Received: 30 June 2021 | Revised: 25 August 2021 | Accepted: 4 September 2021

DOI: 10.1002/ajp.23331

REVIEW ARTICLE

AMERICAN JOURNAL OF PRIMATOLOGY WILEY

Improving rigor and reproducibility in nonhuman primate research

Eliza Bliss-Moreau[1,2] | Rama R. Amara[3] | Elizabeth A. Buffalo[4,5] | Ricki J. Colman[6,7] | Monica E. Embers[8] | John H. Morrison[1,9] | Ellen E. Quillen[10] | Jonah B. Sacha[11,12] | Charles T. Roberts[11] | National Primate Research Center Consortium Rigor and Reproducibility Working Group

[1]California National Primate Research Center, Davis, California, USA

[2]Department of Psychology, University of California Davis, Davis, California, USA

[3]Division of Microbiology and Immunology, Yerkes National Primate Research Center, Atlanta, Georgia, USA

[4]Washington National Primate Research Center, Seattle, Washington, USA

[5]Department of Physiology and Biophysics, University of Washington School of Medicine, Seattle, Washington, USA

[6]Wisconsin National Primate Research Center, Madison, Wisconsin, USA

[7]Department of Cell and Regenerative Biology, University of Wisconsin, Madison, Wisconsin, USA

[8]Division of Immunology, Tulane National Primate Research Center, Covington, Louisiana, USA

[9]Department of Neurology, University of California Davis, Davis, California, USA

[10]Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

[11]Divisions of Pathobiology and Immunology (JS) and Cardiometabolic Health (CR), Oregon National Primate Research Center, Beaverton, Oregon, USA

[12]Vaccine and Gene Therapy Institute, Oregon Health & Science University, Beaverton, Oregon, USA

Correspondence
Charles T. Roberts, Oregon National Primate Research Center, 505 NW 185th Ave., Beaverton, OR 97006, USA.
Email: robertsc@ohsu.edu

Funding information
Office of Research Infrastructure Programs, National Institutes of Health, Grant/Award Numbers: P51OD011107, P51OD011132, P51OD010425, P51OD011106, P51OD011104, P51OD011133, P51OD011092

Abstract

Nonhuman primates (NHPs) are a critical component of translational/preclinical biomedical research due to the strong similarities between NHP and human physiology and disease pathology. In some cases, NHPs represent the most appropriate, or even the only, animal model for complex metabolic, neurological, and infectious diseases. The increased demand for and limited availability of these valuable research subjects requires that rigor and reproducibility be a prime consideration to ensure the maximal utility of this scarce resource. Here, we discuss a number of approaches that collectively can contribute to enhanced rigor and reproducibility in NHP research.

KEYWORDS

biomedical research, data sharing, nonhuman primates, preregistration, quality assurance

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *American Journal of Primatology* published by Wiley Periodicals LLC

1 | INTRODUCTION

Science is in the midst of a major paradigm shift. Multiple scientific disciplines are increasingly facing and addressing, growing concerns about the meaning and impact of their findings. The extent to which scientific research is rigorous—robust and unbiased—and reproducible—able to be repeated biologically in the lab, analytically with the original data, systemically under different conditions, or conceptually at the level of the biological phenomenon—has an impact on almost every facet of modern life. Physical, biological, and social scientists have all begun conversations about how to improve research by creating systems and incentives to promote robustness and transparency in response to the unique challenges faced by each domain. So great is the concern about replication issues in biomedical science that funding agencies around the world have stepped in to create change via a focused effort that outline guidelines for rigor and reproducibility and/or requires their investigators to develop and adhere to procedures to ensure rigor and reproducibility in their research (e.g., the US National Institutes of Health [NIH], https://grants.nih.gov/policy/reproducibility/index.htm; for UK funders see https://acmedsci.ac.uk/policy/policy-projects/reproducibility-and-reliability-of-biomedical-research). The major stakeholders and their respective roles in this effort are illustrated in Figure 1. This issue has recently been addressed by the National Academy of Sciences in a comprehensive report (National Academies of Sciences & Medicine, 2019), which defines reproducibility as a computational matter and replicability as an experimental consistency matter.

Researchers using nonhuman primates (NHPs) face particular challenges in addressing rigor and reproducibility that are of less concern in rodent or in vitro studies. For example, significant concerns have been raised about rigor in psychology research with humans, leading to a massive shift in the norms for sample sizes (Sassenberg & Ditrich, 2019). It is now accepted that sample sizes must be much larger than they once were, and decisions about sample sizes must be made a priori based on the stated study outcomes. Increasing sample size may not be possible for scientists working with NHPs for ethical, logistical, or fiscal reasons. Research carried out with animals, including NHPs, is subject to the "3Rs"—Replacing animals in research when possible, Reducing sample sizes, and Refining techniques (Tannenbaum & Bennett, 2015). These ethical constraints must be balanced with rigor and reproducibility considerations, as underpowered studies waste resources and animals. It should be noted that decisions about sample sizes involve choosing an appropriate sample size that produces reliable data, rather than reducing numbers per se. Achieving the balance between ethical considerations and, in particular, reducing the number of animals used in research, and carrying out rigorous research that is appropriately powered, has appropriate controls, and is reproducible, is particularly challenging in work that involves NHPs—one of the most appropriate animal models for many human disease-related processes (Capitanio & Emborg, 2008; Estes et al., 2018; Phillips et al., 2014). Our goal in this review is to consider the challenges and opportunities to increase rigor and reproducibility in NHP research. To that end, we first briefly discuss NHPs as a model for human health and disease and then discuss themes in rigor and reproducibility that are consistent across the spectrum of scientific disciplines that work with them. We next highlight specific lessons learned from a number of disciplines that use the NHP model at the National Primate Research Centers (NPRCs), with the goal of expanding these best practices to the wider biomedical research community. Rigor and reproducibility in NHP research have been addressed in a recent NIH workshop (https://osp.od.nih.gov/wp-content/uploads/NHP_Workshop_Report_FINAL_20200218.pdf), and some of these considerations with respect to NHP vaccine studies, in particular, have been recently discussed (Prescott et al., 2021).

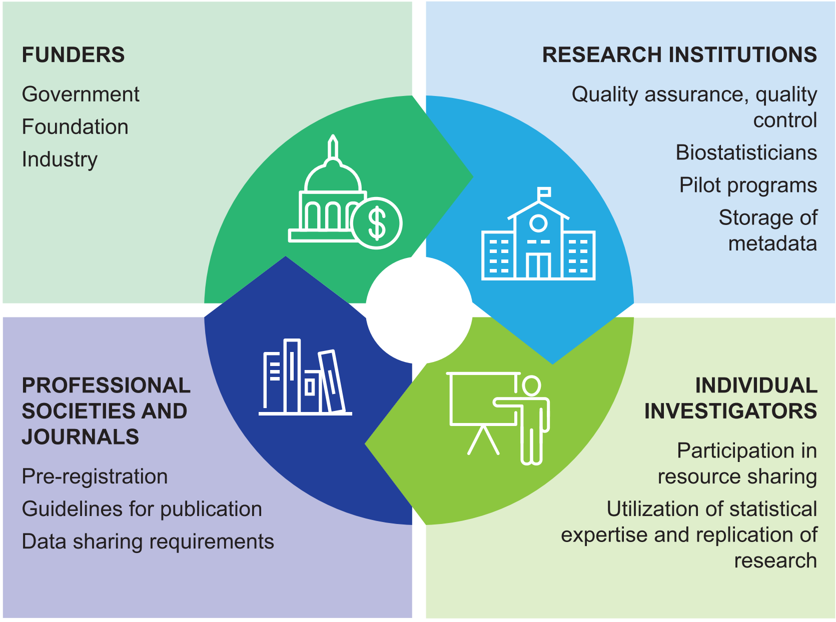
FUNDERS
Government
Foundation
Industry
RESEARCH INSTITUTIONS
Quality assurance, quality control
Biostatisticians
Pilot programs
Storage of metadata
PROFESSIONAL SOCIETIES AND JOURNALS
Pre-registration
Guidelines for publication
Data sharing requirements
INDIVIDUAL INVESTIGATORS
Participation in resource sharing
Utilization of statistical expertise and replication of research

FIGURE 1 Stakeholders and their roles in supporting rigor and reproducibility in nonhuman primate research

2 | NHPS AS BIOMEDICAL RESEARCH MODELS

NHPs represent a small proportion of laboratory animals used in studies of human health and disease. According to the most recent (2018) USDA annual report of animal usage, less than 9% of the 792,000 USDA-covered research animals (i.e., subject to the Animal Welfare Act) used in research in the US were NHPs. It is important to note that laboratory mice (*Mus* sp.) and rats (*Rattus* sp.), fish, and many other laboratory species are not USDA-covered research animals; thus, the proportion of NHPs in the laboratory relative to all laboratory animals is much lower than 9%. In spite of comprising such a small proportion of animal research subjects, NHPs garner significant attention from both regulatory bodies and the public. Their contribution to biomedical science is particularly important based on the extensive similarity in NHP and human physiology and behavior. A full discussion of the specific features of NHPs that make them good models for human health is beyond the scope of this review; a number of recent reviews discuss the similarities between NHPs and humans, highlighting the importance of NHP work for a wide variety of health and disease topics, including cardiac health, genetics (including CRISPR and other genetic engineering approaches), infectious disease (e.g., treatments and vaccine development), immunology, diseases of aging (e.g., cognitive decline, obesity, metabolic disease), respiratory diseases (e.g., asthma), and the neurobiology of psychological disorders (Buffalo et al., 2019; Capitanio & Emborg, 2008; Estes et al., 2018; Miller et al., 2017; Phillips et al., 2014).

In spite of the significant strengths of the NHP model, NHP research faces a number of notable constraints. Most NHPs are long-lived, have long periods of early development during which they are dependent on parental care, and exhibit robust cognitive, affective, and social repertories (Phillips et al., 2014). These features make them good models for humans, but also engender the need for additional protection not typically afforded to many other widely employed animal models (e.g., rodents, zebrafish, *Drosophila*). Sample sizes—typically small—are limited by both ethical and practical constraints, in that their care requires significant and unique infrastructure and resources, including enriched laboratory environments and highly trained personnel.

NHPs used in health-related research develop more quickly than humans, with specific developmental speeds varying by genus and species. For example, macaques develop approximately 3–4 times faster than humans (Machado, 2013; Suomi, 1999), allowing for lifespan studies within an investigator's career. This allows powerful longitudinal studies that can span long periods of time (up to decades) and thus requires significant foresight in the study and long-term care planning. It also means that a single animal may participate in a number of different studies across its lifespan, each of which has the potential to impact the others, thus creating variability in outcomes. Simultaneously, inherent variability in biology and behavior is the hallmark of the NHP model, insofar as NHP models are outbred, rather than inbred like many rodent models. This results in intrinsic "heterogenization," which is increasingly considered a strength in experimental models (Richter, 2017; Voelkl et al., 2020). This strength of the outbred NHP model in terms of translatability to human health and disease requires proper statistical modeling that embraces variability by modeling it appropriately and with potentially larger sample sizes than typically used.

Additionally, as science progresses, so too does our understanding of what features of individuals and their environments may create or sustain individual variation. For example, the norm in research with macaques for decades was nursery rearing (rearing by a human experimenter or with same-aged peers) and housing individuals singly in rooms that included other animals (DiVincenti & Wyatt, 2011; Schapiro et al., 2000). A large body of work now demonstrates that these conditions result in dysregulated biological and behavioral development, which compromise the wellbeing of subjects and also may make them less ideal models for human health and disease (Capitanio et al., 2006; Gottlieb et al., 2013). Although there remain certain instances in which nursery rearing may be desirable (i.e., studies of developmental disorders, infant infectious disease, and generation of specific-pathogen-free animals), this should only be done with compelling justification. At the very least, reporting, and, when possible, controlling for this variation in housing, is important for understanding health and wellbeing outcomes. For example, AJP now requires reporting of the exact social conditions of housing even though such details are recommended and not required by the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines. Similarly, the impact of sex and age on immune system function and response to infection is becoming clearer (Giefing-Kröll et al., 2015; Haberthur et al., 2010; Ingersoll, 2017). However, these and other relevant variables are not always considered when designing studies or documented upon publication. Thus, guidelines for rigor and reproducibility in NHP research must account both for concerns NHPs share with other model organisms—like sex as a biological variable—and other unique features of the NHP model - like inclusion in multiple studies over the lifespan and a relatively outbred genome. Furthermore, given that NHPs are the subjects in a wide variety of studies in disciplines as varied as infectious disease, metabolic disease, neuroscience, and behavior, formulating guidelines that cover multiple disciplines is of the utmost importance.

One potentially rich testing ground for establishing multidisciplinary guidelines for rigor and reproducibility for NHP research is the NPRC consortium, which is composed of the seven NIH-supported NPRCs, and which is, by design, multidisciplinary. Although NHPs are research subjects in a wide variety of laboratories and centers around the globe, the NPRC consortium represents a fairly unique system and environment for carrying out NHP work (nprc.org; NPRCresearch.org). The NPRCs are directly funded by the NIH, originally via congressional legislation in the 1960s that established the "regional primate research centers" (subsequently renamed from *Regional* to *National*), and are coordinated as a network via administrative and scientific processes, allowing for complementary resource development, policy, and integrative science. There are currently seven NPRCs (Yerkes in Georgia; Tulane in Louisiana; Southwest in Texas; and California, Oregon, Washington, and Wisconsin in their respective states). Specific NHP resources and scientific focus vary across centers by design. All NPRCs are united by a mission to improve human health while providing environments that promote NHP well-being and facilitate large-scale, collaborative, interdisciplinary research. NPRC principal investigators are often affiliated with NPRC-associated academic institutions and broadly support both basic and translational research via "affiliate scientist" programs that allow external investigators to carry out NHP work at the centers. In some cases, significant breeding populations at the centers provide NHP resources to laboratories outside the NPRC system as well. In addition to providing unique resources and rich intellectual environments for those carrying out this work, NPRCs are able to advance science and ultimately empirically guide policy around the care and use of NHPs in research, because their focus is almost exclusively NHPs.

3 | TRANSPARENCY AS A PRIMARY WAY TO IMPROVE RIGOR AND REPRODUCIBILITY

For the purposes of this review, when we discuss research characterized by rigor, we are referring to research that is: (a) well-planned with regard to the core questions such that bias is minimized; (b) well-executed with regard to the experimental design; and (c) appropriately analyzed and interpreted without bias. In this way, rigor ensures the robustness of research. When we discuss reproducible research, we are referring to the ability of scientific work to be repeated, yielding the same or biologically consistent outcomes. Efforts to improve rigor and reproducibility are often, if not always, grounded in the assumption that rigorous research is more likely to be reproducible (Munafò et al., 2017). It is important to note that the extent to which research is rigorous and reproducible says little about the extent to which it is translatable (from animal model to human) (Munafo & Davey Smith, 2018). For example, mouse models of sepsis largely converge on a small set of mechanisms that generate this significant health issue, yet research carried out in murine models does not consistently translate to humans (Seok et al., 2013; Stortz et al., 2017). This distinction between research that is evaluated to be rigorous and that can be reproduced and research that translates to humans is similar to the distinction between precision (hitting the same target over and over again) and accuracy (hitting the right target). Improving research generally, and NHP research, in particular, requires addressing both precision (rigor and reproducibility) and accuracy (translatability). The latter is outside the realm of discussion here, with the exception of noting the power of the NHP model as discussed previously. Although the specific approach to improving rigor and reproducibility may ultimately be tailored to individual scientific disciplines (because norms vary with regard to what it means to be robust), a number of themes are relevant to all scientific disciplines and are also relevant to work conducted with NHPs. Ultimately, guidelines for improving rigor and reproducibility, such as those set forth by the NIH and scientific societies, focus on the goal of increasing the robustness of the science and improving the ability to evaluate its strength. Mechanisms that increase transparency and provide pathways to enhance the rigor and reproducibility of research processes include methods, design, and analytical approaches that can be evaluated separately from research outcomes (Collins, 2014; Landis et al., 2012).

4 | PROVIDING ENOUGH DETAIL FOR PROPER EVALUATION

A consistent theme that arises in addressing reproducibility issues has been an inability to replicate published findings due to insufficient detail provided in the original published reports. The need for sufficient detail is driven not only by the goal of reproducibility, but it is also necessary for the scientific community to determine whether the research was conducted with rigor and whether or not the findings are valid. Determining the robustness of research is particularly important in cases where experiments generate null results. The null results of robust experiments may be interpreted as the absence of an effect, whereas the null results of experiments that are not robust may emerge, not because the effect does not exist, but rather because the experiment failed to elucidate the effect. Although this may be the normative level of reporting taught in our laboratories, the reality is that many experimental and analytic details do not make it into print. This may reflect journal reporting standards, biases in peer review, or failure to recognize that certain types of details (e.g., animal experimental rearing and experimental histories) may be critically important, both for shaping the experimental outcomes of a given study and generalizing its findings. At first blush, the solution to solving the transparency issue may be to simply report more detail in methods and data analyses sections and to share data. This raises important secondary questions: What additional details need to be reported, the mechanisms by which it is reported, and the processes by which data should be shared? At least three established efforts that improve transparency can be applied to NHP research: establishing normative protocols for carrying out work, including tracking and reporting relevant details of study design; preregistration; and open science practices that include establishing resources for and carrying out data sharing. We address these three areas below.

4.1 | Establishing normative protocols for experimentation and reporting

One of the historical approaches to addressing issues related to rigor has come from vested parties (e.g., societies and scientific journals) that have developed guidelines for assay or protocol standardization and reporting, sometimes under the definition of resource authentication. Although not specific to NHP research, these guidelines are present in disciplines that use NHPs in research and that are represented at the NPRCs. For example, guidelines on verifying and authenticating antibodies were published both in *Endocrinology* (Gore, 2013) and the *Journal of Comparative Neurology* (Saper, 2005). *Endocrinology* guidelines stipulate that scientists must provide verification that a given antibody binds to a specific target antigen in both experimental and control cases and provides a number of examples of ways this can be done and what must be reported in publications. The *Journal of Comparative Neurology* guidelines stipulate four pieces of information that must be considered and addressed: identifying information for the antibody; information about the preparation of the antibody; information about how the specificity of the antibody was determined; and controls that are present for immunostaining. In a similar vein, the *Journal of Clinical Endocrinology and Metabolism* at the same time instituted a requirement for determination of steroid hormone levels using liquid chromatography-tandem mass spectrometry (Handelsman & Wartofsky, 2013).

Scientific societies and journals can also influence what details get reported in publications by establishing reporting standards for scientific procedures that are common to their communities and then requiring that reporting as a condition of publication or presentations. These reporting standards increasingly take the shape of presubmission checklists that provide specific details about what must be reported and require authors to indicate that they have reported those details or provide specific information about why they are not reporting them. Individual journals such as *Circulation Research* now require the use of checklists to address methodological transparency, to ensure that adequate information about subjects (including animals) is shared, and mandate data sharing (Bolli, 2017). Journals from Cell Press, the family of *Nature* journals, and *BioMed* have developed and adopted reporting checklists specific to their own journals, focused largely on reporting methodological details (Marcus, 2016). Cross-journal efforts also demonstrate the potential for checklists that have a greater normative reach. For example, with the goal of establishing minimum reporting standards in life sciences, the Materials Design Analysis and Reporting (MDAR) Checklist was tested by 13 journals of various scope (http://blogs.nature.com/ofschemesandmemes/2019/10/21/journals-test-the-materials-design-analysis-reporting-mdar-checklist). The MDAR (https://osf.io/bj3mu/) asks authors to report information on antibodies (source, reagents), cell materials, experimental animals (including species, sex, strain, origin), plants and microbes, human participants, step by step study and laboratory protocols, study design (sample size, randomization, blinding, inclusion/exclusion criteria), in-laboratory replication, ethics, attrition, statistics, data availability, and code availability. Piloted with 289 manuscripts across the journals, 80% of authors and editors reported that the checklist was useful, and revisions of the MDAR are being undertaken based on the pilot study. These observations are echoed in empirical studies that demonstrate that publications whose editorial process included checklists, compared to manuscripts at the same journals which did not require checklists, report more methodological details, including those typically deemed necessary to evaluate the robustness of research (Han et al., 2017; NDQIP Collaborative Group, 2019).

Particularly germane to NHP research are the ARRIVE guidelines and checklist developed by the United Kingdom's National Center for the Replacement, Refinement & Reduction of Animals in Research (NC3Rs), and translated from English into nine languages (https://www.nc3rs.org.uk/arrive-guidelines) (Kilkenny et al., 2010). Originally published in 2010, the ARRIVE guidelines were developed in response to a survey that identified serious omissions in reporting of studies that were publicly funded in the United Kingdom and United States; they were developed specifically to fill a gap in other checklists that do not require adequate information specifically related to carrying out live animal research (Kilkenny et al., 2009). Like the checklists described above, the ARRIVE checklist is designed to be used when manuscripts are being prepared for submission although the guidelines cover each section of a standard manuscript. It is important to note that a randomized control trial of the ARRIVE guidelines suggested minimal, if any, improvement in reporting (Hair et al., 2019), demonstrating that changing the norms of reporting in animal science may be particularly difficult (Enserink, 2017). Guided by two randomized control trials (one in collaboration with *PLoS One*) and information provided by users and journal editors, the ARRIVE guidelines were recently updated (https://www.nc3rs.org.uk/revision-arrive-guidelines) (Percie du Sert et al., 2020). The update includes elaboration of what each criterion means and explains the rationale for including it, revision of some of the items, and organization of the original set of items into two sets that vary in priority. Criteria are divided into "essential" and "recommended," and the recent NIH report on enhancing rigor in animal research recommends compliance with the ARRIVE 2.0 guidelines (https://acd.od.nih.gov/documents/presentations/06112021_RR-AR%20Report.pdf). One additional change that might make the ARRIVE guidelines better suited to NHP work would be explicit consideration of the long lifespan of NHPs. A given animal may be a participant in multiple studies over its lifetime, with the data distributed across multiple laboratories and publications; reporting this may be important for understanding variability in experimental outcomes.

The reporting guidelines and checklists discussed above represent efforts to clarify experimental details at the time of publication, an important step towards transparency. Although the guidelines improve our ability to evaluate published science, they do not necessarily improve the integrity of the science that is carried out if they are not used at the time of experimental design and implementation. There has been some movement on the idea that checklists should be used when designing studies, and at least one example exists in the domain of animal research. The Planning Research and Experimental Procedures on Animals: Recommendations for Excellence (PREPARE) guidelines, which have their origins at the Norwegian School of Veterinary Science, propose 15 categories that should be considered when designing a study. These categories of information cover the entire research process, from the initial literature search at the conception of the study to necropsy (Smith et al., 2018). Adopting such guidelines consistently at the experiment preparation phase represents one potentially valuable step forward in ensuring that work that is carried out will be robust.

4.2 | Preregistration

One process that has been offered as a solution to improve transparency and to ensure rigor and reproducibility is preregistration (Nosek & Lakens, 2014; Nosek et al., 2018). The central premise of preregistration is that aspects of experimental design and analysis are documented *before* data collection and/or data analysis occurs. In some cases, these documents are peer-reviewed before the work being carried out (e.g., Part 1 of Registered Reports) to inform how the work is done (Nosek & Lakens, 2014). This allows for the methods and analysis protocol to be evaluated separately from the scientific outcomes. If the protocol is deemed robust, then the results can be published in participating journals regardless of whether there are statically significant effects or null results. For some journals, these documents are reviewed as part of a standard peer-review process.

Preregistration typically occurs using standardized templates that request particular information about experimental design, samples, sample sizes, and their calculation, and intended analyses that are housed on servers that date and time stamp their submission, even if they are not made immediately public. The major champion of preregistration has been the Center for Open Science (https://osf.io/), which offers templates for preregistration (although none specific to animal research at the time of this writing). The forms of preregistration vary in terms of the amount of information that is disclosed and when it is disclosed in the publication process, so the process can be adapted to meet the needs of a wide variety of stakeholders. Preregistration requires scientific planning as well as articulation of that plan in a way that can be evaluated and published. Preregistration templates encourage documentation of which analyses are done in the context of discovery (exploratory) vs those that are designed to test a specific hypothesis (confirmatory). This serves to document what scientists intend to do to prevent HARKING (hypothesizing after results are known; Kerr, 1998) and p-hacking (carrying out statistical analyses until a significant result is found (Simmons et al., 2011)—two processes that have contributed to the reproducibility crisis. Preregistration does not, however, necessarily solve the problem of carrying out poorly theorized, modeled, or informed work (Smaldino, 2019).

4.3 | Data sharing

Data sharing can enhance rigor and reproducibility in a number of ways, including allowing broad scientific communities access to data to carry out independent analyses to confirm published effects and/or build upon established datasets. Models for data sharing exist in multiple scientific domains (e.g., neuroscience: https://www.nwb.org/, and immunology https://www.immuneprofiling.org/hipc/page/show), but are arguably most well-established in genomics. In fact, replication failures were one of the reasons that data-sharing norms were proactively changed in genetics as a mandate from the NIH. The "replication crisis" has been raging in genetic association studies for almost two decades (Hirschhorn & Altshuler, 2002). In 2007, NCI and NHGRI hosted a working group on replication in association studies and developed a list of study details, data issues, methodological disclosures, and deposition requirements that should be considered by authors, reviewers, and journals when evaluating published association studies. Additionally, they set standards for the replication of associations (NCI-NHGRI Working Group, 2007), but replication remains difficult in NHP work due to the challenges of accessing sufficient animals. In 2008, the NIH began mandating the deposition of genetic data in publicly accessible databases (NOT-OD-07-088; https://grants.nih.gov/grants/guide/notice-files/not-od-14-124.html), a requirement enforced by most major journals. The final version of the Genomic Data Sharing Policy was published in 2014 (79 FR 51345; https://www.federalregister.gov/documents/2014/08/28/2014-20385/final-nih-genomic-data-sharing-policy). One of the most popular databases for deposition of basic genotype data, NCBI's dbSNP, closed to nonhuman data in 2017 and, as a result, the European Variation Archive hosted by the European Bioinformatics Institute is now the primary repository of genomic variation for many nonhuman species. However, for rhesus macaques, the macaque genotype and phenotype resource (mGAP; https://mgap.ohsu.edu/) provides searchable access to richly annotated variant data identified using genome-wide sequencing approaches (Bimber et al., 2019). With variant data spanning rhesus macaque populations at each of the NPRCs, the mGAP data resource enables the informed selection of animals based on genetic information, improving the reliability and reproducibility of findings across research disciplines. Gene expression data are frequently archived in NCBI's Gene Expression Omnibus database, which includes more than 18,000 NHP samples. Protein data from any species can be deposited to the UniProt database funded by NIH and the European Molecular Biology Laboratory. The deposition of raw data in these repositories is insufficient to allow for the true replication of findings. This is because, unlike DNA sequence data, which is largely the same whether collected from blood or brain in infants or elderly animals, epigenetic (including methylation, acetylation, micrornas, long noncoding RNAs, and other data types), gene expression, proteomic, and metabolomic data have the additional challenge of being highly sensitive to both tissue type and environmental conditions, including time of day, fasting conditions, etc. NHPs, because of their large body size and phylogenetic similarity to humans, are uniquely suited for the collection of tissue biopsies longitudinally or at necropsy. An excellent example of an NHP tissue bank is the Monkey Alcohol Tissue Research Resource (https://gleek.ecs.baylor.edu/; (Daunais et al., 2014), which provides investigators with tissue samples and related phenotypic measures from NHPs subjected to a standard alcohol consumption protocol. However, in other cases, existing metadata from biorepositories may be insufficient to document differences in collection and storage methods or animal level conditions that are not directly related to experimental parameters but may have a substantial influence on omic data generation.

Beyond individual investigator data sharing, there are ongoing community efforts to improve the utility of NHP omic data, including through the improvement of NHP genome annotations. Accurate and complete reference genomes are essential for any genetic or sequence-derived research. Although the human genome is extremely well-annotated and projects like the 1000 Genomes Project have captured more than 88 million variants (Auton et al., 2015), reference genomes for rhesus and cynomolgus macaques, baboons, vervet monkeys, and marmosets remain relatively poorly annotated, with little understanding of the genetic diversity within these species, as the number of individuals sequenced is in the tens rather than the thousands (Harding, 2017). Significant progress has recently been made with the rhesus macaque genome, with a new build based on >800 animals (Warren et al., 2020). Without an understanding of the variation present in the various species, however, it is difficult to evaluate the effect of individual genetic variants on phenotypes of interest. This recent rhesus genome analysis begins to address the issue of genome diversity (Warren et al., 2020). Furthermore, poorly annotated genomes hinder efforts to include animals of diverse genetic backgrounds in vaccines and other studies to maximize translational potential. Efforts are underway to sequence the genomes of additional animals, with RNA and protein analyses to follow. Due to the small number of NHPs in most studies, the vast majority are underpowered to accurately determine the magnitude of genetic effects on traits of interest. When leveraged correctly, the pedigree structure can help improve statistical power, but both small sample size and small effect size reduce power, such that not only is it more difficult to detect associations, but those associations that are nominally significant have a higher likelihood of being spurious. The issues of false positives and failure to replicate are by no means specific to genetic studies (see Button et al., 2013, for an excellent review in neuroscience), but they have frequently been highlighted in this domain. Sharing genetics data across investigators and centers is a critical first step to ensuring that enough cases are available for analysis and comparisons.

Although genomics has established a standard for the deposition of data into major databases requiring harmonization and long-term storage capacities, depositing data into major databases is only one model for data sharing. For example, so-called clearing houses for sharing data of particular types are becoming more widespread. Building on established models for sharing human neuroimaging data, the PRIMatE Data Exchange (PRIME-DE; http://www.fcon%5f1000.projects.nitrc.org/indi/indiPRIME.html; Milham et al., 2018) provides access to independently collected neuroimaging datasets and information about data quality via the International Neuroimaging Data-sharing Initiative. Furthermore, individual investigators, including those who work with NHPs, are increasingly sharing data associated with specific papers or analyses via databases like Dryad (https://datadryad.org/stash) or project archives on the Open Science Framework (osf.io) or GitHub (github.com). These investigator choices align with the growing consensus among publishers that all data should be shared, with a preference for citable datasets assigned a Digital Object Identifier (Cousijn et al., 2018). LabKey, a laboratory information management system tool, has been used as the foundation for the development of specific colony health databases used by some NPRCs, and allowed scientists at the Wisconsin NPRC to easily share data as it was being generated during the early stages of their Zika virus research, and this approach is being replicated for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) research (https://openresearch.labkey.com/project/home/begin.view). While the norms are slowly changing, ultimately, policies that originate with funding agencies or journals that require data sharing will have a higher likelihood of ensuring that data are made public.

5 | ADDRESSING CHALLENGES IN RIGOR AND REPRODUCIBILITY SPECIFIC TO NHP RESEARCH

One of the major challenges to rigor and reproducibility efforts in NHP research is that the constraints of the 3Rs have been historically misinterpreted in ways that result in studies being carried out with small sample sizes and rarely if ever, replicated. As a result, the robustness of studies may be particularly impacted by features relating to statistical power, associated with both the number of animals tested as well as the number of trials each animal completes (i.e., both within and between individual power), the design of experimental conditions (e.g., including appropriate control conditions), including representative samples (e.g., with regard to age or sex), and the appropriate statistical methods to evaluate those sources of variance. Additionally, small-sample NHP studies benefit from data analysis strategies that evaluate or control for sources of variance because NHPs are genetically variable and articulate the difference between exploratory and confirmatory analyses as defined above, although such methods and clarity around analyses are not necessarily normative. NHP research has, like other domains of science, historically overemphasized using a criterion of $p < 0.05$ as the determining factor for what findings are meaningful or important, with less emphasis on evaluation of raw data, effect sizes, and probability estimates. Another factor that likely contributes to small sample sizes is the level of expertise or training in statistical methods of investigators. Although this is alleviated to some extent by the increasingly common use of biostatisticians as contributors and consultants on research projects, a greater emphasis on training of students and new investigators can also contribute to rigor and reproducibility in NHP research.

6 | REPLICATION EFFORTS CAN BE COLLABORATIVE

The importance of replication of individual studies has been emphasized in many domains of science, particularly behavioral science, and large scientific projects are underway to replicate core findings of a given field. Such efforts are typically considered unfeasible with NHPs because of the practical and ethical constraints associated with replicating NHP studies. Ongoing replication efforts can take the form of both individual laboratories attempting to replicate other laboratories' studies or large group efforts in which a given study (or studies) is replicated across "Many Labs" (Ebersole et al., 2016; Klein et al., 2014, 2018). A recent example is the Center for Open Science's project to replicate murine cancer research findings (https://cos.io/rpcb/and_https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology) and the International Brain Laboratory's virtual laboratory (https://www.internationalbrainlab.com/#home). One major issue that these efforts bring to light is the generalizability of particular findings; i.e., if laboratory A reveals a significant difference between two conditions and laboratory B does not find that difference, what are the features of laboratory A's sample that do not generalize to laboratory B's sample? Factors such as age, race, ethnicity, language, and health status may all play important roles in shaping effects, even if they are not reported as mediators or moderators. Carrying out work across many laboratories in which such factors vary is one approach to increasing research robustness. The “Many Labs” concept has been explicitly imported into behavioral NHP research (Many Primates et al., 2019), but with a specific emphasis on combining data across a number of laboratories and species to answer a core question, either to increase sample size or to test hypotheses about the generalizability of effects across phylogeny.

Recent efforts at NPRCs demonstrate the effectiveness of combing data to unearth effects that would not have been discernable in small studies. For example, following the 2015–2016 Zika virus epidemic in South and Central America, scientists at the NPRCs mobilized to study the virus and its impact on developing fetuses. Research teams across the NPRCs noted fetal mortality following Zika virus infection, but it seemed to vary based on when the fetuses were infected and the small sample sizes at each center precluded drawing conclusions. When those data were pooled (to a total of $N = 50$) across Centers, however, the pattern became clear and statistically significant; i.e., that fetuses infected during the first trimester had a significantly greater chance of dying than those infected later in gestation (Dudley et al., 2018). Currently, a similar effort is underway for studies of SARS-CoV-2 infection, with regular data sharing and evaluation of findings. Efforts like these that combine samples across laboratories and centers are an important step forward because they allow sample sizes to be increased and inherently improve the generalizability of the science because it is being carried out at multiple sites with innate variation across sites with regards to animal care, standard operating procedures (SOPs), etc. These efforts capitalize on an existing NHP model, where small pilot studies are used to develop and ensure the potential effects of proposed interventions or the suitability of specific experimental protocols before larger studies are undertaken. There may be an important role in the context of discovery for pilot studies with small sample sizes (Bacchetti et al., 2011), although such studies may lead to an overestimation of necessary sample sizes which itself has potential ethical implications (Gaskill & Garner, 2020).

Despite these multi-site efforts, attempts to replicate most NHP studies are rare because of practical and ethical constraints in terms of access to animals as already mentioned. Even when sample sizes are large, publishing replication studies can be a major challenge as a result of publication norms across NHP science domains (i.e., nonsignificant effects are often not published), lack of familiarity with processes for publishing replications, and strict editorial guidelines at some of the major journals that publish NHP work. Given that it will likely take time and effort to change these norms, ensuring the robustness of individual NHP studies is crucial, and creating mechanisms by which NHP scientists can carry out pilot work and then build upon others' science via data sharing is critical.

7 | CAPITALIZING UPON ESTABLISHED EFFORTS AND RESOURCES

Significant resources have been invested in generating, implementing, and evaluating the efficacy of checklists, preregistration, and data sharing. NHP scientists need not reinvent the wheel, but certainly need to embrace using the wheels that exist, either through mandated changes at the time of publication (via journals) or at the time of project planning as required or promoted by either granting agencies, institutions, or the NPRCs. An overview of the contributions of various parties to improve rigor and reproducibility, both for NHP research and for biomedical research in general, is shown in Figure 1.

Existing authentication or standardization guidelines and guidelines for validating resources should be identified from the journals and societies that have generated them, centralized, and then integrated into the core workflows at the NPRCs and in other NHP laboratories. In this view, the NPRC system represents a fertile ground for developing and testing such workflows that can then be generalized to NHP research more broadly. General checklists like the MDAR and animal research-specific checklists like PREPARE and ARRIVE are applicable to NHP research and could easily be implemented if journals or institutions began to require them for every NHP study. Existing models of preregistration—particularly those that encourage transparency in methods, clear designation of whether the research is being conducted in the context of discovering (exploratory analyses) or hypothesis testing (confirmatory analyses), and clarity around constraints on the generalizability of the studies—can be imported into NHP research simply by generating a series of NHP specific preregistration templates and then encouraging their use. This would require partnerships with journals to develop “registered reports” formats (where preregistrations are evaluated for robustness separately from the outcomes of the studies), to acknowledge when studies have been preregistered (e.g., with the ‘badge’ system; https://cos.io/our-services/open-science-badges/; as is done at AJP), or to include null results when preregistered experimental design and analysis has been deemed to be robust. Scientists should be encouraged to share their data by depositing it into established databases and depositories, and NPRCs should catalog and index shared data that was generated with their NHP resources. Indeed, the NIH has recently issued a final policy for data management and sharing that specifies the requirements for data generated from NIH funding, and which will go into effect in 2023 (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html).

8 | GENERATING NEW EFFORTS AND RESOURCES

Ultimately, the existing efforts to improve rigor and reproducibility may not account for all of the constraints and benefits of working with NHPs, and because of their collective science and inherent interdisciplinary nature, the NPRCs are in a unique position to lead the way to create new efforts to improve NHP science around the globe. Here we propose two interrelated efforts that are already partially in place, or for which there is established infrastructure at some of the centers, but for which center-wide coordination should be undertaken. All of these cases leverage the resources inherent in the NPRC system but could be easily modified and implemented at other institutions; similarly, putting these efforts into place at NPRCs could also influence how NHP research is carried out at other institutions (e.g., having animal-level metadata that travels with individual animals when they leave the NRPC where they were reared and are transferred to a laboratory at a non-NPRC site).

8.1 | Use of pilot grant programs to provide an opportunity for discovery and collaboration

The NPRCs have established pilot grant programs that provide funds for investigators to carry out studies that have the promise of securing federal funding in the future (https://nprcresearch.org/primate/pilot-programs.php). These funds are often used to bring researchers outside of the NPRCs into NHP research. Such awards can also be used to allow established investigators to carry out novel experiments, develop novel resources for improving science (protocols for standardizing assays, generation of standard datasets that can be shared), or carry out cross-center projects where small samples at individual centers are pooled to address larger questions and evaluate the generalizability of effects across contexts. Such work across centers would be facilitated by the generation of animal level meta-data (e.g., experimental history) and center-based or center-general mechanisms for data sharing. Inherent in this mechanism is the assumption that the use of a minimal number of animals will garner sufficient data to determine if future investment is warranted.

Indeed, numerous successes in animal model development (Burwitz, et al., 2017a, 2017b; Lopez et al., 2014), vaccine testing (Coban et al., 2004; Datta et al., 2017; Petersen et al., 2014), and drug treatment efficacy have arisen from NHP pilot studies. Examples of pilot studies that successfully utilized limited numbers of macaques include a study of five macaques that demonstrated the persistence of the Lyme disease pathogen following a recommended treatment protocol (Embers et al., 2012a) and subsequently devised a strategy for diagnostic test development (Embers et al., 2012b), and a study of four macaques that demonstrated expression of the human NTCP receptor on hepatocytes is sufficient to support full HBV infection of macaques (Burwitz et al., 2017a). Although large sample sizes can be essential for detailed quantitative measures of effect and inter-animal variation, small pilot studies illustrate that experiments that result in a binary black-or-white answer, measure qualitative differences or trends, or uncover new phenomena, can all be achieved with small sample size. Furthermore, some scientific questions can be answered with small sample studies that do not necessitate carrying out subsequent larger studies.

Adaptive trial design (Bauer et al., 2016), in which the data from the initial stages of a study can inform potential adjustments to the overall study design, are increasingly employed in clinical research but are also an option for NHP studies. Such flexible designs can increase the efficiency of cohorts with a limited number of subjects, which is relevant to NHP studies using a limited resource and can also be enhanced by the incorporation of Bayesian analysis techniques (Chevret, 2012) in addition to standard statistical approaches. Although these approaches are yet to be fully implemented in human studies (Pallmann et al., 2018), their introduction into NHP studies deserves serious consideration. In light of the role of NHPs as the penultimate preclinical model in the drug development pathway, the application of aspects of first-in-human study designs (Shen et al., 2019) is also pertinent as, like most NHP studies, these designs usually employ a relatively small number of subjects. In fact, an earlier analysis of the effect of sample size on the outcome of a series of first-in-human dose-escalation studies (Buöen et al., 2003), demonstrated that a sample size of 6–10 subjects was the necessary range to obtain useful data, with less than 6 being insufficient, but increasing the sample size to greater than 10 providing little additional power.

8.2 | Investment in established data sharing and establish mechanisms for sharing colony-wide data

Although broader aspects of data sharing are discussed above, there are certain aspects of data sharing that are specific to NHP colonies. Given that NHPs are a precious resource, it is critical that NHP scientists be willing (or mandated) and easily able to share experimental data and associated metadata without adding significant time or administrative burden. This should be supported across centers so that NPRC investigators are encouraged and have a mechanism to share and pool data. Such efforts can capitalize on established resources where they exist (e.g., genetics, neuroimaging, and neurophysiology databases) but may also require the development of new resources. One of the strengths of the NPRC system is the huge volume of data generated on animals in their colonies, some of whom may only be enrolled in investigators' studies for brief periods of time. This wealth of data about their rearing, health history, and tissue collection and preservation upon their deaths, is a valuable national resource, and making it available to scientists around the globe could help speed scientific discovery and ultimately improve human and NHP health. Furthermore, capitalizing on existing data -whether to determine appropriate sample sizes for new studies, refine methods, or even answer key medical questions without having to involve new animals in experiments- could certainty aid in our goal to address the 3Rs. An excellent example of this approach is the recently announced open-source resource for NHP optogenetics (Tremblay et al., 2020).

The NPRCs, as well as most large NHP laboratories, maintain extensive animal records, including health, genetics, experimental history, and origins (rearing conditions, the breeding facility where they were born, etc.) but this information may not be harmonized, easily searchable, or linked permanently to the individual animal. Stakeholders from various disciplines of NHP research should work together to determine what features may be relevant to their science and the science of others. For example, behavioral scientists increasingly recognize that early rearing conditions can cause permanent variance in an individual's behavior and biology, and present social conditions and changes in social conditions in adulthood exert similar although potentially not as long-lasting impacts. Interdisciplinary perspectives are required to determine what the content of the meta-data should be. Ultimately, these data should be harmonized, searchable, and provided in a flexible format to investigators when they are designing studies, selecting animals, accessing banked biological samples, or purchasing animals from other facilities. Reporting this metadata in the form of supplementary data during publication would also allow scholars to draw conclusions about the generalizability of studies across sites and species. As illustrated by the successful sharing of data related to Zika across NPRCs, LabKey is one potential vehicle for this harmonized meta-data, but the consistent reporting of the data across sites and studies is more important than the software selected.

9 | QUALITY ASSURANCE (QA) PROGRAMS

Variability is inherent in every step of an experimental procedure. However, for those processes that are routinely performed, reduction in variability should be a priority, as this can add validity both to the data acquired and its interpretation and conclusions. Although QA regarding laboratory techniques is often carried out at the level of individual labs (e.g., validating antibodies before their use (Gore, 2013; Saper, 2005), and institutional QA programs exist in many contexts (e.g., pharma and biotechnology companies), NPRC and university-level QA programs are less common. Every NPRC has SOPs and QA programs for practices involving animals. However, the laboratory techniques related to samples derived from the animals are not routinely standardized. One example of an assay that is fraught with reproducibility issues is the ELIspot. This assay is open to subjective quantification and, even with automated systems, evaluation of plates is operator-dependent (Cox et al., 2006; Janetzki et al., 2004). As such, QA programs have been instituted to mitigate variability in these and other assays, exemplified by the Duke University External Quality Assurance Program Oversight Laboratory. The Tulane NPRC (TNPRC) has similarly created a Quality program to assure reproducibility in standard assays. The TNPRC is the Coordinating Center for the COVID-19 response, including the development of the NHP model and testing of vaccines and therapeutics. Initially, steps in the workflow that contribute to potential variability were identified. The implementation included the development of best practices, standard protocols, and oversight by a QA Specialist reporting to the Director of Quality Assurance, administered by the office of the Vice President of Research. Best practices include intellectual honesty and communication of errors, operator training, experiment documentation, safe and organized long-term storage of data, open and efficient dialogue between core laboratories and investigators, and assay supervision by dedicated core facility managers. The standardization of protocols involves testing varying SOPs for intra-assay, inter-assay, and inter-operator variability to ensure reproducible and accurate results and performing rigorous Quality Control (QC) and QA checks on instruments. The use of biological reference controls to test both reproducibility and to ensure confidence in longitudinal data results, and testing of all new reagent lots provided by manufacturers are integral to the QC program as well. The program was initiated with flow cytometry and further applied throughout other core services, such as real-time PCR/RT-PCR and Luminex®-based assays. A similar program exists at the Texas Biomedical Research Institute and is utilized by the Southwest NPRC. Plans are underway to implement QA/QC processes in lab protocols across the NPRCs. Standardization and sharing of these QA/QC procedures across labs will facilitate the robustness of science as well as the ability to conduct multi-site studies and replicate findings.

10 | CONCLUSIONS

Enhancing rigor and reproducibility in the biomedical sciences is truly a collective effort, as outlined in Figure 1, and its collective nature is made possible by both community standards and norms and the individual efforts of each investigator. Greater recognition of the extent to which much published science has not been carried out in a rigorous way and thus slowed progress in basic and translational/health science has led to a science-wide evaluation of how best we can, as a community and as individuals, change the norms in both how we carry out and how we report our science to improve its robustness and other scholars' ability to evaluate it. NHP research is an interesting test case in which to deploy rigor and reproducibility efforts because it is inherently constrained, both ethically and practically, in ways that other types of science are not and it cross-cuts disciplinary boundaries that themselves have their own norms. As a result, what efforts are deployed must be well-tailored to the NHP model and the constraints of carrying out NHP science (e.g., simply increasing the sample size of NHP studies is not an option like it is in some fields), while simultaneously being broad enough to be efficacious across disciplines. Despite these constraints and challenges, the NIH's significant investment in NHP research via the NPRCs creates a unique testing ground for rigor and reproducibility efforts, before they are deployed more generally in NHP science. To that end, our goal in this Perspective is to provide broad consideration and specific examples of how the principles of scientific rigor apply to NHP research, setting the stage for coordinated efforts, initially across NPRCs and then across NHP labs more broadly, to fundamentally improve NHP research.

CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest.

ORCID

Eliza Bliss-Moreau http://orcid.org/0000-0002-0740-5612
Elizabeth A. Buffalo https://orcid.org/0000-0001-6326-9187
Ricki J. Colman http://orcid.org/0000-0001-9706-0525
Monica E. Embers https://orcid.org/0000-0003-4051-7592
John H. Morrison http://orcid.org/0000-0002-8292-0964
Ellen E. Quillen https://orcid.org/0000-0003-2033-8693
Charles T. Roberts http://orcid.org/0000-0003-1756-5772

REFERENCES

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., National Eye Institute, & N. I. H. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bacchetti, P., Deeks, S. G., & McCune, J. M. (2011). Breaking free of sample size dogma to perform innovative translational research. *Science Translational Medicine*, *3*(87), 87ps24. https://doi.org/10.1126/scitranslmed.3001628

Bauer, P., Bretz, F., Dragalin, V., Konig, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine*, *35*(3), 325–347. https://doi.org/10.1002/sim.6472

Bimber, B. N., Yan, M. Y., Peterson, S. M., & Ferguson, B. (2019). mGAP: The macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. *BMC Genomics*, *20*(1), 176. https://doi.org/10.1186/s12864-019-5559-7

Bolli, R. (2017). New initiatives to improve the rigor and reproducibility of articles published in. *Circulation Research*, *121*(5), 472–479. https://doi.org/10.1161/CIRCRESAHA.117.311678

Buffalo, E. A., Movshon, J. A., & Wurtz, R. H. (2019). From basic brain research to treating human brain disorders. *Proceedings of the National Academy of Sciences*, *116*(52), 26167–26172. https://doi.org/10.1073/pnas.1919895116

Buöen, C., Holm, S., & Thomsen, M. S. (2003). Evaluation of the cohort size in phase I dose escalation trials based on laboratory data. *Journal of Clinical Pharmacology*, *43*(5), 470–476. https://doi.org/10.1177/0091270003252243

Burwitz, B. J., Wettengel, J. M., Muck-Hausl, M. A., Ringelhan, M., Ko, C., Festag, M. M., & Sacha, J. B. (2017a). Hepatocytic expression of human sodium-taurocholate cotransporting polypeptide enables hepatitis B virus infection of macaques. *Nature Communications*, *8*(1), 2146. https://doi.org/10.1038/s41467-017-01953-y

Burwitz, B. J., Wu, H. L., Abdulhaqq, S., Shriver-Munsch, C., Swanson, T., Legasse, A. W., & Sacha, J. B. (2017b). Allogeneic stem cell transplantation in fully MHC-matched Mauritian cynomolgus macaques recapitulates diverse human clinical outcomes. *Nature Communications*, *8*(1), 1418. https://doi.org/10.1038/s41467-017-01631-z

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Capitanio, J. P., & Emborg, M. E. (2008). Contributions of non-human primates to neuroscience research. *The Lancet*, *371*(9618), 1126–1135. https://doi.org/10.1016/S0140-6736(08)60489-4

Capitanio, J. P., Mason, W. A., Mendoza, S. P., DelRosso, L., & Robers, J. A. (2006). Nursery rearing and biobehavioral organization. In G. R. In (Ed.) Sackett, G. P. & Elizas, K., *Nursery Rearing of nonhuman Primates in the 21st Century* (pp. 191–213). Springer.

Chevret, S. (2012). Bayesian adaptive clinical trials: A dream for statisticians only? *Statistics in Medicine*, *31*(11-12), 1002–1013. https://doi.org/10.1002/sim.4363

Coban, C., Philipp, M. T., Purcell, J. E., Keister, D. B., Okulate, M., Martin, D. S., & Kumar, N. (2004). Induction of Plasmodium falciparum Transmission-Blocking Antibodies in Nonhuman Primates by a Combination of DNA and Protein Immunizations. *Infection and Immunity*, *72*(1), 253–259. https://doi.org/10.1128/iai.72.1.253-259.2004

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, *505*, 612–613.

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., & Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*, 180259. https://doi.org/10.1038/sdata.2018.259

Cox, J. H., Ferrari, G., & Janetzki, S. (2006). Measurement of cytokine release at the single cell level using the ELISPOT assay. *Methods*, *38*(4), 274–282. https://doi.org/10.1016/j.ymeth.2005.11.006

Datta, D., Bansal, G. P., Grasperge, B., Martin, D. S., Philipp, M., Gerloff, D., & Kumar, N. (2017). Comparative functional potency of DNA vaccines encoding Plasmodium falciparum transmission blocking target antigens Pfs48/45 and Pfs25 administered alone or in combination by in vivo electroporation in rhesus macaques. *Vaccine*, *35*(50), 7049–7056. https://doi.org/10.1016/j.vaccine.2017.10.042

Daunais, J. B., Davenport, A. T., Helms, C. M., Gonzales, S. W., Hemby, S. E., Friedman, D. P., & Grant, K. A. (2014). Monkey alcohol tissue research resource: banking tissues for alcohol research. *Alcoholism, Clinical and Experimental Research*, *38*(7), 1973–1981. https://doi.org/10.1111/acer.12467

DiVincenti, L., Jr., & Wyatt, J. D. (2011). Pair housing of macaques in research facilities: A science-based review of benefits and risks. *Journal of the American Association for Laboratory Animal Science: JAALAS*, *50*(6), 856–863. https://www.ncbi.nlm.nih.gov/pubmed/22330777

Dudley, D. M., Van Rompay, K. K., Coffey, L. L., Ardeshir, A., Keesler, R. I., Bliss-Moreau, E., & O'Connor, D. H. (2018). Miscarriage and stillbirth following maternal Zika virus infection in nonhuman primates. *Nature Medicine (New York, NY, United States)*, *24*(8), 1104–1107. https://doi.org/10.1038/s41591-018-0088-5

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Embers, M. E., Barthold, S. W., Borda, J. T., Bowers, L., Doyle, L., Hodzic, E., & Philipp, M. T. (2012a). Persistence of Borrelia burgdorferi in rhesus macaques following antibiotic treatment of disseminated infection. *PLoS One*, *7*(1), e29914. https://doi.org/10.1371/journal.pone.0029914

Embers, M. E., Hasenkampf, N. R., Jacobs, M. B., & Philipp, M. T. (2012b). Dynamic longitudinal antibody responses during Borrelia burgdorferi infection and antibiotic treatment of rhesus macaques. *Clinical and Vaccine Immunology*, *19*(8), 1218–1226. https://doi.org/10.1128/CVI.00228-12

Enserink, M. (2017). Sloppy reporting on animal studies proves hard to change. *Science*, *357*(6358), 1337–1338. https://doi.org/10.1126/science.357.6358.1337

Estes, J. D., Wong, S. W., & Brenchley, J. M. (2018). Nonhuman primate models of human viral infections. *Nature Reviews Immunology*, *18*(6), 390–404. https://doi.org/10.1038/s41577-018-0005-7

Gaskill, B. N., & Garner, J. P. (2020). Power to the people: power, negative results and sample size. *Journal of the American Association for Laboratory Animal Science: JAALAS*, *59*(1), 9–16. https://doi.org/10.30802/aalas-jaalas-19-000042

Giefing-Kröll, C., Berger, P., Lepperdinger, G., & Grubeck-Loebenstein, B. (2015). How sex and age affect immune responses, susceptibility to infections, and response to vaccination. *Aging Cell*, *14*(3), 309–321. https://doi.org/10.1111/acel.12326

Gore, A. C. (2013). Editorial: Antibody validation requirements for articles published in endocrinology. *Endocrinology*, *154*(2), 579–580. https://doi.org/10.1210/en.2012-2222

Gottlieb, D. H., Capitanio, J. P., & McCowan, B. (2013). Risk factors for stereotypic behavior and self-biting in rhesus macaques (Macaca mulatta): Animal's history, current environment, and personality. *American Journal of Primatology*, *75*(10), 995–1008. https://doi.org/10.1002/ajp.22161

Haberthur, K., Engelman, F., Barron, A., & Messaoudi, I. (2010). Immune senescence in aged nonhuman primates. *Experimental Gerontology*, *45*(9), 655–661. https://doi.org/10.1016/j.exger.2010.06.001

Hair, K., Macleod, M. R., & Sena, E. S. (2019). A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Research Integrity and Peer Review*, *4*, 12. https://doi.org/10.1186/s41073-019-0069-3

Han, S., Olonisakin, T. F., Pribis, J. P., Zupetic, J., Yoon, J. H., Holleran, K. M., & Lee, J. S. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PLoS One*, *12*(9), e0183591. https://doi.org/10.1371/journal.pone.0183591

Handelsman, D. J., & Wartofsky, L. (2013). Requirement for mass spectrometry sex steroid assays in the Journal of Clinical Endocrinology and Metabolism. *Journal of Clinical Endocrinology and Metabolism*, *98*(10), 3971–3973. https://doi.org/10.1210/jc.2013-3375

Harding, J. D. (2017). Genomic tools for the use of nonhuman primates in translational research. *Institute for Laboratory Animal Research Journal*, *58*(1), 59–68. https://doi.org/10.1093/ilar/ilw042

Hirschhorn, J. N., & Altshuler, D. (2002). Once and again-issues surrounding replication in genetic association studies. *Journal of Clinical Endocrinology and Metabolism*, *87*(10), 4438–4441. https://doi.org/10.1210/jc.2002-021329

Ingersoll, M. A. (2017). Sex differences shape the response to infectious diseases. *PLoS Pathogens*, *13*(12), e1006688. https://doi.org/10.1371/journal.ppat.1006688

Janetzki, S., Schaed, S., Blachere, N. E., Ben-Porat, L., Houghton, A. N., & Panageas, K. S. (2004). Evaluation of Elispot assays: Influence of method and operator on variability of results. *Journal of Immunological Methods*, *291*(1-2), 175–183. https://doi.org/10.1016/j.jim.2004.06.008

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and social psychology review: An official journal of the Society for Personality and Social Psychology, Inc*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology*, *8*(6), e1000412. https://doi.org/10.1371/journal.pbio.1000412

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F., Cuthill, I. C., Fry, D., & Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*, *4*(11), e7824. https://doi.org/10.1371/journal.pone.0007824

Klein, R. A., Ratliff, K. A., Vianello, M., Jr., Bahník, R. B. A. Š., Bernstein, M. J., & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., & Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., & Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, *490*(7419), 187–191. https://doi.org/10.1038/nature11556

Lopez, J. E., Vinet-Oliphant, H., Wilder, H. K., Brooks, C. P., Grasperge, B. J., Morgan, T. W., & Embers, M. E. (2014). Real-time monitoring of disease progression in rhesus macaques infected with Borrelia turicatae by tick bite. *The Journal of infectious diseases*, *210*(10), 1639–1648. https://doi.org/10.1093/infdis/jiu306

Machado, C. J. (2013). Maternal influences on social and neural development in macaque monkeys. In Clancy, K. H. K. B. H. & Rutherford, J. N. (Ed.), *Building Babies: Primate Development in Proximate and Ultimate Perspective* (pp. 259–279). Springer.

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., Watzek, J. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLoS One*, *14*(10), e0223675. https://doi.org/10.1371/journal.pone.0223675

Marcus, E. (2016). A STAR Is Born. *Cell*, *166*(5), 1059–1060. https://doi.org/10.1016/j.cell.2016.08.021

Milham, M. P., Ai, L., Koo, B., Xu, T., Amiez, C., Balezeau, F., & Schroeder, C. E. (2018). An open resource for non-human primate imaging. *Neuron*, *100*(1), 61–74. https://doi.org/10.1016/j.neuron.2018.08.039

Miller, L. A., Royer, C. M., Pinkerton, K. E., & Schelegle, E. S. (2017). Nonhuman primate models of respiratory disease: Past, present, and future. *ILAR Journal*, *58*(2), 269–280. https://doi.org/10.1093/ilar/ilx030

Munafo, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature*, *553*(7689), 399–401. https://doi.org/10.1038/d41586-018-01023-3

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

National Academies of Sciences, E., & Medicine. (2019). *Reproducibility and Replicability in Science*. The National Academies Press.

NCI-NHGRI Working Group. (2007). Replicating genotype-phenotype associations. *Nature*, *447*(7145), 655–660. https://doi.org/10.1038/447655a

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

NPQIP Collaborative Group. (2019). Did a change in Nature journals' editorial policy for life sciences research improve reporting? *BMJ Open Science*, *3*(1), 519–521. https://doi.org/10.1136/bmjos-2017-000035

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., & Jaki, T. (2018). Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine*, *16*(1), 29. https://doi.org/10.1186/s12916-018-1017-7

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., & Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biology*, *18*(7), e3000410. https://doi.org/10.1371/journal.pbio.3000410

Petersen, H., Nieves, W., Russell-Lodrigue, K., Roy, C. J., & Morici, L. A. (2014). Evaluation of a Burkholderia pseudomallei outer membrane vesicle vaccine in nonhuman primates. *Procedia Vaccinol*, *8*, 38–42. https://doi.org/10.1016/j.provac.2014.07.007

Phillips, K. A., Bales, K. L., Capitanio, J. P., Conley, A., Czoty, P. W., t Hart, B. A., & Voytko, M. L. (2014). Why primate models matter. *American Journal of Primatology*, *76*(9), 801–827. https://doi.org/10.1002/ajp.22281

Prescott, M. J., Clark, C., Dowling, W. E., & Shurtleff, A. C. (2021). Opportunities for refinement of non-human primate vaccine studies. *Vaccines(Basel)*, *9*(3), 284. https://doi.org/10.3390/vaccines9030284

Richter, S. H. (2017). Systematic heterogenization for better reproducibility in animal experimentation. *Lab Anim (NY)*, *46*(9), 343–349. https://doi.org/10.1038/laban.1330

Saper, C. B. (2005). An open letter to our readers on the use of antibodies. *Journal of Comparative Neurology*, *493*(4), 477–478. https://doi.org/10.1002/cne.20839

Sassenberg, K., & Ditrich, L. (2019). Research in Social Psychology Changed Between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Schapiro, S. J., Nehete, P. N., Perlman, J. E., & Sastry, K. J. (2000). A comparison of cell-mediated immune responses in rhesus macaques housed singly, in pairs, or in groups. *Applied Animal Behaviour Science*, *68*(1), 67–84. https://doi.org/10.1016/S0168-1591(00)00090-3

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., & Host Response to Injury, L. S. C. R. P. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(9), 3507–3512. https://doi.org/10.1073/pnas.1222878110

Shen, J., Swift, B., Mamelok, R., Pine, S., Sinclair, J., & Attar, M. (2019). Design and conduct considerations for first-in-human trials. *Clinical and Translational Science*, *12*(1), 6–19. https://doi.org/10.1111/cts.12582

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A., & Brattelid, T. (2018). PREPARE: guidelines for planning animal research and testing. *Laboratory Animals*, *52*(2), 135–141. https://doi.org/10.1177/0023677217724823

Stortz, J. A., Raymond, S. L., Mira, J. C., Moldawer, L. L., Mohr, A. M., & Efron, P. A. (2017). Murine models of sepsis and trauma: Can we bridge the gap? *Institute for Laboratory Animal Research Journal*, *58*(1), 90–105. https://doi.org/10.1093/ilar/ilx007

Suomi, S. J. (1999). Attachment in rhesus monkeys. In (Ed.) Shaver, J. C. P. R., *Handbook of attachment: Theory, research, and clinical applications* (pp. 181–197). Guilford Press.

Tannenbaum, J., & Bennett, B. T. (2015). Russell and Burch's 3Rs then and now: The need for clarity in definition and purpose. *Journal of the American Association for Laboratory Animal Science: JAALAS*, *54*(2), 120–132. https://www.ncbi.nlm.nih.gov/pubmed/25836957

Tremblay, S., Acker, L., Afraz, A., Albaugh, D. L., Amita, H., Andrei, A. R., & Platt, M. L. (2020). An open resource for non-human primate optogenetics. *Neuron*, *108*(6), 1075–1090. https://doi.org/10.1016/j.neuron.2020.09.027

Voelkl, B., Altman, N. S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., & Wurbel, H. (2020). Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience*, *21*(7), 384–393. https://doi.org/10.1038/s41583-020-0313-3

Warren, W. C., Harris, R. A., Haukness, M., Fiddes, I. T., Murali, S. C., Fernandes, J., & Eichler, E. E. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*, *370*(6523). https://doi.org/10.1126/science.abc6617

How to cite this article: Bliss-Moreau, E., Amara, R. R., Buffalo, E. A., Colman, R. J., Embers, M. E., Morrison, J. H., Quillen, E. E., Sacha, J. B., Roberts, C. T., & National Primate Research Center Consortium Rigor and Reproducibility Working Group. (2021). Improving Rigor And Reproducibility in Nonhuman Primate Research. *Am J Primatol*, 83, e23331. https://doi.org/10.1002/ajp.23331