ACD WORKING GROUP ON ENHANCING RIGOR, TRANSPARENCY, AND TRANSLATABILITY IN ANIMAL RESEARCH

FINAL REPORT

June 11, 2021

Table of Contents

Membership

Chairs

- Barbara Wold, PhD
  Bren Professor of Molecular Biology,
  Caltech
- Lawrence A. Tabak, DDS, PhD
  Principal Deputy Director, NIH

External Members

- Nancy Ator, PhD
  Professor of Behavioral Biology,
  Department of Psychiatry and Behavioral Sciences
  Johns Hopkins School of Medicine
- Lais Berro, PhD
  Postdoctoral Fellow
  University of Mississippi Medical Center
- Eliza Bliss-Moreau, PhD
  Associate Professor, Department of Psychology; Core Scientist, California National Primate Research Center
  University of California, Davis
- Romer A. Gonzalez Villalobos, MD, PhD, FAHA
  Senior Principal Scientist, Cardiovascular & Metabolism Discovery
  Janssen Research and Development, LLC
- F. Claire Hankenson, DVM, MS, DACLAM
  Attending Veterinarian; Director, Campus Animal Resources
  Professor, Pathobiology and Diagnostic Investigation, College of Veterinary Medicine
  Michigan State University
- Véronique Kiermer, PhD
  Chief Scientific Officer
  PLOS
- Keisa Williams Mathis, PhD
  Assistant Professor, Department of Physiology
  University of North Texas Health Science Center
- Sarah Nusser, PhD
  Vice President for Research; Professor of Statistics
  Iowa State University
- Regina Nuzzo, PhD
  Senior Advisor for Statistics Communication
  American Statistical Association
- Eric Prager, PhD
  Associate Director, External Affairs
  Cohen Veterans Bioscience
- F. Daniel Ramirez, MD, MSc
  Cardiac Electrophysiology Fellow
  CHU Bordeaux, IHU Liryc
- Karen Svenson, PhD
  Senior Scientific Program Manager and Research Scientist
  Jackson Laboratory

USG Members

- Brian Berridge, DVM, PhD, DACVP
  Associate Director, National Toxicology Program; Scientific Director, Division of the National Toxicology Program, National Institute of Environmental Health Science, NIH
- Paul Brown, PhD
  Associate Director for Pharmacology and Toxicology
  Office of New Drugs; Center for Drug Evaluation and Research
  FDA

- Janine Clayton, MD
  Director, Office of Research on Women's Health, NIH
- Joshua A. Gordon, MD, PhD
  Director, National Institute of Mental Health, NIH
- Michael Lauer, MD
  Deputy Director for Extramural Research, NIH
- Robyn Lee-Stubbs, MS, CPIA, PStat®
  IACUC Chair/Statistician
  United States Army Medical Research Institute of Chemical Defense
- Glenn Merlino, PhD
  Scientific Director for Basic Research, Center for Cancer Research, National Cancer Institute, NIH
- Shai D. Silberberg, PhD
  Director, Office of Research Quality, National Institute of Neurological Disorders and Stroke, NIH
- Carrie Wolinetz, PhD
  Acting Chief of Staff; Associate Director for Science Policy, NIH

Executive Secretary

- Jordan Gladman, Ph.D.
  National Institutes of Health

Acknowledgments

The co-chairs are most grateful to the entire committee for their continued dedication throughout this effort, despite the remarkable the challenges of the Covid-19 pandemic. Each member brought a unique perspective to the effort – from different vantage points along the career path in Academia, from industry, and from the perspective of journal editors. Within NIH we also benefited from the expertise of both intramural and extramural leaders. We were also fortunate to have good representation from other parts of the USG.

Particular thanks go to Dr. Regina Nuzzo whose communication and editing skills were key to the timely completion of this effort. We are also grateful to Dr. Jordan Gladman, who led the logistical efforts to keep the committee moving forward and provided much help during the editing process. Finally, Dr. Tabak is most grateful to Dr. Barbara Wold, for her extraordinary leadership as co-chair. Dr. Wold was the driving force behind this effort, and I am most grateful to her.

Barbara Wold, PhD

Lawrence A. Tabak, DDS, PhD

Executive Summary

Animals are a precious and essential part of NIH-funded biomedical research. Our working group was charged to advise NIH how it can help researchers improve the rigor, transparency, and reproducibility of NIH research involving animals, paying close attention to the late stages of the translational pipeline that produce new treatments for human health and disease. The overarching goal is to allow all stakeholders to have full confidence in the quality and applicability of research findings from animal studies, and to ensure that animal subjects are used with appropriate consideration of ethics and harm-benefit analysis.

Researchers, funders, publishers, and the public provide an impetus to increase rigor and transparency in animal research. In part, this push has been motivated by recent documented problems in replication and translatability of biomedical research. Yet positive driving forces exist, too, such as the need to keep pace with the rapid progress of science itself, which includes technological advances that allow today's researchers to work with increasingly complex questions and small effect sizes. This progress demands concomitant advances in scientific research methods—the rigor of study design, the handling and analysis of data, and the reporting of results.

Investigators need NIH's support and active participation to increase the rigor and transparency of their research. NIH must obtain and commit sufficient financial resources toward improvements and also effectively use incentives and oversight in the grant application, review, and funding process. It can also uniquely help investigators by identifying and promulgating best practices, investing in strengthening the animal research statistical workforce, and working to educate the scientific community and the public about ongoing challenges and achievements.

Our working group's recommendations follow five themes: 1) improving study design and data analysis; 2) addressing incomplete reporting and questionable research practices; 3) improving selection, design, and relevance of animal models; 4) improving methodological documentation and results reporting; and 5) a crosscutting theme focused on measuring and evaluating the costs and effectiveness of these efforts.

In Theme 1, we recommend that NIH help researchers improve their study design and data analysis by two means: enhancing the statistical training available to students and investigators, and facilitating collaborations between animal researchers and statisticians. We also recommend that NIH find ways for experts to provide feedback on investigators' study plans early in the research process, that it helps applicants share these plans by adding dedicated extra space to NIH research grant applications.

In Theme 2, we recommend that NIH address questionable research practices by raising awareness in the animal research community of the benefits of prospectively documenting study design and analysis plans. We also recommend that NIH launch pilot programs for prospective registration and registered reports to explore their feasibility and utility.

In Theme 3, we recommend that NIH help improve the selection, design, and use of animal models by establishing a framework in which investigators can explain the scientific relevance and rationale behind their chosen animal models. We also recommend that NIH ensure that animal researchers have appropriate venues in which they can exchange best practices for animal models. More broadly, we recommend that NIH fund research programs on comparative animal-human biology and provide adequate research support for larger and long-lived species. Lastly, we recommend that NIH educate the public on the value of animal research for enhancing human health and reducing illness, and we recommend it also create a working group to explore non-animal modeling systems in biomedical research.

In Theme 4, we recommend that NIH help improve study methodology documentation by two means: ensuring that researchers know which factors related to animals' environment can affect research outcomes and thus are critical to document, and providing support for researchers to document the long-term care of larger and long-lived animals. We also recommend that NIH improve complete reporting of results by setting expectations for the use of the ARRIVE 2.0 Essential 10 checklist and the inclusion of statistical measures of uncertainty and effect size.

In Theme 5, we recommend that NIH measure the costs and effectiveness of efforts to improve rigor, transparency, reproducibility, and translatability in animal research. Specifically, we recommend that NIH develop a program to assess the progress in implementing this report's recommendations; conduct and support analyses on elements of rigor and transparency in grant applications and publications; allow applicants to include budget justifications for efforts linked to enhancing rigor, transparency, and reproducibility; and work with scientists who demonstrate the highest levels of transparency and rigor to develop best practices.

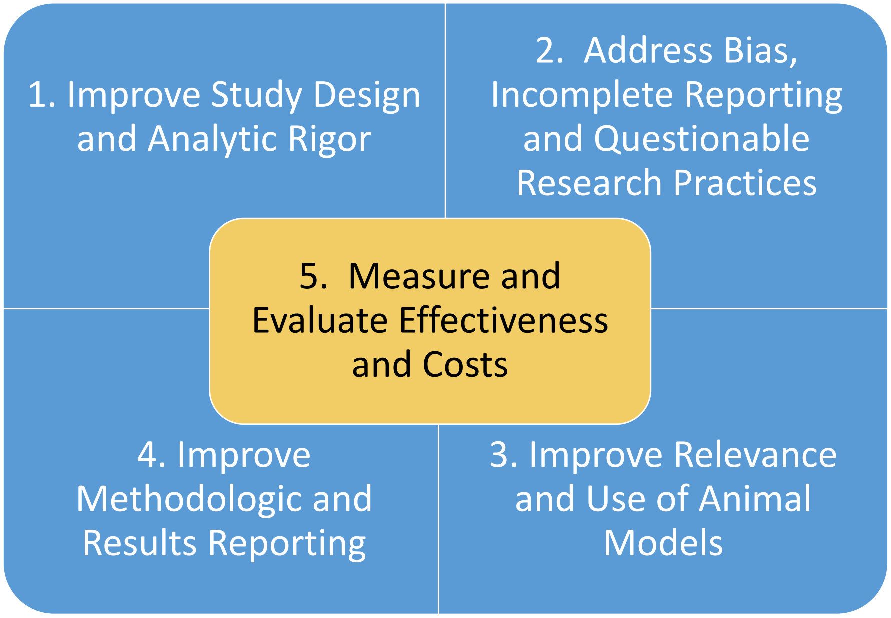

1. Improve Study Design and Analytic Rigor
2. Address Bias, Incomplete Reporting and Questionable Research Practices
5. Measure and Evaluate Effectiveness and Costs
4. Improve Methodologic and Results Reporting
3. Improve Relevance and Use of Animal Models

Figure 1. Themes of the ACD Working Group on Enhancing Rigor, Transparency, and Translatability in Animal Research Report

Introduction

From understanding biological mechanisms to developing state-of-the art vaccines, animal models play a critical role in science [1]. Animal research has accounted for nearly half of all NIH research project grant applications over the past decade, and animal studies serve as a foundation for human clinical trials. Not only do we learn how to prevent, treat, and cure human diseases by studying animals, but often the treatments developed are also used to improve the health of animals themselves.

We study animals because of their comparative relationship to the human species. Moreover, unlike human studies, animal studies allow researchers to carefully control many characteristics that might affect the outcome of experiments, including intrinsic factors such as genetic composition and **extrinsic factors** such as diet or medications. Animal models thereby allow for a more precise understanding of biological factors and provide greater certainty about experimental outcomes.

Despite the numerous successes stemming from animal research, concerning reports over the past decade have described biomedical experiments that fail to replicate or to translate in ways that improve human health. All research is not expected to translate to human treatments, as there is no perfect model. Scientific process is as much about failure as it is success. Yet part of the scientific process is also continual improvement, which includes working to understand what might contribute to unexpected outcomes within animal research.

In many fields of biomedical science, for example, animal researchers are now using more complicated statistical models to capture subtle effects, yet the statistical training and research support provided to investigators has not always kept pace. Likewise, although problems with questionable research practices and publication bias plague all researchers, solutions suitable for the animal research community are not in widespread use. Design and selection of animal models has become increasingly difficult, and it has been shown across multiple disciplines that extrinsic sources of variability from experimental conditions—such as lighting intensity, noise, vibrations and even the animal's microbiome—could contribute to difficulties in fully replicating experimental results.

The ILAR *Guide to the Care and Use of Animals in Research* [2] reflects that "[u]sing animals in research is a privilege granted by society to the research community with the expectation that such use will provide either significant new knowledge or lead to improvement in human and/or animal well-being." That expectation can be met only if studies are rigorous, well designed, transparently reported, and use an animal model relevant to the condition of interest when intended for human translation. Following these practices will help ensure that the appropriate number of animals are involved to obtain reliable and reproducible results. Unreliable results from animal research can lead the scientific community astray, slow the progress of medical discovery, waste animals and other valuable resources, and lead the general public to lose trust in the scientific mission.

NIH has taken several steps in recent years to improve the scientific rigor of the research it supports, including releasing policies to enhance reproducibility of its supported research through rigor and transparency, increasing the focus of rigor in grant review, and establishing an ACD working group in response to the 21st Century Cures Act [3, 4]. Yet animal research requires additional attention for several reasons. First, animal researchers are required to consider the ethical imperative of the "3Rs" of **Replacement***,* **Reduction***,* and **Refinement**; these principles impel researchers to use sound study design and contemporary methods for improved animal welfare and to include only the appropriate number of animals necessary to answer a specific scientific question. Second, animal research faces unique challenges, such as species-specific animal husbandry and the appropriateness of selecting animal systems to model human biology and disease for translation. Third, even small improvements in animal research rigor can have large impact on the funding of science, since animal studies serve as the foundational knowledge base for most human trials. Costs of research failures increase as research advances from early discovery-stage studies to mature translational animal studies that precede first-in-human trials. These costs include the misdirection of scientific effort, NIH dollars spent, and the well-being of human participants in follow-on patient trials. The resulting cost savings could then be re-invested in research at all stages from discovery to translation, thus further improving the return on investment.

Other organizations are also addressing reproducibility issues in animal research. In 2014, for example, the National Academies held a workshop on "The Missing "R": Reproducibility in a Changing Research Landscape"[5], in which researchers discussed facets of animal-based research that could contribute to irreproducible results and provided including perspectives on improving experimental planning, design, and execution; the importance of reporting all methodological details; and efforts to establish harmonization principles of reporting on the care and use of animals in research studies.

Improving rigor and reproducibility will have costs—in both dollars and opportunity—as well as benefits to science and society. Yet the amounts on both sides of the balance sheet are difficult to predict. Likewise, it will not be simple to adjust policies and practices in a data-driven way, but it is critical to do so. Managing the cost-benefit balance at the micro-level (such as individual research grants) and at the macro-level (such as across the diverse NIH animal research portfolio) will be an ongoing challenge shared by NIH and its research community.

Rigorous science and its comprehensive reporting are primarily in the hands of researchers—in how we design, execute, and report our experiments, in how we conduct peer review of papers and grants, and in how we teach the next generation of scientists about rigor and transparency. The professional and societal benefits of conducting rigorous and transparent research are plentiful. Yet for individual animal researchers the opportunity to elevate 3Rs can be a balancing act that often comes with obstacles in the form of added costs on "do-more-for-less" budgets, statistical training deficits, manpower constraints, misaligned incentives, and other challenges. NIH is uniquely positioned to support initiatives for improved scientific outcomes, building upon efforts it has already undertaken to improve rigor, transparency, and reproducibility, and taking into account work done by outside organizations, including the National Academies, the National Centre for the Replacement, Refinement, and Reduction of Animals in Research, American Physiological Society, Society for Neuroscience, FASEB, and other scientific societies.

Working Group Charge

The Working Group was charged by the NIH Director to:

- Identify gaps and opportunities to improve the rigor, reproducibility, translational validity, and transparency of studies involving animal models, including:
  - selecting animal models that are most appropriate for the scientific question

    - strengthening experimental design and statistical analysis, including appropriate statistical power and definition of endpoints
    - achieving appropriate transparency of methodological measures
    - considering how the conditions in which animals are housed and bred affect experimental outcomes
    - enhancing processes to incorporate rigor and transparency into grant application, peer review process, and manuscript publication
- Evaluate how animal models of human disease are currently developed, validated, and accepted into routine use, and how this process could be improved.
- Assess the current state of science for validating alternative models to animal research.
- Consider the benefits and burdens of registering animal studies that aim to lead to first in human trials (e.g., preregistration of the experimental plan).
- Model the financial implications of potential changes in the average costs of grants using animal models, the number of studies funded, or the need to develop multi-lab consortia to achieve appropriate statistical power.

A Framework and Vocabulary to Discuss Rigor and Reproducibility

In our working group discussions, we quickly found that terminology and vocabulary played an important role in our shared understanding and framing. In this section and in Appendix 1 we share the results from some of those discussions. While our charge did not specify a definition for “animal,” the major impetus for establishing the working group was to improve translation of animal research to human therapies. With this in mind, we focused our deliberations and emphasized recommendations that would most directly contribute to enhanced translatability. Although some recommendations and principles in the report could be applied to all biomedical research, our scope is “animals,” which we defined to include any nonhuman vertebrate or cephalopod species.

NIH defines **scientific rigor** as the strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation, and reporting of results. Through the rigorous application of the scientific method, researchers can acquire knowledge that helps understand living systems and enhance health. **Transparency**, the accessibility of information, is essential for rigorous science and reproducibility.

“Reproducibility” and “replicability” are often used interchangeably to refer to sufficient consistency or robustness of experimental results when repeated by either the original researchers or, more often, by researchers who were not involved in the original experiments. However, these terms are also sometimes used to refer to distinct concepts. For example, we see both “reproducibility crisis” [6] and “replication crisis” [7] used to describe a growing concern among researchers, funding bodies, and the public that some of the scientific literature is fundamentally flawed or unreliable, and we hear of a “Replication Project” [8] and “Reproducibility Projects” [9] designed to explore the prevalence of this problem. Despite the rapidly growing number of publications dealing with issues of “reproducibility” [10], and a consensus study by the National Academies of Sciences, Engineering, and Medicine [11] this term still lacks a standardized definition across the biomedical community. It was therefore necessary to clarify these terms and adopt a common vocabulary for the working group.

From the multiple definitions we reviewed, one set in particular resonated with our members. For our discussions and this report, we elected to follow terminology from the conceptual framework of reproducibility proposed by Goodman, Fanelli, and Ioannidis [10].

- **Methods/methodological reproducibility** refers to researchers providing enough detail about study procedures and data so that the same procedures could, in theory or in actuality, be exactly repeated. Transparency is a key component of methods reproducibility.
- **Results reproducibility** refers to researchers obtaining the same results when they conduct an independent new study with procedures as close to the original study as possible. This definition is similar to others' definitions of "replicability."
- **Inferential reproducibility** refers to researchers drawing qualitatively similar conclusions, or making knowledge claims of a similar strength, from either an independent replication of a study or a reanalysis of the original study. This is part of the process by which a scientific field decides which research claims or effects are to be accepted.

When assessing results reproducibility, it can be difficult for researchers to define what is required for results to be considered sufficiently similar. Some investigators use statistical significance (that is, a p-value less than some threshold, typically 0.05) to determine whether a study has successfully replicated previous results. Yet this black-and-white approach to reproducibility can cause problems. For example, if a study captures an experimental effect with a p-value of 0.05, there is only a 50% chance that even a perfect replication study will also reach statistical significance [12] (Appendix 2). Expectations about "successful replications" should therefore be tempered. Perhaps more important than narrowly assessing replication is for researchers to assess the rigor of studies, consider the magnitude of their effects within the context of the field, and evaluate their cumulative evidence.

Similarly, when researchers evaluate inferential reproducibility, they must incorporate prior knowledge of experimental techniques and the subject matter at hand. If all results and analyses in a study were not completely reported, or if questionable research practices are suspected, then researchers should be less willing to accept its research claim. In general, when undisclosed flexibility in study procedures increases, inferential reproducibility decreases. When experimental plans are documented and adhered to, and when study results are reported completely, then inferential reproducibility increases.

Biomedical research is often conceptualized as spreading along a continuum: the earliest discovery research works to illuminate how biological systems work, and later preclinical research shifts to focus on knowledge to translate to human clinical trials. Human clinical trials often rely on applying results from late-stage **preclinical** animal studies; their success depends upon the quality of the preclinical research, so there is a cost when animal research fails to reproduce. Moreover, the costs stemming from low rigor and reproducibility grow as we move along the translational chain and radiate forward into therapeutic development in pharma and industry partners. Conversely, there are costs to insisting on high levels of rigor and reproducibility in early discovery stage research that may not be productive or beneficial. As part of our deliberations, we considered the unique needs and challenges of rigorous animal research across the stages of scientific pipeline.

The concept of **generalizability,** which an NSF report [13] refers to as "whether the results of a study apply in other contexts of populations that differ from the original one," is important to consider in the context of reproducibility and translation. In this report we are concerned with whether results generalize among animal studies and especially whether they generalize from an animal system to humans. We also recognize that there are different types of research using animals, which are generally designed and conducted in fundamentally different ways, and our methodological and inferential expectations of them should differ accordingly. Recognizing and embracing the distinction and co-dependence between **exploratory** ("hypothesis-generating") and **confirmatory** ("hypothesis-testing") research could also be beneficial in many ways. Journals, funding bodies, and peer reviewers should ensure that researchers who conduct exploratory experiments are not penalized for doing so. These researchers should feel encouraged to fully and transparently report experimental shortcomings (e.g., if experiments were underpowered or unblinded) and to appropriately temper the reach and strength of their conclusions, rather "sell" results as being more certain or being of immediate translational potential than the evidence supports.

Meeting these unique challenges will require changes and investments from multiple angles. The itemized themes outlined below will highlight areas of opportunity for solutions and success.

Theme 1: Improve Study Design and Data Analysis

Biomedical advances are largely founded on prior discoveries attained through experiments that are well-designed and carefully conducted and analyzed. When studies are rigorously designed in this way, they minimize the risk of unconscious researcher biases, control for random variability and confounders, and provide a reliable estimate of experimental effects. Key measures used to minimize the risk of bias include randomly and blindly assigning samples to comparison groups; conducting experiments and analyzing the results in a blinded fashion; and prospectively defining criteria for inclusion or exclusion of samples from analysis. Randomization serves to control for confounders by reducing the probability that comparison groups are not well balanced. Prospectively defining a data analysis plan is also a fundamental aspect of good study design. Establishing a data analysis plan serves to refine the experimental design and assures that questionable research practices associated with overly flexible analysis are mitigated (which we discuss in more detail in Theme 2).

The inherent variability among biological samples typically necessitates the use of multiple samples per group to achieve a reliable estimate of the parameters being measured. Optimal sample sizes minimize the risk of chance observations by reducing the probability that true relationships/effects will be overlooked or will lead to false-positive "findings." Estimating an appropriate sample size is often complicated by insufficient information on the variability among the biological samples and expected effect sizes, and by the complexity of the experimental design. Developing mature confirmatory designs requires antecedent and often stepwise exploratory work to refine these parameters. As such, it is imperative that scientists designing experiments or evaluating the work of others possess appropriate statistical skills and/or have access to statistical consultation.

Good study design and good data analyses are also essential from an ethical standpoint. Poorly designed studies can lead to mistaken results that lead the scientific community astray, slowing the progress of science. Such endeavors waste valuable resources and, importantly, might lead to unjustified use of animals or the need to repeat studies and wase animal life.

The following four recommendations under this theme address ways to help researchers improve their study designs, sample size estimations, and analysis plans, as well as help reviewers and NIH assess the quality of such plans in manuscripts and grant applications.

Recommendation 1.1: NIH should improve and expand statistical training for animal researchers.

SUB-RECOMMENDATION 1.1A: NIH SHOULD PARTNER WITH OTHER ORGANIZATIONS TO DEVELOP MODERN AND INNOVATIVE STATISTICS CURRICULA RELEVANT TO ANIMAL RESEARCHERS.

SUB-RECOMMENDATION 1.1B: NIH SHOULD DEVELOP STATISTICAL RESOURCES SPECIFICALLY FOR ANIMAL RESEARCHERS.

SUB-RECOMMENDATION 1.1C: NIH SHOULD REQUIRE STATISTICAL TRAINING FOR TRAINEES CONDUCTING ANIMAL RESEARCH AND STRONGLY ENCOURAGE IT FOR TEAM MEMBERS INVOLVED IN STUDY DESIGN AND DATA ANALYSIS.

Animal researchers need strong statistical skills to produce research that is rigorous, reproducible, cost-effective, and ethical. To acquire these statistical skills, researchers need effective training. There are many dimensions to effectiveness in statistical education, but it is especially important that resources for animal researchers directly address the unique challenges and needs of doing research particularly with animal models. Training resources for animal researchers should also be easily accessible, engaging, and useful for audiences of different levels, from undergraduate students to principal investigators.

Current statistical training for animal researchers often falls short of these ideals, however. Curricula for research methods and statistics courses in graduate programs are not always modern, engaging, or relevant for animal researchers. Training requirements for NIH trainees vary greatly between programs. Resources for staying abreast of new statistical tools and methods for researchers beyond the trainee level are not easily accessible or centralized.

We recommend that NIH play a direct role in improving and expanding statistical training for animal researchers. A first step is to partner with other organizations (e.g., American Statistical Association) to develop innovative and effective statistical curricula especially suitable for animal researchers. From this, specific resources should be developed and disseminated. NINDS already has a planned educational platform on the principles of rigorous research; we recommend taking advantage of this opportunity and developing for this platform effective statistical resources specifically designed for animal researchers at various levels. Other platforms, such as the NIGMS Clearinghouse for Training Modules to Enhance Data Reproducibility [14], can also be leveraged to enhance training. After these resources have been developed, we recommend requiring statistical training for NIH trainees conducting animal research, and we recommend strongly encouraging it for team members who are involved in study design and data analysis.

RECOMMENDATION 1.2: NIH SHOULD FACILITATE COLLABORATION BETWEEN STATISTICIANS AND ANIMAL RESEARCHERS.

SUB-RECOMMENDATION 1.2A: NIH SHOULD EXPAND RESEARCH COLLABORATIONS BETWEEN STATISTICIANS AND ANIMAL RESEARCHERS.

SUB-RECOMMENDATION 1.2B: NIH SHOULD FUND TRAINING FOR STATISTICIANS ON DOMAIN-SPECIFIC SUBJECT MATTER AND ON CHALLENGES FACED BY ANIMAL RESEARCHERS.

SUB-RECOMMENDATION 1.2C: NIH SHOULD INCREASE ANIMAL RESEARCHERS' ACCESS TO STATISTICAL CONSULTING THROUGH FUNDING OPPORTUNITIES.

SUB-RECOMMENDATION 1.2D: NIH SHOULD INCENTIVIZE RESEARCH IN STATISTICAL METHODS FOR ANIMAL STUDY DESIGN AND ANALYSIS.

Collaborations and cooperation between statisticians and empirical scientists are part of the culture in many scientific fields, including clinical research, social science, and agricultural science. This is not widespread practice in animal research, however. We recommend that NIH help change preclinical research culture by facilitating interactions between animal researchers and statisticians.

Research collaborations: Collaborations between empirical scientists and statisticians can have a synergistic effect: their interactions drive research advances in multiple areas, suggest new areas of research, and enhance both fields. We recommend NIH find ways to facilitate research collaborations between animal researchers and statisticians through funded research programs. For example, institutions could establish short-term "idea incubator" residential educational programs that bring together junior statisticians and junior animal researchers where they are trained in issues at the interface of their fields, learn how to collaborate with each other, and build research relationships.

Training for statisticians: When statisticians gain domain-specific scientific knowledge, they are able to communicate in the language of the scientific collaborator and better understand the underlying biological considerations in statistical approaches. We recommend that NIH fund training specifically designed for statisticians working with animal researchers. For example, training resources could be developed for graduate students on topics such as particular animal models, biological and scientific knowledge in specific fields, experimental design for animal studies, and applied statistics consulting skills.

Access to consulting: When empirical scientists can consult applied statisticians, the additional methodological expertise can increase the rigor and effectiveness of their studies. We recommend that NIH increase animal researchers' access to statistical consulting through funding opportunities. For example, seed funding could be provided for an institution to create a position specifically for an animal-research statistical consultant, or a faculty joint appointment between Statistics and College of Veterinary Medicine. Additionally, graduate students in applied statistics could receive funding that would require statistical consulting projects with animal researchers.

Methodological research: We recommend that NIH incentivize statistical methodology research specifically focused on animal study design and analysis. For example, research could be funded around small-sample techniques, methods for controlling for confounders common in animal experiments, strategies for randomization and blinding in animal experiments, and experimental designs and data analysis tools designed for specific animal models.

RECOMMENDATION 1.3: NIH SHOULD ADD A SINGLE PAGE TO THE NIH GRANT APPLICATION RESEARCH STRATEGY SECTION THAT IS SOLELY DEDICATED TO THE DESCRIPTION OF CRITICAL ELEMENTS OF STUDY DESIGN, INCLUDING INCLUSION/EXCLUSION CRITERIA, SAMPLE-SIZE ESTIMATION, DATA ANALYSIS PLAN, BLINDING AND RANDOMIZATION, TO REDUCE THE RISK OF BIAS AND CHANCE OBSERVATIONS. THIS PAGE WOULD BE IN ADDITION TO THE CURRENT RESEARCH STRATEGY PAGE LIMIT AND WOULD APPLY TO VERTEBRATE AND CEPHALOPOD STUDIES.

Journals and funders are increasingly requiring or encouraging researchers to report critical elements of their study design and analysis in reports and applications, often by using tools such as the ARRIVE guidelines and journal-specific checklists. Their goal is to increase transparency and reduce the risk that published results are affected by undisclosed biases, random variability, or errors in experimental design, laboratory protocols, or data analysis. Unfortunately, investigators often fail to fully report all factors in their experimental design and analysis, and these journal efforts are continuing to evolve to improve and facilitate compliance [15, 16]. However, a major limitation of a journal's checklist is that it comes at the end of a study when it is too late to improve the design with a blinding and randomization plan or good sample size estimation.

Both the quality and character of future studies and the ease of using reporting checklists at the time of publication could be improved with two changes in approach by NIH. First, when investigators are fully aware of reporting recommendations and standards early in the research process, such as during the grant application process, they will naturally incorporate better design elements from the beginning. In this direction, a recent report from the CSR Advisory Council [17] argues for increased consideration of rigor and reproducibility in all research in grant applications. Our working group concurs, and we also identified additional needs for animal studies.

Second, an important lesson from studies of checklist successes and failures is that when fewer but more important items are recommended or required, investigators are more likely to fully report. Long reporting lists create burdens for investigators and hamper the ability of reviewers and editors to adequately evaluate or enforce compliance. We therefore concluded that the best path forward will be to include into the grant application Research Plan a succinct set of rigor items that are specifically tailored for animal use and which have already been found to improve reporting of blinding, randomization, and sample size calculations for *in vivo* studies.

Leveraging the recommendations from the CSR Advisory Council and acknowledging that short lists have increased value, we recommend that NIH have grant applicants who are doing vertebrate and cephalopod research explicitly address a specific list of study design elements as part of the Research Strategy. We also recommend that NIH provide supporting information to fully explain what is expected in the additional page (Appendix 3).

We recommend that these study design elements include:

1) **Inclusion and exclusion criteria**: criteria that will be used to include or exclude animals during data collection and analysis
2) **Sample-size estimation**: how planned sample sizes for each group will be derived
3) **Data analysis plan**: what statistical tests will be used and how outliers will be defined
4) **Blinding**: who will be aware of group allocation at different stages of the experiment
5) **Randomization**: how allocation of experimental units to control and treatment groups will be done.

We feel that the additional specificity required to satisfactorily address these elements of study design requires an additional, separate, dedicated page in the grant application. This additional page would be used solely for describing methods supporting specifically named elements and not for expanding upon or adding any other details of the Research Strategy. Applicants would use this page to summarize design strategies and further direct reviewers to sections in the Research Strategy where additional detail is provided. If any of these elements is not appropriate to the proposed studies, researchers would use this space to say that and explain why.

This dedicated single page would have several benefits. Most importantly, it would provide opportunities for researchers to sufficiently explain how these study design elements will improve the success of their study, which can subsequently be reported at the publication stages [15]. Given the multiplicity and diversity of study designs typically included in grant applications, this additional page would allow the investigator to give sufficiently detailed descriptions of bias-reducing measures that may not be permitted by the current space limitations. Furthermore, in support of the CSR's Advisory Council and in line with Recommendation 1.4, reviewers at all stages of the process will be able to assess these elements and determine whether they have been adequately addressed, and how they should affect funding or oversight. It would also enable NIH to retrospectively monitor the effectiveness of these measures across grants. Finally, this additional space would help to level the playing field for young investigators or those new to later stage preclinical work, many of whom do not have the option, often exercised by established investigators to parsimoniously refer to rigor elements in a body of their own published work.

RECOMMENDATION 1.4: NIH SHOULD EVALUATE WHERE IN THE PRE-STUDY RESEARCH PROCESS EXPERTS COULD ASSESS THE QUALITY OF STUDY DESIGN AND DATA PLANS, THEN IMPLEMENT PILOT STUDIES OF ASSESSMENT AT THE MOST PLAUSIBLE STAGE(S).

A system of guardrails and quality checks in the research process can help ensure that researchers' study designs, data analysis plans, and statistical results are as strong and accurate as possible. We recognize the pressing need to have these "statistical guardrails" as well as the equally pressing obligation not to create an untenable burden for researchers, institutions, and NIH. Therefore, we recommend that NIH evaluate where in the pre-study research process statistical experts could assess the quality of study design and data plans, including the stages listed below, in order to balance enhanced rigor with acceptable burden. Expert assessment at multiple points in the research process might be appropriate. NIH should launch pilot studies to evaluate the effectiveness and assess the burdens of using quality checkpoints. Before implementation, a plan should be created to collect data and evaluate the efficacy and impact of the interventions.

(i) *Researcher study design stage*: Grant applications with animal studies could be strongly encouraged to use the NC3R's Experimental Design Assistant tool [18, 19] and include its flowchart in the vertebrate animal section of the application.

(ii) *Grant peer review stage*: All study panels with animal studies could include at least one trained reviewer who can evaluate the application's statistical elements, including study design and analysis plans. This would be aided by Recommendation 1.3, to create a separate page containing elements of study design and data analytic plans

(iii) *Grant post-peer review stage*: As an alternative to (ii), a statistical review panel composed of applied statisticians can be formed to evaluate proposals with animal studies that have received the highest scores in the previous peer review stage. This would also be aided by Recommendation 1.3.

Theme 2: Address Incomplete Reporting and Questionable Research Practices

The reliability of published research findings depends on the use of rigorous scientific practices designed to minimize unconscious research bias. Unfortunately, the use of "questionable research practices" has been documented in many fields, including the absence of bias-controlling measures in experimental design [20-22], small and underpowered sample sizes, selective outcome reporting and outcome switching, p-hacking, and hypothesizing after the results are known (HARKing). These questionable research practices all increase the likelihood that study results and conclusions are unreliable.

Likewise, the reliability of published studies depends on the full reporting of research processes and outcomes. For example, when there is a lack of transparency in reporting how a study was conducted and analyzed, or in reporting whether key parameters and procedures were decided before or after the data were collected, other researchers cannot fully interpret the study's results and evaluate its robustness [23, 24]. Similarly, when positive results receive priority for publication (a documented problem known as publication bias), the evidence base used to inform preclinical and clinical studies is weakened [25-27].

Strong motivation therefore exists to improve research and reporting practices, including the testing and adoption of approaches in which animal researchers plan and commit to experimental and analytic steps before they collect experimental data. The two recommendations under this theme address ways NIH can improve research and publication practices by addressing the problems of incomplete reporting and questionable research practices.

RECOMMENDATION 2.1: NIH SHOULD LAUNCH A CAMPAIGN TO RAISE AWARENESS AND UNDERSTANDING OF PROSPECTIVELY DOCUMENTING STUDY DESIGN AND ANALYSIS PLANS.

One approach to improving research rigor and transparency is prospective registration, in which researchers specify the details of their study design, analytic plan, and primary outcomes *before* data are collected, and either register these details publicly (in a repository) or keep a time-stamped record under embargo until the study results are known. Prospective registration offers a way of retrospectively assessing bias control in study design by comparing published results with original plans. A publication process known as Registered Reports enables a journal to peer-review the study protocol before data collection and to offer a guarantee of publication regardless of the study results. In principle, this process offers benefits to researchers, helps improve study design before data collection, and protects against publication bias.

Prospective registration: Creating a permanent record of a study design, analytic plan and primary outcome before the data are collected. Prospective registration allows retrospective assurance against selective reporting and outcome switching. The registered research plan can be embargoed for a limited time to protect intellectual property (IP) and when the registration is published it allows to identify and mitigate publication bias. Researchers who prospectively register their studies demonstrate their intent to prioritize rigorous reporting. Prospectively registered studies do *not* undergo peer review at the time of registration.

Registered Reports: Journal article type in which the detailed study protocol is submitted for peer review before the data are collected. Upon successful review, the journal guarantees publication of the article regardless of the study findings. Like prospective registration, Registered Reports mitigates selective reporting and outcome switching. In addition, peer review of the protocol can help improve experimental design, for example by ensuing that sample size and inclusion/exclusion criteria are predetermined, and that proper measures are taken to mitigate experimental and analysis bias. Researchers who publish with this review workflow demonstrate their intent to prioritize rigor, and they also secure a path to publication regardless of findings, which mitigates publication bias.

Prospective registration has been successfully adopted in a number of research areas, including clinical research, economics [28], psychology [29], and systematic reviews in health and social care [30]. Despite implementation challenges, prospective registration has improved the accountability of clinical trials and transparency in their reporting [31-33]. Yet, prospective registration is not in widespread use across all research disciplines, particularly in animal studies. Furthermore, awareness and understanding about what prospective registration means and its assumed benefits for research reliability are largely lacking in the animal research community. This lack of adoption and understanding in turn limits our ability to assess the efficacy of prospective registration in animal studies, demonstrate its benefits, measure its costs, and understand and mitigate its potential unintended consequences.

We recommend that NIH initiate a program to raise awareness and understanding of prospective registration methods and how they support universal goals of rigor and transparency. Key elements to communicate through this program include:

- Clear definitions of and distinction between prospective public (or embargoed) registration in a repository and Registered Reports, since these two mechanisms of prospective registration are likely candidates for potential pilot and intervention studies.
- Clear articulation of the benefits of prospective registration and Registered Reports for researchers and their fields.
- Mitigation measures (e.g., embargo periods) to be considered for legitimate concerns such as the protection of intellectual property and the minimization of risks to researchers, including if ongoing studies are discoverable by anti-animal research groups who engage in threatening behavior.
- Awareness of prospective registration as a means to bring benefit to hypothesis-testing studies, particularly focusing on *in vivo* studies that are intended to directly inform clinical trials (i.e., towards the clinical end of the pre-clinical spectrum), as this is the context in which prospective registration can be most easily understood, justified, and applied.

The campaign should begin with a strong message from NIH leadership, for example in the form of a published Commentary, to help raise the stakes and articulate the benefits sought through prospective registration.

RECOMMENDATION 2.2: NIH SHOULD DEVELOP AND IMPLEMENT A PILOT PROGRAM TO GENERATE DATA ON AND EVALUATE THE EFFECTS OF SOLUTIONS THAT INVOLVE THE PROSPECTIVE DOCUMENTATION OF STUDY DESIGN AND ANALYSIS PLANS IN PRECLINICAL ANIMAL STUDIES.

> **SUB-RECOMMENDATION 2.2A: NIH SHOULD DEVELOP AND INCENTIVIZE PROJECTS THAT GENERATE DATA ON THE IMPACT OF PROSPECTIVE REGISTRATION AND REGISTERED REPORTS.**
>
> **SUB-RECOMMENDATION 2.2B: NIH SHOULD SET UP A DEDICATED PROGRAM TO EVALUATE THE DATA GENERATED FROM THE PROJECTS RECOMMENDED IN 2.2A AND GUIDE FUTURE ADOPTION OF PROSPECTIVE REGISTRATION PRACTICES IN PRECLINICAL ANIMAL STUDIES.**

Because adoption of prospective registration in animal studies has been limited thus far, much remains unknown about its specific benefits, its cost impact to individual researchers and the research enterprise, and its potential unexpected consequences. A path to broad adoption will require a stronger evidence base. However, based on the current lack of both understanding and appetite in the community, we believe that specific incentives will be required to generate sufficient data. We therefore recommend that NIH:

(i) identify funding vehicles to incentivize prospective registration and Registered Reports for appropriate studies within a funded research proposal, and
(ii) develop an evaluation program to study the potential impact of these interventions and formulate a recommendation on future adoption of prospective registration.

Prospective Registration Projects

In order to gather data regarding the feasibility and the utility of prospective registration to improve rigor and transparency in animal research specifically, we recommend that NIH support a prospective registration platform (possibly by partnering with existing registries) to allow any NIH grantee and applicant to prospectively register a research and analysis plan. If prospective registration were required, it would apply to hypothesis-testing studies that are typically a subset of studies described in a multi-year grant or are central to translational projects. Thus, to gather as much data as possible during an initial trial, we specifically recommend that:

- NIH leadership should issue a call to action to all NIH grantees, incentivizing grantees to use the platform for prospective registration of studies within pre-existing grants (with the option to embargo public release).
- Funding Opportunity Announcements (FOAs) specifically targeted at translational projects should serve as a testing ground for another prospective registration program. For a subset of such grants that have one or more Aims in which the proposed therapeutic is tested in animal models, NIH could require prospective registration of these specific studies. Prospective registration could be required as the first milestone in the study and described in a "Prospective registration plan" to be submitted with the grant application; the following milestone would be the association of the prospective registration with the publication of the study results.

Registered Reports Projects

The Registered Report publication format not only directly addresses publication bias but also presents strong benefits for researchers in the form of peer review of the proposed study design and a guaranteed publication regardless of study outcome. We recommend that this mode of publication be considered, where relevant, in NIH's efforts to generate data. For example, NIH could encourage researchers to consider the Registered Report mode of publication, and it could incentivize its use by allowing evidence of an accepted "in-principle" Registered Report (i.e., post protocol peer-review but prior to publication of results) to be cited as an outcome in progress reports and grants, and to influence reviewing and funding deliberations. Linking published Registered Reports to their grant IDs would allow the evaluation program to retrospectively evaluate how effective this publication format can be for reducing bias in the conduct and publication of studies.

Evaluation program

Prior to initiating a vehicle for gathering data on outcomes associated with prospective registration, we recommend that NIH establish a dedicated standing committee or task force to:

- Establish goals for collecting evidence related to prospective registration,

- Specify or create funding programs that provide both a vehicle for data collection and approaches to analyzing outcomes,
- Track the status of prospective registration plans submitted through annual NIH Research Performance Progress Reports, and
- Publish a report of the outcomes of Prospective Registration Projects and Registered Reports Projects, specifying potential benefits and burdens of prospective registration at the NIH level, identifying funding programs for future prospective registration projects (assuming the data justifies further action), and clearly stating the future directions of the NIH prospective registration program.

Theme 3: Improve Selection, Design, and Relevance of Animal Models

Evolutionary conservation of mammalian biology is a fundamental feature supporting the wide use of animals as models for normal and diseased human biology. Animal models enable researchers to study biological and pathobiological processes at scales and under conditions untenable in humans. Our ability to recapitulate human-like functional and morphologic phenotypes provides significant confidence in their human relevance. Yet human biology is complex, integrated, and individually variable. Because animal biology does not perfectly recapitulate humans, many animal models therefore imperfectly represent likely human outcomes.

A relatively small number of animal species have become staples in biomedical research representing varying degrees of human genetic, physiologic, and behavioral homology. Non-human primates generally represent the pinnacle of human-relatedness, though their use is minimized as much as possible due to significant ethical sensitivities. Rodents, particularly mice, represent the most common mammals used in biomedical research due to their small size, lower cost, and suitability for genetic manipulation. Other animal models including zebrafish and invertebrates such as fruit flies and nematode worms, for example, have also been useful for elucidating basic physiological processes.

Despite many similarities, differences in molecular, cellular, and organ systems level physiology exist across species. Improving our approaches to animal model design and selection will require a multi-disciplinary approach involving comparative animal scientists (representing animal biology), basic biomedical researchers (representing the questions and mechanism) and clinicians (representing the human context and outcomes of interest). When the scientific goal is direct translation to humans, an animal model should be characterized to ensure that fundamental elements of human biology are reasonably represented, and that modulation would reflect a human response. Animal model selection should be rationalized by an evidence-based description of the human relevance of the model within the context of the experimental intent [34, 35].

High rates of failure in the development of novel therapeutics whose progression to human clinical trials was supported by animal studies has prompted concern that animal models of the human condition are more imperfect than presumed [36-38]. That concern has instigated a useful and critical reflection on how and why we use animals in research as models for either humans or for particular aspects of human biology or behavior. Design and selection of animal models is one area of consideration and we have made the following recommendations to strengthen the process.

Recommendation 3.1: NIH should establish a framework for rationalizing the scientific and, when appropriate, translational (human) relevance of an animal model and its selection. This framework

SHOULD BE EMPLOYED AS PART OF THE JUSTIFICATION FOR ANIMAL USES IN GRANT APPLICATIONS AND INCLUDED IN ETHICAL REVIEW PROCESSES AND IN JOURNAL REPORTS.

Designing or choosing an animal model is not a straightforward undertaking. For many disorders, mechanisms of human disease cannot easily be translated [34, 39, 40]. In other cases, non-human species may not possess the biological features that are the target of modeling. Yet there is no established, standardized approach for selecting an animal model, and there is considerable variability in the approaches used to justify the use of a particular model in grant applications, journal articles, and other fora.

We recommend that NIH establish a framework for researchers to use to explain and justify the scientific relevance of their animal model and, when appropriate, its translatability to human processes and/or diseases. Such a framework would let applicants effectively justify their selected animal models and explain why the animal is appropriate for the specific scientific questions of interest. A standard framework would also help reviewers to evaluate the relevance of the animal models, which is especially important when the research seeks to help create new therapies, medical procedures, or diagnostics. With such a framework, reviewers would be better able to evaluate the scientific and public health impact of grant applications.

A useful framework could include a ‘points to consider’ document or a ‘question-based analysis’ that is provided with the supporting guidance for grant applications. It could include prompts to provide a brief summary of the known human biology or pathobiology, how that biology is represented in the animal, and historical experience with extrapolating outcomes from the model to humans. A few authors have proposed evidence-based frameworks that could guide these efforts [41-44].

RECOMMENDATION 3.2: NIH SHOULD ESTABLISH OR IDENTIFY VENUES FOR THE EXCHANGE OF INFORMATION RELATED TO ANIMAL MODEL DESIGN AND CHARACTERIZATION, STUDY DESIGN, AND GENERAL BEST PRACTICES.

While there are numerous avenues supporting and encouraging explication and exploration within specific scientific fields of inquiry or disease areas, there is little intellectual space devoted to understanding and promulgating the principles of proper design of animal models *per se*. The biomedical research community could benefit from dedicated pathways and processes for information exchange, training, and data collation and curation specifically related to the design and characterization of animal models. These various activities would be aimed at generating, maintaining, and disseminating a knowledge base focused on issues of rigor, reproducibility, and translatability in animal models. These activities would facilitate the exchange of data and experimental approaches that would promote the evaluation of disease models. Suitable venues could include journals, websites, or data platforms, and might involve pre-competitive consortia with private partners. The NCATS-funded Clinical and Translational Science Awards provide an existing infrastructure that could be leveraged.

RECOMMENDATION 3.3: NIH SHOULD WORK TO IMPROVE THE DESIGN OF ANIMAL MODELS THROUGH THE FUNDING OF FOCUSED RESEARCH PROGRAMS THAT ENHANCE UNDERSTANDING OF COMPARATIVE HUMAN-ANIMAL BIOLOGY.

A valuable target of support by NIH would be the funding of specific, cross-disciplinary scientific efforts that aim to develop and compare animal models for their utility in answering scientific and translational questions. These efforts would aid many fields by ensuring that comprehensive and unbiased processes exist to evaluate models. These efforts to develop, characterize, and validate animal models of disease would also provide useful support and resource to disease-focused researchers. Spread across multiple laboratories and enabling collaboration across disciplines, these programs could also enhance reproducibility and enable effective and efficient training of the next generation of scientists. One mechanism for these efforts could be to employ centers of excellence (COE) charged with providing leadership, best practices, research, support and/or training with a specific focus area. Phenotypic and molecular characterizations as well as the outcomes of model validation studies could be shared in public resources such as those suggested in Recommendation 3.2. Continuing education courses in animal model design and use could be developed and provided.

RECOMMENDATION 3.4: NIH SHOULD PROVIDE ADEQUATE RESEARCH SUPPORT FOR LARGER AND LONG-LIVED NON-RODENT-SPECIES WHEN JUSTIFIED.

SUB-RECOMMENDATION 3.4A: NIH SHOULD CREATE POLICY TO ACCOMMODATE LONGER TIME FRAMES AND HIGHER BUDGETS FOR LARGER AND LONG-LIVED NON-RODENT-SPECIES.

SUB-RECOMMENDATION 3.4B: NIH SHOULD CONTINUE TO DEVELOP NATIONAL RESOURCES TO PRODUCE LARGE AND LONG-LIVED ANIMALS.

If grant funding needs to remain within the R01 budgets of $500,000 (the limit per year without traversing an additional approval process), then the cost of acquiring and supporting large and long-lived animals, particularly nonhuman primates, will limit longitudinal studies and will limit the ability to conduct experiments with sample sizes that provide sufficient statistical power. In studies of aging, for example, a 5-year funding period is insufficient for measurement of aging-related changes or pathological consequences of aging-related disorders. Therefore, NIH should provide both funding strategies and grant durations related to the species and experimental question being studied in larger and long-lived animals.

Often, discussions regarding the use of animals in research include a tacit argument that the use of small animals (or those with short lifespans) will save financial resources. In this view, the cost of each experiment or grant is the primary unit of measurement and the lesser cost of smaller animals is evident. An alternative framework is to consider the costs to answer a particular research question or human health problem. In this view, a single large nonhuman primate study may cost substantially more than a mouse study but, if it can more effectively answer the research question and solve the human health challenge, it represents a better investment of resources.

The need to train young investigators in the care and use of larger and long-lived animals is crucial to using such species in translational or pre-clinical studies but this is also resource intensive. Particularly important is training in how to choose the animal model that is most pertinent to the human disorder or disease of concern rather than vice versa. For example, this issue has been considered in relation to cardiology research [45] as well as for selecting optimal species for efficacy assessment in drug development [42, 44].

Gaps in funding are particularly problematic for research with larger and long-lived animals. It is expensive to maintain such colonies, difficult to restart them, and almost impossible to start one *de novo,* in the absence of a significant infusion of funds and institutional support (e.g., space, human resources support). Therefore, policies that ensure that facilities for such animals can maintain their colonies and retain staff with the relevant expertise across any funding hiatus are needed. Strategic investments in existing resources (e.g., the National Primate Centers [46]) and/or generating new models of support for both large centers and smaller laboratories/programs are needed. Stagnant budgets to create and support national non-human primate breeding programs, for example, have significantly handicapped such centers, leading, in some cases (including COVID-19), to national shortages of experimental animals

Recommendation 3.5: NIH should educate the public on the value of animal research, including the important roles of long-lived, non-rodent mammals for translation to improved human health and disease.

Animals are essential in many areas and have produced medical advances that enhance the lives of both humans and the animal species originally studied. Animal research allows scientists to study animals throughout the entire life cycle and across generations within a manageable timeframe. Laboratory conditions allow researchers precise control over the animals' environment and control experimental variables. Additionally, research with larger and long-lived animals, including non-human primates, has enabled the discovery and development of treatments and interventions for a variety of diseases. These research models have been instrumental to significant scientific and medical advances including, but not limited to, deep brain stimulation to treat Parkinson's disease and experimental vaccines to prevent Ebola, polio, and COVID-19. NIH should take the strongest stances possible in supporting the need for research that uses animals properly by taking steps to educate the public on the value of this research regarding its translational relevance to cures for human disease. An additional focus should be on explaining the specific value of larger and long-lived animals.

There is also a need to improve the awareness and dissemination of existing resources and research networks provided to scientists using different species and working in different research areas to maximize translatability of research findings, especially for work with larger and long-lived animals. Initiatives such as the sharing of non-published large animal data (example: Open Source for Nonhuman Primate Optogenetics) and creation of large databases (example: National NHP DNA Bank) should be expanded and promoted. As well, current initiatives for optimizing animal and tissue sharing need to be expanded and promoted across institutions and beyond National Primate Research Centers.

Recommendation 3.6: NIH should charter a high-level Working Group on 'non-animal modeling systems in biomedical research' to complement the activities and recommendations of this ACD Working Group.

Human cell-based modeling has also been a staple of biomedical research for over 50 years where 2D *in vitro* cultures of immortalized human cells have been the norm. These systems have enabled important investigation of human biology and disease at the cellular level without the need for animal studies, but they also have significant limitations depending on the question being asked. However, the rapid progress in our ability to model more complex elements of human biology in human-derived and *in vivo*-relevant modeling systems (e.g., human-derived microphysiological systems, organoids) is newly enabling study at the more complex tissue and organ level [47-49]. Albeit still reductionist in overall biological complexity, these modeling systems enable more efficient (e.g., time, cost) mechanistic investigation that could precede and complement animal studies or even replace them, depending on the intent of the research. Alternatives to animal models may be considered when a mechanistic question is amenable to a less complex modeling platform or when a translationally relevant animal model is lacking. The working group we envision would ideally be constituted by experts with experience in both developing and applying computational, *in vitro,* and *in vivo* modeling systems. The group could examine similar questions regarding rigor, reproducibility, transparency, and relevance in the use of these systems, which likely share some considerations with animal studies but also come with their own sources of variance and uncertainty.

Theme 4: Improve Methodological Documentation and Results Reporting

To achieve results reproducibility, methodological reproducibility must first be achieved. Researchers must provide enough details about the study details and procedures so that independent investigators can repeat the study exactly. Methodological documentation and standards (e.g., ARRIVE [50, 51], CONSORT [52-56], MDRA [16]) for reporting results have been developed by the research community; these standards emphasize reporting critical elements in study design, sample size, inclusion and exclusion criteria, randomization, blinding, outcome measures, statistical methods, experiential animals, experiential procedures, and results. However, these standards are not currently enforced or required earlier on in the life of the study (e.g., grant applications or study design). Strengthening these elements across the life of a study, from planning to execution and publication, will result in a higher quality knowledge base and will better inform future research.

RECOMMENDATION 4.1: NIH SHOULD EXPECT THAT KEY SUPPORTING DATA REPORTED ON ANIMAL RESEARCH SUBMITTED IN SUPPORT OF GRANT APPLICATIONS WILL INCLUDE MEASURES OF QUALITY AND UNCERTAINTY FOR REPORTED ESTIMATES AND AN INTERPRETATION OF EFFECT SIZES WITHIN THE CONTEXT OF THE FIELD.

For statistical results to be understood and properly interpreted, quantitative estimates must be accompanied by a measure of their quality or uncertainty, such as the standard deviation, confidence interval, credible interval, or posterior distribution of the cited estimate. Quantitative estimates are also most useful to other researchers when they are placed in the proper context, for example, by using standardized effect sizes or interpreting reported effect sizes within the context of the field. We recommend that NIH expect this information for key supporting data reported on animal research.

RECOMMENDATION 4.2: NIH SHOULD EXPECT ALL VERTEBRATE AND CEPHALOPOD ANIMAL RESEARCH TO INCLUDE THE ARRIVE 2.0 ESSENTIAL 10 AT THE PUBLICATION STAGE.

The ARRIVE guidelines 2.0 were published during this working group’s activities [50, 51]. In this update to the ARRIVE guidelines, authors improved the clarity of the original guidelines and prioritized them into two groups: the ARRIVE Essential 10, and the ARRIVE Recommended Set. The Essential 10 describes the information that the authors describe as “the basic minimum to include in a manuscript, as without this information, reviewers and readers cannot confidently assess the reliability of the findings presented.” This list of information includes specific details about comparison groups and the experimental units, sample sizes, inclusion and exclusion criteria, randomization, blinding, outcome measures, statistical analysis methods, animals used, experimental procedures, descriptive statistics, and effect sizes and confidence intervals. We recommend that NIH expect that all manuscripts reporting on vertebrate and cephalopod research supported by NIH include the ARRIVE Essential 10.

RECOMMENDATION 4.3: NIH SHOULD ENCOURAGE AND SUPPORT WORK TO BETTER UNDERSTAND, MONITOR, RECORD, AND REPORT IMPORTANT EXTRINSIC FACTORS RELATED TO ANIMAL CARE THAT MAY IMPACT RESEARCH RESULTS.

SUB-RECOMMENDATION 4.3A: NIH SHOULD PROVIDE EDUCATION ABOUT THE IMPORTANCE OF EXTRINSIC FACTORS TO THE RESEARCH COMMUNITY, PROVIDE A METHOD TO REPORT SUCH FACTORS AND INCENTIVIZE PILOT STUDIES TO FURTHER IDENTIFY WHICH-EXTRINSIC FACTORS ARE IMPACTFUL TO REPRODUCIBILITY.

SUB-RECOMMENDATION 4.3B: NIH SHOULD ESTABLISH A TASK FORCE TO IMPLEMENT THE CATALOGING OF EXTRINSIC FACTORS AS DATA FROM PILOT STUDIES ARE GATHERED.

SUB-RECOMMENDATION 4.3C: NIH SHOULD DEDICATE FUNDS FOR CONTROLLED RANDOMIZED TRIALS TO TEST THE EFFECT OF POTENTIALLY HIGH-VALUE EXTRINSIC FACTORS IDENTIFIED FROM PILOT STUDIES AND TASK FORCE RECOMMENDATIONS.

Obtaining reproducible results in animal studies requires researchers to undertake a critical evaluation of all aspects of the study design that have an impact on experiences of the animals themselves. To reproduce animal experiments requires that researchers document and share critical information inherent to the animals (e.g., species, strain, and sex) but also those extrinsic factors of the animals' environment (e.g., ambient temperature, microbiota, lighting levels) that systematically influence the experimental outcomes.

Much remains unknown about exactly which extrinsic factors in animal studies should be documented, with what temporal granularity documentation is needed, and how documentation should be shared for a study to be methodologically reproducible. Further, no systematic characterization of the effects of various extrinsic conditions on different biological factors has been conducted. Without a better understanding of the impact of extrinsic factors on specific animal models or diseases, research outcomes will continue to suffer from unexplained differences both between and within laboratories and studies. We therefore recommend that NIH dedicate funds for the investigation into of high-value extrinsic factors.

To raise awareness and understanding of the importance of extrinsic animal care factors, we recommend that NIH commit to educating investigators on appropriate animal data recording and reporting, which can occur in partnership with animal resource experts at their institutions. Investigators should be aware of how to obtain environmental and husbandry data points, and to present impacts of deviations to standard housing and husbandry variables or standard operating procedures when necessary. Importantly, so as not to overburden investigators, much of the data regarding animal environmental conditions are currently tracked within existing animal program records (e.g., AAALAC program descriptions, daily housing room checklists) and can be provided to research teams by those involved in the delivery of animal care, including attending veterinarians or other animal care staff, IACUC administrators and grants and research program leadership. With consistent access to records of extrinsic factors, investigators can retain key information for their data files, and if deviations from expected outcomes occur, these can be explored, addressed, and reported in research findings, per the ARRIVE 2.0 guidelines [50, 51].

One potential avenue for the disclosure of relevant extrinsic factors might be within the NIH Research Performance Progress Reports (RPPRs). NIH should consider how reporting key extrinsic factors in RPPRs, especially in a computable format, might aid in compiling information regarding the importance of these factors in experimental outcomes and/or facilitate sharing of these data more broadly. NIH should maintain a website or similar cloud-based resource to standardize the documentation of extrinsic factors and permit sharing of extrinsic factor data to the scientific community.

We also recommend that NIH develop a taskforce or working group to critically assess which extrinsic factors should be cataloged, identify how information could be stored and retained, and develop and fund the necessary technology for investigators to report these extrinsic factors so that they are computable and harvestable for data analysis. In addition, materials to educate and incentivize investigators about what factors should be reported and recorded, how to handle changes in conditions over the course of the award, and how and why to report this information in publications should be provided.

The taskforce should explicitly include end-users from the community of laboratory animal veterinary specialists, preclinical animal researchers who represent different animal systems, scientific publishers, and experts in machine learning and artificial intelligence to develop tools for assessing metadata. The charge of the task force should include the following:

- Determine how data are to be recorded, formatted, summarized, stored securely, and accessed/mined;
- Identify benefits, costs, and possible burdens involved in recording extrinsic factors;
- Assess what measures of impact can and should be systematically recorded;
- Review and advise incentives and enforcement mechanisms to ensure accurate reporting.

As a future goal, we recommend that NIH dedicate funding to study the potential impact of animal housing and husbandry environment on different conditions, such as behavior and the microbiome. As part of this project, NIH should:

- Establish goals for collecting data and how outcome variables are influenced by animal facility environmental factors (e.g., noise, vibrations, light, temperature);
- Specify or create funding programs that provide both a vehicle for data to be collected and approaches to analyzing data from the outcomes, with the resulting information published for public access.

RECOMMENDATION 4.4: NIH SHOULD PROVIDE SUPPORT FOR DOCUMENTING LARGER AND LONGER-LIVED ANIMALS' LONGITUDINAL EXPERIMENTAL, MEDICAL, AND HUSBANDRY HISTORIES.

> **SUB-RECOMMENDATION 4.4A: NIH SHOULD FORMALIZE FUNDING MECHANISMS TO LONGITUDINALLY RECORD AND MANAGE ANIMAL-LEVEL EXPERIMENTAL, MEDICAL, AND HUSBANDRY HISTORY METADATA FOR LARGER AND LONGER-LIVED ANIMALS.**
>
> **SUB-RECOMMENDATION 4.4B: NIH SHOULD IDENTIFY MINIMAL ANIMAL-LEVEL EXPERIMENTAL, MEDICAL, AND HUSBANDRY HISTORY METADATA THAT WOULD BE LONGITUDINALLY RECORDED.**
>
> **SUB-RECOMMENDATION 4.4C: NIH SHOULD ENCOURAGE THE SHARING OF ANIMAL-LEVEL EXPERIMENTAL, MEDICAL, AND HUSBANDRY HISTORY.**

With long-lived species, it is often the case that the group of people working with a particular animal will not remain constant across the animal's lifetime. It is therefore essential to encourage and support good record-keeping for each animal. The community of researchers using larger and long-lived species must coalesce around the goal of establishing standards for maintaining useful records not only of the experimental histories of each animal but also of their medical histories and of the locations and characteristic of their housing. Medical records for large long-lived species generally are already being kept by veterinarians and often by researchers as well. These records should remain with each animal even if transferred. NIH can assist by bringing stakeholders together to identify minimal animal-level experimental, medical, and husbandry history metadata that would be longitudinally recorded. This additional record-keeping improves methodological reproducibility and provides additional context when interpreting results. This documentation should be financially supported through NIH funding. Finally, NIH should communicate to researchers who use large animals the importance in sharing animal level experimental, medical, and husbandry history metadata. This documentation is in the hands of the researchers and veterinarians who work with the animals.

Theme 5: Measure the Costs and Effectiveness of Efforts to Improve Rigor, Transparency, Reproducibility, and Translatability

The recommendations in this report have the potential to reshape the way we use animal models in biomedical research over the coming decade. Taken together they comprise a roadmap, with numerous recommendations structured for staged implementation (Appendix 4). We expect that the process of implementation will evolve and adjust in midcourse, informed by rigorous evaluation of both costs and effectiveness across different parts of the diverse NIH animal research portfolio.

While uncertainty is inherent in science, one certainty is that with few exceptions modern rigorous science requires financial support. As is true in many areas of life, scientists underestimate costs and time. A recent effort to test replicability of experiments reported in 50 high-impact science papers needed to be scaled back by more than a third because costs turned out to be much higher than expected [57]. Likewise, most changes in the oversight and conduct of animal research outlined in the four preceding Themes have cost implications, with increased financial costs to NIH being easy to imagine in concept yet difficult to predict in numbers. The changes recommended in this report will also bring opportunity costs. That is, given a finite NIH budget, increased animal costs needed to improve rigor will likely create opportunity costs for non-animal portions of the NIH enterprise. And within animal research, the financial costs of greater rigor applied in pre-clinical animal studies could lead to fewer grants or, alternatively, to grants with diminished scope.

We anticipate that counterbalancing these immediate increased costs will be longer-term savings. This will stem from outcomes such as increased scientific success rates and an improved efficiency resulting from building upon more rigorous findings. For example, if rigor is enhanced for research moving to the clinical research sphere, it is reasonable to expect there will be fewer failures of translation. These cost savings will open new scientific opportunities, and the savings from improved efficiency can be reinvested. As with costs, however, the magnitude of these savings is difficult to predict quantitatively.

NIH will therefore need to devise and implement ways to quantify costs, in dollars and in research opportunity, and to measure savings and success (e.g., the extent of improvement in rigor) for the changes it implements in pursuit of improved rigor. This perceived need prompted us to conduct preliminary integrative analyses of scientific rigor and grant economics by leveraging an existing set of over 4000 published research papers that had been coded for measures of rigor [58]. These initial analyses suggested that enhanced rigor, especially sample size estimation, may indeed be associated with increased grant costs (Appendix 5). This pilot study also highlighted challenges with obtaining the kinds and amounts of data needed, including their computability, and drew attention to interpretive difficulties in discriminating between correlative and causal associations of rigor signatures with costs.

Our overall conclusion is that NIH should proactively design and lead evaluation of actions it takes in pursuit of rigor and use the results to hone its oversight practices and policies. Key activities will include identifying critical computable data and providing the means to obtain it, making specific plans for analysis, and appropriately evolving its ongoing actions based on interim results. While it is important for NIH to lead, we do not imagine that evaluation is a job that NIH can or should do alone. The research community should participate in the evaluative process from its design phase through the end analyses.

RECOMMENDATION 5.1: NIH SHOULD DEVELOP AN EVALUATION PROGRAM TO ASSESS THE PROGRESS IN IMPLEMENTING THE REPORT RECOMMENDATIONS, THEIR EFFECTS ON NIH AND THE RESEARCH COMMUNITY, AND CHALLENGES THAT ARISE IN IMPLEMENTING RECOMMENDATIONS.

Measuring improvements in quality is not a trivial activity. Decisions made by the NIH in support of the field of reproducibility research will have broad and far-reaching consequences. We were mindful of this and balanced our recommendations with a strong directive to NIH that implementation be staged, thoughtful, and periodically evaluated. As pilot studies recommended in this report or independently undertaken are completed and recommendations are implemented, NIH must be committed to ongoing review of the efforts. Detailed evaluation plans should be developed to, as much as possible, ensure that policies and strategies are having the intended effects. Considering that implementation of some recommendations may take significant time and resources, transparency and accountability is essential. It may be necessary for NIH to adjust course when implementing a recommendation outlined in this report. When such cases arise, NIH should use the accumulated evidence and work with stakeholders to modify strategies to achieve the desired goals.

RECOMMENDATION 5.2: NIH SHOULD EXTERNALLY SUPPORT AND INTERNALLY CONDUCT ANALYSES ON ELEMENTS OF RIGOR AND TRANSPARENCY IN GRANT APPLICATIONS AND PUBLICATIONS TO EXAMINE THEIR FINANCIAL COSTS, OPPORTUNITY COSTS, AND IMPACT ON PORTFOLIO BALANCE

> **SUB-RECOMMENDATION 5.2A: NIH SHOULD IDENTIFY AND COLLECT COMPUTATIONALLY EXTRACTABLE INFORMATION FROM GRANT PROPOSALS AND REPORTS ON POTENTIALLY IMPORTANT VARIABLES, INCLUDING PUBLICATION METRICS, METHODOLOGICAL RIGOR, FUNDING, INVESTIGATOR CAREER STAGE, INVOLVEMENT OF STATISTICIANS, EXPERIMENTAL DESIGN DESCRIPTIONS, NUMBERS AND SPECIES OF ANIMALS AND CONDUCT EXTENSIVE ANALYSES ON THESE DATA.**
>
> **SUB-RECOMMENDATION 5.2B: NIH SHOULD ALLOW APPLICANTS TO INCLUDE TEXT IN THE BUDGET JUSTIFICATION SECTION ON HOW PROJECTED ANIMAL BUDGETS ARE LINKED TO EFFORTS TO ENHANCE TRANSPARENCY, RIGOR, AND REPRODUCIBILITY.**

NIH should support research on the resource implications of replication of animal research and of the implementation of measures to enhance rigor, transparency, and translatability. Through its evaluation of efforts aimed at improving rigor, NIH should build a foundation of evidence and analysis methods to enable ongoing monitoring and evolution of its rigor and reproducibility practices. Overall, NIH should monitor trends and changes in the nature, amount, and costs of its animal research model portfolio. Additional attention should be paid to changes in portfolio balance (e.g., between hypothesis generating and hypothesis testing work; or between discovery, pre-clinical and clinical work.) In addition, NIH and/or others should be charged to specifically design and conduct analyses on sets of publications whose level of rigor has been or could be systematically assessed to extend and report how the level of rigor and other signatures of scientific quality are associated with other pertinent variables such as investigator career stage, involvement of statisticians, experimental design descriptions in grant proposals, use of registration.

As a starting point, NIH should define what additional data on animal research it needs to collect and what high priority questions it should ask to quantify financial costs and opportunity costs associated with higher levels of rigor. We discussed several entry points listed below. We note that a standing working group composed of all stakeholders with appropriate technical expertise could be chartered to aid NIH in designing and executing its evaluations:

- Sample Size: More rigorous prospective sample size estimation will likely lead to increases in the number of required animals. Previous work has shown that scientists rarely report sample size estimation, and many, if not most, experiments may be substantially underpowered [59, 60]. If each experiment requires more animals, grant costs may increase, especially if the number of experiments per grant remains unchanged. There may need to be a culture shift at all levels in which each grant is expected to report fewer experiments.
- Conduct of experiments: Increasing rigor may require investments of more time and financial resources devoted to changes in the set-up, conduct, and monitoring of experiments to including more detailed reporting and/or control of extrinsic factors (including feed, microbiome, housing, climate); greater use of pre-application scientific review (see Recommendation 1.2) ; more intensive preparation and review of documents; increased use of multi-center protocols; and a greater use of contract research organizations. NIH, or its appointed study group, should assess what data could/should be obtained to objectively measure these effects.
- Model Selection: There may be a shift toward developing and using multiple animal systems, each designed to better answer a specific set of scientific questions or to model specific aspects of a human disease. Additionally, as researchers focus on selecting models that answer specific research questions, especially for translation to humans, we may see greater use of larger animals, including non-human primates. For example, the American Heart Association reviewed the different causes of heart failure in various animal systems and recommended a shift away from rodent models after it found that critical features of heart failure in patients did not align with the current rodent model [34]. NIH should gather data on the number, costs, and effectiveness of adding new animal models, retiring others, and of employing non-animal models.
- Infrastructure: Development of infrastructure to support quality studies, prospective study design and analysis documentation (e.g., prospective registration), and enhanced data management, curation, storage, and sharing are expected to initially increase the cost of animal use. These costs should be evaluated relative to possible gains in efficiency that may balance the infrastructure investment over time.
- Spillover effects: There may be other effects outside of the context of specific grant applications and awards. Some scientists may choose to abandon certain areas of research because they and their institutions perceive new rigor requirements as overly restrictive, burdensome, or expensive. Institutions that are less well-resourced may make policy decisions to de-emphasize research involving animal models, leading to a distortion that narrows investigator diversity and institutionally concentrates which scientists are able to do animal research.

- Other challenges: At both the micro- and macro-level, it is difficult to determine costs linked with specific experiments. Papers describing NIH-funded science often reference support from multiple sources, including more than one grant and more than one funding body. While publication and grant databases include a great deal of administrative data and metadata, other types of valuable information may be hard to glean and may require either manual curation or sophisticated text-based machine-learning approaches.

As these recommendations are implemented across the NIH it will be important to allow, and even encourage, applicants to include text in the Budget Justification section to address how projected budgets are linked to efforts to enhance transparency, rigor, and translatability. Additional justification can include costs associated with enhanced transparency, rigor, and translatability. This could induce justification of statistical personnel or services, the costs associated with replication, and transparent documentation and prospective registration. The explanations should focus on how each budget item is required to achieve enhanced transparency, rigor, and reproducibility.

RECOMMENDATION 5.3: NIH SHOULD IDENTIFY SCIENTISTS WHO DEMONSTRATE THE HIGHEST LEVELS OF TRANSPARENCY AND RIGOR TO HELP DEFINE ENTERPRISE BEST PRACTICES.

During our deliberations, we heard from several presenters on the science of rigor. As part of many presentations we saw examples of effective practices they have adopted to improve rigor and enhance transparency in their laboratories (Appendix 6). NIH should identify scientists at all career stages and across scientific disciplines who consistently demonstrate the highest transparency and rigor. They should be recognized explicitly for their contributions to best practices, and their lessons and best practices should be promulgated by NIH.

Conclusion

Successfully coupling strong scientific rigor and transparency with human curiosity and imagination is foundational to high-quality science. The frontiers of biomedical science are advancing rapidly, driven by breakthroughs that include facile gene editing, big data mining, Artificial Intelligence, and high-resolution imaging. Such progress is transforming animal research and presenting more opportunities for translation into improved human health outcomes. Yet the ability to interrogate increasingly complex animal biology comes with formidable challenges - both long-standing and new - to improve and more fully implement the best practices of the scientific method. This equally requires that researchers respect the ethical imperatives of studying animal subjects. We therefore strongly urge that NIH and the biomedical community coalesce around the shared mission to improve the rigor and transparency of research by partnering to test, implement, evaluate, and refine recommendations made here. The roadmap we outline aims to spark multiple actions, some of which can begin almost immediately. Overall, it is intended to guide a process that will play out over the coming decade to deliver lasting improvements. NIH must regularly communicate with the wider scientific community and the general public about the successes, obstacles, surprises, and evolution of these efforts. Our working group concludes its efforts with optimism that it is the right moment for NIH and the research community to embark on this joint endeavor with commitment, care, and necessary resources. We believe that all stakeholders will be rewarded with an unprecedentedly robust chain of new knowledge and improved health outcomes.

References

1. NIH. *Impact of NIH Research - Our Stories*. [cited 2021; Available from: https://www.nih.gov/about-nih/what-we-do/impact-nih-research/our-stories.
2. National Research Council (U.S.). Committee for the Update of the Guide for the Care and Use of Laboratory Animals., Institute for Laboratory Animal Research (U.S.), and National Academies Press (U.S.), *Guide for the care and use of laboratory animals*. 2011, National Academies Press,: Washington, D.C. p. xxv, 220 p.
3. *H.R.34 - 21st Century Cures Act*. 12/13/2016 [cited 2021; Available from: https://www.congress.gov/bill/114th-congress/house-bill/34.
4. NIH. *RIGOR AND REPRODUCIBILITY*. [cited 2021; Available from: https://www.nih.gov/research-training/rigor-reproducibility.
5. Use, I.f.L.A.R.R.o.S.a.W.i.L.A. *The Missing "R": Reproducibility in a Changing Research Landscape Workshop In Brief*. 2015 [cited 2021; cientific progress is achieved by robust experiments that generate reliable and reproducible results to be used with confidence by the research community. Recent publications have drawn attention to an apparent and concerning prevalence in the number of peer-reviewed studies that cannot be reproduced, particularly those containing data from experiments using animals and animal models. At this workshop1, researchers from around the world explored the many facets of animal-based research that could contribute to irreproducible results, including perspectives on improving experimental planning, design, and execution; the importance of reporting all methodological details; and efforts to establish harmonization principles of reporting on the care and use of animals in research studies. What follows is a factual summary of the presentations and discussions at the workshop.]. Available from: https://www.nap.edu/read/21835/#slide1.
6. Baker, M., *1,500 scientists lift the lid on reproducibility.* Nature, 2016. **533**(7604): p. 452-4.
7. Loken, E. and A. Gelman, *Measurement error and the replication crisis.* Science, 2017. **355**(6325): p. 584-585.
8. Behaviour, N.H. *Social Sciences Replication Project*. [cited 2021; Available from: https://www.nature.com/collections/nfkchhxllx.
9. *OSF reproducibility project*. [cited 2021; Available from: https://osf.io/search/?q=reproducibility%20project&page=1.
10. Goodman, S.N., D. Fanelli, and J.P. Ioannidis, *What does research reproducibility mean?* Sci Transl Med, 2016. **8**(341): p. 341ps12.
11. National Academies of Sciences Engineering and Medicine (U.S.). Committee on Reproducibility and Replicability in Science, et al., *Reproducibility and replicability in science*. A consensus study report of the National Academies of Sciences, Engineering, Medicine. 2019, Washington, DC: National Academies Press. xxi, 234 pages.
12. Goodman, S.N., *A comment on replication, p-values and evidence.* Stat Med, 1992. **11**(7): p. 875-9.
13. Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, B., and Economic Sciences, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. 2015.
14. NIH/NIGMS. *Clearinghouse for Training Modules to Enhance Data Reproducibility*. [cited 2021; Available from: https://www.nigms.nih.gov/training/pages/clearinghouse-for-training-modules-to-enhance-data-reproducibility.aspx.
15. *Did a change in Nature journals' editorial policy for life sciences research improve reporting?* BMJ Open Science, 2019. **3**(1): p. e000035.

16. Macleod, M., et al., *The MDAR (Materials Design Analysis Reporting) Framework for transparent reporting in the life sciences.* Proc Natl Acad Sci U S A, 2021. **118**(17).
17. *Recommendations for Simplifying R01 Review Criteria: Report from the Clinical Trials Criteria Working Group to the CSR Advisory Council*. 2021 [cited 2021; Available from: https://public.csr.nih.gov/sites/default/files/2021-04/SRC_WG2_workgroup_report_final.pdf.
18. du Sert, N.P., et al., *The Experimental Design Assistant.* Nat Methods, 2017. **14**(11): p. 1024-1025.
19. Percie du Sert, N., et al., *The Experimental Design Assistant.* PLoS Biol, 2017. **15**(9): p. e2003779.
20. Crossley, N.A., et al., *Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach.* Stroke, 2008. **39**(3): p. 929-34.
21. Hirst, J.A., et al., *The need for randomization in animal trials: an overview of systematic reviews.* PLoS One, 2014. **9**(6): p. e98856.
22. Holman, C., et al., *Where Have All the Rodents Gone? The Effects of Attrition in Experimental Research on Cancer and Stroke.* PLoS Biol, 2016. **14**(1): p. e1002331.
23. Macleod, M.R., et al., *Risk of Bias in Reports of In Vivo Research: A Focus for Improvement.* PLoS Biol, 2015. **13**(10): p. e1002273.
24. Landis, S.C., et al., *A call for transparent reporting to optimize the predictive value of preclinical research.* Nature, 2012. **490**(7419): p. 187-91.
25. Sena, E.S., et al., *Publication bias in reports of animal stroke studies leads to major overstatement of efficacy.* PLoS Biol, 2010. **8**(3): p. e1000344.
26. Wieschowski, S., et al., *Publication rates in animal research. Extent and characteristics of published and non-published animal studies followed up at two German university medical centres.* PLoS One, 2019. **14**(11): p. e0223758.
27. Kimmelman, J. and J.A. Anderson, *Should preclinical studies be registered?* Nat Biotechnol, 2012. **30**(6): p. 488-9.
28. Swanson, N., et al., *Research Transparency Is on the Rise in Economics.* AEA Papers and Proceedings, 2020. **110**: p. 61-65.
29. Nosek, B.A. and S. Lindsay, *Preregistration Becoming the Norm in Psychological Science*, in *Observer*. 2018, Association for Psychological Science.
30. Miguel, E., et al., *Social science. Promoting transparency in social science research.* Science, 2014. **343**(6166): p. 30-1.
31. Zarin, D.A., et al., *Update on Trial Registration 11 Years after the ICMJE Policy Was Established.* N Engl J Med, 2017. **376**(4): p. 383-391.
32. Kaplan, R.M. and V.L. Irvin, *Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time.* PLoS One, 2015. **10**(8): p. e0132382.
33. Ioannidis, J.P., A.L. Caplan, and R. Dal-Re, *Outcome reporting bias in clinical trials: why monitoring matters.* BMJ, 2017. **356**: p. j408.
34. Houser, S.R., et al., *Animal models of heart failure: a scientific statement from the American Heart Association.* Circ Res, 2012. **111**(1): p. 131-50.
35. Maeso-Diaz, R., et al., *New Rat Model of Advanced NASH Mimicking Pathophysiological Features and Transcriptomic Signature of The Human Disease.* Cells, 2019. **8**(9).
36. Begley, C.G. and J.P. Ioannidis, *Reproducibility in science: improving the standard for basic and preclinical research.* Circ Res, 2015. **116**(1): p. 116-26.
37. Prinz, F., T. Schlange, and K. Asadullah, *Believe it or not: how much can we rely on published data on potential drug targets?* Nat Rev Drug Discov, 2011. **10**(9): p. 712.
38. Begley, C.G. and L.M. Ellis, *Drug development: Raise standards for preclinical cancer research.* Nature, 2012. **483**(7391): p. 531-3.

39. Cavaillon, J.M., M. Singer, and T. Skirecki, *Sepsis therapies: learning from 30 years of failure of translational research to propose new leads.* EMBO Mol Med, 2020. **12**(4): p. e10128.
40. Dyson, A. and M. Singer, *Animal models of sepsis: why does preclinical efficacy fail to translate to the clinical setting?* Crit Care Med, 2009. **37**(1 Suppl): p. S30-7.
41. Denayer, T., T. Stöhr, and M. Van Roy, *Animal models in translational medicine: Validation and prediction.* New Horizons in Translational Medicine, 2014. **2**(1): p. 5-11.
42. Ferreira, G.S., et al., *Correction: A standardised framework to identify optimal animal models for efficacy assessment in drug development.* PLoS One, 2019. **14**(7): p. e0220325.
43. Everitt, J.I. and B.R. Berridge, *The Role of the IACUC in the Design and Conduct of Animal Experiments that Contribute to Translational Success.* ILAR J, 2017. **58**(1): p. 129-134.
44. Ferreira, G.S., et al., *A standardised framework to identify optimal animal models for efficacy assessment in drug development.* PLoS One, 2019. **14**(6): p. e0218014.
45. Janssen, P.M.L. and M.T. Elnakish, *Modeling heart failure in animal models for novel drug discovery and development.* Expert Opin Drug Discov, 2019. **14**(4): p. 355-363.
46. NIH/ORIP. *National Primate Research Centers Consortium*. 2021]; Available from: https://orip.nih.gov/resource-directory/national-primate-research-centers.
47. Marx, U., et al., *Biology-inspired microphysiological systems to advance patient benefit and animal welfare in drug development.* ALTEX, 2020. **37**(3): p. 365-394.
48. Low, L.A. and D.A. Tagle, *Microphysiological Systems (Tissue Chips) and their Utility for Rare Disease Research.* Adv Exp Med Biol, 2017. **1031**: p. 405-415.
49. Ewart, L., et al., *Navigating tissue chips from development to dissemination: A pharmaceutical industry perspective.* Exp Biol Med (Maywood), 2017. **242**(16): p. 1579-1585.
50. Percie du Sert, N., et al., *Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0.* PLoS Biol, 2020. **18**(7): p. e3000411.
51. Percie du Sert, N., et al., *The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research.* PLoS Biol, 2020. **18**(7): p. e3000410.
52. Moher, D., et al., *The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials.* JAMA, 2001. **285**(15): p. 1987-91.
53. Moher, D., et al., *The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials 2001.* Explore (NY), 2005. **1**(1): p. 40-5.
54. Moher, D., et al., *The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials.* BMC Med Res Methodol, 2001. **1**: p. 2.
55. Moher, D., et al., *The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials.* Ann Intern Med, 2001. **134**(8): p. 657-62.
56. Moher, D., et al., *The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials.* J Am Podiatr Med Assoc, 2001. **91**(8): p. 437-42.
57. Science, C.f.O. *Reproducibility Project: Cancer Biology (RP:CB) Overview*. [cited 2021; Available from: https://www.cos.io/rpcb.
58. Ramirez, F.D., et al., *Journal Initiatives to Enhance Preclinical Research: Analyses of Stroke, Nature Medicine, Science Translational Medicine.* Stroke, 2020. **51**(1): p. 291-299.
59. Button, K.S., et al., *Power failure: why small sample size undermines the reliability of neuroscience.* Nat Rev Neurosci, 2013. **14**(5): p. 365-76.
60. Ramirez, F.D., et al., *Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation.* Circ Res, 2017. **120**(12): p. 1916-1926.
61. Begg, C., et al., *Improving the quality of reporting of randomized controlled trials. The CONSORT statement.* JAMA, 1996. **276**(8): p. 637-9.

62. Turner, L., et al., *The influence of CONSORT on the quality of reporting of randomised controlled trials: an updated review.* Trials, 2011. **12**(1): p. A47.
63. Kline, R.B., *Beyond significance testing : statistics reform in the behavioral sciences*. 2013, Washington, DC: American Psychological Association. xi, 349 p.
64. Fisher, L.D., *Advances in clinical trials in the twentieth century.* Annu Rev Public Health, 1999. **20**: p. 109-24.
65. Fitzpatrick, B.G., E. Koustova, and Y. Wang, *Getting personal with the "reproducibility crisis": interviews in the animal research community.* Lab Anim (NY), 2018. **47**(7): p. 175-177.
66. Fitts, D.A., *Ethics and animal numbers: informal analyses, uncertain sample sizes, inefficient replications, and type I errors.* J Am Assoc Lab Anim Sci, 2011. **50**(4): p. 445-53.
67. Unger, E.F., *All is not well in the world of translational research.* J Am Coll Cardiol, 2007. **50**(8): p. 738-40.
68. van der Worp, H.B., et al., *Can animal models of disease reliably inform human studies?* PLoS Med, 2010. **7**(3): p. e1000245.
69. Clayton, J.A. and F.S. Collins, *Policy: NIH to balance sex in cell and animal studies.* Nature, 2014. **509**(7500): p. 282-3.
70. Zucker, I. and A.K. Beery, *Males still dominate animal studies.* Nature, 2010. **465**(7299): p. 690.
71. Ramirez, F.D., et al., *Sex Bias Is Increasingly Prevalent in Preclinical Cardiovascular Research: Implications for Translational Medicine and Health Equity for Women: A Systematic Assessment of Leading Cardiovascular Journals Over a 10-Year Period.* Circulation, 2017. **135**(6): p. 625-626.
72. Clayton, J.A., *Studying both sexes: a guiding principle for biomedicine.* FASEB J, 2016. **30**(2): p. 519-24.
73. Reardon, S., *A mouse's house may ruin experiments.* Nature, 2016. **530**(7590): p. 264.
74. *Troublesome variability in mouse studies.* Nat Neurosci, 2009. **12**(9): p. 1075.
75. Crabbe, J.C., D. Wahlsten, and B.C. Dudek, *Genetics of mouse behavior: interactions with laboratory environment.* Science, 1999. **284**(5420): p. 1670-2.
76. Nagy, T.R., et al., *Effect of group vs. single housing on phenotypic variance in C57BL/6J mice.* Obes Res, 2002. **10**(5): p. 412-5.
77. Meakin, L.B., et al., *Male mice housed in groups engage in frequent fighting and show a lower response to additional bone loading than females or individually housed males that do not fight.* Bone, 2013. **54**(1): p. 113-7.
78. Febinger, H.Y., et al., *Effects of housing condition and cage change on characteristics of sleep in mice.* J Am Assoc Lab Anim Sci, 2014. **53**(1): p. 29-37.
79. Sorge, R.E., et al., *Olfactory exposure to males, including men, causes stress and related analgesia in rodents.* Nat Methods, 2014. **11**(6): p. 629-32.
80. Vesterinen, H.M., et al., *Meta-analysis of data from animal studies: a practical guide.* J Neurosci Methods, 2014. **221**: p. 92-102.
81. Hooijmans, C.R., et al., *Meta-analyses of animal studies: an introduction of a valuable instrument to further improve healthcare.* ILAR J, 2014. **55**(3): p. 418-26.
82. Glasziou, P., et al., *Reducing waste from incomplete or unusable reports of biomedical research.* Lancet, 2014. **383**(9913): p. 267-76.
83. Freedman, L.P., I.M. Cockburn, and T.S. Simcoe, *The Economics of Reproducibility in Preclinical Research.* PLoS Biol, 2015. **13**(6): p. e1002165.
84. Scott, S., et al., *Design, power, and interpretation of studies in the standard murine model of ALS.* Amyotroph Lateral Scler, 2008. **9**(1): p. 4-15.
85. Horn, J., et al., *Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review.* Stroke, 2001. **32**(10): p. 2433-8.

86. Macleod, M.R., et al., *Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality.* Stroke, 2008. **39**(10): p. 2824-9.

87. Lazic, S.E. and L. Essioux, *Improving basic and translational science by accounting for litter-to-litter variation in animal models.* BMC Neurosci, 2013. **14**: p. 37.

Appendices

Appendix 1 – Glossary

Confirmatory Research: hypothesis-testing research to test the validity of an a priori hypothesis

Early-Stage Preclinical Research: Animal research to understand the basis of human biology, disease or disorders and develop interventions. This NIH usage is inclusive, and we note that industry uses the term more narrowly to mean research focused on assessing the efficacy of candidate therapeutics.

Exploratory Research: hypothesis-generating research, research to clarify the exact nature of the problem to be solved

Extrinsic Factors: factors that have a direct impact on the experience of the research animal during the course of experimental phases such as housing, husbandry, handling, feed, water, bedding, enrichment, caging type, light cycles, etc.

Generalizability: whether the results of a study apply in other contexts of populations that differ from the original one

Inferential reproducibility is achieved when researchers draw similar conclusions, or make knowledge claims of a similar strength, from either an independent replication of a study or a reanalysis of the original study. This is part of the process by which a scientific field decides which research claims or effects are to be accepted as true.

Late-Stage Preclinical Research: Research using animals to find out if a treatment is likely to be efficacious. Often done immediately before testing in humans.

Methods/methodological reproducibility is achieved when researchers provide enough detail about study procedures and data so that the same procedures could in theory or, in actuality, be exactly repeated. Transparency is a key component of methods reproducibility.

Prospective Registration: Creating a permanent record of a study design, analytic plan and primary outcome before the data are collected. Prospective registration allows retrospective assurance against selective reporting and outcome switching. The registered research plan can be embargoed for a limited time to protect intellectual property and when the registration is published it allows to identify and mitigate publication bias. Researchers who prospectively register their studies demonstrate their intent to prioritize rigorous reporting. Prospectively registered studies do *not* undergo peer review at the time of registration.

Publication bias: A form of bias in which the outcome of a study influences the decision to publish its results, resulting in the prioritization of positive results and large effects, over null or negative results. Despite the availability of a range of journals and publishing outlets that welcome studies with null and negative results, publication bias is documented, indicating that researchers' behavior and incentive systems contribute to its occurrence.

Reduction: Appropriately designed and analyzed animal experiments that are robust and reproducible, and truly add to the knowledge base

Refinement: Advancing animal welfare by exploiting the latest *in vivo* technologies and by improving understanding of the impact of welfare on scientific outcomes

Registered Reports: Journal article type in which the detailed study protocol is submitted for peer review before the data are collected. Upon successful review, the journal guarantees publication of the article regardless of the study findings. Like prospective registration, Registered Reports mitigates selective reporting and outcome switching. In addition, the protocol peer review can help improve experimental design, for example by ensuing that sample size and inclusion/exclusion criteria are predetermined, and that proper measures are taken to mitigate experimental and analysis bias. Researchers who publish with this review workflow demonstrate their intent to prioritize rigor, and they also secure a path to publication regardless of findings, which mitigates publication bias.

Replacement: Accelerating the development and use of models and tools, based on the latest science and technologies, to address important scientific questions without the use of animals

Results reproducibility is achieved when researchers conduct an independent new study with procedures as close to the original study as possible and obtain the same results. This definition is similar to others' definitions of "replicability."

Scientific Rigor: the strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation, and reporting of results.

Translation: Applying results from preclinical research, usually via late-stage preclinical animal studies, to justify, design and execute trials in humans

Transparency: accessibility of information

Appendix 2 – Outcomes of High Rigor and Managing Expectations of Statistical Analysis

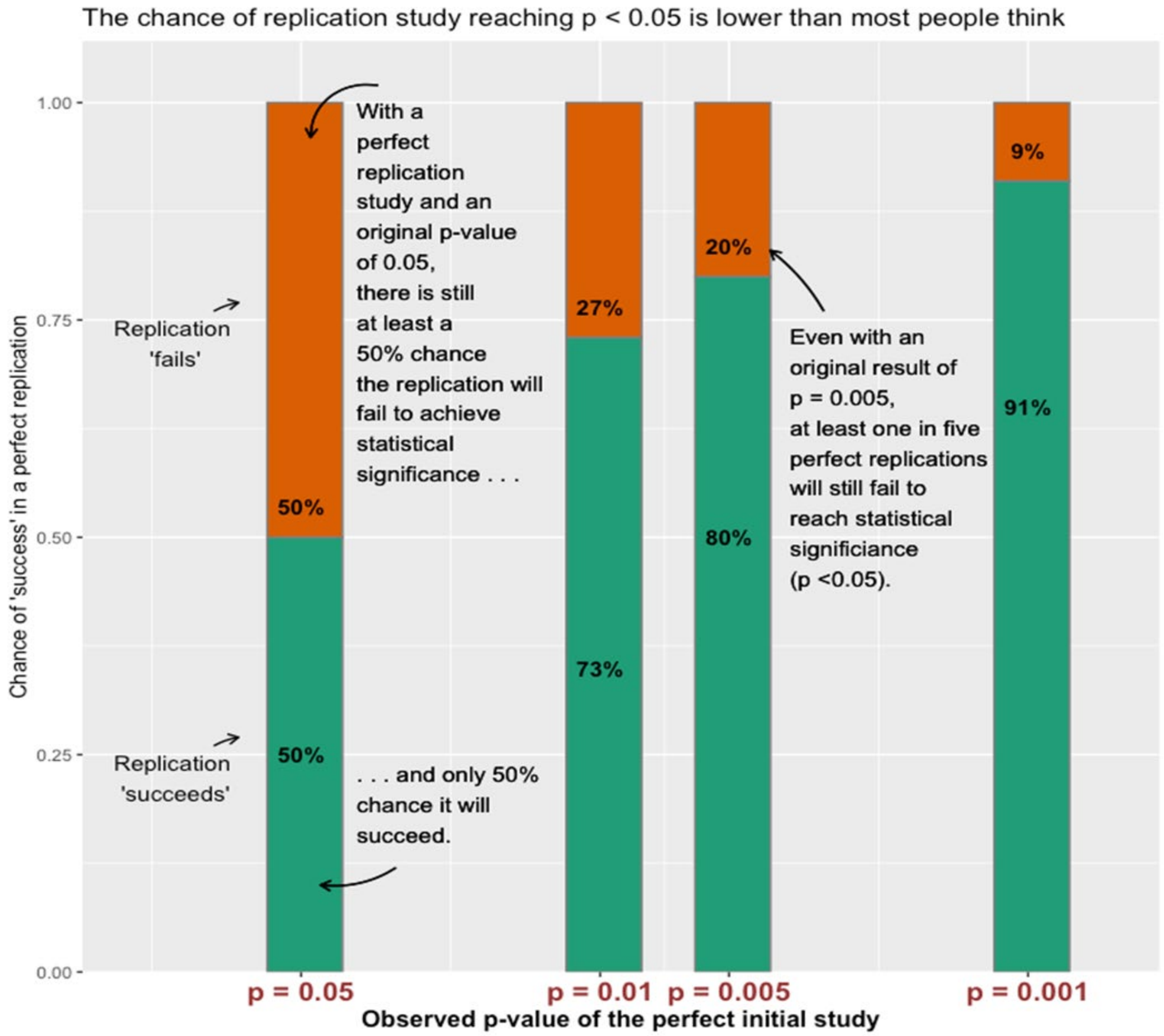


Figure 2. Statistical significance is not enough to judge reproducibility. Given a statistically significant initial study, the chance of a replication "succeeding" (another statistical significance; p < 0.05) is surprisingly low.

Much of good general statistical practice has been refined over the decades, particularly in clinical research [61-64]. Yet preclinical animal research has unique statistical needs and challenges, many of which have not been fully addressed by the statistical and animal research communities. For example, due to ethical considerations and resource constraints, animal researchers are under great pressure to a minimal number of animals in their studies. Yet studies with smaller sample sizes generally have lower statistical power, which leads to more false negatives (overlooking true effects) and more false positives (incorrectly declaring an effect to be statistically significant). To achieve adequate statistical power with the minimum number of animals, which is essential for confirmatory and late-stage preclinical experiments, researchers need to use appropriate experimental design and sample size calculations, but many animal researchers report lacking the knowledge or impetus to do so [65, 66].

Furthermore, although researchers may have more genetic and environmental control over animals than with humans in clinical research, they must still deal with the animal-to-animal variations that occur naturally and from induced injury and disease; this makes sample size calculations crucial for ensuring that experimental effects can be estimated with enough precision to be informative [67, 68]. Similarly, the preferential use of male animals has been a longstanding issue in animal experiments [69-71], which has led to disadvantaging women by skewing our understanding of diseases and the development of potential therapies to disproportionately favor males. The NIH instituted the Sex as a Biological Variable (SABV) in 2016, requiring inclusion of male and female subjects, which increased overall animal use and sample sizes intentionally to aid in translation and applicability to human disease; researchers and their funders must balance the importance of considering sex as a biological variable with its potential impact on statistical power and the ethical use of both sexes of animals in modeling human disorders [72].

Sources of confounding can also arise at virtually every stage of an animal experiment [73-75], such as room temperature [73-75], group housing [76, 77], lighting [73], cage environmental stability [78], and even the sex of the experimenter [79].[76, 77], lighting [73], cage environmental stability [78], and even the sex of the experimenter [79].[80, 81], lighting [73], cage environmental stability [78], and even the sex of the experimenter [79].[76, 77], lighting [73], cage environmental stability [78], and even the sex of the experimenter [79]. Preclinical animal research also lacks established analytic tools, such as the ‘minimally important difference’ common in clinical trials or meta-analytical techniques that can handle the diversity of species and variations in animal study designs [80, 81].

Using animals to model human physiology and disease comes with well-known limitations. Yet even the best animal models are ineffective if the studies employing them also have improper experimental design, inappropriate data analysis, inadequate results reporting, or other statistical problems [36, 82-86]. One analysis, for example, found that 91% of reviewed studies in a particular field failed to use the correct experimental unit in their analyses, leading to incorrect claims that would be impossible to replicate [87]. Another analysis estimated that problems in study design, data analysis, and statistical reporting account for more than 50% of study errors that lead to irreproducibility in U.S. research [83]. [36, 82-86]. One analysis, for example, found that 91% of reviewed studies in a particular field failed to use the correct experimental unit in their analyses, leading to incorrect claims that would be impossible to replicate [87]. Another analysis estimated that problems in study design, data analysis, and statistical reporting account for more than 50% of study errors that lead to irreproducibility in U.S. research [83].

Appendix 3 – Additional Context for Recommendation 1.3

Proposed information to guide applicants in addressing each critical study design elements:

The following items are critical to ensure rigorous research by minimizing risk of bias and fostering transparency. Describe how these items will be addressed for each proposed study design in the application, alongside unambiguous descriptions of animals used, where applicable. These items are also expected to be reported appropriately when communicating study results.

1. **Inclusion/exclusion criteria:** *Describe the criteria that will be used for inclusion or exclusion of samples or animals during the experiments and for data used in analysis.*

2. **Sample-size estimation:** *Provide planned sample sizes for each group and how they were derived.*

3. **Data analysis plan:** *Describe plans for data analysis, including statistical methods as appropriate, designed to answer the proposed scientific questions.*

4. **Blinding:** *Describe measures planned to blind the investigators during group allocation, the conduct of the experiment and the analysis, where applicable. If none taken and blinding is not appropriate to the study design, provide justification.*

5. **Randomization:** *Describe methods planned for random allocation to comparison groups and strategies for random sample processing and collection of data where applicable. Provide rationale if a randomization scheme is not used.*

Resources to support successful completion of additional page

A landing page of FAQs regarding rigor and reproducibility currently exists at nih.gov and the above resources could be added to this site. However, it is the recommendation of the Subcommittee that this page be updated and linked to additional resources for preparation of grant applications.

To support applicants and facilitate satisfactory completion of details required in the additional page, we recommend that resources additionally describing these items be made readily available to applicants on the NIH Grants & Funding portal. Such resources could include:

- **NIH Rigor and Reproducibility resources.** https://www.nih.gov/research-training/rigor-reproducibility/resources

- **NIH Principles and Guidelines for reporting preclinical research**. https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

- **The Experimental Design Assistant (EDA)**

https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2003779

https://www.nc3rs.org.uk/experimental-design-assistant-eda

- **The ARRIVE Guidelines 2.0**

Percie du Sert, N., Hurst, V., Ahluwalia, A. et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biol* 18(7): e3000410 (2020). https://doi.org/10.1371/journal.pbio.3000410

Percie du Sert, N., Ahluwalia, A., Alam, S. et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLoS Biol* 18(7): e3000411 (2020). https://doi.org/10.1371/journal.pbio.3000411

- Landis, S., Amara, S., Asadullah, K. et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490, 187–191 (2012). https://doi.org/10.1038/nature11556

- Chambers, Karen, Andy Collings, Chris Graf, Veronique Kiermer, David T. Mellor, Malcolm Macleod, Sowmya Swaminathan, et al. 2019. "Towards Minimum Reporting Standards for Life Scientists." MetaArXiv. April 30. doi:10.31222/osf.io/9sm4x (including reporting information framework: https://osf.io/xfpn4/ and elaboration document: https://osf.io/xzy4s/

- Nature journals reporting summary: https://www.nature.com/documents/nr-reporting-summary-flat.pdf - As part of Nature Research reporting requirements: https://www.nature.com/nature-research/editorial-policies/reporting-standards#reporting-requirements

Appendix 4 – Report Recommendation Implementation Gantt Chart

As part of our recommendation planning, we developed a possible timeline for the NIH to consider when implementing our recommendations.

Actions	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022	Q3 2022	Q4 2022	Q1 2023	Q2 2023	Q3 2023	Q4 2023	Q1 2024	Q2 2024	Q3 2024	Q4 2024	Q1 2025	Q2 2025	Q3 2025	Q4 2025	Q1 2026	Q2 2026	Q3 2026	Q4 2026	Q1 2027	Q2 2027	Q3 2027	Q4 2027	Q1 2028	Q2 2028	Q3 2028	Q4 2028
Theme 1: Improve Study Design and Data Analysis																															
Recommendation 1.1																															
Sub-recommendation 1.1a			Development																												
Sub-recommendation 1.1b			Developmen	Funding	5-year development project																										
Sub-recommendation 1.1c		Planning (Phase 1 - current systems)			Phased Implementation (phase 1 - current systems)																		Plan (phase 2 - new systems)			Phased Implementation (phase 2 - new systems)					
Recommendation 1.2																															
Sub-recommendation 1.4a		Planning/Budgeting		Pilot FOA				FOAs				FOAs				FOAs															
Sub-recommendation 1.4b		Planning/Budgeting		Pilot FOA				FOAs				FOAs				FOAs															
Sub-recommendation 1.4c		Planning/Budgeting		Pilot FOA				FOAs				FOAs				FOAs															
Sub-recommendation 1.4d		Planning/Budgeting		Pilot FOA				FOAs				FOAs				FOAs															
Recommendation 1.3		Plan		Announce	Phased Implementation																										
Recommendation 1.4		Evaluate				Recs																									
Theme 2: Address Incomplete Reporting and Questionable Research Practices																															
Recommendation 2.1	Report	Planning		Phased Campaign - Education, Awareness													Phased Campaign - Education, Awareness & outcomes														
Recommendation 2.2		Planning/Budgeting		FOAs for Pilots				FOAs for Pilots			multiyear projects and evaluation													Phase 2 Implementation							
Sub-recommendation 2.2a		Planning/Budgeting		FOAs for Pilots				FOAs for Pilots			multiyear projects and evaluation													Phase 2 Implementation							
Sub-recommendation 2.2b		Planning/Budgeting		FOAs for Pilots				FOAs for Pilots			multiyear projects and evaluation													Phase 2 Implementation							
Theme 3: Improve Selection, Design, and Relevance of Animal Models																															
Recommendation 3.1		Planning	Development				Announcemer	Phased Implementation																							
Recommendation 3.2		Planning	Development				Implementation																								
Recommendation 3.3		Planning/Budgeting		Pilot FOA				Multiyear Projects; Evaluation; Planning and Budgeting for Phase 2												FOAs				Multiyear Projects							
Recommendation 3.4		Plan		Announce	Phased Implementation																										
Sub-recommendation 3.4a		Plan		Announce	Phased Implementation																										
Sub-recommendation 3.4b		Plan		Announce	Phased Implementation																										
Recommendation 3.5		Report		Phased Campaign - Education, Awareness																											
Recommendation 3.6		Plan	Announce	Deliberations					Recs																						
Theme 4: Improve Methodological Documentation and Results Reporting																															
Recommendation 4.1		Plan		Announce	Phased Implementation																										
Recommendation 4.2	Report	Announce																													
Recommendation 4.3																															
Sub-recommendation 4.3a		Plan		Phased Campaign - Education, Awareness																											
Sub-recommendation 4.3b		Plan	Announce	Deliberations								Recs																			
Sub-recommendation 4.3c						Planning/Budgeting		FOAs; Planning/Budgeting for Phase 2				FOAs; Multiyear Projects; Evaluation				Multiyear Projects; Evaluation				Multiyear Projects; Evaluation				Multiyear Projects; Evaluation							
Recommendation 4.4																															
Sub-recommendation 4.4a		Planning	Development				Phased Implementation																								
Sub-recommendation 4.4b		Plan and Deliberations		Announce	Phased Implementation																										
Sub-recommendation 4.4c	Report	Announce																													
Theme 5: Measure the Costs and Effectiveness of Efforts to Improve Rigor, Transparency, Reproducibility, and Translatability																															
Recommendation 5.1		Planning	Development				Phased Evaluation																								
Recommendation 5.2																															
Sub-recommendation 5.2a		Planning	Development				Internal and External Analyses																								
Sub-recommendation 5.2b		Planning	Development				Announcemer	Phased Implementation																							
Recommendation 5.3		Analysis and Development of Best Practices																													

Appendix 5 – Preliminary Appendix Report: Potential Financial Ramifications of Working Group Recommendations

Preliminary Appendix Report: Potential Financial Ramifications of Working Group Recommendations

Michael Lauer, Robyn Lee-Stubbs, Sarah Nusser, Regina Nuzzo, Eric Prager

2021-05-23

Introduction[1]

While uncertainty is inherent in science, one certainty is that with few exceptions modern rigorous science requires financial support. As is true in many areas of life, scientists underestimate costs and time. A recent effort to replicate the experiments reported in 50 high-impact science papers was scaled back because costs turned out to be much higher than expected.

The subcommittee engaged in 3 sets of activities:

- Developed a set of possible financial implications of substantive changes to the way that NIH funds and oversees research involving animal models.
- Conducted two *preliminary* analyses of scientific rigor and grant economics by leveraging data from publications in Circulation Research and Stroke of over 5000 published research papers that had been coded for measures of rigor. This *preliminary report* is based on the latter publication since it included a larger sample with publications occurring over a wider range of years.
- Developed recommendations.

Possible Financial Implications

Changes in the funding, oversight, and conduct of research involving animal models may lead to increased costs to the NIH (and therefore opportunity costs for the entire NIH-funded enterprise) fora number of reasons:

- Numbers of animals: Rigorous prospective sample size estimation will likely lead to substantial increases in the number of required animals. Previous work has shown that scientists rarely report sample size estimation and that many, if not most, experiments may be substantially underpowered. If each experiment requires more animals, grant costs may increase, especially if the number of experiments per grant remains unchanged. In other words, there may need to be a shift in expectations whereby each grant (as well as each paper) reports fewer experiments,but each experiment will be properly powered.
- Type of animals: There may be a shift towards greater use of larger animals (including non-human primates) and/or a greater diversity of species used for each scientific question.For example, the American Heart Association described existing animal models for different causes of heart failure and recommended that there be a shift away from rodent models.

[1] Authors are listed in alphabetical order

- Conduct of experiments: Changes in the set-up, conduct, and monitoring of experiments include greater levels of control of living conditions (including feed, microbiome, housing, climate); greater use of pre-IACUC scientific review; more intensive preparation and review of documents; increased use of multi-center protocols (such as the Intervention Testings Program of the National Institute on Aging); greater use of contract research organizations; and development of infrastructures needed to support replication studies, pre-registration, and enhanced data management, curation, storage, and sharing.
- Spillover effects: There may be other effects that will occur outside of the context of specific grant applications and awards. Some scientists may choose to abandon certain areas of research because they and their institutions perceive new rigor requirements as overly restrictive, burdensome, or expensive. Institutions that are less well resourced may make policy decisions to de-emphasize research involving animal models, leading to an increased concentration of research resources, a trend that some see as concerning. On the other hand, it can be argued that poorly designed (including underpowered) research is inherently wasteful and unethical. Furthermore, greater rigor in pre-clinical research should leader to fewer efforts in the clinical research sphere to study the effects of candidate agents or devices that should never have been studied in the first place.
- Other challenges: At both the micro- and macro-level, it is difficult to determine costs linked with specific experiments. Many papers describing NIH-funded science mention support from more than one source – including more than one grant and more than one funding body (e.g. support may come from NIH, non-profit foundations, industry, and/or academic institutions). While publication and grant data-bases include a great deal of administrative data and meta-data, other types of valuable information may be hard to glean, requiringeither manual curation or sophisticated text-based machine-learning approaches.

Preliminary Analyses

F Daniel Ramirez, a member of the ACD Working Group, kindly provided us with data basedon an analysis of rigor of 4162 cardiovascular articles published over 18 years. The articles were selected from 3 journals (*Stroke*, *Nature Medicine*, and *Science Translational Medicine*) that had implemented steps in 2011 to enhance rigor and from 2 "control" journals. All papers described experiments involving nonhuman mammals. Ramirez provided us with data for each paper, including whether there was reporting of randomization, blinding, sample size estimation, and sex used.

The data we received did not include paper ID numbers (e.g. Scopus or Pubmed). We were able to match 4001 papers (or 96%) with Pubmed ID numbers. We used the NIH SPIRES data base to link 2348 papers to 6073 unique NIH grants. We captured the complete history of each grant, including total years of funding up to the year of publication, total grant dollars, and total dollarsduring the most recent year of funding. All funding values were corrected for inflation using the BRDPI (2019 reference value).

In our effort to gain insights into possible fiscal ramifications of rigor recommendations that the ACD WG might make, we leverage 3 types of data:

- Article metrics: relative citation ratio (or "RCR"), and the approximate potential to translate (or "APT"), namely the likelihood that a paper will be cited in a guideline or clinical trial publication.
- Methodological rigor as coded by Ramirez: randomization, blinding, and sample size estimation, as well as reporting of sex.

- Grant measures: number of grants cited, grant-years of prior support, total grant funding up to publication (in inflation-adjusted dollars), most recent grant funding, and use of certain specific mechanisms (U, G, M, P, and Training).

We use these data to consider three possible desirable outcomes:

- Research conducted and reported with methodological rigor: we use data from Ramirez
- Research that is translatable – at least in some instances: we use APT data from the NIH Office of Portfolio Analysis.
- Research that is influential – though this may be particularly controversial: we use RCR data from the NIH Office of Portfolio Analysis.

NIH Funding and Measures of Article Rigor, Influence, and Translatability

We were able to find Pubmed ID numbers for 2348 NIH-funded papers and 1653 funded entirely from other sources.

The NIH Office of Portfolio Analysis has developed a tool by which one can map a set of papers according to the "Triangle of Biomedicine" of Weber. The "Triangle" uses Medical Subject Heading terms to classify papers according to axes of molecular/cellular, animal, or human focus.

Figures 1 and 2 show the triangles for NIH and non-NIH funded papers. As expected, all papers gravitate towards the animal corner. NIH and non-NIH papers appear to be similar by this scheme.

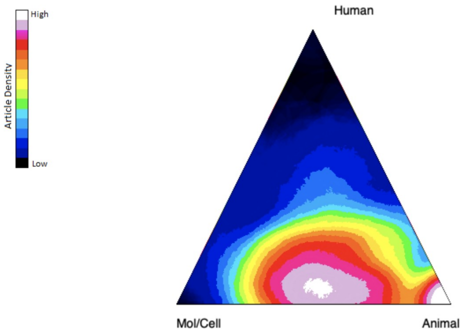


Figure 1: Triangle of Biomedicine for 2348 NIH-funded Papers from Ramirez et al

Table 1 shows article, influence, translatability, and rigor measures according to source of funding.We considered papers to be of higher influence if they had a relative citation ratio (RCR) greater than 3 and to be of higher translatability if the approximate potential to translate was greaterthan 50%. We define rigor points as the number of rigor measures (randomization, blinding, power) described, and define a paper as "More Rigorous" if at least one of these rigor measures is present. We considered articles to be "most rigorous" if all 3 measures of rigor – randomization, blinding, and sample size estimation – were present. NIH-funded papers were more likely to be published in general journals (i.e., *Nature Medicine* and *Science Translational Medicine*) and to use mouse models, and were less likely to use rat models. Influence measures were higher for NIH-funded papers, while translatability measures were similar. NIH-funded papers were less likely to report randomization and blinding, but were more likely to report sample size estimation. Even so, only 11% of NIH-funded paper reported formal sample size estimation.

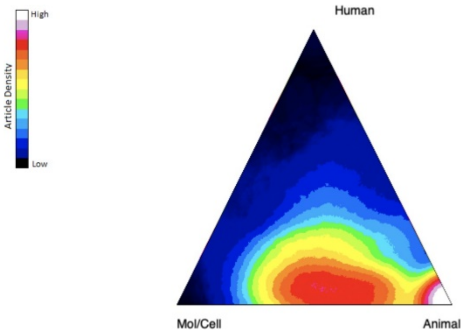


Figure 2: Triangle of Biomedicine for 1653 Non-NIH-funded Papers from Ramirez et al

Table 1: Characteristics of Articles According to Source of Funding

Characteristic / Funding	NIH Funded		Not NIH Funded
Total N (%)	2348 (58.7)		1653 (41.3)
Model	Mouse	1680 (71.6)	1003 (60.7)
	Rat	328 (14.0)	393 (23.8)
	Combination	127 (5.4)	119 (7.2)
	Primate	52 (2.2)	29 (1.8)
	Rabbit	53 (2.3)	26 (1.6)
	Dog	42 (1.8)	30 (1.8)
	Pig	42 (1.8)	27 (1.6)
	Other	24 (1.0)	26 (1.6)
Journal	Stroke	465 (19.8)	518 (31.3)
	Nat Med	562 (23.9)	326 (19.7)
	Sci Transl Med	520 (22.1)	281 (17.0)
	Circulation	422 (18.0)	322 (19.5)
	Circ Res	379 (16.1)	206 (12.5)
Intervention Journal	Yes	1547 (65.9)	1125 (68.1)
Nat Med or Sci Transl Med	Yes	1082 (46.1)	607 (36.7)
Relative Citation Ratio	Median (IQR)	2.9 (1.8 to 5.4)	2.5 (1.5 to 4.5)
Higher Influence	Yes	1142 (48.6)	682 (41.3)
Approximate Potential to Translate	Median (IQR)	0.5 (0.2 to 0.8)	0.5 (0.2 to 0.8)
Higher Translatability	Yes	1126 (48.0)	769 (46.5)
Randomization	Yes	729 (31.0)	581 (35.1)
Blinding	Yes	935 (39.8)	725 (43.9)
Sample Size Estimation	Yes	258 (11.0)	114 (6.9)
Rigor Points	0	1080 (46.0)	701 (42.4)
	1	744 (31.7)	546 (33.0)
	2	394 (16.8)	344 (20.8)
	3	130 (5.5)	62 (3.8)
More Rigorous	Yes	1268 (54.0)	952 (57.6)
Most Rigorous	Yes	130 (5.5)	62 (3.8)

Article Metrics, Rigor Measures, and Grant Characteristics Among NIH-funded Papers

Table 2 shows article metrics, rigor, and grant measures for the 4001 papers according to animal model. As might be expected, papers reporting on primate models and combination models were associated with greater grant costs.

In Tables 3, 4, and 5 we break down article metrics and grant measures according to the specific type of rigor measure. Of note, sample size estimation was associated with greater grant costs (Table 5). Figure 3 shows distribution of most recent grant dollars according to whether there was sample size estimation. Power calculation was the least common measure of rigor, while it appeared to associated with greater costs. The box plots suggest show, as might be expected given the sample imbalance, a wider distribution and a higher mean for papers without a power calculation.

Table 2: Characteristics of Articles and Associated Grants According to Animal Model

Characteristic / Model		Mouse	Rat	Combination	Primate	Rabbit	Dog	Pig	Other
Total N (%)		1680 (71.6)	328 (14.0)	127 (5.4)	52 (2.2)	53 (2.3)	42 (1.8)	42 (1.8)	24 (1.0)
Relative Citation Ratio	Median (IQR)	3.1 (1.8 to 5.7)	2.4 (1.6 to 3.8)	3.8 (2.3 to 6.2)	3.0 (2.0 to 5.8)	2.1 (1.2 to 3.1)	2.0 (1.5 to 3.6)	2.6 (1.7 to 4.6)	2.4 (1.7 to 4.3)
Higher Influence	Yes	863 (51.4)	117 (35.7)	79 (62.2)	26 (50.0)	15 (28.3)	14 (33.3)	18 (42.9)	10 (41.7)
Approximate Potential to Translate	Median (IQR)	0.5 (0.2 to 0.8)	0.5 (0.2 to 0.8)	0.8 (0.5 to 0.8)	0.8 (0.5 to 0.8)	0.5 (0.2 to 0.8)	0.5 (0.5 to 0.8)	0.5 (0.2 to 0.8)	0.5 (0.2 to 0.8)
Higher Translatability	Yes	816 (48.6)	131 (39.9)	73 (57.5)	33 (63.5)	25 (47.2)	20 (47.6)	19 (45.2)	9 (37.5)
Intervention Journal	Yes	1060 (63.1)	260 (79.3)	95 (74.8)	50 (96.2)	35 (66.0)	10 (23.8)	23 (54.8)	14 (58.3)
Nat Med or Sci Transl Med	Yes	874 (52.0)	36 (11.0)	90 (70.9)	47 (90.4)	8 (15.1)	3 (7.1)	17 (40.5)	7 (29.2)
Randomization	Yes	471 (28.0)	122 (37.2)	58 (45.7)	14 (26.9)	19 (35.8)	13 (31.0)	21 (50.0)	11 (45.8)
Blinding	Yes	660 (39.3)	156 (47.6)	51 (40.2)	7 (13.5)	21 (39.6)	9 (21.4)	20 (47.6)	11 (45.8)
Sample Size Estimation	Yes	192 (11.4)	19 (5.8)	24 (18.9)	5 (9.6)	6 (11.3)	2 (4.8)	8 (19.0)	2 (8.3)
Rigor Points	0	791 (47.1)	132 (40.2)	47 (37.0)	31 (59.6)	26 (49.1)	28 (66.7)	14 (33.3)	11 (45.8)
	1	548 (32.6)	107 (32.6)	39 (30.7)	16 (30.8)	13 (24.5)	5 (11.9)	13 (31.0)	3 (12.5)
	2	248 (14.8)	77 (23.5)	29 (22.8)	5 (9.6)	9 (17.0)	8 (19.0)	9 (21.4)	9 (37.5)
	3	93 (5.5)	12 (3.7)	12 (9.4)	0 (0.0)	5 (9.4)	1 (2.4)	6 (14.3)	1 (4.2)
More Rigorous	Yes	889 (52.9)	196 (59.8)	80 (63.0)	21 (40.4)	27 (50.9)	14 (33.3)	28 (66.7)	13 (54.2)
Number of Grants	Median (IQR)	4.0 (2.0 to 6.0)	3.0 (2.0 to 3.0)	4.0 (2.0 to 6.0)	4.0 (2.0 to 6.0)	2.0 (2.0 to 4.0)	2.5 (2.0 to 4.0)	3.5 (2.0 to 5.0)	2.0 (2.0 to 4.0)
Grant-Years of Funding	Median (IQR)	28.0 (13.0 to 50.0)	16.0 (7.0 to 28.0)	26.0 (10.5 to 51.0)	23.0 (9.8 to 48.0)	20.0 (8.0 to 25.0)	12.5 (9.0 to 26.0)	19.0 (11.5 to 37.8)	12.0 (5.0 to 27.0)
Total Grant Funding (2019 $Million)	Median (IQR)	15.5 (4.4 to 48.0)	4.1 (1.8 to 12.0)	18.6 (5.4 to 40.7)	56.5 (9.2 to 150.4)	7.2 (2.8 to 15.2)	11.0 (1.9 to 17.8)	9.0 (3.8 to 22.6)	4.4 (1.4 to 7.2)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	1.9 (0.6 to 5.4)	0.6 (0.3 to 1.2)	2.3 (0.6 to 5.4)	8.1 (1.2 to 16.8)	0.8 (0.4 to 2.6)	0.8 (0.3 to 3.4)	1.1 (0.4 to 2.6)	0.5 (0.3 to 0.9)
P, U, M, or G Mechanism	Yes	975 (58.0)	140 (42.7)	76 (59.8)	42 (80.8)	32 (60.4)	20 (47.6)	20 (47.6)	4 (16.7)
Training Grant	Yes	381 (22.7)	21 (6.4)	28 (22.0)	10 (19.2)	4 (7.5)	5 (11.9)	5 (11.9)	1 (4.2)

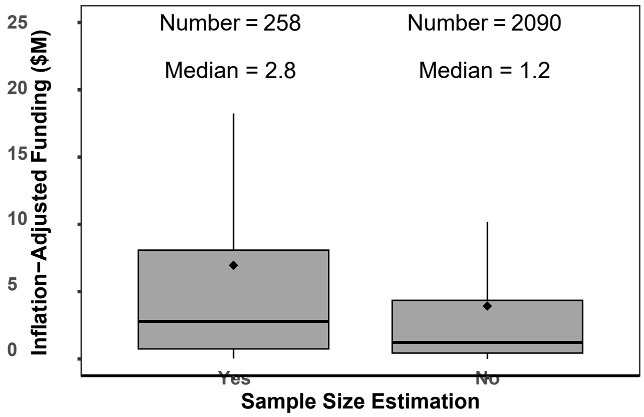


Figure 3: Distribution of recent grant funding according to whether there was sample size estimation

In Table 6 we compare paper and grant history measures according to whether the paper appears to have a higher translatability, which we define as a value > 50%. Articles with higher translatability had higher relative citation ratios, higher rates of randomization, higher rates of sample size estimation, and higher grant costs. In Table 7 we compare article, and grant history measures according to whether the paper appears to have a higher level of influence, which we define as a relative citation ratio (RCR) >3. Higher influence papers had higher rates of sample size estimation and higher grant costs.

Article and Grant Measures According to Sex

Table 8 shows article metrics, rigor, and grant measures according to the sex of animals describedin the papers. Papers that reported on experiments involving males and females had higher grantcosts.

Exploratory Regression Analyses

Table 9 shows exploratory logistic regression analyses of 4 paper-based outcomes, namely randomization, blinding, sample size estimation, and reporting of any of these 3 types of rigor (what we've called "More Rigorous"). None of the models explain the outcomes well. There may be a weak association between grant funding and higher odds of sample size estimation.

At this point, we should keep in mind some important limitations:

- The data are limited to cardiovascular papers published from 2000 to 2017.
- Few papers reported on sample size estimation.
- Other measures of interest likely require manual curation.
- Grant acknowledgments are not standardized.
- While sample size calculation appears to be associated with increased costs (as might be expected), grant costs explain little of the variance.
- These analyses should only be considered as *preliminary*; they present an example of what might be possible, but should not be construed as suggesting or reaching any conclusions.

Table 3: Characteristics of Articles and Associated Grants According to Whether There was Randomization

Characteristic		Randomization	No Randomization
Total N (%)		729 (31.0)	1619 (69.0)
Relative Citation Ratio	Median (IQR)	2.9 (1.8 to 5.5)	2.9 (1.8 to 5.3)
Higher Influence	Yes	356 (48.8)	786 (48.5)
Approximate Potential to Translate	Median (IQR)	0.8 (0.2 to 0.8)	0.5 (0.2 to 0.8)
Higher Translatability	Yes	378 (51.9)	748 (46.2)
Intervention Journal	Yes	615 (84.4)	932 (57.6)
Nat Med or Sci Transl Med	Yes	416 (57.1)	666 (41.1)
Number of Grants	Median (IQR)	3.0 (2.0 to 5.0)	3.0 (2.0 to 6.0)
Grant-Years of Funding	Median (IQR)	25.0 (12.0 to 46.0)	25.0 (11.0 to 44.0)
Total Grant Funding (2019 $Million)	Median (IQR)	13.1 (4.1 to 53.1)	11.3 (3.3 to 34.1)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	1.5 (0.5 to 5.7)	1.3 (0.4 to 4.3)
P, U, M, or G Mechanism	Yes	427 (58.6)	882 (54.5)
Training Grant	Yes	140 (19.2)	315 (19.5)

Table 4: Characteristics of Articles and Associated Grants According to Whether There was Blinding

Characteristic		Blinding	No Blinding
Total N (%)		935 (39.8)	1413 (60.2)
Relative Citation Ratio	Median (IQR)	2.8 (1.8 to 5.2)	3.0 (1.8 to 5.5)
Higher Influence	Yes	438 (46.8)	704 (49.8)
Approximate Potential to Translate	Median (IQR)	0.5 (0.2 to 0.8)	0.5 (0.2 to 0.8)
Higher Translatability	Yes	443 (47.4)	683 (48.3)
Intervention Journal	Yes	687 (73.5)	860 (60.9)
Nat Med or Sci Transl Med	Yes	425 (45.5)	657 (46.5)
Number of Grants	Median (IQR)	3.0 (2.0 to 6.0)	3.0 (2.0 to 6.0)
Grant-Years of Funding	Median (IQR)	24.0 (10.0 to 44.0)	25.0 (11.0 to 46.0)
Total Grant Funding (2019 $Million)	Median (IQR)	10.8 (3.3 to 36.7)	12.5 (3.7 to 41.1)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	1.1 (0.4 to 4.2)	1.6 (0.5 to 5.1)
P, U, M, or G Mechanism	Yes	501 (53.6)	808 (57.2)
Training Grant	Yes	169 (18.1)	286 (20.2)

Table 5: Characteristics of Articles and Associated Grants According to Whether There was SampleSize Documentation

Characteristic		Sample Size Estimation	No Sample Size Estimation
Total N (%)		258 (11.0)	2090 (89.0)
Relative Citation Ratio	Median (IQR)	3.6 (2.1 to 6.4)	2.8 (1.7 to 5.3)
Higher Influence	Yes	152 (58.9)	990 (47.4)
Approximate Potential to Translate	Median (IQR)	0.8 (0.5 to 0.8)	0.5 (0.2 to 0.8)
Higher Translatability	Yes	153 (59.3)	973 (46.6)
Intervention Journal	Yes	239 (92.6)	1308 (62.6)
Nat Med or Sci Transl Med	Yes	204 (79.1)	878 (42.0)
Number of Grants	Median (IQR)	4.0 (2.0 to 6.0)	3.0 (2.0 to 5.0)
Grant-Years of Funding	Median (IQR)	32.0 (17.0 to 58.0)	24.0 (11.0 to 44.0)
Total Grant Funding (2019 $Million)	Median (IQR)	28.2 (7.2 to 80.3)	10.8 (3.3 to 35.2)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	2.8 (0.8 to 8.1)	1.2 (0.4 to 4.4)
P, U, M, or G Mechanism	Yes	171 (66.3)	1138 (54.4)
Training Grant	Yes	66 (25.6)	389 (18.6)

Table 6: Characteristics of Articles and Associated Grants According to Potential to Translate

Characteristic		Higher Translatability	Lower Translatability
Total N (%)		1126 (48.0)	1222 (52.0)
Relative Citation Ratio	Median (IQR)	5.1 (3.2 to 8.1)	2.0 (1.3 to 2.9)
Higher Influence	Yes	866 (76.9)	276 (22.6)
Randomization	Yes	378 (33.6)	351 (28.7)
Blinding	Yes	443 (39.3)	492 (40.3)
Sample Size Estimation	Yes	153 (13.6)	105 (8.6)
More Rigorous	Yes	634 (56.3)	634 (51.9)
Intervention Journal	Yes	836 (74.2)	711 (58.2)
Nat Med or Sci Transl Med	Yes	695 (61.7)	387 (31.7)
Number of Grants	Median (IQR)	4.0 (2.0 to 6.0)	3.0 (2.0 to 5.0)
Grant-Years of Funding	Median (IQR)	27.0 (12.0 to 50.0)	23.0 (10.0 to 41.0)
Total Grant Funding (2019 $Million)	Median (IQR)	16.1 (4.1 to 51.1)	9.6 (3.1 to 29.8)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	1.9 (0.6 to 6.1)	1.1 (0.4 to 3.9)
P, U, M, or G Mechanism	Yes	674 (59.9)	635 (52.0)
Training Grant	Yes	242 (21.5)	213 (17.4)

Table 7: Characteristics of Articles and Associated Grants According to Level of Influence

Characteristic		Higher Influence	Lower Influence
Total N (%)		1142 (48.6)	1206 (51.4)
Relative Citation Ratio	Median (IQR)	5.5 (4.0 to 8.3)	1.8 (1.2 to 2.3)
Higher Translatability	Yes	866 (75.8)	260 (21.6)
Randomization	Yes	356 (31.2)	373 (30.9)
Blinding	Yes	438 (38.4)	497 (41.2)
Sample Size Estimation	Yes	152 (13.3)	106 (8.8)
More Rigorous	Yes	614 (53.8)	654 (54.2)
Intervention Journal	Yes	842 (73.7)	705 (58.5)
Nat Med or Sci Transl Med	Yes	733 (64.2)	349 (28.9)
Number of Grants	Median (IQR)	4.0 (2.0 to 6.0)	3.0 (2.0 to 5.0)
Grant-Years of Funding	Median (IQR)	28.0 (12.0 to 52.8)	23.0 (10.0 to 39.0)
Total Grant Funding (2019 $Million)	Median (IQR)	17.7 (4.3 to 52.5)	8.8 (3.0 to 28.5)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	2.3 (0.6 to 6.4)	1.0 (0.4 to 3.6)
P, U, M, or G Mechanism	Yes	684 (59.9)	625 (51.8)
Training Grant	Yes	261 (22.9)	194 (16.1)

Table 8: Characteristics of Papers and Associated Grants According to Sex of Animals

Characteristic / Sex		Male	Female	Both	Unclear
Total N (%)		962 (41.0)	575 (24.5)	474 (20.2)	337 (14.4)
Relative Citation Ratio	Median (IQR)	2.6 (1.7 to 4.5)	2.9 (1.7 to 5.4)	3.3 (1.9 to 6.2)	3.8 (2.0 to 7.1)
Higher Influence	Yes	406 (42.2)	273 (47.5)	265 (55.9)	198 (58.8)
Approximate Potential to Translate	Median (IQR)	0.5 (0.2 to 0.8)	0.5 (0.2 to 0.8)	0.8 (0.5 to 0.8)	0.8 (0.5 to 0.8)
Higher Translatability	Yes	420 (43.7)	268 (46.6)	243 (51.3)	195 (57.9)
Intervention Journal	Yes	613 (63.7)	333 (57.9)	334 (70.5)	267 (79.2)
Nat Med or Sci Transl Med	Yes	269 (28.0)	273 (47.5)	297 (62.7)	243 (72.1)
More Rigorous	Yes	549 (57.1)	238 (41.4)	283 (59.7)	198 (58.8)
Number of Grants	Median (IQR)	3.0 (2.0 to 5.0)	4.0 (2.0 to 6.0)	4.0 (2.0 to 6.0)	4.0 (2.0 to 6.0)
Grant-Years of Funding	Median (IQR)	22.0 (10.0 to 40.0)	27.0 (11.0 to 48.0)	27.0 (11.2 to 48.8)	28.0 (12.0 to 46.0)
Total Grant Funding (2019 $Million)	Median (IQR)	8.9 (2.9 to 25.5)	12.8 (4.0 to 39.8)	19.5 (5.2 to 52.2)	17.0 (3.7 to 65.4)
Most Recent Grant Funding (2019 $Million)	Median (IQR)	1.0 (0.4 to 3.8)	1.5 (0.5 to 5.1)	2.5 (0.6 to 6.3)	1.8 (0.5 to 7.5)
P, U, M, or G Mechanism	Yes	489 (50.8)	315 (54.8)	309 (65.2)	196 (58.2)
Training Grant	Yes	152 (15.8)	135 (23.5)	97 (20.5)	71 (21.1)

Table 9: Exploratory Logistic Regression Models for Measures of Rigor

	Dependent variable:			
	Randomized	Blinded	Sample Size Estimation	More Rigorous
	(1)	(2)	(3)	(4)
Number of Grants	−0.278	0.270	−0.050	0.115
	(0.256)	(0.244)	(0.380)	(0.238)
Grant Years	−0.053	0.011	0.268	−0.052
	(0.189)	(0.179)	(0.292)	(0.175)
Most Recent Funding	0.420^{***}(0.114)	$−0.244^{**}$	0.576^{***}(0.172)	0.122
		(0.105)		(0.103)
U, M, G, or P Grant	−0.132	0.004	−0.156	−0.074
	(0.130)	(0.120)	(0.200)	(0.118)
Intercept	$−0.591^{***}$	$−0.529^{***}$	$−2.500^{***}$	0.190
	(0.186)	(0.177)	(0.298)	(0.173)
Observations	2,348	2,348	2,348	2,348
R^2	0.011	0.005	0.030	0.002
χ^2 (df = 4)	17.660^{***}	8.813^{*}	35.902^{***}	2.900
Note:	*p<0.1; **p<0.05; ***p<0.01			

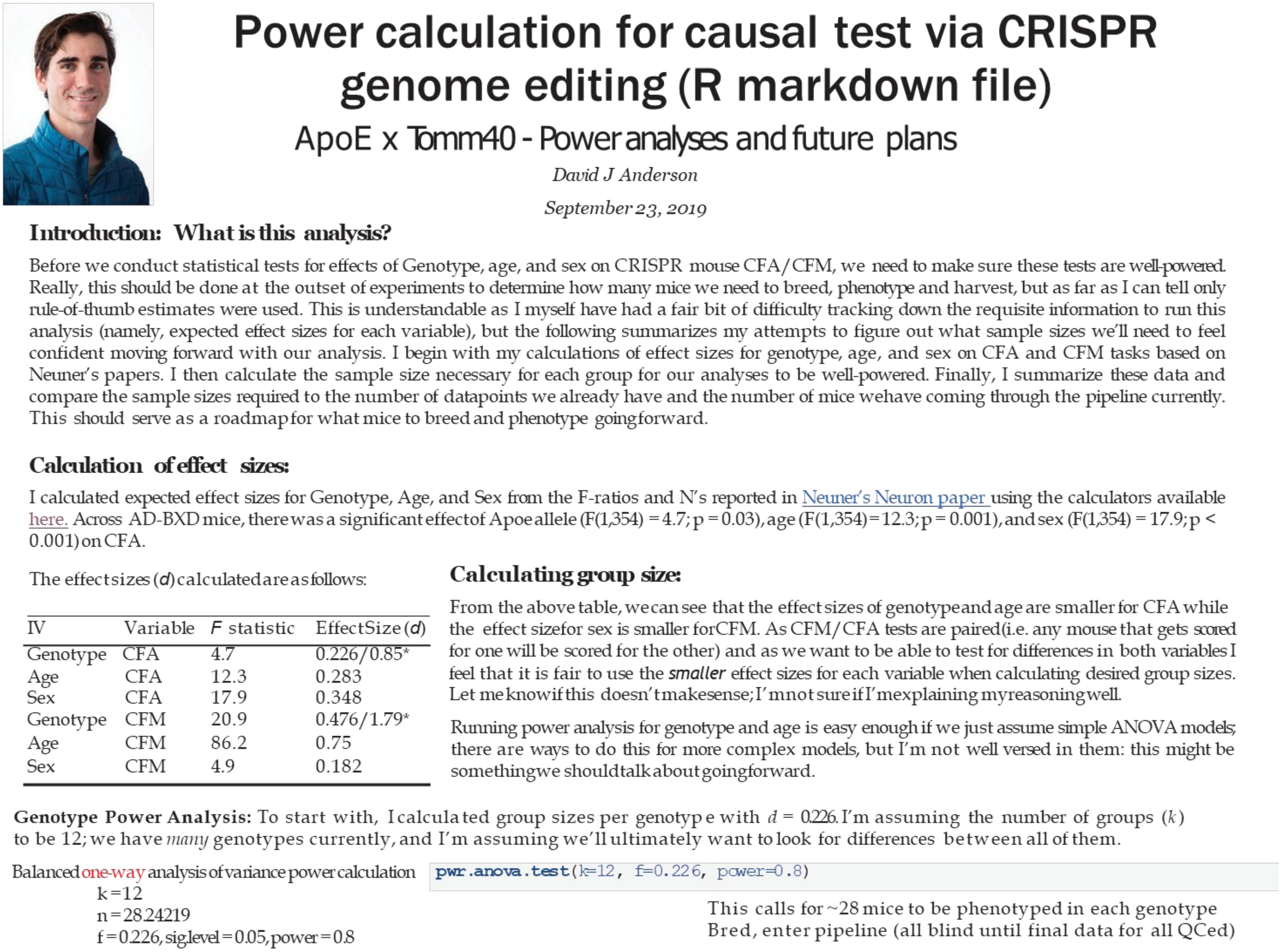
Recommendations

The subcommittee identified a number of possible financial ramifications of changes to the way in which NIH funds and oversees animal research. Informed by those possibilities, we linked methodological rigor data from Ramirez with data on article metrics and grant history. Exploration of these descriptive data suggest that higher levels of rigor do not necessarily imply higher costs, with the exception of sample size estimation. Higher levels of translatability may be associated with higher levels of rigor and higher costs.
The subcommittee offers the following recommendations:

- NIH should develop an evaluation program to assess the progress in implementing recommendations, their effects on both the NIH and the external research community, and challenges that arise in implementing recommendations.
- NIH should conduct and support analyses on the elements and impact of rigor and transparencyin grant applications and publications to better understand the impact of measurable factors on quality, cost and rigor in NIH-funded research.
- NIH should allow applicants to include text in the Budget Justification section on how projectedanimal budgets are linked to efforts to enhance transparency, rigor, and translatability.
- NIH and/or other analysts should consider studies of scientists who demonstrate the highest levels of transparency and rigor; they may offer lessons that could be applied enterprise-wide.

Appendix 6 – Example of Effective Practices to Enhance Rigor and Transparency

The following power calculation was an example provided by Catherine Kaczorowski, Ph.D. whose research laboratory at The Jackson Laboratory focuses collaborative and interinstitutional research program that uses a multidisciplinary approach to identify early causative events underlying 'normal' nonpathological age-related memory decline and Alzheimer's dementia. As part of a presentations from Dr. Kaczorowski we saw examples of effective practices they have adopted to improve rigor and enhance transparency in their laboratories.



Power calculation for causal test via CRISPR genome editing (R markdown file)

ApoE x Tomm40 - Power analyses and future plans

David J Anderson

September 23, 2019

Introduction: What is this analysis?

Before we conduct statistical tests for effects of Genotype, age, and sex on CRISPR mouse CFA/CFM, we need to make sure these tests are well-powered. Really, this should be done at the outset of experiments to determine how many mice we need to breed, phenotype and harvest, but as far as I can tell only rule-of-thumb estimates were used. This is understandable as I myself have had a fair bit of difficulty tracking down the requisite information to run this analysis (namely, expected effect sizes for each variable), but the following summarizes my attempts to figure out what sample sizes we'll need to feel confident moving forward with our analysis. I begin with my calculations of effect sizes for genotype, age, and sex on CFA and CFM tasks based on Neuner's papers. I then calculate the sample size necessary for each group for our analyses to be well-powered. Finally, I summarize these data and compare the sample sizes required to the number of datapoints we already have and the number of mice wehave coming through the pipeline currently. This should serve as a roadmap for what mice to breed and phenotype goingforward.

Calculation of effect sizes:

I calculated expected effect sizes for Genotype, Age, and Sex from the F-ratios and N's reported in Neuner's Neuron paper using the calculators available here. Across AD-BXD mice, there was a significant effectof Apoe allele (F(1,354) = 4.7; p = 0.03), age (F(1,354) = 12.3; p = 0.001), and sex (F(1,354) = 17.9; p < 0.001) on CFA.

The effectsizes (*d*) calculated are as follows:

IV	Variable	*F* statistic	EffectSize (*d*)
Genotype	CFA	4.7	0.226/0.85*
Age	CFA	12.3	0.283
Sex	CFA	17.9	0.348
Genotype	CFM	20.9	0.476/1.79*
Age	CFM	86.2	0.75
Sex	CFM	4.9	0.182

Calculating group size:

From the above table, we can see that the effect sizes of genotypeand age are smaller for CFA while the effect sizefor sex is smaller forCFM. As CFM/CFA tests are paired(i.e. any mouse that gets scored for one will be scored for the other) and as we want to be able to test for differences in both variables I feel that it is fair to use the ***smaller*** effect sizes for each variable when calculating desired group sizes. Let me know if this doesn't makesense; I'mnot sure if I'mexplaining myreasoning well.

Running power analysis for genotype and age is easy enough if we just assume simple ANOVA models; there are ways to do this for more complex models, but I'm not well versed in them: this might be somethingwe shouldtalkabout goingforward.

Genotype Power Analysis: To start with, I calculated group sizes per genotype with $d = 0.226$. I'm assuming the number of groups (k) to be 12; we have *many* genotypes currently, and I'm assuming we'll ultimately want to look for differences between all of them.

```
pwr.anova.test(k=12, f=0.226, power=0.8)
```

```
Balanced one-way analysis of variance power calculation
          k = 12
          n = 28.24219
          f = 0.226, sig.level = 0.05, power = 0.8
```

This calls for ~28 mice to be phenotyped in each genotype
Bred, enter pipeline (all blind until final data for all QCed)