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.

The ARRIVE 2.0 Essential 10: Guidance for NIH-sponsored Research

Speakers:

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

Broadcast Date: June 15, 2023

Slide 1: The ARRIVE 2.0 Essential 10: Guidance for NIH-sponsored Research

>>Nicolette Petervary: Good afternoon. I'm Dr. Nicolette Petervary, part of the NIH Office of Laboratory Animal Welfare. Today is Thursday, June 15th, 2023, and I'm pleased to welcome you and our speaker to our webinar today, titled "The ARRIVE 2.0 Essential 10: Guidance for NIH-sponsored Research." There are just a few housekeeping details before we get started.

- If you have questions throughout the webinar, please enter them in the Q&A box. The Q&A box does allow questions to be submitted anonymously, and the chat will also be enabled for this webinar. Dr. Reynolds will be taking questions at the end of the webinar, and if the question is a little more nuanced or context-specific, we'll forward the question to her after the webinar, and then we'll pin the question and answer to the end of the transcript. We'll monitor the chat as 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, but they do need to be processed for 508 compliance compatibility before posting, and this can take a few weeks, so please do bear with us.

Okay, and now let's get started with an introduction for Dr. Reynolds. 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 coauthor of the ARRIVE 2.0 Revised Guidelines and ARRIVE 2.0 Explanation and Elaboration (E&E) document. She was awarded the 2021 UK Animals in Science Education Trust (ASET) 3Rs Prize, and together with collaborators Maggie Hull 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 is a published advocate for the improvement of animal-based research through application of statistically-based experimental design principles and quality improvement strategies. But before we begin with the presentation, we do have a few poll questions for attendees to see how much you know about our topic today. Can we please have the poll posted? So go ahead and answer the poll questions, and once we have enough answers in, we will see the results.

Okay, so we have a mixed bag here. We have a lot of people who have heard of the ARRIVE 2.0 or other reporting guidelines but maybe not so many people who have experience in writing up your research using them. That's really good to know. And then about the journals... There are a lot of people who don't know if the journals you submit most frequently to endorse the ARRIVE 2.0 Reporting Guidelines, so that will be your homework after today's webinar. Okay. Terrific. Attendees, once you've looked at the final results, please feel free to close the poll window, and we can continue. And just a quick update before we begin, Dr. Reynolds does not currently have camera capabilities, so you won't see her during the presentation. Welcome, Dr. Reynolds. The floor is yours.

>>Penny Reynolds: Thank you, Nicolette. Thank you for the very kind introduction. So, best practices in scientific research, just like everything else, are constantly evolving as more and better evidence becomes available, so that means that the norms and standards for what constitutes quality research are also changing, and what might have been appropriate a few years ago probably won't meet expectations today. The problem for most of us is that it's really hard to keep up with all of these changes.

Slide 2: Outline: Arrive 2.0 Essential 10

So today I'm going to talk about why the NIH is now recommending the ARRIVE Essential 10. I'll explain why it's now a feature and why it's so important to them and why it should be important to you as well. I will then do a deep dive into each one of the 10 items that are featured in these guidelines. I will explain what they are, what they mean, and why they're important, and then finally I'll conclude by showing a couple of tips and tricks on how to make these work for you in writing up and design of your research.

Slide 3: I. Why the Essential 10 (and why ARRIVE 2.0)? *(Skipped)*

Slide 4: NEW!! NOT-OD-23-057 10 Feb 2023

>>Nicolette Petervary: Dr. Reynolds, could you speak up a little bit, because some folks are saying that the volume isn't quite loud enough?

>>Penny Reynolds: Oh, I'm sorry. Just a moment.

>>Nicolette Petervary: No worries.

>>Penny Reynolds: Let's see. Trying to find the volume here ... we're going to have to stop screen sharing for a moment… I can't see the controls. Oh, I see. That was the problem there. Okay. Can everyone see the screen now?

>>Nicolette Petervary: Yes, we can.

>>Penny Reynolds: Okay, and can you hear me?

>>Nicolette Petervary: I can hear you. If any of the attendees ... Yes, we have thumbs-up.

>>Penny Reynolds: Yes? Oh, good, yes, I see that. Okay, so a few months ago back in February, NIH produced a Notice, which is encouraging the use of the ARRIVE Essential 10 Checklist in all publications featuring animal-based research involving vertebrates and cephalopods. Now this is a recommendation; it's not mandatory. However, when NIH suggests something, it's probably a good idea to pay attention, especially since the funding climate doesn't show any signs of improving any time soon.

Slide 5: What are the ARRIVE guidelines?

So what are the ARRIVE guidelines? Well, ARRIVE is an acronym standing for "**A**nimal **R**esearch: **R**eporting of **I**n **V**ivo **E**xperiments", and what these guidelines are is what has been agreed on by international consensus as best practice for reporting animal-based research. The whole goal is to improve it so it's more useful and has a longer shelf life. So the entire theme is this increased emphasis on rigorous, well-described methodology and shifting the emphasis away from sexy, splashy results which may have no substance to them.

Slide 6: ACD-NIH recommendations 11 June 2021

Now part of the reason for the NIH recommendation coming as it did was in response to the Advisory Committee to the Director of NIH (ACD), who made a report back in June 2021. They submitted recommendations in five domains for enhancing rigor, transparency, and translatability of animal research. What featured quite largely in there was that the ARRIVE 2.0 guidelines should be followed at all stages of the entire research process, not just for the writing stage, which is what they are actually recommended for; but, they suggested that if these elements are strengthened across the entire research life cycle, the resulting knowledge base and quality of research as a whole will be better and better able to inform future research.

Slide 7: Main theme

So the theme is that good science must be done for sure, but it must not only be *done,* it must be *seen* to be done. So high-quality science is both valid and reliable, and the only way to ensure that is to have sound methodology. Methodological elements are good experimental design; minimization of bias; appropriate, well-conducted statistical methods; and reporting which is clear, thorough, and honest.

Slide 8: Why is it necessary?

So why is it necessary? Well, unfortunately, it's because the vast majority of papers published, even recently, do not report even the most basic items necessary to assess the quality of research. [P]robably the best example of this was a massive text-mining study by Menke. et al., that came out several years ago where they did text-mining on over a million and a half papers. Of those, over 51,000 specifically reported details for animal-based studies in 1 year, and what they found was actually rather appalling because there were major reporting gaps for all items which are most relevant to study quality. Randomization, blinding, and even sample size and sample-size justification, as you can see, were very poorly reported (mostly under 30 percent) and even things incredibly basic like what the test animal actually was in a way that identifiability could be verified and the sex of animals were far below 100 percent reporting compliance, so obviously work needs to be done.

Slide 9: Reproducibility issues in preclinical research

Why these are reproducibility issues is for three reasons. First of all, poor reporting is a proxy for poor experimental design and conduct. Now it could be argued and has been argued: "Well, just because something isn't written down doesn't mean that it wasn't done." However, there have been several studies, of which the Macleod, et al., study [that] came out in 2015 was probably the most rigorous, [that] would suggest that, no— there is truth in the old adage that if it's not written down, it didn't happen. [S]tudies that did not report doing a proper experimental design or conducting the experiment properly didn't do it. [The] second thing that it indicates and why it's a threat to reproducibility is researcher complacency; that methodological illiteracy in terms of the method standards that we have previously identified seems to be the norm rather than the exception. Finally, the checks and balances which are supposed to ensure that poor-quality papers don't get published are clearly broken. We found that [the] peer review and editorial processes are not stringent, they're not sufficient, and many times, even if they're endorsed, they're certainly not enforced.

Slide 10: Why should we care?

So why should we care? Well, we should care because poor-quality research is wasted research. It doesn't matter how many publications are out there. If the question is irrelevant, incoherent, if the methods are inappropriate and inadequate, if the data are inaccessible and unreliable, then results are unusable. There are two implications for this.

Slide 11: 1. Wasted research is a major loss of investment

Poor-quality research is a major loss of money, and large funding agencies, such as the NIH and the Department of Defense, are understandably irritated by pouring hundreds of millions if not billions of dollars into research and getting no return on their dollar. So this old paper from Friedman actually broke down the economics of lack of this reproducibility in animal-based research. The numbers are a bit dated, but the principle is the same: that more than half of studies are essentially not reproducible and therefore a waste, and at least two-thirds of this has to do with study design, data analysis, and reporting. More recent analyses suggest that actually some of these proportions have shifted, that the reporting of biological reagents and reference materials has improved considerably. However, the reporting of methodology and lab protocols has gotten substantially worse.

Slide 12: Wasted Research is an ethical issue

The second major effect of wasted research is ethical, and that relates to the collateral costs. First of all, hundreds of millions if not billions of animals are wasted in research which is noninformative, and thousands of human beings on the other end of the research pipeline are injured and even killed because of overhyped research which is not properly conducted. These facts have been weaponized by some of the groups that are trying to stop animal use in research, and it has to be admitted that they do have a point. Just a few of the examples of preclinical research which has gone on to have extremely deleterious effects for the practice of human medicine include nearly all ALS drugs. Most of the positive drugs have been shown in animal trials [to] actually be an artifact of poor study design, mostly confounded with litter effects. [The] famous study of sildenafil used to treat fetal growth restriction directly led to the death of 10 neonates, or 11 neonates rather. A TB vaccine shown to be promising in nonhuman primates turned out to harm children, and of course the recent dust-up about amyloid-beta as [a] causative agent for Alzheimer's had been shown to be actually fraudulent. [The] jury is still out on whether or not all of that line of research has been wasted. We'll have to see.

Slide 13: Enter ARRIVE

So back in 2009, concerned about the lack of reporting standards in [the] preclinical research area, the NC3Rs of the UK with funding from NIH/OLAW conducted a systematic survey and review of all government-funded, published research using animals. The results were so dismal that they came out with the first standardized reporting guidelines, the ARRIVE guidelines, in 2010.

Slide 14: Enter Arrive *(continued)*

However, even though well over 1,000 journals have now officially endorsed these guidelines, they still don't seem to have had sufficient traction, and reporting standards are still universally quite poor. So the Second International Working Group was convened in 2017 with the goal of revising, updating, and streamlining these guidelines to make them more useful and more user-friendly so that researchers would be more inclined to use them. Now this was a multi-year, international, collaborative effort, and it's characterized by extremely rigorous and strict methodology.

Slide 15: I2C2: International, iterative, collaborative, consensus-driven

So just so you don't go away with the idea that it's just some random thoughts on how you ought to write a paper by some random collection of people, this was incredibly, incredibly rigorous. Development of the guidelines alone took over 2 years, via a working group comprised of 28 members from seven countries representing a diversity of stakeholders and expertise. After the guidelines had been worked out as a first pass, they were sent out for an external consensus process involving Delphi procedures, 73 participants from 19 countries, again representing diverse stakeholders, and so it went through an iterative process where each of these participants would rate it, review it, comment, [and] send it back to the working group who proceeded to revise it [and] send [it] out again until there was consensus. And in the final phase, they were road-tested by publicly posting the guidelines on public forums, such as arXiv (arXiv.org). Manuscripts were reviewed by people who were willing to collaborate, and again, it went back to the working group for further revision.

Slide 16: Product

So the end result was two big products. First was the checklist itself. It's two-tiered. The Essential 10 is [a] list of the minimum information required for assessing rigor and reproducibility. These are the items which convey internal validity of a study. Now the second tier is called the Recommended 11. This is a bit of an unfortunate name because people assume that “recommended” means “optional.” They're not. It's just [that] this information is required for assessing the context of the specific study and the generalizability of the study. It does not contribute directly to the internal validity of the study, which is what I'm going to focus on today. Along with the checklist is an Explanation and Elaboration (E&E) document, so this is an extremely hefty 80-page user's manual which explains every item, all 21 items in depth, what they mean, why they're useful, what they're for, and then it provides practical working examples from the published literature on how to report those items.

Slide 17: II What are the ‘Essential 10’? (*Section title slide)*

Slide 18: ‘Essential Ten’ reproducibility items

So what are the Essential 10? Well, this is what the checklist looks like. Again, it's an expert consensus on the priority information needed to communicate the internal validity of the study. These are the best-practice items universally agreed on for reliability, validity and reproducibility, and the items are arranged not in rank order of importance but in workflow order, so this reflects the natural flow of an experimental process.

Slide 19: 1. Study Design

[The] first one is study design. What the guidelines say is, for each experiment, at a minimum, describe the groups [and] the controls. If you don't have a control group, explain why and explain what your experimental unit is.

Slide 20: What it means

So what it means [is] what are you comparing? And this is a formal statistical structuring of your predictor variables. What is being compared? The experimental unit is your unit of analysis. It *can* be the individual animal, but it *might not* be, and I will get on to this in a second.

Slide 21: Key idea: Statistically-based designs control variation

The key idea is that a statistical study design is a method of controlling variation. It is the statistical model for your research. Some examples are randomized complete block design, which controls external sources of variation, repeated measures design and paired designs, which control variation both between and within animals,

Slide 22: Factorial design are the best for multiple inputs

and [the] factorial design, which is useful for when you have multiple inputs. In this example, you want to compare the effects of drug and sex. A factorial design allows you to simultaneously evaluate multiple inputs. It can identify interactions, synergisms, [and] antagonisms. It shares the sample size across multiple input variables, so as a result, it saves time, money, and it's extremely animal-sparing.

Slide 23: What is an experimental unit (EU)?

The other aspect of study design is identification of the experimental unit, and this is defined as the smallest division of experimental material such that any two units receive different treatments in the actual experiment. Now that's a bit of a mouthful. So what it means is, what is the smallest amount which can receive a complete intervention? For example, in this case, the experimental unit is not the mouse. It's the mouse flank. Each one can receive a separate treatment. So the mouse is actually a type of block effect. In the randomized complete block design, blocking is on [the] cage [level], so each mouse can receive a separate intervention, so the experimental unit in this case is the mouse, and you've got four mice in each group. However, this [*speaker demonstrates to a different example]* is an example of pseudo-replication, which is actually unfortunately very common. Suppose you wanted to administer a drug or a control drug or placebo substance in the drinking water, so that means all animals in the same cage would get the same treatment. In this case, the cage is the experimental unit, not individual mice, so instead of having a sample size of eight per group, you really only have a sample size of two per group, so this is to be discouraged.

Slide 24: Why it matters

Why does it matter? It matters because study design is the backbone of good research. It details how your data are collected [and] what data are collected. It determines the statistical analyses for sure and also how the results are to be interpreted. As a result, it increases power, reduces noise, and increases the information you can get out of the study. It also reduces animal numbers. The problem here is that a design cannot be imposed after data are collected, so if you've collected your data [and] you haven't thought about what the design is, you see this on the checklist ("oh, forgot to put it in")— it's too late. You didn't have a design.

Slide 25: Why it matters (*continued)*

So it matters because this is the single biggest obstacle to improving the quality of research overall; that people don't understand what the study design is. Often I see it conflated with a method of analysis such as a t-test or an analysis of variance. The clue is in the name. Those are methods of *analysis* of the data, which are predicated on the assumption you have an underlying design to begin with.

The other shocking thing is that these sorts of statistically-based designs, the ones that I've mentioned before like randomized block designs and factorial designs, have been around for almost 100 years, but most studies in the literature don't have anything reported. They certainly weren't done, and this is because [in] introductory statistical service courses, most of them are simply not taught.

Slide 26: Why it matters (*continued)*

And it really matters because if a study is not designed, it's grossly inefficient and highly wasteful. What we see is that a lot of studies are more or less organized by what the investigator calls groups or cohorts. This is not only a mistaken and completely misunderstood use of these concepts, they're also incorrect. So what this means is that the study is not designed; it means your statistical methods downstream are going to be misused, and therefore you will miss any true effects if there [are] any to be detected. And logistically, it wastes a tremendous amount of animals.

Slide 27: 2. Sample size

Number two is sample size. What the guidelines request is that you specify the exact number of experimental units, the total number of experimental units in each experiment, and the total number of animals used. Remember "experimental unit" is not necessarily the same as the animal. And then it asks you to justify how you arrived at the sample size you did.

Slide 28: What it means

So sample size is the number of experimental units per group. It's important because you need to track the numbers through the study. They need to add up, and you need to see if there's any losses. But it's also part of numbers justification. Are the numbers of animals used in the study or the experimental units used in the study— are they adequate to answer the research question in the first place? So are the numbers feasible? Are they verifiable, and are they ethical?

Slide 29: Why it matters

And it matters because sample size is, [of] course, the number one reproducibility item. It's the most basic statistic you can ever hope to find, so that's why it's number one for reproducibility. If you don't know what the sample size is, you cannot possibly assess any of the results or even the appropriateness of any statistical analyses methods that we used.

But it's also the number one defining ethical principle for any use of animals. Remember the 3Rs: minimal harm for maximum scientific value. [In] most of the book published by Russell and Burch on the 3Rs, this is what they really talk about most of the time, [the] thinking being that if the sample size is too large, it wastes those extra animals, but if a sample size is too small and the study is underpowered, it wastes *all* of them. Unfortunately, [in] the majority of published studies and more than 95 percent (it's probably closer to 98 percent), they neither justify the numbers that they used (n), or even report the numbers in such a way you can actually figure out how many animals were used in the first place.

Slide 30: These are not justifications

Often when I'm reviewing the literature and animal protocols, I'll see several kinds of justifications which are not justifications at all, and the first one of these, the most common, is what I call "magic number syndrome:" they pick [a] sample size that worked in previous studies. "Based on our previous publications, "This number is sufficient to obtain statistically significant results," or, "What everyone else does." These ... No— that's not correct. Secondly, some people just make it all up because, oh, well, they'll just pick some number out of the air because "you never know what might happen," or "we don't know how many animals we will require because it's an exploratory study." My personal favorite of all time is someone who requested 91,386,777 mice for a 3-year project without any sort of rationalization at all except that they had spreadsheets. Number three is what I call the "passive-aggressive excuse." This is the quiet part out loud, which nobody really verbalizes. However, Fitzpatrick, et al., did a survey of over 1,000 IACUCs and several thousand people associated with them across the United States, and one of the most common responses, especially from senior investigators, is they don't see the need for power calculations for sample-size justification at all. They concluded that it's just a necessary evil to satisfy both the IACUC and reviewers of journal articles, so that's completely missing the point of what the statistical sample-size justification is supposed to do.

Slide 31: 3. Inclusion and Exclusion Criteria

Number three, along the lines of sample size, are the inclusion/exclusion criteria. What were the reasons that you used to disqualify or include animals in their data? Did you specify these criteria up front? If not, say so. For each experimental group, explain why animals or data points were included, and for each analysis, report the exact sample size in each group.

Slide 32: What it means

It means that you need consistent *a priori* criteria for including or disqualifying both the animals and their data. Now, inclusion criteria are just any key features of the target population you use to answer the research question and should be addressed by the animals or experimental units in your sample. Exclusion criteria are those features which will interfere with the study goals, which you also decide on ahead of time. So, for example, the animal was sick or an instrumentation failed, so you wouldn't be able to collect the data that you needed to measure the outcomes of the experiment. Those are reasonable exclusion criteria as long as they're defined ahead of time.

Slide 33: Why it matters

It matters because not only do you need to define the subject pool for obtaining the best positive data so that your sample is truly representative of the defined study population, it also minimizes the bias that results from arbitrary decisions as to whether or not to include or exclude data. [On] more than one occasion, I've gone into a lab where an experiment was being conducted and overheard the project leader saying, "Oh, well, this animal doesn't seem to be doing so well. We're not going to include it," or "We'll make it a control." That's cherry-picking data and results. It's dishonest. It is borderline unethical, and it is beginning to skirt research misconduct, which is kind of harsh, but at the very best, all you're doing is producing a highly biased and nonrepresentative set of results.

Slide 34: 4. Randomisation

Number four is randomization. In the checklist it says, "Did you use it?" If it was, describe how you did it. What's the method?" And randomization can also be used to minimize the effects of potential confounders, like the location of the cage or the order in which the animals were processed. If you didn't do it, say why.

Slide 35: What it means

So what is randomization? This is probably the statistical term that is most confused and most conflated with layperson's definitions of the term. Random does not mean haphazard, ad hoc, [or] unplanned. It doesn't mean you just sort of arbitrarily pick a cage and arbitrarily yank a mouse out of the cage by the tail and say, "That's random." No, it's a formal technical process based on a method of probability and assessment of assigning your interventions to the experimental units. It's also a probabilistic-based method of assigning order to the process. The best way of doing it is by computer algorithm because it's unbiased, and also you can use it to provide an audit trail for your methods, so you need to explain what the method is and the particular algorithm that you used. Many programs, such as R and SAS, will provide randomization schedules for you.

Slide 36: Why it matters

It matters because randomization is the number one item for validity. If sample size is the number one element for reproducibility, this is the one for validity. Randomization minimizes systematic bias, and that's what you'll see most often cited in the literature. What people don't understand is that most of the basic statistical hypothesis tests are predicated on the fundamental assumption that randomization was performed. If it's not performed, your statistical hypothesis tests are actually invalid. You really don't know what it is that your results are being compared to, and there are no good reasons not to randomize. Most of the time when I'm chatting to people after the fact and they finally let on they haven't randomized, it's because they didn't know. Well, that's really not a good enough reason. Most of the— a*ll* of the basic statistical hypothesis tests that are in textbooks explicitly state that there has to be random allocation underlying the test to make it valid.

Slide 37: 5. Blinding

Sorry, I'm losing my voice a bit.

Number five is blinding. This is a question of logistics. Who was aware of group allocation when: during the allocation of interventions to the experimental units, during the conduct of the experiment, during assessment of the results, and during the data analysis?

Slide 38: What it means

Now "blinding" is kind of an old-fashioned term. It's a bit biased and ableist and sort of discriminatory, so I actually prefer the more descriptive term "allocation concealment". What it is, is that you're hiding from some or all of the personnel involved in the experiment which treatment was received by which subject or experimental unit. This is logistics. You can't do allocation concealment after the data are collected, although it can be imposed at any or all stages, preferably all four: the personnel assigning treatments to the experimental units, during the performance of the experiment, certainly during the assessment phase [for] the people who are evaluating the outcomes, and during the interpretation phase [by] whoever your data analyst is.

Slide 39: Why it matters

It matters because this minimizes the cognitive biases that are always present. You may not even know that you have cognitive biases, but it's especially critical for outcomes where any sort of subjective evaluation is required, like histology or assessing behavior or clinical progress of an animal. The tendency is to be biased in favor of whatever intervention that you prefer, say a test over control. You really want your test to work, so you'd be more inclined to judge results favorably if you knew which treatment it had already received.

Slide 40: 6. Outcome Measures

Number six is outcome measures. It clearly defines what it was that you were measuring, and if you have a hypothesis test study, which most people do, you need to specify a primary outcome measure.

Slide 41: What it means

What are outcome measures? Well, it's the dependent or the response variable. This is anything that's specific, that's measurable, that assesses the effects of the intervention. The primary outcome is the one that you have prioritized to be most important relative to the central hypothesis. This is important because this is the one in which studies are powered off of. Other variables, which there's usually a lot of other things that people measure, those are nice to know but they have a lower priority.

Slide 42: Why it matters

It matters because the study is not only *powered* off the primary outcome, it's also *interpreted* off the primary outcome. Now many researchers want to measure lots and lots of things in order to maximize as much information as possible [as] can be obtained from each animal, and that's a good thing to do. However, if the outcomes are not prioritized, first of all, the study tends to be overly large and unfocused, so it's really difficult to interpret. The biggest source of bias here is the temptation to chase significance or cherry-pick the results. So [if] one of your results, a key result, was not statistically significant … usually the temptation is to go haring off after something else you measured, which did show a small p-value. Well, this leads to a lot of false positives and certainly noninformative research.

Slide 43: 7. Statistical Methods

Number seven: statistical methods. What the guidelines are requesting is to provide details of what you did and the software that you used, and when you do use statistical methods, you have to assess whether or not the assumptions underlying each method were actually met and what you did if the assumptions weren't met.

Slide 44: What it means

What it means: if you do a survey of any sort of research literature for very long, you'll see that the statistical methods sections of which all papers usually have something tend to be copied almost exactly from one paper to another. But statistical methods have to be bespoke. They should not be boilerplate. What did you do for *this* study and was it appropriate for *this* study? Your study design and your statistical methods have to be aligned. Methods will be dictated by the hypotheses you wish to test. The methods follow from the study design and the variables that were measured, and it should go without saying (although people don't seem to know this) that statistical methods should also be best practice. A lot of the things that I see are [haven't] really changed much for about 80 to 100 years. There are numerous statistical methods out there now which are actually much better, easier to use, and will enable you to get much more information out of your data.

Slide 45: Analyses: Definitely not bespoke

However, with a couple of colleagues, I've done several surveys in different research domains in multiple animals for close to about 1,000 papers. That's what these data represent, and they find that analyses are definitely not bespoke. Most people tend to prefer t-tests because they do lots of two-group comparisons. Rat and swine people tend to use mostly one-way ANOVA. Were the methods appropriate? Virtually nobody's were. Nobody specified a design. Hardly anyone ever reported sample size. Between 75 and 90 percent [of the] studies surveyed reported time dependencies in the data (so they're actually looking at repeated measures over the same animal) but almost nobody accounted for them. Almost 90 percent were trying to measure multiple factors simultaneously, but nobody accounted for those, and almost everyone (well, everybody— 100 percent) reported P-values that were both orphaned (that means they had no statistical context) and they were inexact. They were “less 0.05”, “less than 0.01.”

Slide 46: Why it matters

It matters because your statistical methods have to be valid, and these are essential for interpreting what your results mean. However, what's been found by numerous other statisticians is that most of these errors that are made in the published literature are serious enough to completely invalidate results, and moreover the errors are not because of advanced mathematics because “everything is too difficult to understand.” They're most basic, most fundamental application of methods that are taught.

Slide 47: 8. Experimental Animals

Number eight: experimental animals. Details of the animals used, which are appropriate for the species, species, strain, substrain, sex, age, weight if relevant, and then provide information on their provenance.

Slide 48: What it means

So what this means is that you have to describe: Who is in the study? What are their characteristics? It is two domains here, and it’s what veterinarians call the signalment. You need details about the animals used, [the] species, age, strain and weight, that kind of thing, but also you need verifiable identification as to the source, their strain numbers, stock numbers, [and] the vendor. It matters because this is really important information for assessing study validity. Is the sample appropriate? Is it representative? And can the results be extended? So when you see signalment information, that's analogous to the sort of demographic data we have to report for human clinical trials. Again, is the sample representative? The source, and especially for mice, is really, really essential because [animals] from different vendors or even different facilities from the same vendor, or different sublines can show totally different responses, and one of the problems we have in trying to chase down anomalies in the literature is finding out where the mice came from. This frightening example I've shown to the right here shows one study done with some knockout Black 6 strains. They had this knockout strain that they wanted to compare to two wild-type lines. One of them was just the regular Jackson Black 6 line, but the other was an NIH subline called NJ, and they got completely different results from using the different wild-type controls. So reporting the source and reporting all aspects of the signalment, and especially the strain and stock lines, is extremely important information to include.

Slide 50: 9 Experimental procedures

Number nine is experimental procedures. What did you do? When did you do it? How did you do it and why? Pretty straightforward.

Slide 51: What it means

Now most people are afraid of getting too wordy, although they're usually quite good about explaining the experiment itself, so flow diagrams like the one shown on the left are actually quite useful for reducing the verbiage. But what is also included here is not just the experiment itself; we also need information on what happened before the experiment [and] after the experiment and termination procedures (especially euthanasia).

Slide 52: Why it matters

It matters because every manipulation you do can affect the results. Most people are pretty good about reporting the direct technical aspects, [such as] the molecular, the lab work and so on, but the information about what was done to the animals tends to be very poorly reported. That includes things like husbandry and handling, habituation training that you did to minimize stress, the disease and injury model is sometimes very vague, surgery, monitoring, sampling; and most disturbingly, drugs, analgesia, anesthesia and welfare care, and even euthanasia are very, very poorly reported indeed. Usually less than 10 percent have enough detail so we can figure out what was going on.

Slide 53: 10. Results

And finally, number 10, for each experiment, including if it was independently replicated, report your summary statistics, measure of variability, and the effect size of the confidence interval. Now this is the one section which is explicitly statistical.

Slide 54: What it means

So what it means is that, first of all, you need to quantify the details and characteristics of [what] was studied— [the] summary data for your sample: so details of the signalment, any baseline, preintervention, clinical, lab characteristics, mean, standard deviations, a sample size and a measure of variation such as a standard deviation or a continuous data, interquartile range if you expressed it as a median. You do not report standard error of the mean— that is a population sample; it's not appropriate for describing a sample. Sorry. Standard error of the mean is for a population. It's not for a sample.

Slide 55: What it means *(continued)*

You also need to describe the results in quantitative terms, [a] summary of the major results for each group: the difference in size that's applicable to the results of your hypothesis tests and some population-based measure of precision. So that is your sample size per group, the point estimate (which is a mean or mean difference), and a measure of variation (which in this case is confidence intervals).

Slide 56: Why it matters

It matters because it's the results of your hypothesis tests, which are not only used to interpret your data but are also used to make inferences about the larger population from which you've been studying a sample, which hopefully is representative. So descriptive statistics like your mean and your standard deviation or counts and percentages will summarize the properties of that sample. What confidence intervals do is provide interpretable information about the population. They describe if you’re comparing say, two or three different groups, they'll describe the size of the difference, [and] the direction of the difference. Was it bigger or smaller, and how much uncertainty [was] there associated with the observed effect? The uncertainty is measured by confidence intervals.

It's important to note here that this section does not state anything about p-values. This is because p-values have *no* clinical or biological meaning. They are an artifact of the sample itself, its sample size and the amount of variation in the sample, and when they're reported on their own with no statistical context, you can't use them *at all* in any way to summarize how useful, how actionable, or even how important the results are. P-values don't tell you that.

Slide 57: III. Making the Essential 10 Work for You *(Section title slide)*

Slide 58: Why should ARRIVE be used?

So finally, making these Essential 10 items work for you. Well, as the ACD report suggested, ARRIVE can be used at any stage in the research cycle, not just in writing up your results. It can be used to design experiments, to identify and record information that you might otherwise have missed, [and] to report all the information in your manuscript, but you can prioritize and organize lots of complex information, and as a review, make sure that all the information has been included not only in your manuscript but also in other manuscripts you may be given as part of your external peer-review duties.

Slide 59: 1. During planning and protocol development

So during planning, you can build quality into your study right upfront before data ... before the study even begins because it takes the guesswork out of determining what you need to include for [a] high-quality study. You can't report what wasn't done, obviously, but if it's left out and there's a lot left out, that's not a good look.

Slide 60: 2. During manuscript writing

During the writing of the manuscript itself, it takes the guesswork, as I said previously, out of organizing all this information you might have generated and prioritizing the most important ones. So by identifying the items that are most associated with reliability, validity, and reproducibility, [it] makes your papers and grants much easier to write and much easier to review. It should be noted that if you've planned these items into your study from the get-go and you report them accordingly, the study will be more likely to be flagged up as high quality and is more likely to be funded and published.

Slide 61: 3. After publication

Number three, after publication, it gives it a much longer shelf life. Hopefully the data shouldn't end with just succeeding and getting a paper published and then having nobody use it ever again. What you want are data that actually can contribute to databases and contribute to systematic reviews because it's high quality, and it's reported in enough detail so that other people can use it, and therefore, this is what it is that reliably will inform the progress of future research.

Slide 62: FAQ and common misunderstandings

Now I want to quickly address the remaining few minutes some of the frequently asked questions and concerns. It should be clear that the ARRIVE guidelines simply tell you *what* to report (what you *did* do) and justify what you *didn't* do. Most of the problems occur because researchers get a bit confused about conducting research and reporting research, and they conflate the two.

Slide 53: FAQ: Won't these guidelines stifle creativity?

One of the questions I get most asked ... asked most often is (actually, it's usually stated) is that people are afraid it's going to stifle their creativity. Well, no, it's not. This is actually a bit silly because guidelines don't prescribe any research topic. You can do whatever you want. What the guidelines do is just help you report your methods and results clearly so that other people can use them, and they can be clearly understood and evaluated. If the information is missing, the article is going to be useless.

Slide 64: FAQ: "What if I *don't* do these items?"

What if you don't do these items? Well, the reporting can still be complete even if you didn't do it. If you didn't do it, say so, and in fact, some of them might not be possible. For example, [if] you're doing experiments on evaluating a new procedure or a device, well obviously you cannot blind the allocation there. But if key reproducibility items are not performed, you need to report it honestly [and] justify that it's scientifically warranted, like "it's not possible to conceal the allocation for a procedure or device." But list it as a study limitation, and for goodness' sake, don't lie about it. If you didn't do it because you didn't do it, you have to say you didn't do it, but don't try to make something up.

Slide 65: FAQ: "What if I just say I did all that?"

This one is the quiet part spoken out loud, which a few researchers have been unwise enough to say to me: "but what if I just say that I did it anyways?" Well, this is certainly questionable research practice, and research misconduct is a continuum, and if you persist in engaging in bad practices, well, that *will* lead to misconduct. At the very, very best, what this is communicating is that it's too much bother to find out and incorporate best practices, so [it] could be judged as being irresponsible and even negligent. However, if you deliberately misrepresent and distort what you've done to make it look better than it is, that's scientific fraud, and this is a very serious thing and you're a bad person. So it really is in everyone's best interest, including yours as a researcher, to get up to speed with what the requirements are because even if you're not telling a lie, being ignorant and incompetent is really not a good look.

Slide 66: Examples of box-ticking

So some examples of people who tried to do this to say, "I did it," without really doing it comes up in these various red-flag items which show up in published research. For example, "Experiments were performed according to the National Health guidelines and ARRIVE guidelines," "The animal experimental protocol was in accord with the ARRIVE guidelines," "Experiments were conducted in compliance with ARRIVE guidelines." No— ARRIVE is for disclosure. Remember, it doesn't tell you what to do. It's not study-specific conduct. It doesn't dictate experimental protocols. It doesn't mandate experimental protocols.

Slide 67: Red flag claim 2

This is the second type of red flag claim I've seen, "We followed ARRIVE guidelines for the care of animals," "All efforts were made to minimize the number of animals used and the suffering of animals in accordance with the ARRIVE guidelines," "This protocol was performed in accordance with the Welfare Act, the Guide, and the ARRIVE guidelines." No— ARRIVE is not a statement of compliance with ethical care and use standards. In fact, I went back to each of these papers, and not a single one of them actually had a verifiable ethical animal care and use protocol number assigned with it, so they couldn't even be verified.

Slide 68: An informal test

So I did another informal test of one journal’s articles in the last 7 or 8 months where they explicitly claimed that they had complied with ARRIVE guidelines. Six out of nine of them said it was procedural. Two out of nine of them said it was ethical oversight. One correctly said that their article was reported in accordance with the guidelines, which was correct, and two of them even provided checklists. However, when I looked through all of them, not a single one had identified a design, [and] even only two even identified the groups that were studied. Only three clearly identified a sample size that could be verifiable. Nobody justified it. Five claimed that randomization was performed, but none of them provided a method, and astoundingly, two of the checklists marked that item as nonapplicable except they used hypothesis tests, so that was clearly incorrect. Randomization is applicable. They probably just didn't do it and marked it as not (N/A). Outcomes: nobody identified any at all [of the] things they measured. That had to be figured out from looking at figures and tables. No one identified a primary outcome, and again, in [the] checklist, when they said, "Was the primary outcome identified?" they marked that as not applicable. Nine used statistical methods, none of which were appropriate. Everybody reported orphan inexact p-values, and all of them reported a positive study. That means they found statistically significant difference in favor of the test intervention over a control.

Slide 69: Concluding thoughts

So that was quite disturbing and depressing.

So in conclusion, what this is all telling us is that to incorporate the Essential 10 into your research means it's not going to be business as usual, either for researchers, ourselves, or for the people who review them.

Slide 70: Implementation will be disruptive

Researchers are going to have to get new skills, and the biggest single one is experimental design. They will certainly need to learn more relevant and updated statistical analysis methods, but to be totally fair, this is really not anybody's fault. Where I place the blame is on introductory statistical service courses, which don't have very good, relevant instruction and [are] certainly not up-to-date. So what researchers can certainly do is pressure people who are offering these courses to make instruction in methods that they need, which will actually address their research needs. And secondly, grant and journal reviews have to do their due diligence. We can't just read and evaluate either grants or papers because of splashy, sexy, flashy research. We need to prioritize sound methodology over small p-values. And it should be noted that if you follow checklist standards for reviewing papers, it will certainly minimize the time you spend reviewing because [if] they don't have their methodology in place, you don't have to bother with their discussion because it's obviously really not in accordance with what they did.

Slide 71: Summary

So in summary: in order to have good science, we need reliable, valid, and clearly and honestly reported information. The quality of research depends on how valid the experiments are, and the reporting guidelines will help us get there. So if you know what the essentials are for good reporting, you use the checklist, and you use the explanation and elaboration document; these will enable you to build in quality from the very beginning of your study before you start even collecting data or use animals, and that means the fewer animals are used, the less research is going to be wasted.

Slide 72: Where to find ARRIVE 2.0 guidelines

So where to find the ARRIVE guidelines: I've posted links to the website, which is your one-stop shop for everything you need. The checklist, the overview, and the E&E documents can be found in PLoS Biology, and also there is simultaneous release in several other journals of checklists, so they're freely available.

Slide 73: Acknowledgements

And I want to finally thank Dr. Nathalie Percie du Sert of the NC3Rs who led this really massive effort to revise and update the guidelines and the massive organization efforts it took the NC3R staff and ARRIVE 2.0 Working Group colleagues.

Slide 74: Questions? Thank you for your attention

For now, I'd be happy to entertain any questions.

>>Nicolette Petervary: Thank you, Dr. Reynolds. We are at 2:00 pm now on the dot, so we ... I'm afraid we don't have any time for questions right now, but don't worry. We will append… We've taken note of these questions, and we will append answers to the end of the transcript when it comes out. We are very appreciative of this incredible information. Thank you, Dr. Reynolds, and please be aware that our next OLAW online seminar will be sometime in the fall with a topic to be determined. Thanks very much for attending. I apologize that we have to cut it early, but there are other events being scheduled right now, so we will see you at the next one. Thanks very much.

Questions

These questions were collected from the chat, Q&A, and email after the session and provided to the speaker. The responses represent the speaker's comments and opinions.

1. **If publications do not have enough information for translatability and reproducibility, how/why is it being published?**

   Peer reviewers and journals share in the responsibility. One of the reasons for the 2017 updating of the ARRIVE guidelines was that although >1000 journals actively endorse the guidelines (and describe them in their Instructions to authors), very few were actually enforcing them. Even so, reporting is still very poor & incomplete. On the other hand, journals that mandate checklist submission by authors, and those using the Cell Press STAR Methods, have seen major improvements in reporting completeness.

2. **The percentage of causes of not reproducibility discussed in the presentation are very specific. How were they determined?**

   See the following paper for a detailed breakdown.

   Freedman LP et al. (2015). The Economics of Reproducibility in Preclinical Research. PLoS Biol 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165

   The paper is now somewhat dated and better information is available, notably from the Reproducibility in Cancer Biology project:

   Errington TM et al. (2021). Reproducibility in Cancer Biology: Challenges for Assessing Replicability in Preclinical Cancer Biology. eLife 10:e67995. https://doi.org/10.7554/eLife.67995).

   The Freedman paper estimates that 10% of irreproducibility could be attributed to poor protocol descriptions and approximately 60% to poor study design and data analysis. The Errington team found that none of 193 experiments described in the original papers were described in enough detail to allow them to design protocols that would enable experiments to be repeated, and authors of >32% did not respond to requests for clarification. Statistical analyses for >50% were poorly reported.

3. **Can you clarify what is appropriate "Randomization"?**

   Randomization is a formal statistically-based process that needs to have a precisely defined, associated methodology and a predesignated experimental design. It refers to the probability-based assignment of treatments (intervention or control) to the experimental units and also to sequence order of processing and obtaining measurements. "Appropriate" means that the randomization process was performed so that each experimental unit has an equal and defined probability of receiving a particular treatment. There are different methods for conducting randomization depending on the study design, e.g., completely randomized, block randomization, stratified randomization, cluster randomization.

A reproducible randomization algorithm is scripted and performed in a good quality statistical programming language with a defined seed, for example not in Microsoft Excel. When it is important to balance groups across covariates (such as sex, body weight, and/or age), then block randomization is far superior to simple randomization.

4. **If researchers think that animals are "randomly" housed upon receiving from commercial vendors, does this suffice the randomization criteria?**

No. Housing assignment has nothing to do with allocation of experimental treatments to experimental units which is the context for the guideline item. The assumption that animals were "randomly" housed at the vendor makes no sense. Used this way, what is really meant is haphazard or unplanned allocation to caging, resulting from a poor understanding of randomization, basic statistics, and basic experimental conduct.

When animals are brought in to be entered into an experiment a good practice is to randomize cage position (using a computer randomization program) to minimize the possibility of some sort of confounding effect of cage rack location that may occur on animal response (although this does not seem to be a problem with ventilated closed caging systems). This does not mean that there is no need to randomize any further; it is still necessary to randomize treatment to experimental units, order of experimentation, etc.

5. **I have reviewed a protocol (looking at whether or not a drug they're testing is safe) that stated they would need x% more mice to test to account for morbidity. As an IACUC or administrative reviewer, how can we respond to help promote better design?**

This is where the 3Rs kicks in. The researcher needs to explain where they got that percent estimate from and what measures are they putting in place going forward to reduce these losses. For example, you can ask the following questions:

- What is the expectation of a given rate of morbidity which is not related to a desired experimental endpoint and is it related to the procedures? Is morbidity due to instrumentation?
- Is morbidity an expected experimental endpoint? If so, are the humane endpoints strictly defined and strictly enforced?
- Does morbidity and attrition indicate significant adverse side effects or toxicity? If so, how will they address these before continuing?
- What is the investigator doing to train lab members and standardize and improve technique?

6. **Can we somehow extrapolate the ARRIVE guidelines to evaluating protocols submitted to an animal welfare committee or IACUC? What do you think the role of the IACUC should be when it comes to reviewing protocols, in terms of ARRIVE 2.0?**

The most we can do is lead the horse to water, but it is heartbreaking when the PIs insist on their ways. Many of the ARRIVE 2.0 items, especially those in the Recommended list (Second Eleven), can be addressed though routine protocol descriptions. Reproducibility items already reviewed by IACUCs include agents, drugs, housing and husbandry, humane and study endpoints, analgesia, euthanasia, nominal personnel skill sets, etc.

With respect to the Essential 10, this is more difficult. However, the low hanging fruit is sample size justification. Sample size and numbers reporting & justification are still extremely poor in the published literature. This should not be a burden for investigators as this is required by IACUC, and the IACUC can certainly ask for more rigorous justification. Justifications such as sample size is based on “judgment,” “intuition,” “what everyone else does,” “industry standard,” “based on my experience,” “what is needed for statistical significance,” or “what is needed for publishable results” should never be permitted.

Sample size calculations are essential to reduce the number of animals wasted in both underpowered and unnecessarily large experiments. Appropriate and rigorous animal numbers review on the part of the IACUC would go far to improving quality. Power calculations are not always necessary, and other essentials are not necessarily highly mathematical. I identified these in my article:

Reynolds PS (2021). Statistics, statistical thinking, and the IACUC. Lab Animal 50: 266-268, 2021. https://doi.org/10.1038/s41684-021-00832-w

I. Outcomes:
   1. What is being measured? Is the outcome clearly defined, specific, measurable, quantifiable?
   2. What are the units of measurement?
   3. How often will it be measured? one time point per animal; many time points per animal?

II. II Study design:
   1. What is the experimental unit?
   2. How many experimental ‘groups’?
   3. What are the control groups?
   4. Can the number of groups be reduced?

III. Sample size:
   1. Are numbers verifiable (did they show their work)?
   2. Feasible (enough people, time, money to do the work)?
   3. Ethical (how many animals will be wasted)?

7. **Has NIH incorporated ARRIVE guidelines in solicitations? Are awards based on alignment with the guidelines?**

Applicants are encouraged to use the Essential 10 but it is not a requirement at this time. However, rigor and appropriate study design is considered throughout peer review, and a study that is not proposed with rigor in mind is likely not to be funded.

8. **You said that Standard Error of the mean is not appropriate for comparing groups. Shouldn't that be OK if your hypothesis is that the treated groups are separate representative populations? These are the error bars shown in many data graphs.**

   The standard error of the mean is a measure of the precision of a population estimate and is not appropriate for describing variation of an observed sample for a continuous variable; use the standard deviation instead. When reporting results, standard error of the mean is not recommended for indicating the precision of an estimate because it is essentially a 68% confidence interval for the population. The 95% confidence intervals should be used instead. One criticism that has been suggested is that investigators prefer standard error of the mean for sample data because it makes the data look less variable and more precise. A common mistake is graphs that display means of sequential measurements with "error bars" (usually denoting standard error of the mean) displayed at each time point. This practice has been strenuously criticized because it ignores the underlying structure of the data and often does not account for either missing observations and dropouts, or within-subject correlation. Many papers also report performing a one-way ANOVA on each time point which is completely incorrect.

9. **Oftentimes when some of this information is included in the publication in detail (i.e. husbandry, drug doses, surgical methods, etc.) the editors delete the information as being superfluous and unnecessary. For instance, if mice are maintained in IVC caging with cage changes every 14 days, does that really need to be stated? And does this information take away from the actual results that are the emphasis of the research?**

   Integrating the different pieces of information involved in ARRIVE 2.0 isn't particularly compatible with the traditional manuscript format. Some journals are asking for the ARRIVE checklist to be completed, which is fine, while others are asking for a narrative response to be included in the methods section. Are there any sample templates for a coherent narrative answering the ARRIVE questions?

   It is a common misconception that inclusion of all the recommended information (recall this information is *essential* for assessing study validity) adds a lot of extra wording. Not true! The ARRIVE 2.0 Explanation and Elaboration document gives examples for each reporting item which show that long tedious descriptions are not at all necessary. If there is a lot of information required for understanding the experimental and animal use protocol but it exceeds page limits, then it can always be included as an online supplementary file. However, you do need to use some common sense in reporting details and prioritize accordingly. Humane care and welfare, drugs (agent, dose, concentration, route, manufacturer), and palliative care measures should always be reported in detail. "Standard of care" for housing, handling and husbandry can vary considerably between institutions, facilities, and colony type, (e.g., breeding colonies will have different cage change schedules than non-breeders, so these details should be reported with sufficient detail to allow assessment).

10. **What role does peer review have in ensuring these essentials are implemented from the very beginning?**

More stringent and critical first-pass review would incentivize investigators to produce better papers and better research. Reviewers should give priority to well-executed and well-documented studies. If design and analysis are statistically sound, a paper should not be rejected because of ‘negative’ results. Reviewers must be far more critical and willing to reject non-compliant manuscripts. On the other hand, poor design and execution of animal experiments should not be condoned merely because a paper is topical or ‘interesting,” or originates in the laboratory of an “opinion leader”.

Reviewers are the gatekeepers for preventing the publication of sloppy or dishonest science. The role of editors and reviewers is to ensure that animals have been treated ethically and humanely, and that the research is of sufficient scientific quality that animals were not ‘‘wasted’’ or made to suffer needlessly. This means that all aspects of study quality (animal models and husbandry, experimental design, performance, analysis, and presentation of results) should be considered when evaluating manuscripts.

However, peer review is the end of a long process and all this is downstream after the animals have been used. To avoid needless animal waste and suffering, investigators must be reminded that rigor and reproducibility must be built in (and guideline compliance begins) in the planning phases of an experiment, well before animals are entered into trials, and much earlier than the publication (or rejection) phase of the research cycle.