ANNUAL REVIEWS

Annual Review of Animal Biosciences

Advocating for Generalizability: Accepting Inherent Variability in Translation of Animal Research Outcomes

F.C. Hankenson,[1] E.M. Prager,[2] and B.R. Berridge[3]

[1]Division of Laboratory Animal Medicine, Department of Pathobiology, School of Veterinary Medicine and University Laboratory Animal Resources, University of Pennsylvania, Philadelphia, Pennsylvania, USA; email: fclaire@upenn.edu

[2]Research Program Management, Regeneron Pharmaceuticals, Inc., Tarrytown, New York, USA; email: eric.prager@regeneron.com

[3]B2 Pathology Solutions LLC, Cary, North Carolina, USA; email: brberridge@b2pathologysolutions.com

Annu. Rev. Anim. Biosci. 2024. 12:391–410

The *Annual Review of Animal Biosciences* is online at animal.annualreviews.org

https://doi.org/10.1146/annurev-animal-021022-043531

Keywords

animal research, extrinsic factors, generalizability, reproducibility, translation, variability

Abstract

Advancing scientific discovery requires investigators to embrace research practices that increase transparency and disclosure about materials, methods, and outcomes. Several research advocacy and funding organizations have produced guidelines and recommended practices to enhance reproducibility through detailed and rigorous research approaches; however, confusion around vocabulary terms and a lack of adoption of suggested practices have stymied successful implementation. Although reproducibility of research findings cannot be guaranteed due to extensive inherent variables in attempts at experimental repetition, the scientific community can advocate for generalizability in the application of data outcomes to ensure a broad and effective impact on the comparison of animals to translation within human research. This report reviews suggestions, based upon work with National Institutes of Health advisory groups, for improving rigor and transparency in animal research through aspects of experimental design, statistical assessment, and reporting factors to advocate for generalizability in the application of comparative outcomes between animals and humans.

1. INTRODUCTION

Visibility around the importance of rigor and reproducibility in animal research has increased significantly over the last 15 years. Beginning a decade ago, research support organizations, including the Institute for Laboratory Animal Research (ILAR) and the National Centre for the Replacement, Refinement & Reduction of Animals in Research, published guidance to aid in animal research expectations (1) and documentation of experimental design features (2). Shortly after these publications were released, the Director of the US National Institutes of Health (NIH) coauthored an overview of how the NIH intended to restructure aspects of the animal research enterprise to improve reproducibility and translatability (3). One of the most comprehensive scientific societies, the Federation of American Societies for Experimental Biology (FASEB), advocated for these NIH initiatives toward improved reproducibility in their report (4). FASEB articulated that enhancements in scientific design and reporting would be iterative and take due time as an evolving process; however, FASEB detailed the extraordinary number of stakeholders impacted by these attempted improvements and the long-term benefits that would be achieved by adopting these practices into research design methodology and protocol execution. Specifically, beyond the investigators and research staff conducting the animal work, other key professionals include program officers, grant reviewers, professional societies, animal facility personnel, veterinary care staff, regulatory oversight personnel, and journal editors, all of whom provide critical and unique perspectives on animal science and medicine. As recently as 2019, the National Academies of Sciences, Engineering, and Medicine released additional recommendations from their own congressionally mandated study on reproducibility and replicability in science (5).

Following the signing of the 21st Century Cures Act into law in late 2016 [42 U.S.C. 201 (2016)], Section 2039 required the NIH Director to convene a working group under the Advisory Committee to the Director (ACD) to develop policies to enhance the rigor and reproducibility of NIH-funded scientific research. In continuation of these efforts, in 2019, Dr. Collins organized another ACD Working Group (WG) on Enhancing Rigor, Transparency, and Translatability in Animal Research (6), in which the authors of this review participated as members, with shared experiences and outcomes described in further detail below. The intent of the recent NIH ACD WG was to address specifically, at a high level, how NIH might improve the value, rigor, and reproducibility of animal studies. Unfortunately, the vocabulary terms that surround this topic are at times poorly differentiated and confusing, despite efforts to standardize and provide a uniform lexicon (4, 7, 8). To this end, the authors have included a glossary as part of this review, which has been pulled from the literature and supported by the NIH ACD WG (see the sidebar titled Glossary of Terms). Per the NIH, rigor intends to ensure robust and unbiased experimental design, whereas transparency entails sharing of detailed methodology, statistical analysis, broad interpretation, and reporting of results (9). If research results can be reproduced in general, potentially across multiple institutions and laboratories and by differing research scientists, it will validate original findings and further support scientific progression onto the next phase of the research, particularly for translation to human clinical trials and application.

Reproducibility of animal studies requires stakeholders (as described above from FASEB) to undertake a critical evaluation of all aspects of the research plan, including the study design, the relevance of the animal model to the question or hypothesis, and the impact on and experiences of the animals within their environment. The research community has witnessed the evolution of certain factors—once unknown or treated as irrelevant—to be critically important. For example, experiments on both sexes of animals are deemed essential to reveal how this biological variable might influence research outcomes in myriad ways (10). Similarly, disclosure of environmental parameters (called extrinsic factors, such as specifics of animal housing and husbandry details) and rodent strains and genotypes has been explicitly recommended by animal science

GLOSSARY OF TERMS

Blinding: concealment of group allocation from one or more individuals involved in a clinical research study
Extrinsic factors: housing, husbandry, handling, feed, water, bedding, enrichment, caging type, light cycles, etc., that have a direct impact on the research animal's experience during the course of experimental phases
Generalizability: how well the results of a study apply in other contexts, situations, and populations
Inferential reproducibility: 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; part of the process by which a scientific field decides which research claims or effects are to be accepted as true
Methods/methodological reproducibility: ability to obtain consistent results using the same inputs, steps, methods/code, and conditions
Publication bias: form of bias in which the outcome of a study influences the decision to publish its results, resulting in 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
Randomization: process of assigning participants to treatment and control groups, assuming that each participant has an equal chance of being assigned to any group
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 improving understanding of welfare's impact on scientific outcomes
Replacement: accelerating the development and use of models and tools, based on the latest science and technologies, to address important scientific questions without using animals
Replicability: getting consistent or duplicated results when using the same procedures or when asking the same scientific question but where new data are collected; ability to independently obtain consistent results
Results reproducibility: production of corroborating results in a new study, having followed the same methods as the original study
Scientific rigor: strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation, and reporting of results
Statistical analysis: process of collecting and analyzing data to identify patterns and trends and inform decision making
Therioepistemology: study of how knowledge is gained from animal research
Translation: applying results from preclinical research, usually via late-stage preclinical animal studies, to justify, design, and execute trials in humans
Transparency: a range of open practices including registering studies, sharing data, publicly reporting findings, and other means to increase information accessibility
Validity: extent to which results measure what they are supposed to measure

texts (11), as well as by ILAR and FASEB. In FASEB's guidance, the suggestion was made to have investigators include animal facility staff in discussions of relevant aspects of study design and development of procedural checklists to facilitate the review of animal care variables and to denote study-specific variations (4, 12, 13). Comprehensive reporting of all experimental factors cannot guarantee perfect reproducibility of findings, nor should it. However, given the inherent variability in how studies are designed and executed, reporting out the extensive details that comprise animal care will improve the quality of animal research and lead to more robust and translatable outcomes (14). Ultimately, transparency, reproducibility, and translation from animals to humans will help to expand access to scientific findings and accelerate biomedical discoveries. This review advocates

that the generalizability of an animal study (i.e., the ability to apply the outcomes to a variety of situations and populations) is a product of its transparency, reproducibility, and translational relevance and should therefore be a priority for ongoing biomedical discoveries.

2. CONTRIBUTING FACTORS TO LOW REPRODUCIBILITY

A lack of both methodological rigor and transparent reporting contributes to significant impediments in reproducibility across many areas of scientific (biomedical) research and, in turn, reduces one's ability to apply data outcomes to different settings or applications (7, 15–17). Correspondingly, one charge of the NIH ACD WG (6) was identifying gaps and opportunities to improve the rigor, reproducibility, translational validity, and transparency of studies involving animal models by strengthening experimental design, statistical analyses, and data and methodological reporting. In this section, our goal is not to repeat why experimental design factors contribute to irreproducible research but rather, first, to consider the primary challenges and opportunities that the NIH ACD WG identified to increase rigor and reproducibility in animal research (6) and, second, to expand the recommended outcomes of the NIH ACD WG for scientific investigators to adopt a generalizability mindset.

2.1. Experimental Design Factors

Experimental design is the process of generating specific combinations of attributes and levels (factors that comprise the research objects and their possible values or outcomes) in response to specific questions or hypotheses while being aware of alternative approaches (18). Variability is inherent to outcomes in animal research; however, improperly designed and poorly documented study protocols often lead to misunderstood results that slow progress, use animals without statistical justifications, and waste precious resources. For example, though they are critical to experimental design, manuscript publications rarely include a full account of all experimental procedures. Indeed, in 2010, a review of 100 articles published in *Cancer Research* revealed that only 28% of papers reported randomization, and only 2% of papers reported that observers were blind to treatments; further, no publication described how animal numbers were determined (19, 20). Similarly, before the implementation of requirements to report study design elements (e.g., randomization, blinding, sample size estimation) across three major journals in 2011, less than 33% of studies reported randomization, less than 47% reported blinding, and less than 6% reported sample size estimation. The number of articles reporting these factors increased significantly after journal interventions were implemented (21, 22). Yet despite advances in required reporting by journals, fewer than 50% of published studies include rigor criteria such as randomization, blinding, and sample size estimation (23).

Pursuing generalizable research relies upon rigorous study designs and clear reporting to mitigate biases, which can—either intentionally or unintentionally—distort results and lead to incorrect or false conclusions (24). Bias may be introduced during subject selection and assignment, data analysis, and interpretation, as well as at the publication stage. Yet, by developing strong study design and statistical training skills, scientists can be made aware of sources of bias and can work to minimize these effects and ensure the ultimate outcomes are accurate. Across differing study types (e.g., exploratory and confirmatory studies) (25), bias can be reduced and transparency increased by randomly assigning samples to comparison groups; conducting experiments in a blinded fashion; prospectively determining the sample size necessary to achieve a sufficiently (predetermined) powered study; and reporting all details of protocols and study designs, including inclusion or exclusion criteria, prospective statistical approaches, variations to the original study design, housing, and husbandry parameters, and how missing data are to be handled (see **Table 1**; 16, 26–28).

Table 1 Experimental factors to include in the design, execution, and reporting of animal research

Experimental design and materials
1. Describe the procedures used to conduct the experiments in sufficient detail so that others might independently reproduce the results. Be sure to include what was done, how it was done, what was used, when and how often, and where.
a. Is the study exploratory or confirmatory, cross-sectional or longitudinal, and/or a within-, between- or mixed design?
b. Describe the intervention type and the primary, secondary, and tertiary endpoints.
c. Describe biological (e.g., sex, age, strain) and technical (e.g., housing/husbandry factors, batch, experimenter) variables that could potentially influence research outcomes.
d. Include the sequence and timing in which tests will occur and describe counterbalancing procedures (if any).
e. Describe any deviations made from your original design.
f. Clearly describe stopping rules for data collection and rules for invalid data.
g. Report all failed manipulations.
h. If multiple facilities are included, describe similarities, differences, and how the study will attempt to control for potential confounding factors.
i. (Optional) Include a timeline diagram or flowchart to illustrate any complex study design.
2. Include a precise description of how you prepared and used methodological tools and procedures.
a. Describe machines/tools used for assessments and how validity and reliability of samples were assured.
b. Describe the materials used and how they were prepared and cite the original protocols used.
c. Report Research Resource identifiers (RRIDs), PubChem Chemical Identifiers (CIDs), and available validation data, including statements of authentication and possible contamination.
d. Include full citations, links, and references to key biological resources (e.g., antibodies, organism, cell line, plasmid, oligonucleotides).
e. Describe and report where source material, software project/tools, and code can be found (include active DOI links from an open science repository).
f. Describe staff training and measurement logistics.
g. Describe tools and/or procedures used to mitigate attrition or missing data.
Experimental subjects
3. Include the following for either human or animal research:
a. The total number of subjects in each experiment, including the number of animals (or participants) and sex and age at the start of the experiments.
b. Describe how the number of subjects was arrived at and provide details of any sample size calculation, including power analysis for each set of experiments. If no power analysis was performed, either describe how sample sizes were obtained to ensure at least 80% power or document alternative strategies to ensuring sufficient sample was collected.
c. Indicate the number of independent replications of each experiment, when applicable, including how often each experiment was performed. Reporting should be sufficient to distinguish independent biological data points and technical replicates.
d. Describe blinding and randomization procedures. If not applicable, describe why.
e. Describe subject inclusion/exclusion procedures.
f. Address sex as a biological variable and other potential variables.
g. Include a statement of ethical compliance.
h. Describe the animal source, species, and strain used; breeding strategies; inbred and characteristics of transgenic animals; and how this model achieves construct, face, and predictive validity.
i. Address animal housing information, including housing paradigm, cage/pen space requirements (cage size, animal density, brand, cage component, materials, opacity), room temperature and relative humidity, light cycle/duration, intensity and color, bedding substrate (e.g., corncob, paper chips, straw, wood shavings), environmental complexity and enrichment type, diet (type, source, supplements, feeding method/frequency, method of preparation, water quality, type, and supply), veterinary and supportive treatments, chemicals and methods used for sanitation of housing area, and any other key factors.
j. Completely describe husbandry information, including travel to performance site (e.g., vendor source information, environmental factors, intracage/crate measurements, ventilation type) and performance site husbandry (e.g., acclimation period to new facility, intracage measurement, cage change frequency, air exchange or flow rates in cage/room, enclosure density, cleaning chemicals and other sanitation practices, and any deviations to standard husbandry conditions).

(*Continued*)

Table 1 (*Continued*)

Statistics
4. Provide details of the statistical methods used for each analysis, including the following:
a. Describe the primary and secondary outcomes, including which experiments are considered exploratory, whether data are independent (no subjects/specimens are related to each other), or whether conditions are nonindependent or paired.
b. Provide inclusion and exclusion criteria for data, including how outliers and missing data were controlled, whether data were missing at random, and how data would change if eliminated observations were included.
c. Provide a statement of the statistical test used for each relevant figure and panel presented and justification/rationale for each statistical test.
i. Describe methods used to assess whether data will meet assumptions for the specific statistical approach and what will be done if assumptions are not met.
ii. Define test as one- or two-sided and significance threshold and include a definition of center, dispersion, and precision measures (e.g., mean, median, SD, effect size, confidence intervals) and exact value of *N* for each relevant figure and panel presented.
iii. If analyses include a covariate, report the results of the analysis without the covariate.
iv. Report adjustments for transformation and multiple comparisons and provide statements about the factors tested, how post hoc tests were chosen, and post hoc comparisons.
d. Fully report statistics (including exact value of *N*, degrees of freedom, test value, and exact p-value when >0.001).
i. In addition to reporting p-value, which provides only a qualitative statement of whether something is significant, report effect sizes and confidence intervals to provide quantitative assessments of magnitude of the effect and certainty.
e. Include the following additional analyses, if applicable:
i. Analysis of counterbalanced data to ensure no object bias exists
ii. Evaluation of consistency of data from control groups across time
iii. Between-cohort analysis and comparisons to other assays in analogous domains
iv. Disaggregated data for males and females
f. Include software, packages, and libraries (including version) used for analyses.
g. Ensure graphical representation of data is consistent with statistical approach and experimental design and is depicted appropriately.

As a primary outcome from the NIH ACD WG, it was recommended that critical elements of the experimental design and analysis should be reported in study designs and grant and funding renewal applications, in addition to rigorously reporting these factors, as outlined in the ARRIVE 2.0 guidelines, in publications. The original 2010 ARRIVE guidelines were revised after a global survey of animal research literature revealed the need for significant improvement in data reporting to accomplish the goals of reproducibility. The updated guidelines, and their Explanation and Elaboration sections, are the result of an extensive collaborative effort across the scientific community (27). The NIH has reinforced suggestions from the recent ACD WG report with the issuance of NOT-OD-23-057 (February 2023), which encourages use of the Essential 10 checklist in all publications reporting vertebrate animal and cephalopod research.

Compliance with NIH reporting expectations could be achieved by prospectively registering studies to create a permanent record of study designs, analytic plans, and primary outcomes. Alternatively, research protocols can be submitted to journals as a Registered Report (29). This format has been expanding rapidly across journals, with more than 275 journals now offering this type of submission, benefitting researchers whose proposed study designs undergo peer review (30). Additionally, the NIH ACD WG recommended expanding training programs for animal researchers to include reviews on study design and data analytic plans, to encourage investigators conducting animal research to expand their statistical knowledge base by developing relevant curricula for trainees and to encourage trainees conducting animal research to augment their understanding of appropriate study design.

2.2. Inclusion and Exclusion Criteria and Missing Data

Given the complexity of executing animal studies, decisions are necessary as to when to include or exclude specific animals or data points. Explanations for how these criteria are established should be defined and deduced a priori but in many cases are not considered until after the study has commenced or even concluded. In other words, investigators should consider what parameters permit an animal or data point to be included in the final study analysis and what factors must be considered to exclude such information. Without a priori consideration, investigators might fall prey to the influence of excluding data when they do not fit the research story or because they do not align with expected data trends. Understandably, data removal and omission can lead to potential irreproducibility and the inability to generalize study findings.

Data are removed from a published report, referred to as missing data, in two primary ways: through (*a*) dropping outliers, a practice in many laboratories that causes outcomes to become disproportionally large and may lead to the statistically significant effects (28, 31), and (*b*) excluding specific animals or cohorts based upon health reasons or because the treatment might adversely affect a certain animal in a sample (26). To overcome this obstacle, during the design of a study (and subsequently in reporting how data inclusion was defined as part of the final publication), investigators should determine criteria for data use and consider, with input from statistical consultants, whether to perform estimation analyses to limit data misinterpretation and better explain any data that are "missing at random" (32–36). **Table 1** is included as an aid to assist in reminders about data disclosures to mitigate potential sources of bias and promote improved preclinical research outcomes and other key research factors; this aid incorporates key elements from other research guidance, like ARRIVE 2.0, the CONSORT checklist, and PREPARE (13, 23, 26, 27, 37, 38).

2.3. Randomization and Blinding

Improving research accuracy and reducing selection bias require investigators to comprehensively describe the participants or animal subjects and mitigate opportunities to adversely influence observations. A recent meta-analysis revealed that as many as 75% of animal studies do not report any type of randomization or blinding (see the sidebar titled Glossary of Terms), which can taint resulting interpretations (39). As dictated by the NIH ACD WG, a recommendation was made to specifically address critical study design elements, including randomization and blinding, in grant applications (6). In turn, journals are supportive of these endeavors and more commonly require authors to provide details. Further, advances in artificial intelligence software will assist journals with determining when key elements are reported (23). Common randomization techniques can be implemented readily into study designs, including (*a*) simple randomization, which randomizes subjects based on a single sequence of random assignments; (*b*) block randomization, which randomizes participants/subjects into groups of equal sample sizes; (*c*) stratified randomization, which controls and balances baseline characteristics and how potential covariates influence the dependent variable; and (*d*) covariate adaptive randomization, which assigns participants to a specific treatment group by considering covariates and previous participant assignments (40, 41). Although there are advantages and challenges to each unique randomization strategy, it is critical to remember that because animal modeling typically uses smaller sample sizes, block randomization is often superior for between-subject designs because it achieves balance in the allocation of subjects to treatment arms (42). Different randomization approaches should be assessed if a study uses within-subject or crossover designs (43). Regardless of how randomization is determined, one should maintain a consistent approach throughout the study to avoid introducing potential biases.

Though rarely implemented in preclinical research (one study estimates that only approximately 14% of studies report any attempt toward blinding), blinding in preclinical research is

crucial to protect against performance and detection bias and is essential to mitigate effect overestimation (44). Indeed, nonblinded outcome assessments can contribute up to 45% inflation of effect sizes (27). Similar to randomization, there are three levels of blinding: (*a*) In assumed blinding, experimenters have access to the group or treatment codes but do not know how those codes correspond between the groups until the end of the study, reducing the risk of bias; (*b*) partial blinding occurs in situations where blinding cannot be implemented for the entire experiment; and (*c*) full blinding—the preferable approach—requires complete allocation concealment from the beginning to the end of the experiment (43). Despite blinding's critical role in reducing bias throughout the conduct of the experiment and during assessment and analysis (45), it may not always be possible; in this case, experimenters would be expected to articulate why blinding was not incorporated and outline how performance bias was mitigated, if not eliminated.

2.4. Power Analysis and Sample Size Determination

Recent evidence indicates that only ~12.5% of rodent studies are sufficiently powered to provide evidence against a null hypothesis (46). In other words, the actual power of studies is much lower than the planned power, which can be due to sample sizes being selected without robust estimates for minimally important effects (47–50). Of the 3Rs, the concept of reduction is highlighted in experimental design to ensure that the most conservative number of animals is used to answer the research question appropriately. The impact is such that sampling errors can significantly impact the precision and interpretation of results; ethically, this may lead to a waste of animals and other resources and will likely increase unreliability and reduce the generalizability of research results. The sample size is the number of experimental units (e.g., animals, cells, or clusters) that are randomly and blindly assigned to the differing comparison and control groups (51, 52). To overcome sampling errors and determine an appropriate sample size, investigators should conduct a priori power analyses, which will reliably determine how many animal subjects are needed to result in the smallest effect size of interest. Importantly, investigators should report in grant applications and publishable manuscripts as to how the sample size was calculated, including details about services and software used for the analysis to help improve comprehension by other research groups and reduce uninterpretable research practices.

In determining a sample size, three factors are needed, including the alpha value (traditionally reported to be 0.05), the β value (traditionally reported as 0.20, or 80% power), and the effect size. Although the former two variables are often understood, deriving an effect size often causes investigators and trainees confusion. The authors hope to clarify that the effect size is a number (e.g., the difference between the mean of two groups) that expresses the magnitude of the outcome relevant to a specific research question and can be determined in multiple ways. [A powerful resource that details the calculation of effect sizes using differing approaches is cited herein (53).] Usually, the higher the effect size, the larger or stronger the difference or relationship between variables. Once these variables are known, power analysis software, such as G*Power, pwr R package, or jamovi, can assist in determining an appropriate sample size (54–59). If a study is insufficiently powered, investigators can also increase power without increasing sample size by using fewer factor levels; having a more focused hypothesis test; binning continuous variables; and using a factorial arrangement, which occurs when all levels of one factor co-occur with all levels of another factor (60).

2.5. Statistical Approaches

Another variable that can contribute to the inability to expand scientific results into a generalizable mindset is statistical error, which involves the application of incorrect or suboptimal tests or

reporting errors in the p-value (61–63) or, as discussed above, through insufficiently powered studies (47, 51, 61, 64, 65). Issues are compounded by inadequate or selective reporting in journals or grant applications regarding what statistical approach(es) were applied and whether assumptions were tested. Because of a need to publish results biased toward a desired effect, and because publishers prioritize positive results (66), investigators might choose their analytical approach after seeing the data or performing multiple analyses to choose the most favorable outcome, referred to colloquially as "p-hacking" (67). A lack of transparency in reporting how studies are analyzed and whether parameters and data analysis approaches are decided a priori or after data are collected will affect the ability of investigators and readers to appropriately interpret findings, evaluate the robustness of the data, and design future studies based upon those results (26, 39).

The NIH ACD WG highlighted poor statistical practices, whether due to lack of understanding, inadequate training, or intentional data bias, as an area of needed intervention. In fact, the first themed recommendation to the NIH Director was to improve statistical training for scientists. Unfortunately, statistical manipulations are entrenched in preclinical research, including examples of investigators continuing to add samples/subjects until a significant p-value is achieved (68). The p-value measures whether an observed result can be attributed to chance (a qualitative answer), yet it does not address the more important interpretation, which is the magnitude of an effect. Rather than focusing on the probability (or p-value), the NIH ACD WG also recommended reporting quantitative estimates that are accompanied by measures of uncertainty (i.e., effect size and confidence interval). The effect size, when combined with confidence intervals, serves as a better representation of the evidence by providing a range of plausible values within which point estimates (e.g., mean differences) may lie, allowing others to calibrate an interpretation of the data (69). Further, this approach allows for a more accurate comparison across studies and sample-size planning of future studies and can be optimized for use in meta-analyses. These meta-analyses can be used to assess data from independent data sets and enable a more generalizable understanding of research outcomes, while simultaneously mitigating potential publication bias (70).

2.6. Extrinsic Factors to Consider Within Animal Facilities

As described in the Introduction, animal-specific factors, particularly those once unknown or thought to be irrelevant, are increasingly recognized as impactful, if not critically important, to study outcomes. Indeed, one multi-laboratory study investigating confounding effects of the laboratory environment and a gene-by-environment interaction found that despite rigorous standardization of housing conditions and study protocols, between-laboratory variations led to a significant interaction between genotype and laboratory (71). Similarly, a lack of standardization of specific protocols (e.g., timing, severity/invasiveness of procedures, housing conditions) and differences in equipment (7, 72, 73) also contributed to irreproducible research.

The NIH ACD WG recommended to the Director that investigators should be required to report extrinsic factors, such as aspects of animal handling, housing, husbandry, transportation to animal housing facilities, and ambient environmental parameters (e.g., temperature, light, humidity, noise) (74). Furthermore, variables such as enrichment, social housing, experimenter sex/gender, and experimenter handling and refinements in these practices impact behavioral and experimental outcomes (75) and should be disclosed in publications.

Most animal research facilities are equipped with specialized housing rooms and cage systems that minimize potential confounding variables influenced by the environment. The facilities are maintained by qualified and dedicated personnel who adhere to well-conceived and regulated operating procedures. Controlling, to the best of one's ability, both the microenvironment (the

animal's primary enclosure, including temperature and humidity) and the macroenvironment (the physical conditions surrounding the microenvironment, including room lighting and air quality) is essential to maximize animal well-being and the quality of research data obtained from the animals (76–79).

Recent ILAR publications provide summary overviews of important extrinsic factors like the room light cycle (80); social and behavioral factors (81); and the animal microbiota (82), which is strongly influenced by rodent chow formulations and can also vary wildly by brand and nutritional composition. Drinking water provided to laboratory animals is rarely considered in the experimental design, but recent evidence indicates that water source, microbial and chemical contaminants, and purification methods can result in potential experimental variability (83). In addition, a growing area of study in laboratory animal sciences and regulatory oversight is the evaluation of enrichment substrates to better promote animal-specific behaviors (84). No matter the species involved, bedding and enrichment items can affect behavior and physiological parameters and should be disclosed in protocols and publications to aid reproducibility efforts (85, 86).

Attention to sources of unanticipated noise and vibration within animal facilities warrants consideration for their potential adverse impacts on animal health. There are known sources of noise, like cage washers, loading docks, and large animal housing spaces; however, noise from smoke-detection devices, lighting systems, computer and equipment alarms, HVAC, and laminar flow hoods also can cumulatively impact animals and may or may not be at the frequency where certain animal species are affected (87–89). Noise exposure influences virtually every area of biomedical research, from the immune system, to the development of tumors and heart disease, to typical circadian rhythms. Noise and vibration are sources of stress and can lead to a cascade of physiologic responses in animals (87). Even at relatively low intensities, such noise can be damaging to research animals and humans alike; therefore, procedural recovery areas are best kept to low, undisturbed sound volumes until animals can be returned to typical housing environments.

Animal housing, handling, and husbandry will never be standardized fully across institutions, because other external factors (e.g., rotation of animal researchers and care personnel, building and facility age, HVAC, weather/seasonal changes) will always be present. The NIH ACD WG advised that specific subcategories of the ARRIVE 2.0 guidelines be disclosed when providing details in animal study designs, grant submissions, and progress reports. Specific vivarium factors are outlined in Item #8 (Experimental Animals) within the ARRIVE Essential 10, Item #15 (Housing and Husbandry) from the ARRIVE Recommended Set, and Item #16 (Animal Care and Monitoring) from the ARRIVE Recommended Set (12, 27).

The recognition of the importance of extrinsic factors has emerged as its own discipline, classified as therioepistemology (90). Importantly, because human conditions of disease are not standardized as they are studied, it is unrealistic to expect that animal conditions should therefore be identical. Although documentation of extrinsic factors and animal husbandry and environment may be disregarded as a perceived regulatory burden, much of the data regarding the animal environment and husbandry are available within existing animal program records (e.g., AAALAC International program descriptions and daily housing room checklists). These data are readily provided to research teams by those involved in the delivery of animal care, including Attending Veterinarians or other animal care personnel, Institutional Animal Care and Use Committee administrators, and grant and research program leadership. With consistent access to records of environmental factors, investigators can review and retain important experimental information for their data files; in particular, deviations from expected outcomes can be explored, addressed, and reported in research findings.

3. TRANSLATIONAL RELEVANCE AS KEY SUPPORT FOR GENERALIZABILITY

The centuries-long relationship between humans and animals is important to our mutual well-being and survival (91, 92). An essential part of that relationship is the reliance on animals as research subjects to expand our understanding of human and animal health and disease, to characterize potential hazards of substances in our environment, and to support the discovery and development of innovative medicines and therapies. Our ability to interrogate biology and disease in animals has grown considerably with the advent of genome sequencing; the development of genetic modification techniques; and the application of a growing portfolio of noninvasive cellular, biochemical, protein, and imaging biomarkers.

Although the NIH ACD WG recognized that no model system perfectly represents the complexity of the human system, there is significant conservation across the biological spectrum between many mammals and humans at the gene, cell, tissue, and organ system level with the recapitulation of most major processes at both the molecular and macromolecular levels. Accordingly, many animal species represent reasonable surrogates of human biology that can be leveraged to better understand humans. Some of these experimental systems are induced by intentionally altering environmental conditions (e.g., giving animals a high-fat diet or exposing them to unique stimuli like altered circadian light/dark cycles) or administering agents that cause injury or alter physiological norms (e.g., intratracheal administration of the chemotherapeutic bleomycin to induce pulmonary fibrosis, intraperitoneal administration of pentylenetetrazole to induce seizures, or topical administration of house dust mite allergen to induce atopic dermatitis). Others are spontaneous or induced genetic modifications that are perpetuated through intentional breeding.

Animal research has expanded significantly as an outcome of the Human Genome Project and the idea that the human genome provides insights into the causes of human disease as well as novel pharmaceutical targets for mitigating those diseases (93). This convergent concept of a genetic basis for disease, coupled with our ability to genetically modify animals, has prompted significant standardization in the field of basic animal research and even efforts to humanize animals (e.g., replace a mouse gene with a corresponding human orthologous gene sequence) to improve their human relevance (94). Accordingly, the focus on animal modeling systems has narrowed to a smaller number of animal species, of which rodents, particularly mice, have emerged as the most commonly used (95). Though mice are considered relevant biological surrogates for humans, the motivations for their experimental prominence have less to do with their biological relevance than their relative (easy-to-handle) size, cost, fecundity, and susceptibility to genetic modification.

Many fundamental aspects of mammalian biology are well conserved between laboratory animal species and humans. A variety of animal species generally represent the major organ systems and the basic physiology typically evaluated in human patients. Even so, there is variability in gene expression and physiologic concordance to humans among common laboratory animal species. For example, the lack of understanding of the etiology and pathogenesis of important classes of diseases, such as neuropsychiatric disorders, hampers the ability to model them effectively in animals. This is also complicated by the frequent lack of a representative morphologic phenotype to replicate and the challenges in relating animal behaviors to human behaviors that are often the primary manifestation of disease in humans. Accordingly, and for this area of biology and class of diseases, there is a frequent need to study nonhuman primates (NHPs) as the pinnacle of human relevance (96–98). Despite the exquisite biological relevance of NHPs, there are significant societal and logistical challenges to using these species to the extent that rodents are used in research studies. Even so, the NIH ACD WG recognized the continuing need for NHPs and other large animal species (dogs, cats, pigs, sheep) to support certain areas of research (e.g., neuropsychiatric,

immunologic, and some infectious diseases) and recommended that appropriate funding support be allocated for those needs (6).

Recognizing the unique phenotypic features of each laboratory animal species is critically important for interpreting outcomes and understanding the implications of these features in translating the model to the human condition. Comparative biologists and pathologists are attuned to the high heart rate of rats and mice, the lack of a gallbladder in the rat, the unique histologic prominence of Purkinje cells in the mini-pig, and the cuboidal parietal epithelium of the glomerular Bowman's capsule in the male mouse. In addition, there are known background pathologies in these species, like multifocal cardiomyocyte necrosis and mononuclear cell infiltrate in rodents and mini-pigs, chronic progressive nephropathy in rodents, and vascular necrosis in the coronary arteries of dogs. ILAR recently published a series of articles summarizing common background lesions in a broad spectrum of animal species used in research (99, 100). Accordingly, animal model selection and characterization will benefit from multidisciplinary partnerships between the investigators defining the biomedical research questions and the comparative scientists who understand the animals being studied to continue to refine model selection and interpret data outcomes.

There are many reasons why an animal study might not accurately or precisely model what will happen in human patients (i.e., it may not be generalizable to a patient population of interest). Despite substantial conservation of fundamental biology, as well as efforts to rigorously standardize and control experimental studies, biology is complex and variable in animals and humans. Individual variability in basal biology and responses to perturbation is a reality in both animals and humans, leading some investigators to advocate for representing that variability in animal studies (101). That fundamental principle presented itself clearly with the evolution of the COVID pandemic, during which individuals had widely disparate responses to infections from the many variants of SARS-CoV-2 that emerged. Accordingly, it is important to understand the zone of applicability or validity for the selected model and the study in which it will be used, relative to the hypothesis being tested or the disease/disorder under question. Key contributors to translational failures include lack of sufficient concordance of the biology or pathobiology of interest to humans, procedural failure to control for experimental bias, and translationally weak study designs.

Physiological differences among species likely influence how well they model the human condition. The comparative physiology of the cardiovascular system is a good example. Basal heart rates are highly variable across species, and are dependent on housing environment and ambient temperatures, ranging from approximately 450–750 bpm in laboratory mice to 250–400 bpm in rats, 70–120 bpm in dogs, and 50–90 bpm in humans (102). In addition to differences in basal heart rate, the approaches used to measure that heart rate can influence its generalizability. Common approaches to measuring heart rate in laboratory animals include manual or anesthetized restraint using electrocardiogram leads on a closed chest for short-duration measurements (closely aligned with a human clinical assessment) and surgical implantation of a telemetered device for longer-duration assessments (excellent for continuous measures at high resolution without the distraction of restraint, but not similar to the human experience). In addition to functional differences in parameters like heart rate, there are also meaningful anatomic differences between species to consider. A relevant example is coronary collateral circulation, which is variable but well developed in humans, as it is in dogs, but much less so in rodents (103). Collateral coronary circulation capacity significantly influences the outcomes of myocardial ischemia from coronary artery occlusion. Accordingly, the American Heart Association recognized the usefulness of rodent models but recommended that large animal models also be used to support the development of human heart failure therapies due to their more relevant cardiac biology and pathobiology (104). This presents a potential conflict between our interest in conducting translationally relevant animal research and our strong commitment to the judicious use of animals that often includes a

default to rodents over dogs or NHPs. This ethical conflict likely contributes to the conduct of studies that are not optimized for their generalizability to humans. As noted above, the NIH ACD WG recognized specific needs for large animal (nonrodent) models for some areas of research and advised the NIH to consider that unique need in their funding support for those areas.

Another potential source of translational weakness is differences in the ways pathobiology or disease is represented or experimentally induced in experimental animals relative to how it manifests in human patients. Not unexpectedly, diseases with simple etiologies (e.g., single–gene mutation diseases or infectious disease) are often modeled with better human fidelity than more complex diseases, like many of the chronic progressive diseases of contemporary interest to drug developers (e.g., chronic heart failure, chronic progressive renal disease, chronic lung disease, and progressive neurological diseases). Adding to this complexity is a frequent lack of understanding of the etiologies of many of these disease conditions in humans, so investigators are often left replicating the morphologic phenotype without replicating the pathogenesis (i.e., disease initiation and progression). Examples of this include the use of bleomycin to induce pulmonary fibrosis in rodents in a week or two, when a similar condition develops over years in human patients (105). Yet another example is the dextran sulfate sodium model of chemically induced inflammatory bowel disease, which is often immune mediated in humans (106). Both models morphologically reflect organ-specific responses to injury that are present in human disease progression, though they are induced in ways that do not reflect the pathogenesis of the human disease, which is likely to influence the translation of animal to human outcomes.

Though rigorous standardization of studies may facilitate more efficient decision making and use of fewer animals, it significantly undermines the translational relevance of those studies, because human biology and pathobiology are anything but standardized. The conflict between these two fundamental issues becomes more acute as the interest in understanding the influence of individual genetic variability and personalized medicine increases. The Collaborative Cross and Diversity Outbred mouse populations are attempts to better represent the genetic diversity of human populations. These stocks are used to investigate gene–phenotype associations and also model the potential dynamic range of a biological response to a stimulus or insult in a genetically diverse population. Studies using outbred mouse colonies can be effective at supporting those investigations; however, animal cohorts will likely require larger group sizes, and there may be challenges in data interpretation (i.e., due to the variability of the results) (107, 108).

Several considerations and approaches can improve the translational relevance of an animal study. In addition to adopting as common practice the inclusion of the criteria described in Section 2, one must (*a*) clearly define the primary aim of the study, hypothesis, or question (i.e., the context for assessing the relevance of the model and guiding the study design); (*b*) consider the human biology and pathobiology of interest; (*c*) appropriately characterize the animal model for its reflection of that biology and pathobiology; and (*d*) design the study to optimally represent the human clinical context. Structured approaches to evaluating an animal model can be useful in assessing its human relevance and representing the translational strengths of the outcomes from studies using that model. The NIH ACD WG recommended that the NIH establish a framework or guidance for rationalizing the scientific and appropriate translational relevance of animal models. It was thought that this would support rigorous assessment of model selection and transparency of model strengths and weaknesses. As the ARRIVE Guidelines have improved the methodologic transparency of animal studies, some frameworks are emerging that could support the NIH ACD WG's recommendation. For example, Storey et al. (109) recently outlined a framework called the Animal Model Quality Assessment. In brief, this framework considers features consistent with many of those described herein, including the biological relevance of the animal, the level of understanding of the human disease, how well the human disease is modeled

in the animal, the history of pharmacologic response in the animal model, and the reproducibility of the phenotype (110, 111).

Disease modeling brings with it additional challenges, because assessing the translational relevance of an animal model of disease requires the requisite understanding of the human disease. There are many human diseases for which the salient clinical features are well-defined. These features generally include assessments of organ function, clinical pathologic parameters (e.g., hematology, serum chemistry, urine chemistry), assessments of behavior or sense of well-being, and macroscopic morphology. Histopathologic characterization of autopsy or biopsy material may also be available for specific target-organ diseases (e.g., liver, kidney, intestine) but is less common. Many of the more complex human diseases of interest today are not well characterized etiologically (i.e., their specific cause is unknown), nor is there reliable characterization of their early stages (e.g., subclinical phases of disease initiation and progression) or molecular pathogeneses. These critical gaps make it difficult to ensure that these features are represented in an animal model, which creates a bias toward replicating morphologic phenotype over etiology and pathogenesis. Ultimately, efforts to improve the translational relevance of animal modeling will require better characterizations of human disease to optimize concordance (112).

The NIH ACD WG recognized the importance of rigorous animal model design and characterization and the difficulties in putting that burden on individual investigators. To this end, it was recommended to establish venues for the exchange of pertinent animal model information, as well as funding of research centers to support design and characterization efforts (6).

4. CONCLUSIONS AND FUTURE EFFORTS

As the NIH and other research organizations continue to invest in basic and biomedical research, ongoing (career-long) training and education will be critical for scientific stakeholders to strengthen experimental design and analysis, select relevant animal models to address the questions of interest, and optimize animal welfare to mitigate the impact on experimental outcomes. Animal models and the studies in which they are involved are critical surrogates for human biology and disease, as well as the most ethical alternative to human experimentation, given the limited availability of nonanimal alternatives that could fully replace animal use. In the spirit of generalizability, as mentioned previously, outcomes and interpretations of animal studies can apply across a variety of situations and populations, thus serving to reproduce scientific efforts and foster innovative ideas and discoveries.

The NIH ACD WG expressed a shared foundational agreement, supported by the NIH Director, that animal studies contribute to significant findings and breakthroughs in both basic and translational research. Enhancing transparency and openness around the rationale and importance of biomedical investigations will help scientists, as well as the public, to engage in meaningful exchanges about how animals contribute to our society. In their final report, the NIH ACD WG emphasized the value of open-source methods for sharing findings and data. Openness is the foundation upon which institutions will be able to promote understanding (when, why, and how animals are used) and garner continued public support for the necessity of animals for medical outcomes that benefit human, as well as animal, health.

SUMMARY POINTS

1. Animal studies contribute to significant findings and breakthroughs in both basic and translational research.

2. Reproducibility of research findings cannot be guaranteed due to extensive inherent variables in attempts at experimental repetition.
3. Stakeholders within the scientific community should advocate for generalizability in the application of data outcomes to ensure a broad and effective impact on the translation of animal models to human research.
4. Investigators should report extrinsic factors, as these are the ambient environmental parameters have direct impact on the experience of research animals, and therefore can impact experimental outcomes.
5. Openness is the foundation upon which institutions will be able to promote understanding and garner continued public support for the necessity of animals for medical outcomes that benefit human, as well as animal, health.
6. Advancing scientific discovery requires investigators to embrace research practices that increase transparency and disclosure about materials, methods, and outcomes.
7. Preparatory checklists, like those provided in ARRIVE 2.0, CONSORT, and PREPARE, provide prompts about data disclosures that will help to mitigate potential sources of bias and promote reproducibility and improved research outcomes.
8. As research organizations continue to invest in basic and biomedical research, continued training and education will be critical for scientific stakeholders to strengthen experimental design and analysis, select relevant animal models to address the questions of interest, and optimize animal welfare to mitigate the impact on experimental outcomes.

DISCLOSURE STATEMENT

E.M.P. contributed to this article in his personal capacity and the views expressed do not necessarily represent the views of Regeneron Pharmaceuticals, Inc. The consultancy practice of B.R.B. supports technology companies that provide support to animal research. The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review; all authors participated as invited working group members for "Enhancing Rigor, Transparency, and Translatability in Animal Research" to the National Institutes of Health Advisory Committee to the Director from 2019 to 2021.

ACKNOWLEDGMENTS

The authors would like to thank their colleagues on the National Institutes of Health Advisory Committee to the Director Working Group, particularly the leadership of Drs. Barbara Wold and Lawrence A. Tabak.

LITERATURE CITED

1. Natl. Res. Counc. 2011. *Guide for the Care and Use of Laboratory Animals: Eighth Edition*. Washington, DC: Natl. Acad. Press
2. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLOS Biol*. 8:e1000412
3. Collins FS, Tabak LA. 2014. Policy: NIH plans to enhance reproducibility. *Nature* 505:612–13
4. FASEB (Fed. Am. Soc. Exp. Biol.). 2016. *Enhancing Research Reproducibility: Recommendations from the Federation of American Societies for Experimental Biology*. Rockville, MD: FASEB

5. Natl. Acad. Sci. Eng. Med. 2019. *Reproducibility and Replicability in Science*. Washington, DC: Natl. Acad. Press
6. Wold B, Tabak L. 2021. *ACD working group on enhancing rigor, transparency, and translatability in animal research*. Final Rep., Adv. Comm. Dir., Natl. Inst. Health, Bethesda, MD. **https://acd.od.nih.gov/documents/presentations/06112021_ACD_WorkingGroup_FinalReport.pdf**
7. Goodman SN, Fanelli D, Ioannidis JP. 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8:341ps12
8. Grant S, Wendt KE, Leadbeater BJ, Supplee LH, Mayo-Wilson E, et al. 2022. Transparent, open, and reproducible prevention science. *Prev. Sci.* 23:701–22
9. NIH Cent. Resour. Grants Fund. Inf. 2023. *Enhancing reproducibility through rigor and transparency*. **https://grants.nih.gov/policy/reproducibility/index.htm**
10. Clayton JA. 2018. Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiol. Behav.* 187:2–5
11. Beynen A, Gärtner K, van Zutphen L. 2003. Standardization of animal experimentation. In *Principles of Laboratory Animal Science*, ed. LFM van Zutphen, V Baumans, A Beynen, pp. 103–10. Amsterdam: Elsevier
12. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, et al. 2020. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *Exp. Physiol.* 105:1459–66
13. Smith AJ, Clutton RE, Lilley E, Hansen KEA, Brattelid T. 2018. PREPARE: guidelines for planning animal research and testing. *Lab. Anim.* 52:135–41
14. Macleod M, Mohan S. 2019. Reproducibility and rigor in animal-based research. *ILAR J.* 60:17–23
15. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, et al. 2017. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ. Res.* 120:1916–26
16. Prager EM, Chambers KE, Plotkin JL, McArthur DL, Bandrowski AE, et al. 2019. Improving transparency and scientific rigor in academic publishing. *J. Neurosci. Res.* 97:377–90
17. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1:0021
18. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Muhlbacher A, et al. 2013. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value Health* 16:3–13
19. Hess KR. 2011. Statistical design considerations in animal studies published recently in cancer research. *Cancer Res.* 71:625
20. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. 2016. Reproducible research practices and transparency across the biomedical literature. *PLOS Biol.* 14:e1002333
21. Ramirez FD, Jung RG, Motazedian P, Perry-Nguyen D, Di Santo P, et al. 2020. Journal initiatives to enhance preclinical research: analyses of *Stroke*, *Nature Medicine*, *Science Translational Medicine*. *Stroke* 51:291–99
22. Kousholt BS, Praestegaard KF, Stone JC, Thomsen AF, Johansen TT, et al. 2022. Reporting quality in preclinical animal experimental research in 2009 and 2018: a nationwide systematic investigation. *PLOS ONE* 17:e0275962
23. Menke J, Roelandse M, Ozyurt B, Martone M, Bandrowski A. 2020. The Rigor and Transparency Index quality metric for assessing biological and medical science methods. *iScience* 23:101698
24. Simundic AM. 2013. Bias in research. *Biochem. Med.* 23:12–15
25. Dirnagl U. 2016. Thomas Willis Lecture: Is translational stroke research broken, and if so, how can we fix it? *Stroke* 47:2148–53
26. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, et al. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 490:187–91
27. Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, et al. 2020. Reporting animal research: explanation and elaboration for the ARRIVE guidelines 2.0. *PLOS Biol.* 18:e3000411
28. Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. 2010. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLOS Biol.* 8:e1000344

29. Kiyonaga A, Scimeca JM. 2019. Practical considerations for navigating Registered Reports. *Trends Neurosci*. 42:568–72
30. Chambers CD, Tzavella L. 2022. The past, present and future of Registered Reports. *Nat. Hum. Behav.* 6:29–42
31. Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, et al. 2013. Evaluation of excess significance bias in animal studies of neurological diseases. *PLOS Biol*. 11:e1001609
32. Altman DG. 2009. Missing outcomes in randomized trials: addressing the dilemma. *Open Med*. 3:e51–53
33. Rubin LH, Witkiewitz K, Andre JS, Reilly S. 2007. Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water. *J. Undergrad. Neurosci. Educ.* 5:A71–77
34. Lane P. 2008. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharm. Stat.* 7:93–106
35. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med. Res. Methodol.* 17:162
36. Kang H. 2013. The prevention and handling of the missing data. *Korean J. Anesthesiol.* 64:402–6
37. Schulz KF, Altman DG, Moher D, CONSORT Group. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLOS Med.* 7:e1000251
38. Han S, Olonisakin TF, Pribis JP, Zupetic J, Yoon JH, et al. 2017. A checklist is associated with increased quality of reporting preclinical biomedical research: a systematic review. *PLOS ONE* 12:e0183591
39. Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, et al. 2015. Risk of bias in reports of in vivo research: a focus for improvement. *PLOS Biol*. 13:e1002273
40. Altman DG, Bland JM. 1999. How to randomise. *BMJ* 319:703–4
41. Kang M, Ragan BG, Park JH. 2008. Issues in outcomes research: an overview of randomization techniques for clinical trials. *J. Athl. Train.* 43:215–21
42. Efird J. 2011. Blocked randomization with randomly selected block sizes. *Int. J. Environ. Res. Public Health* 8:15–20
43. Bespalov A, Wicke K, Castagné V. 2019. Blinding and randomization. In *Good Research Practice in Non-Clinical Pharmacology and Medicine*, ed. A Bespalov, M Michel, T Steckler, pp. 81–100. Handb. Exp. Pharmacol. 257. Cham, Switz.: Springer Open
44. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, et al. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLOS ONE* 4:e7824
45. Karp NA, Pearl EJ, Stringer EJ, Barkus C, Ulrichsen JC, Percie du Sert N. 2022. A qualitative study of the barriers to using blinding in in vivo experiments and suggestions for improvement. *PLOS Biol*. 20:e3001873
46. Bonapersona V, Hoijtink H, RELACS Consort., Sarabdjitsingh RA, Joëls M. 2021. Increasing the statistical power of animal experiments with historical control data. *Nat. Neurosci.* 24:470–77
47. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14:365–76
48. Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafo MR. 2017. Low statistical power in biomedical science: a review of three human research domains. *R. Soc. Open Sci.* 4:160254
49. van Zwet EW, Goodman SN. 2022. How large should the next study be? Predictive power and sample size requirements for replication studies. *Stat. Med.* 41:3090–101
50. Nord CL, Valton V, Wood J, Roiser JP. 2017. Power-up: a reanalysis of 'power failure' in neuroscience using mixture modeling. *J. Neurosci.* 37:8051–61
51. Lazic SE, Clarke-Williams CJ, Munafo MR. 2018. What exactly is 'N' in cell culture and animal experiments? *PLOS Biol*. 16:e2005282
52. Krzywinski M, Altman N. 2013. Power and sample size. *Nat. Methods* 10:1139–40
53. Calin-Jageman RJ. 2018. The new statistics for neuroscience majors: thinking in effect sizes. *J. Undergrad. Neurosci. Educ.* 16:E21–E25
54. Bakker M, Veldkamp CLS, van den Akker OR, van Assen M, Crompvoets E, et al. 2020. Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE* 15:e0236079

55. Bartlett J. 2021. *Introduction to power analysis: a guide to G*Power, jamovi, and Superpower*. **https://osf.io/zqphw**
56. Faul F, Erdfelder E, Buchner A, Lang AG. 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41:1149–60
57. Kelley K, Preacher KJ. 2012. On effect size. *Psychol. Methods* 17:137–52
58. Champely S, Ekstrom C, Dalgaard P, Gill J, Weibelzahl S, et al. 2020. *pwr: basic functions for power analysis (1.3–0)*. **https://CRAN.R-project.org/package=pwr**
59. Bartlett J, Charles S. 2021. Power to the people: a beginner's tutorial to power analysis using jamovi. PsyArXiv. **https://doi.org/10.31234/osf.io/bh8m9**
60. Lazic SE. 2018. Four simple ways to increase power without increasing the sample size. *Lab. Anim.* 52:621–29
61. Nuijten MB, Hartgerink CH, van Assen MA, Epskamp S, Wicherts JM. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* 48:1205–26
62. Greenberg L, Jairath V, Pearse R, Kahan BC. 2018. Pre-specification of statistical analysis approaches in published clinical trial protocols was inadequate. *J. Clin. Epidemiol.* 101:53–60
63. Kahan BC, Forbes G, Cro S. 2020. How to design a pre-specified statistical analysis approach to limit p-hacking in clinical trials: the Pre-SPEC framework. *BMC Med*. 18:253
64. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14:1105–7
65. Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. 2018. Why we need to report more than 'Data were Analyzed by t-tests or ANOVA.' *eLife* 7:e36163
66. Wieschowski S, Biernot S, Deutsch S, Glage S, Bleich A, et al. 2019. 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* 14:e0223758
67. Chan A-W, Hróbjartsson A, Jørgensen KJ, Gøtzsche PC, Altman DG. 2008. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ* 337:a2299
68. Reynolds PS. 2022. Between two stools: preclinical research, reproducibility, and statistical design of experiments. *BMC Res. Notes* 15:73
69. Durlak JA. 2009. How to select, calculate, and interpret effect sizes. *J. Pediatr. Psychol.* 34:917–28
70. Calin-Jageman RJ, Cumming G. 2019. The new statistics for better science: Ask how much, how uncertain, and what else is known. *Am. Stat.* 73:271–80
71. Crabbe JC, Wahlsten D, Dudek BC. 1999. Genetics of mouse behavior: interactions with laboratory environment. *Science* 284:1670–72
72. Lasseter HC, Provost AC, Chaby LE, Daskalakis NP, Haas M, Jeromin A. 2020. Cross-platform comparison of highly sensitive immunoassay technologies for cytokine markers: platform performance in post-traumatic stress disorder and Parkinson's disease. *Cytokine X* 2:100027
73. Ioannidis JP. 2005. Why most published research findings are false. *PLOS Med.* 2:e124
74. Prager EM, Bergstrom HC, Grunberg NE, Johnson LR. 2011. The importance of reporting housing and husbandry in rat research. *Front. Behav. Neurosci.* 5:38
75. Larrieu T, Cherix A, Duque A, Rodrigues J, Lei H, et al. 2017. Hierarchical status predicts behavioral vulnerability and nucleus accumbens metabolic profile following chronic social defeat stress. *Curr. Biol.* 27:2202–10.e4
76. DeMarco G, Makidon P, Suckow M, Hankenson F. 2022. *ACLAM position statement on reproducibility*. Position Statement, Am. Coll. Lab. Anim. Med., Chester, NH. **https://www.aclam.org/media/83cf63c9-75ee-4271-a70a-7b83fcd401bc/S2Cx9g/ACLAM/About**
77. Hogan M, Norton J, Reynolds R. 2018. Environmental factors: macroenvironment versus microenvironment. In *Management of Animal Care and Use Programs in Research Education and Testing*, ed. R Weichbrod, G Thompson, J Norton, pp. 461–77. Boca Raton, FL: CRC Press
78. Hasenau JJ. 2020. Reproducibility and comparative aspects of terrestrial housing systems and husbandry procedures in animal research facilities on study data. *ILAR J.* 60:228–38
79. Lee VK, David JM, Huerkamp MJ. 2020. Micro- and macroenvironmental conditions and stability of terrestrial models. *ILAR J.* 60:120–40

80. Hanifin JP, Dauchy RT, Blask DE, Hill SM, Brainard GC. 2020. Relevance of electrical light on circadian, neuroendocrine, and neurobehavioral regulation in laboratory animal facilities. *ILAR J.* 60:150–58
81. Whittaker AL, Hickman DL. 2020. The impact of social and behavioral factors on reproducibility in terrestrial vertebrate models. *ILAR J.* 60:252–69
82. Franklin CL, Ericsson AC. 2020. Complex microbiota in laboratory rodents: management considerations. *ILAR J.* 60:289–97
83. Kurtz DM, Feeney WP. 2020. The influence of feed and drinking water on terrestrial animal research and study replicability. *ILAR J.* 60:175–96
84. Pritchett-Corning KR. 2020. Environmental complexity and research outcomes. *ILAR J.* 60:239–51
85. Van Loo PL, Mol JA, Koolhaas JM, Van Zutphen BF, Baumans V. 2001. Modulation of aggression in male mice: influence of group size and cage size. *Physiol. Behav.* 72:675–83
86. Kingston SG, Hoffman-Goetz L. 1996. Effect of environmental enrichment and housing density on immune system reactivity to acute exercise stress. *Physiol. Behav.* 60:145–50
87. Turner JG. 2020. Noise and vibration in the vivarium: recommendations for developing a measurement plan. *J. Am. Assoc. Lab. Anim. Sci.* 59:665–72
88. Reynolds R, Garner A, Norton J. 2020. Sound and vibration as research variables in terrestrial vertebrate models. *ILAR J.* 60:159–74
89. Reynolds RP, Kinard WL, Degraff JJ, Leverage N, Norton JN. 2010. Noise in a laboratory animal facility from the human and mouse perspectives. *J. Am. Assoc. Lab. Anim. Sci.* 49:592–97
90. Gardenier JS, Resnik DB. 2002. The misuse of statistics: concepts, tools, and a research agenda. *Account. Res.* 9:65–74
91. Franco NH. 2013. Animal experiments in biomedical research: a historical perspective. *Animals* 3:238–73
92. Kinter LB, DeHaven R, Johnson DK, DeGeorge JJ. 2021. A brief history of use of animals in biomedical research and perspective on non-animal alternatives. *ILAR J.* 62:7–16
93. Rood JE, Regev A. 2021. The legacy of the Human Genome Project. *Science* 373:1442–43
94. Zhu F, Nair RR, Fisher EMC, Cunningham TJ. 2019. Humanising the mouse genome piece by piece. *Nat. Commun.* 10:1845
95. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
96. Tarantal AF, Noctor SC, Hartigan-O'Connor DJ. 2022. Nonhuman primates in translational research. *Annu. Rev. Anim. Biosci.* 10:441–68
97. Tramacere A, Iriki A. 2021. A novel mind-set in primate experimentation: implications for primate welfare. *Anim. Model. Exp. Med.* 4:343–50
98. Natl. Acad. Sci. Eng. Med. 2023. *Nonhuman Primate Models in Biomedical Research: State of the Science and Future Needs*. Washington, DC: Natl. Acad. Press
99. Cooper TK, Meyerholz DK, Beck AP, Delaney MA, Piersigilli A, et al. 2022. Research-relevant conditions and pathology of laboratory mice, rats, gerbils, guinea pigs, hamsters, naked mole rats, and rabbits. *ILAR J.* 62:77–132
100. Helke KL, Meyerholz DK, Beck AP, Burrough ER, Derscheid RJ, et al. 2021. Research relevant background lesions and conditions: ferrets, dogs, swine, sheep, and goats. *ILAR J.* 62:133–68
101. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, et al. 2020. Reproducibility of animal research in light of biological variation. *Nat. Rev. Neurosci.* 21:384–93
102. Winter AL, ed. 2016. *MSD Veterinary Manual*. Rahway, NJ: Merck & Co.
103. Jamaiyar A, Juguilon C, Dong F, Cumpston D, Enrick M, et al. 2019. Cardioprotection during ischemia by coronary collateral growth. *Am. J. Physiol. Heart Circ. Physiol.* 316:H1–H9
104. Houser SR, Margulies KB, Murphy AM, Spinale FG, Francis GS, et al. 2012. Animal models of heart failure: a scientific statement from the American Heart Association. *Circ. Res.* 111:131–50
105. Liu T, De Los Santos FG, Phan SH. 2017. The bleomycin model of pulmonary fibrosis. *Methods Mol. Biol.* 1627:27–42
106. Hoffmann M, Schwertassek U, Seydel A, Weber K, Falk W, et al. 2018. A refined and translationally relevant model of chronic DSS colitis in BALB/c mice. *Lab. Anim.* 52:240–52

107. Hackett J, Gibson H, Frelinger J, Buntzman A. 2022. Using the collaborative cross and diversity outbred mice in immunology. *Curr. Protoc.* 2:e547
108. Saul MC, Philip VM, Reinholdt LG, Chesler EJ. 2019. High-diversity mouse populations for complex traits. *Trends Genet.* 35:501–14
109. Storey J, Gobbetti T, Olzinski A, Berridge BR. 2021. A structured approach to optimizing animal model selection for human translation: the Animal Model Quality Assessment. *ILAR J.* 62:66–76
110. Ferreira GS, Veening-Griffioen DH, Boon WPC, Moors EHM, Gispen-de Wied CC, et al. 2019. A standardised framework to identify optimal animal models for efficacy assessment in drug development. *PLOS ONE* 14:e0218014
111. Wendler A, Wehling M. 2012. Translatability scoring in drug development: eight case studies. *J. Transl. Med.* 10:39
112. MacRae CA. 2019. Closing the 'phenotype gap' in precision medicine: improving what we measure to understand complex disease mechanisms. *Mamm. Genome* 30:201–11