've been doing the same thing over and over again, why can't we keep on doing that? Because a statistically-based study design requires far fewer resources for the amount of information you get. It's more efficient, it saves you money, it gives you more information, and that benefit spares animal use. You're not wasting animals on non-informative studies, and you're using far fewer than you might have thought necessary from the old traditional ways of doing things. But the second thing which is often overlooked is that the common statistical tests that you rely on also in their turn rely on a correct design. If the design is bad or if there's no real design, your statistical tests are actually invalid. That means they're giving you false positives or false negatives, not because the science was bad, it's because the design was bad. And again, quality can't be rescued by analyses done after the data are collected. Those problems are built-in.

Slide 22: Design Basics: Key Components of Internal Validity
So now consider the design basics, the key components to make sure your results are reliable and true. *Bias minimization* is the first one. I'm just presenting this as a need-to-know rather than as the means of assessing a protocol because we don't assess those aspects. But I will cover in more detail *design elements* and *sample size justification* because those definitely are within the purview of the IACUC.

Slide 23: 1. Bias Minimization: Randomization
So *bias minimization*. Bias minimization methods have nothing to do with sample size. A large sample size doesn't mean your study is biased or not biased.

[Randomization is] a method of ensuring that there's no systematic differences in the treatment assignment, say, like, all the controls are processed first, followed by all the tests; or this animal doesn't look like it's doing so well, let's make it a control. No, randomization is a statistical method for ensuring that all of your interventions have an equal probability being assigned to the experimental units, such as a mouse. It also determines the order of processing and measurements. So there's no time dependencies baked in, which might affect your results.

“Random” is not haphazard, it's not ad hoc, it's not unplanned, and it's not alternating. Those are the layperson's definition of random. It's a formal probabilistic process, which probably should be— the randomization scheme needs to be generated by some sort of computerized algorithm. [SAS and R as references]. So again, it prevents both systematic and selection bias in your results, but it also ensures the validity of inferential tests. Randomization is the cornerstone of all inferential hypothesis tests

that we perform. If it's not randomized, your tests are invalid. That's pretty frightening because a lot of people don't know about this.

Slide 24: 2. Bias minimization: Allocation and Sequence Concealment (Sometimes Known as "Blinding," "Sealed Envelope")

The second method of bias minimization is allocation and sequence concealment. This is also known as "blinding" or "the sealed envelope method." Now, this method of bias minimization is operational, because it has to be built into your standard operating procedures. They are methods that conceal which treatment was received by which subject. So suppose you're going to give one of two drugs to your study animals. You would make sure that they were coded so that you couldn't interpret or figure out what [each one] was: they'd be covered in foil, so you [couldn't] see what the fluid was, that kind of thing. It minimizes conscious or unconscious bias in not only the allocation of interventions to your experimental units, but also in how they're assessed and how they're interpreted. And this is especially important for subjective outcomes like histology, which is where most people report doing blinding, but also things like behavioral assessments. That's especially important. It's not always possible to blind every single person involved in the study. Ideally, all of the operators, all the assessors, and all the analysts should be blinded. Sometimes that's just not possible. So it behooves you to use the highest level that you can. Sometimes it might only be possible to blind the analyst. Definitely, if it's a subjective outcome, you want to blind the assessors.

Slide 25: Design Skeleton

But now we're going to get to where the rubber meets the road, and that's the design skeleton, or also the research question skeleton. So what it does is break the research question into its five basic components. There should always be sufficient information in the protocol to evaluate every single one of these items, no matter how complicated it is. The acronym is very simple, and I've shamelessly stolen this from clinical research: the acronym is PICOT, P-I-C-O-T; participants- study population or test platform, I is for intervention. C is for comparators and controls, O is for outcome, and T is for the timeframe of the study.

Slide 26: Think of it as a System Diagram

So you want to think about your research question and also your statistical study design as a systems diagram. You have your patient platform (your mice, your mouse sample) you're going to allocate either an intervention or a control or some sort of comparator to [the measurement unit]. You're going to measure the outcomes, and it's all constrained within a certain timeframe with a designated start and stop time. Who does what to what and what happens? And that's how you assess causality (cause and effect).

Slide 27: P = Animal Model

So starting with P, the first question to ask is, are animals needed at all? People are often disappointed when they ask me "What's the minimum sample size for my study?" and I say, zero. But this is something we really need to ask more. Training: animals don't need to be obtained for purpose all the time. You can learn skills on simulators, on carcasses, culls; it doesn't have to be a lot of money, but it will save money and spare animals. The second red flag item to ask for is why that animal? There is a principle known as the August Krogh principle from a physiologist in the 1930s. And he said the choice of the animal model should be most appropriate to the study question being asked. Don't say, "It's because it's the lowest phylogenetic representative or it's the least sentient." Those arguments were debunked in 1823, so 200 years ago.

The issue is not whether or not they're sentient, or they can reason, the issue is can they suffer? And it's very clear that they can. So you must justify the model scientifically based on the features of the anatomy, physiology, behavior and genome, which are most tailored to answer your scientific question. Postulated effects or mechanisms of action need to be incorporated, and also a reasonable assessment of any competing models which are obtained from the study of the literature. You also need to ask whether or not the test conditions, especially for the disease management model? How is that induced? Is it spontaneous? Or is it induced by some sort of mechanical procedure?

Slide 28: Model Choice Affects Results and Interpretation
Now, poor choice of model will affect everything. This, the top figure is a fairly old one from NIH just showing the same three strains of mice. [A given strain] can either be a reasonable comparator or complete outgroup, depending on which portion of the genome is targeted. So you really have to know your animal. The second figure is from a really terrifying study, which concluded there's no such thing as a black six (C57BL/6) mouse because it's really dependent on the choice of control strain or substrain, or even what facility with the same vendor that the mice were derived from. And as you can see from this figure, much to the horror of the investigators who were reporting it, you get completely contradictory results from exactly the same type of experiment. So, model choice has to be justified, and it has to be well reasoned.

Slide 29: I = Interventions
Number two is interventions. This is just the treatment or whatever is administered to the study subjects. Generally, because this is somebody's pet brainchild, it's going to be fairly well-described. It's just what the investigator plans to do to the subject or for the subject: a test, a procedure, a therapy, and so on. All you need to do is identify what it is, how much, and how often.

Slide 30: C = Comparators and Controls
Comparators and controls. These are what are used to establish cause and effect between the intervention and the response. So, it's another method of distinguishing your signal from any noise. There are seven types of controls. I'm not going to go into them in detail here, but I do want to flag up a couple ones where there's some fairly profound misunderstanding.

Slide 31: C = Comparators and Controls (continued)
And those are historical controls, shams, and what I call redundant controls. Now people sometimes suggest that historical controls should be used because it saves on animal use. This is actually a false economy, because historical controls aren't valid for hypothesis tests against new data because they can't be randomized, at all. So, they're also implicitly assuming that the animals and conditions obtained from the previous condition are identical in every respect to the animals in the new condition. And unless they meet very stringent [requirements], what's called "Pocock criteria," that every single experimental condition, reagents, personnel, animals, vendors, handling, husbandry, yada, yada, yada on into the night— are identical, they're not going to be valid. So historical controls should definitely be discouraged. It's a waste of animals, and it doesn't tell you what you really want to know. It's really essentially comparing apples with bowling balls.

Shams: There are some occasions when shams are appropriate, these need to be scientifically justified. However, if it's an iatrogenic procedure which is well-established, the endpoints are highly predictable and spontaneous recovery is not possible, then having a group of animals that are untreated is not only wasteful, it's also unethical. So that includes things like acute spinal cord injury, acute hemorrhagic shock, and many sepsis models. We know they're going to die, so why-- you don't need that.

Then there's redundant controls. And this is a little more difficult, it should be flagged up as something to ask about, but it's not something that an IACUC really has control over. A lot of researchers were trained in statistical courses which assume the two groups or two-arm model where you're comparing a control against some specified intervention, is the norm. This is certainly the case for human clinical trials where you're trying to establish efficacy, but it is not appropriate for exploratory animal-based trials. Then what happens is you've got lots and lots and lots of little two-arm studies, each with its own control. So that's essentially wasting all those animals, because it's very inefficient, you don't get a very high precision of your estimates, and it can't detect interactions between competing factors. So, this is where designs like factorial designs would be especially important for animal-based studies. And again, I don't have time to go into it here, unfortunately.

Slide 32: O = Outcomes

Outcomes: Outcomes are the answers to the research questions. These are the responses that you're measuring, and these are the variables which drive sample size calculations, and they determine the methods. So, this is really where the rubber hits the road. When you're reading a protocol, the experimental outcomes have to be relevant to the research question. They have to be clearly defined, they have to be specific, and they have to be measurable. What's being measured? How will it be measured? How often will it be measured?

Slide 33: Outcomes: What, How, When, How Often?

You really want to watch out for things that are really vague and undefined, like “survival” or “function” or “protection.” Those can't be measured, so they have no meaning. If they gave a specific metric associated with it like protection, meaning “reduction by 75% of the bacterial load in the gut,” that's something you can work with. But just little vague statements, you can't. Another one is: How will they know the experiment worked? And so, they have to have some sort of defined success criteria, [not] saying “it's better” or “it's improved.” Again, that's not a standard of comparison. So, they have to give some concrete differences [e.g.,] “We predict that if this drug works the way we expect it will, we should see a 25% increase [in] blood pressure,” or whatever the outcome metric was— CD4 counts, white blood cell counts [etc.], you should see an elevation or a reduction, whatever. But it's got to be quantifiable and measurable.

The third thing about outcomes is that there's always going to be some technical methods associated with them. So, the things you want to ask is how are they obtaining these measurements? Is it invasive or non-invasive? Is a lot of instrumentation required? Are multiple surgeries required? Is a lot of extra sampling required? And then there's the 3Rs considerations: how much pain and distress is associated with it, and what alleviation or welfare plans are already in place?

Slide 34: Humane Endpoints

Now, humane endpoints come into this because these will affect your choice and your measurement of your outcome variables and will also dictate the timeframe. I prefer to call them “welfare indicators” rather than “humane endpoints” because people get confused with experimental endpoints. But these are one or more behavioral or physical criteria, and the more the better, that determine the point at which the animal is going to be removed from the study to minimize its pain, its suffering and its distress— usually by euthanasia, [but] it doesn't have to be. But the indicators need to be predetermined, objective, easily recognized, and people have got to stick to them, and they've got to be really rigorous about this. So, these are just as important as the experimental outcomes that are already being measured. And there's plenty of published guidelines for specific research models; I just included a few here. I've also provided a resource page with Dr. Petervary. So that should be available on the website.

Slide 35: T = Experiment Time Frame
And then finally, the experimental timeline. This is all the timeframe and the associated tasks that are involved in the entire experiment. So, you should be able to track a whole animal all the way through. That includes:

- Pre-experiment preparation, like the amount of time it needs for acclimation after it comes from the vendor: How long are you going to habituate it to handling?
- The instrumentation phase, [for example] if you're doing a surgery: How long are you actually monitoring it after the intervention and after the- for the follow up period?
- And what is the point at which welfare indicators say the experiment should be terminated?

Now, the question to ask: Is that post-experiment period actually long enough for the effect to show up, especially if there are delayed responses?

Slide 36: Design Skeleton
So, here's a little example. An investigator wants to study effects of a new drug thought to improve motor function in rats with osteoarthritis. They're going to make the measurements once a week for three weeks. So, you break it down into your PICOT statement:

- Your study population is rats with osteoarthritis. Okay, why rats? What strain of rats? How did they create osteoarthritis? Was it spontaneous or induced? How did they do it?
- The intervention is the new drug. What is that drug? Why did they choose it? How is it formulated? How is it going to be administrated? What's the drug dose? What's the concentration?
- Comparators: they didn't say, [so you should] flag it up. What are you going to use to know if your experiment works?
- [The] outcome is motor function. I'm going to go into this in a bit more detail in a minute, because you'll see how important it is to define how the rest of your experiment is performed.
- [The] timeframe is three weeks. Okay, so that's good enough, once a week for three weeks. So, is that only three samples? What's the justification for three weeks, and what are the endpoints?

Slide 37: Example: What is the Outcome?
Now [with] “motor function,” you see how vague this is because you can have multiplicity of metrics. So, I just picked two. One is a non-invasive one: just measuring step length. Well, that's easy to do, and the animal could be measured many times. But if you're taking, say, peak forces of an isolated muscle, well, that's not only highly invasive, it's also terminal. So that's going to affect not just the course of the rest of the experiment, but also your sample size because obviously, you'll need many more animals for a cross-sectional study than you would for a repeated measures study.

Slide 38: Screening Animal Numbers: Why Does it Matter?
Now, the third leg of this design assessment is screening animal numbers.

Slide 39: Animal Numbers
Now, the IACUC is tasked with determining if numbers are appropriate, justifiable, and ethical. But what are justifiable numbers? Well, in a lot of discussions, they always seem to bring up power calculations. Those are the gold standard, but unfortunately 99 times out of 100, and I'd say even 99.9 times out of 100, they're misapplied. Most of the time, they're incorrect. They use assertions which are absolutely unverifiable. And most of the time, they're gamed: people will just perform calculations so they can

get their favorite number- five per group, eight per group [etc.]. Or it's to tick a box. In 2018, Fitzpatrick and collaborators did a survey of over 1,000 IACUCs across the United States, and the number one answer was [that] power calculations are just a necessary evil to satisfy the oversight committee and the reviewers. That's the wrong attitude. Power calculations are essential part of your experimental design and also, they will drive the integrity of the results. But another problem with power calculations is they ignore logistic constraints.

Slide 40: Instead Consider 'Right-Sizing' the Experiment
So instead of going by power calculations, I would like to consider "right-sizing" the experiment. It means sample sizes are not only statistically justifiable, they're also operationally and ethically justifiable. So, the numbers have to be sufficient to address the research question— not too few, not too many. And they have to ensure that animals are not wasted. So, the three questions that you can ask without knowing any math whatsoever are: Are the numbers feasible, verifiable, and ethical?

Slide 41: Feasible?
Feasible: do the numbers match the lab capabilities? Do they have enough trained, competent personnel to do the work? You don't want five technicians who have just taken their first AALAS online training session. You want people who know what they're doing. Do they have enough money, space, time, budget, or are the numbers completely ridiculous? This was usually the case with most protocols. We usually get requests in thousands, tens of thousands, and hundreds of thousands. My two personal favorites are a million mice. And the number one on the leaderboard was 91,386,777 mice for a three-year protocol.

Now, you don't have to do any more assessment. I mean, this is just clearly ridiculous. But what it communicates is, number one, no planning. And number two, it's a type of gaming the system, which I particularly dislike, because if the IACUC is foolish enough to approve these numbers, that means they [the researchers] can waste any number of animals in futile, non-informative experiments which go nowhere. And it's part of a syndrome, which Joe Garner of Stanford calls the "mice as disposable furry test-tubes syndrome."

Slide 42: Verifiable
Second, are the numbers verifiable? Does the researcher show their work? And if they don't, you do the math. How many experiments? How many groups? How many animals per group? And what's the expected loss? And just multiply it out. Quite simple.

Slide 43: These Statements Cannot be Verified
These statements, which are frequently used in protocols as numbers justification are not verifiable and therefore not admissible. "Based on our previous publications." Well, I've always looked them up when they supply them, they never provide a justification. It's just the same number they've always used. "In our experience" - we don't know what your experience is. "This number is sufficient to obtain statistically-significant results." Well, that's a big red flag because it means they don't know what P-values mean. [A small P-value] could be a false positive. "They might have unforeseen problems." "It's unknown how many animals we require because it's exploratory." "It's what everyone else does". "It's the industry standard." No. [The] major take-home here is that sample size justification is bespoke, it's not copied piecemeal from other publications. [Numbers] have to be aligned and customized to the study that's being considered right now. So, what are the facts, the inputs, what are the outputs that are being measured?

Slide 44: Ethical?
And finally, is it ethical? Do they have- can they reduce the numbers? Generally, they can. Can they use better designs and methods? Something we can suggest. Do we have welfare checks and mitigation

plans in place? Something we certainly need to check out. And then there's the collateral losses. So, what are they going to do with the animals that they might be generating but not actually using? What is the end-use disposition? And this is especially true when researchers are allowed to do their own in-house breeding. For example, we had a PI who requested 400 mice for a certain genotype, and that was appropriate. And it was actually feasible and it was justifiable. But in the generating of that genotype, they had produced well over 1,200 animals of an unwanted genotypes which were just going to be euthanized without ever being used. So, we should encourage more thinking through these things and finding mitigation plans, maybe the animals could be transferred to another protocol, for example. Maybe they don't need 400 mice.

Slide 45: Example: Conventional vs. Statistical Study Design
So just in the few minutes we've got left, I want to show an example of how a conventional plan works versus a statistically-based study design.

So, this is a trial of vaccine efficacy in mice. They wanted to do a bunch of two-group comparisons, or what they call one-way ANOVAs on six strains, three doses, three dosing intervals, three age classes, in both sexes, which is all perfectly appropriate. And these are questions which needed to be asked. They requested five mice per group because everybody else does. And so, if you do the simple math, it works out to be about 1,620 mice.

Slide 46: Why is this a bad "design"?
Okay, this is not really a design, it's actually not very good. Even if you could process 40 mice per day, that's still a month's worth of personnel time that has to be involved, so it's unmanageable. There [are] large sources of variation through time, through space, through different technicians being segued in and out of the study, that's going to swamp any true experimental signals.

The second thing is a big sprawling experiment like this with lots of two-one comparisons is going to miss every important result, the most important drivers compared to each other. It'll miss synergism, [and] it'll miss the optimum best response. It's probably unfeasible. I would always ask: Do they have enough trained people to process? But it's almost certainly wasteful, because many investigators order animals in bulk just to save on shipping costs. But the end result is that the animals age out of the study before they can be used, so then it's simply sad. Also, it makes for a very strong temptation to discard any non-statistically significant results that may come up out of the study and only publish the ones that give very small P values. That is questionable research conduct. And in some places like the Netherlands, that's now rising to the standard of research misconduct.

Slide 47: Compare with Appropriately-Designed Screening Experiments
So this is a screening design, which I generated in about two minutes in SAS JMP Pro, where you can assess all four factors: dose, interval, age, and sex, separately for each strain. You simply code the minimum, central, medium and the maximum dose or info or age with a series of numbers. Sex, of course, is a binary variable so it's either zero for females, one for males, whatever you want. You get 18 runs, you perform those in random order, nine males, nine females, 18 mice per strain— you don't have to replicate necessarily, because the center points here provide the variance estimates for estimating the statistical significance of the main effects. If you are concerned about biological variability differences between individual mice, then you could certainly replicate it. But as it stands, with six strains and 18 runs, you only need 108 mice rather than 1,600.

Slide 48: Data Visualization Identifies Most Important Effects
And then it's very easy to analyze; you just use regression and all those main factors that are two-way interactions. Plot it out in what's called a half-normal plot, and you can see instantly what are the most significant driving factors. In this experiment, there were no interactions we had to worry about. It was

just three factors which were [of] primary importance. So then you can design a second definitive experiment with only those important factors, greatly saving on time and iterating to a true solution.

Slide 49: Summary
So now I'll quickly wrap up.

Slide 50: Take Homes
So three major take-homes:

- That an appropriate statistically-based study design is the fundamental tool not only for translation, but also for the 3Rs.
- Statistics is a process, it's not a "thing" and it's certainly not just analysis. Design comes before inference, and the data quality comes before analysis.
- And there's three design basics that you have to consider for your own research, if not for anyone else's. Bias minimization, again, IACUCs don't check that, but the design skeleton with five elements and the numbers with the three screens are certainly things that we can assess.

Slide 51: For Both Reviewers and Researchers
And for both of us, reviewers and researchers, this will really require a culture change; we need to become more familiar, all of us, with best practice standards which had been highlighted in our reporting guidelines like ARRIVE 2.0. But you can't report what you haven't performed. So if people understand what the best practice standards are, they're more likely to incorporate in their own research.

We, as a whole- NIH and various institutions- really need to facilitate and promote best research practices by encouraging the training in statistical design of experiments. Most of them are out of reach, or they're just not feasible for most researchers. We need to evaluate protocols and publications that come to us for review in line with current best practice standards, not an imaginary standard that happened 35 years ago. And we need to commit ourselves to continuing education in how best scientific and welfare practices are continually evolving, because none of these are static.

Slide 52: Parting Thoughts
So about, well, golly, over 40 years ago, one of the world's leading applied statisticians, Doug Altman, said that poorly designed experiments and misuse of statistics is not only irresponsible and negligent, it's also unethical. And we can see that that's certainly the case for animal-based research. Animals suffer and they're killed, for what? If the studies are non-informative, it's kind of simply a waste. And on the other end of the translation pipeline, humans are harmed or killed if preclinical research results are misleading. So it's part of the duty of care of all of us to minimize harms and promote best practice. So we need to be evangelists for that.

Slide 53: Questions? Thank You for Your Attention
So that's all I had to say. And I guess what I'll do is stop sharing and hand it over to--

>>*Nicolette Petervary*: Oh, Dr. Reynolds, please continue because we've got some questions that we received before the webinar that are on the slides.

>>*Penny Reynolds*: Oh, this?

Slide 54: Question 1

> *Nicolette Petervary*: Yes. That's right. That's it. But first, I'd like to— I think some of the ones I see in the Q&A box, I'll be able to address with some of these questions that we received. But there is one that I'd like you to answer, and it came in around slide 38 of your presentation: When you speak of outcomes, are outcomes the dependent response variable or something else?...

> *Penny Reynolds*: Yes.

> *Nicolette Petervary*: …If something else, please explain.

> *Penny Reynolds*: No, it's the dependent response variable. And that's what's so confusing with the way statisticians talk and the way normal people talk, because we talk about "treatments" and "outcomes". Well, a "treatment" to, say, a clinician, means anything you do, whereas we're talking about specific input variables; and outcome, again, I'm talking about response variables. So, yes, I'm talking about response variables.

> *Nicolette Petervary*: Thank you. Yes, I would certainly appreciate statistics being more plain English language. We also had a terrific comment that compares scientific training to medieval apprenticeship.

> *Penny Reynolds*: [LAUGHTER]. That's so true.

> *Nicolette Petervary*: And, you know, it's no wonder that we're in the situation we're in because there's got to be a great deal of variability when people learn that way. So we certainly have our work cut out for us. But the first question I have for you that was submitted in advance was: What practical suggestions do you have for ways that IACUCs can positively impact the design of experiments in the initial stages, before the grant proposal, to include more robust statistical approaches without overstepping or micromanaging? And then what types of outreach or processes have you tried that have been effective?

> *Penny Reynolds*: Well, that's the thing. I mean, IACUC is like the Wild West, and everybody does it differently. And it really depends on the composition, doesn't it? And the leadership of the IACUC chair. What we've done is that we've provided resources for researchers, of course they complain they can never find them, and to be honest, it's really hard to navigate the websites, but that's one thing you can do is provide these resources. What we've done is provide links to the AALAS module on experimental design, which is very good. We've also created our own little handout on sample size justification, which is currently being revised. And so usually, individual reviewers are very good about directing researchers to those resources.

But I think if people [don't] focus more on just sort of the exciting, promising, you know, razzle-dazzle things that the research is promising and focus more on the design skeleton, like what are the elements? And just ask the questions of the reviewer to clarify it, I think we could make significant headway. Because a lot of proposals are approved, and the researcher really doesn't have- I mean, it's just going to be trial and error. And that's what really wastes more animals than anything else, in futile experiments that go nowhere because there's no real plan. So yeah, starting by asking questions, but we've got to be more proactive in providing the appropriate resources.

>>*Nicolette Petervary*: Thank you. If you can advance to the next slide.

Slide 55: Question 2
Can you provide some examples of when your IACUC has felt the need to reach out to the PI regarding study design and animal numbers (besides the 91 million) and when reviewing a protocol? How have you broached these conversations in a productive way? And what was the outcome?

>>*Penny Reynolds*: Well, the 91 million one was actually, [LAUGHS] I mean, they provided all the spreadsheets to prove it. So they were quite indignant when that was queried. So it was just by being patient, no, we weren't going to approve it, because it just wasn't feasible, and it wasn't justifiable, and it wasn't appropriate. So they did revise it down to a scope of work, which was more appropriate given the staff that they had on hand. Now, that hasn't stopped them in continuing to ask for outrageous numbers, but at least they're not in the realms of the hundreds of millions. They're now in the range of the tens of thousands. So it's just being really, really patient, saying, "No, we can't approve it, because you haven't shown that you can actually do the work." If you've only got one person, and you're asking for half a million mice, that's just not feasible.

So our previous IACUC chair also asked for people to design what they called a decision tree. So if this first set of experiments works out the way your research question predicted, then what do you intend to do? If they don't work, then what's your plan B? And so by just getting them to map out the scope of work that way, which is actually appropriate scientific practice because that's why we test hypotheses, that actually makes it a lot more manageable. But I think we've got to- we also have to get away from thinking that IACUC protocols are just a wish list of things you want to accomplish. It's supposed to be a blueprint for what you can realistically hope to achieve.

>>*Nicolette Petervary*: That's a great suggestion. And I like that decision tree process, just, you know, introducing PIs to that who you might be struggling with is a great idea. And then we have--

>>*Penny Reynolds*: Especially if it's exploratory.

>>*Nicolette Petervary*: No, no, go ahead. Yeah, exactly, especially if it's exploratory.

Slide 56: Question 3 (1)
All right. And this last question, I think this will somewhat answer the question from the anonymous attendee: Why can the IACUC not evaluate experimental design? And so this question for OLAW is: There's a lot of overlap between the IACUC's role in evaluation of the science and the scientific review groups. What balance should IACUCs strike when evaluating the statistics behind animal numbers requested in protocols, especially if the protocol is submitted to the IACUC at the just-in-time phase?

Slide 57: Question 3 (2)
And if you'll proceed to the next slide, I will refer you back to OLAW FAQ D.2. [*Correction: Should refer to OLAW FAQ D.12.*] So it's not so much that the IACUCs are not supposed to evaluate experimental design, but the IACUCs' mandate is to evaluate animal care and use in the context of appropriate welfare and they evaluate elements of the experiment through that lens. So IACUCs are not expected to conduct peer review, but they are expected to consider the US Government Principles, which Dr. Reynolds went through. And other PHS Policy review criteria refer to sound research design, the rationale for involving animals and scientifically valuable research. So all of those are right there on the OLAW website.

And because there is overlap between scientific review and IACUC review, it's a nuanced question. And I think Dr. Reynolds gave a lot of really good examples on how to approach this and eventually get increased buy-in from investigators as you start becoming more curious about these experimental design elements.

Slide 58: Question 3 (3)
So once again, this slide is just showing how if you're not doing this, and a study can't meet these basic criteria, it's an essentially a waste of the animals, which Dr. Reynolds also pointed out. So although the primary focus of the scientific review group is merit, scientific merit, and the primary focus of the IACUC is animal welfare, they do overlap. And so scientific review groups are perfectly free to raise questions about animal welfare, and the same is true for IACUCs. They can question the scientific rationale or necessity for a procedure.

Slide 59: Next Webinar: Summer 2023 Topic TBD
And with that, I'm so sorry that we are over. We have some more questions about observational studies that are very nuanced. We will forward those to Dr. Reynolds and send them off with the final transcript. And you will also get that list of resources that Dr. Reynolds mentioned so that you can present those to your IACUC and to your PIs. But please stay tuned for our next webinar in the summer with the topic to be determined, and thank you.

###
ADDITIONAL QUESTIONS

Question: Most of the examples are for experimental studies mentioned involve manipulation or drugs. What about purely observational studies? How do statistics come into play there?

>*Penny Reynolds*: Design of observational studies began to evolve about the same time as the design of modern clinical randomized controlled trials (RCTs) in the 1930s and 1940s. One of the chief architects was Sir Austin Bradford Hill. The Bradford Hill criteria list the minimum requirements for establishing a causal relationship between observed factors and a disease, so they are common in epidemiological studies. The biggest problem with observational studies is bias from the effects of confounders and selection bias which are minimized in RCTs by randomization and blinding. Expectations for the design and reporting of observational studies (case series, case control, cohort) can be found in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement. However, observational study designs are not appropriate for laboratory rodent studies which are usually experimental, not epidemiological. They are really only appropriate for large veterinary studies or veterinary clinical chart reviews where the goal is to establish a link between certain risk factors and a disease.