Reproducibility Crisis in Science: A Discussion of the Disregard of ARRIVE Guidelines and Other Shortfalls of Pre-Clinical Research Reporting

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation	Knox, Christopher. 2020. Reproducibility Crisis in Science: A Discussion of the Disregard of ARRIVE Guidelines and Other Shortfalls of Pre-Clinical Research Reporting. Master's thesis, Harvard Extension School.
Citable link	https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365054
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Reproducibility Crisis in Science: A Discussion of the Disregard of ARRIVE Guidelines and Other Shortfalls of Pre-clinical Research Reporting

Christopher J. Knox

A Thesis in the Field of Biotechnology Management

for the Degree of Master Liberal Arts in Extension Studies

Harvard University

May 2020

Abstract

Science is moving forward at an unmatched pace in today's society with technological advances allowing scientists to complete research which only a decade ago would have seemed like something out of a science-fiction novel. As economic, technological and computational advances allow us to design and apply these new tools toward medical advancement and innovation, are our foundations in the scientific method, methods documentation, and experimental design in the academic research environment being implemented to the fullest of their potential? Lack of reproducibility surrounding the preclinical research is trending. Problems with reproducibility in have become a recurrent announcement that produced paper retractions after paper retraction, the reason behind these retractions is vast including but not exclusive to, poor study design, improper statistical analysis (underpowered studies) and misleading or omitted instruction in the methods section (missing key procedural instructions and/or using unsuitable reagents). We can only hope that the assumption that most of these irreproducible studies are being reported in error but without malice intent.

Nevertheless, the data are wrong and resources were and are being wasted. Under the assumption that the issues with the unreliable research come from poor quality data and not unreliable scientists, we have to ask ourselves what we can do to improve our reports and ensure that what we are reporting is true. This translational failure has become very troubling for executives, scientists, investors, and taxpayers. There has to be a better system for proper reporting study findings and methods. Once implemented this

system should help alleviate the financial, time and trust which currently presents a significant issue.

Herein, we will use data collected through an anonymous survey which will allow us to gain a better understanding of current best practices, knowledge, and implementation of the ARRIVE method, reporting best practices. Moreover, this survey may allow us to understand why the ARRIVE method is not gaining traction by investigators during reporting. Finally, we will discuss potential areas in which peer-review journals can help improve reporting, (ex. instituting a set of study design questions that must be answered before manuscript publication or guaranteed publication acceptance following the publication of study design to help report both positive and negative data outcomes. Academic science is paramount to the development of novel scientific approaches, as such, we must ensure that the data produced and reported are of the highest possible quality. Further, we must seal potential gaps in reporting the key elements as defined by the ARRIVE method helping to ensure the highest possible quality reporting.

Dedication

I am honored to dedicate this to my family. Without their support I would not have been able to pursue or complete this project and degree. To my parents who have always sacrificed their own needs to allow their children greater opportunities. Next, to my grandparents, who laid the foundations for their grandchildren, pushing them to always strive toward greater academic achievement. Finally, and most importantly, I dedicate this to my wife and son for their support and sacrifice throughout this entire process!

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

The irreproducibility crisis is not novel, and has been discussed in the literature over the last decade. The below will help to carve the literature to the focus of this thesis. Which is primarily the relationship between the irreproducibility crisis and academic pre-clinical work. It should be noted that the irreproducibility crisis affects both the scientific and biomedical communities alike and further affects every part of the drug development process from bioinformatics analysis to discovery to preclinical to clinical studies. Delaying time to market of novel potentially lifesaving or at least life prolonging drugs, and missing the potential to better quality of life for patients across numerous disease areas. This report is by no means meant to be an all-encompassing report of the recent history and development of methods to help resolve the irreducibility in scientific literature. Yet, it is our hope to help uncover potential irreproducibility issues related to the pre-clinical area, as these directly proceed clinical trial research and are essential for the FDA investigational new drug (IND) process. Furthermore, once we uncover potential pre-clinical issues we will propose resolution techniques for these problem areas.

A Brief History of the Irreproducibility Crisis

Science has been experiencing rapid economic and technological growth. As with any system the foundations on which this growth happens need to be firm. If the foundation is lacking stability only minimal outside pressures are required for the system to bend or break. This example may very well be the cause of the irreproducibility crisis in science today. So how did it begin? Which foundational areas are lacking? And what does the future hold?

One of the most significant reports on this current irreproducibility crisis came from Prinz et al in 2011. (Prinz, Schlange, & Asadullah, 2011) In their manuscript entitled; Believe it or not: how much can we rely on published data on potential drug targets? Prince reported on the major hurdles they had with irreproducible science in the areas of oncology, women's health, and cardiovascular disease. They stated that when attempting to reproduce published data during their target validation programs that only 20-25% of the reported results were in fact reproducible. This notion that potentially only 25% of accepted peer-review publications in these areas were, in fact, reproducible started to raise an alarm. As is usually the case when such a striking claim against the institution is made there was obvious doubting and questioning in regards to the validity of Prinz et. al's report.

The Prinz et. al report put in motion a movement to understand if, in fact, their notorious claims against peer-review and authors who had published in these areas were true or not. Following the publication by Osherovich in 2011 entitled Hedging Against Academic Risk, another report claiming that academia was not to be trusted when it came to reproducibility, research scientists would have a much harder time concluding that the

Prinz report was an outlier report. Osherovich's work promoted and confirmed that all of the irreproducible issues were coming from the academic realm and that companies needed to make sure they were ear-marking funds for modification and/or null results when trying to translate academic reports into their labs. (Osherovich, 2011) Furthermore, this report helped to fuel the fire of the irreproducibility crisis and in the end created more questions than it was able to answer. At this point, irreproducibility was becoming a "hot topic" which investigators would try their best to ignore.

Unfortunately following the publication by Begley in 2012, which stated that when trying to replicate 53 of the top tier paramount research reports in oncology they were only able to replicate 6, investigators and the research world could no longer turn a blind eye to this issue. (Begley & Ellis, 2012) Irreproducibility was, in fact, a problem in these areas of academic pre-clinical research. Once the scientific community agreed that the irreproducibility problem was real they accepted these reports. The next step became to start understanding where the irreproducibility problem stemmed from and further to take corrective action to alleviate this issue in the future. Mina Bissell unknowingly revealed a key problem area of published research work as she was trying to deliver a short defense of the academic manuscript in question.

Mina's defense paper, which spoke to the fact that there is no "crisis", entitled; "Reproducibility: The risks of the replication drive". Bissell tried to defend the fact that there is not a large irreproducibility issue and that it is the reproducing author's fault for not contacting the original authors for more details (Bissell, 2013). Moreover, many times following a literature review of her own work her own newly hired post-docs had trouble reproducing prior work, based solely on their understanding from reading her

manuscript alone. Further, she stated that her new post-docs will almost always have to review the laboratory notebooks for the specific protocol's "fine details". Additionally, she added that published works are almost always missing some information related to the materials and methods sections and that people who are trying to reproduce such cutting edge research would need more training directly from the lab in question. Moreover, as these highly complicated scientific methods are not always fully described by many in her field. That additional "training" is required to be able to reproduce her lab's published works. We understood this as the reader having access to the omitted details of the materials and methods section to be able to reproduce her works, something which is undesirable for peer-reviewed published science.

The revelation during Mina's justification that there was almost always: "missing method information in her publications", became very concerning to the research community. Mina's ease in disclosing this statement on the record, would lead us to believe that this type of omission is accepted or at least not faux pa in her research community. Thus, we could infer that many other researchers from her community are committing the same type of omissions in their publications. Again it should be stressed that this type of omission should never be acceptable in a scientific publication, unless specific legal intellectual property disclosures are noted for the non-disclosure. Moreover, a statement like this should be a red-flag for a potential key problem related to irreproducibility crisis.

This common theme of "errors of omission" is shared by her fellow authors who were having trouble with transparency or correct data reporting and irreproducibility of their works, and often used the "errors of omission" excuse as their deflection tool during

the defense of their work or the work of their peers. Many reports started to surface discussing the reasoning behind these false reports and trying to elucidate the main limitations from these irreproducible publications, and mainly blamed flexibility in study design, outcomes measure, statistics, and/or blamed investigator bias as a key problematic elements (Ioannidis, 2005; Kilkenny et al., 2009; Perrin, 2014; Tsilidis et al., 2013). Yet, even as more and more reports these authors in question tend to keep falling back to the defense that is was not flexibility in study design, outcomes measure, statistics, or investigator bias as a key element in this problem, but that they did not report correctly on these key elements in the peer-reviewed publications for one reason or another. An excuse that the research community accepted most of the time for the authors to avoid having to retract their works.

Steen et al moved to confirm the increase in the number of fraudulent reports and helped to defend the assumption of authors forgetting or omitting the data unintentionally. In this study, Steen reviewed 742 retraction notices between 2000 and 2010 via PubMed and further investigated the reasons behind these articles being retracted. Following data collection on the reasoning behind these retractions, the authors categorized those reasons into specific areas to better define their etiology. Steen et. al concluded and confirmed the assumption that the main reason for retraction was in fact due to error or omission 73.5% and not due to fraud 26.6 % (Steen, 2011). So this work presented clear confirmation that nearly ¾ of retractions were due to errors of omission and under the assumption that some of these retractions stem from others not being able to reproduce and question the work. Moreover, they stated that it would not be irrational to correlate these retraction results with the reason works were being reported

irreproducible. Leading the research community to investigate further as to why authors are omitting key details of their studies in their peer-reviewed published research findings. The omission of these materials and methods, which would prove to be paramount to the reproduction, would be deemed completely unacceptable. Further, if in fact most of these works are reproducible because of errors of omission corrective action to rectify these issues by simply including the proper details during report should prove surmountable. This corrective action would help promote the increased transfer of knowledge and the quality of reporting.

So now that the scientific community had acknowledged the irreproducibility crisis in these areas of science and seemed to understand its etiology, the time had come for the community to cure this festering issue once and for all, but where would they start?

Who created the ARRIVE guidelines?

The ARRIVE guidelines were created by a group of researchers in 2010. That group was comprised of members from the NC3Rs, which is a UK based collaboration dedicated to improving animal research science. The NC3Rs was started in 2004 with the primary goal is improving protections of animal in research science by implementing the 3R's. Which are; 1. Replacement 2. Reduction 3. Refinement. These guidelines for animal research designed to improve animal studies by; replacement- using thoughtful consideration on whether or not this research can be done without the use of animals. If this is not possible, the next the scientist should consider; reduction- which focuses on the

statistical study design and data analysis to make sure that studies that are performed on animals are robust and thoughtful, getting the most data with the least animal use. Once you have answered these first two guidelines on improving the humanity of animal research the final of the 3R's is refinement- which guides researchers to always be completing proper reviews of the most current methods and exploiting the advances in all aspects of their animal science research. As to be sure the data obtained from these animals is of maximum value.

The 3R's remind scientists that animal life should be treated with the utmost respect and that their use in research should be completed in a way to maximize data yield and have the research completed in the most humane way possible. Although this set of guidelines might seem self-evident and common practice for all those who might be performing research on animals, it surprisingly took more than 50 years from the initial creation of these guidelines for the 3R's committee to be formed to properly implement them. It is now common practice for all Institutional Review Boards to have a question in all animal research proposals in consideration of the 3R's, making sure that all research studies utilizing animals are doing so in a humane way.

Following their successfully structuring and rollout of the 3Rs and the fact that the irreproducibility crisis was becoming a hot topic in research, this group thought it was well suited to solve the "error of omission" problem of the crisis. The NC3r's group met initially to discuss, and fully understand the issue. After a plethora of meetings and in partnership with many from the research community, a set of publication guidelines for the author's to ensure that they would avoid "errors of omission" were released in their 2010 publication.

What are the ARRIVE guidelines?

These publication guidelines set forth to be a structure for the minimum required reportable information for all preclinical animal studies across all areas of science. The NC3Rs committee hoped to get out in front of the problem when they created this guidance system for authors across all scientific fields. As a committee, they felt they had created an all-encompassing system that would systemically cure the "errors of omission" associated with the replication crisis. Their system was comprised of 20 key areas that they deemed to be essential to ensuring nothing was omitted. Moreover, they not only provided the 20 topics, but they also gave specific inclusion criteria on why each of these items was essential (Kilkenny, Browne, Cuthill, Emerson, Altman, et al., 2010). The group called the method ARRIVE (Animal Research: Reporting of In Vivo Experiments) (Figure 1) and published this guideline in 2010.

The ARRIVE guideline touched on many points that were reported to be common "errors of omission" by authors and stated that reporting of these key study characteristics should be a requirement for all authors before manuscript publication to ensure that their works were sound and reproducible. All of the 20 areas and a brief description of each follows, to emphasize how obvious reporting on these areas can be but also how easily one can forget to report on them during manuscript preparation. (Figure 1a and 1b).

The ARRIVE Guidelines Checklist

Animal Research: Reporting In Vivo Experiments

Carol Kilkenny[1], William J Browne[2], Innes C Cuthill[3], Michael Emerson[4] and Douglas G Altman[5]

[1]The National Centre for the Replacement, Refinement and Reduction of Animals in Research, London, UK, [2]School of Veterinary Science, University of Bristol, Bristol, UK, [3]School of Biological Sciences, University of Bristol, Bristol, UK, [4]National Heart and Lung Institute, Imperial College London, UK, [5]Centre for Statistics in Medicine, University of Oxford, Oxford, UK.

	ITEM	RECOMMENDATION	Section/ Paragraph
Title	1	Provide as accurate and concise a description of the content of the article as possible.	
Abstract	2	Provide an accurate summary of the background, research objectives, including details of the species or strain of animal used, key methods, principal findings and conclusions of the study.	
INTRODUCTION			
Background	3	a. Include sufficient scientific background (including relevant references to previous work) to understand the motivation and context for the study, and explain the experimental approach and rationale. b. Explain how and why the animal species and model being used can address the scientific objectives and, where appropriate, the study's relevance to human biology.	
Objectives	4	Clearly describe the primary and any secondary objectives of the study, or specific hypotheses being tested.	
METHODS			
Ethical statement	5	Indicate the nature of the ethical review permissions, relevant licences (e.g. Animal [Scientific Procedures] Act 1986), and national or institutional guidelines for the care and use of animals, that cover the research.	
Study design	6	For each experiment, give brief details of the study design including: a. The number of experimental and control groups. b. Any steps taken to minimise the effects of subjective bias when allocating animals to treatment (e.g. randomisation procedure) and when assessing results (e.g. if done, describe who was blinded and when). c. The experimental unit (e.g. a single animal, group or cage of animals). A time-line diagram or flow chart can be useful to illustrate how complex study designs were carried out.	
Experimental procedures	7	For each experiment and each experimental group, including controls, provide precise details of all procedures carried out. For example: a. How (e.g. drug formulation and dose, site and route of administration, anaesthesia and analgesia used [including monitoring], surgical procedure, method of euthanasia). Provide details of any specialist equipment used, including supplier(s). b. When (e.g. time of day). c. Where (e.g. home cage, laboratory, water maze). d. Why (e.g. rationale for choice of specific anaesthetic, route of administration, drug dose used).	
Experimental animals	8	a. Provide details of the animals used, including species, strain, sex, developmental stage (e.g. mean or median age plus age range) and weight (e.g. mean or median weight plus weight range). b. Provide further relevant information such as the source of animals, international strain nomenclature, genetic modification status (e.g. knock-out or transgenic), genotype, health/immune status, drug or test naïve, previous procedures, etc.	

The ARRIVE guidelines. Originally published in *PLoS Biology*, June 2010[1]

Figure 1a:*(Kilkenny, Browne, Cuthill, Emerson, & Altman, 2010) Arrive Guidelines: items 1-8*

Housing and husbandry	9	Provide details of: a. Housing (type of facility e.g. specific pathogen free [SPF]; type of cage or housing; bedding material; number of cage companions; tank shape and material etc. for fish). b. Husbandry conditions (e.g. breeding programme, light/dark cycle, temperature, quality of water etc for fish, type of food, access to food and water, environmental enrichment). c. Welfare-related assessments and interventions that were carried out prior to, during, or after the experiment.	
Sample size	10	a. Specify the total number of animals used in each experiment, and the number of animals in each experimental group. b. Explain how the number of animals was arrived at. Provide details of any sample size calculation used. c. Indicate the number of independent replications of each experiment, if relevant.	
Allocating animals to experimental groups	11	a. Give full details of how animals were allocated to experimental groups, including randomisation or matching if done. b. Describe the order in which the animals in the different experimental groups were treated and assessed.	
Experimental outcomes	12	Clearly define the primary and secondary experimental outcomes assessed (e.g. cell death, molecular markers, behavioural changes).	
Statistical methods	13	a. Provide details of the statistical methods used for each analysis. b. Specify the unit of analysis for each dataset (e.g. single animal, group of animals, single neuron). c. Describe any methods used to assess whether the data met the assumptions of the statistical approach.	
RESULTS			
Baseline data	14	For each experimental group, report relevant characteristics and health status of animals (e.g. weight, microbiological status, and drug or test naïve) prior to treatment or testing. (This information can often be tabulated).	
Numbers analysed	15	a. Report the number of animals in each group included in each analysis. Report absolute numbers (e.g. 10/20, not 50%[2]). b. If any animals or data were not included in the analysis, explain why.	
Outcomes and estimation	16	Report the results for each analysis carried out, with a measure of precision (e.g. standard error or confidence interval).	
Adverse events	17	a. Give details of all important adverse events in each experimental group. b. Describe any modifications to the experimental protocols made to reduce adverse events.	
DISCUSSION			
Interpretation/ scientific implications	18	a. Interpret the results, taking into account the study objectives and hypotheses, current theory and other relevant studies in the literature. b. Comment on the study limitations including any potential sources of bias, any limitations of the animal model, and the imprecision associated with the results[2]. c. Describe any implications of your experimental methods or findings for the replacement, refinement or reduction (the 3Rs) of the use of animals in research.	
Generalisability/ translation	19	Comment on whether, and how, the findings of this study are likely to translate to other species or systems, including any relevance to human biology.	
Funding	20	List all funding sources (including grant number) and the role of the funder(s) in the study.	

Figure 1b:*(Kilkenny, Browne, Cuthill, Emerson, & Altman, 2010) Arrive Guidelines items 9-20*

This comprehensive guideline had a mostly positive reception but would lay dormant from 2010 to about 2013 (Collins & Tabak, 2014). This period included the release of the above-discussed reports and 2 crucial additional reports which warrant mention, as they played a significant role in the ARRIVE guidelines push for adoption.

ARRIVE Guidelines from Production to Adoption

The first of these final two paramount papers were authored by Begley in 2013. Following the release of Begley manuscript entitled; "Drug Development: Raise Standards for Preclinical Cancer Research.", plethora of researchers were reaching out to Dr. Begley wanting him to disclose which studies were referenced in this 2012 report. Begley stated that these investigators reaching out to him did not want to waste their time and resources working from a foundational level which had already been proven to be fundamentally bad and irreproducible. Yet, due to the fact that Begley had signed non-disclosure agreements with many of the groups whom he had reported on during the creation of his initial publication, he was rendered unable to share this information. As such, he felt the next best thing was to create a novel manuscript, for which he does his best to describe the flaws of the prior irreproducible reports, affirming that all the irreproducible reports shared 6 common flawed areas. Moreover, Begley stated that if investigators reread the prior work keying on these areas, they should be able to parse through the reports to validate the efficacy or reliability of all 53 in question.

The topics he discusses include: 1. Were experiments performed blinded? 2. Were basic experiments repeated? 3. Were all the results presented? 4. Were there positive and

negative controls? 5. Were reagents validated? 6. Were statistical tests appropriate? (Begley, 2013) Begley discussed how each of these categories from the report were areas identified as issues within clinical research years ago and should have been corrected by the time of these reports. Moreover, the ARRIVE guideline published almost 3 years before his initial review includes all of his key points and 14 more, which would have corrected these issues before they happened. Following the release of Begley's 2013 report, the ARRIVE guidelines started to move closer and closer to being discussed as the answer to the "error of omission" issue and brought them closer to the forefront of the irreproducibility discussion.

This "error of omission" issue was reinforced when Fang and Casadevall 2011 report resurfaced during this time. Fang and Casadevall manuscript entitled; "Retracted Science and the Retraction Index", included the discussion of how there was an increase of retractions which could be attributed to errors of omission. They went further as they discussed how this rise in retractions could potentially be attributed to the fact that authors are trying to deceive the readership while obtaining the required publication index for advancement. An accusation that moved further to blame the investigators for having malice intent when omitting specifics during their publications. (Fang & Casadevall, 2011; Grimes, Bauch, & Ioannidis, 2018; Lawrence, 2003).

Fang and Casadell's also reported a retraction index, which essentially adds up the retractions errors found in each journal article and by giving each a numerical score based on their scale of unacceptable mistakes. Moreover, they provided a detailed description of the creation of their retraction index, please refer to (Fang & Casadevall, 2011) for the full description as the fine details are too lengthy to include in this report.

Once the score was created for each manuscript the index score was then correlated to the impact factor of the journals which these manuscripts were published in, which did include some of the top-tier journals. Their plot clearly shows a strong correlation between higher impact factor journals and their retractions index.

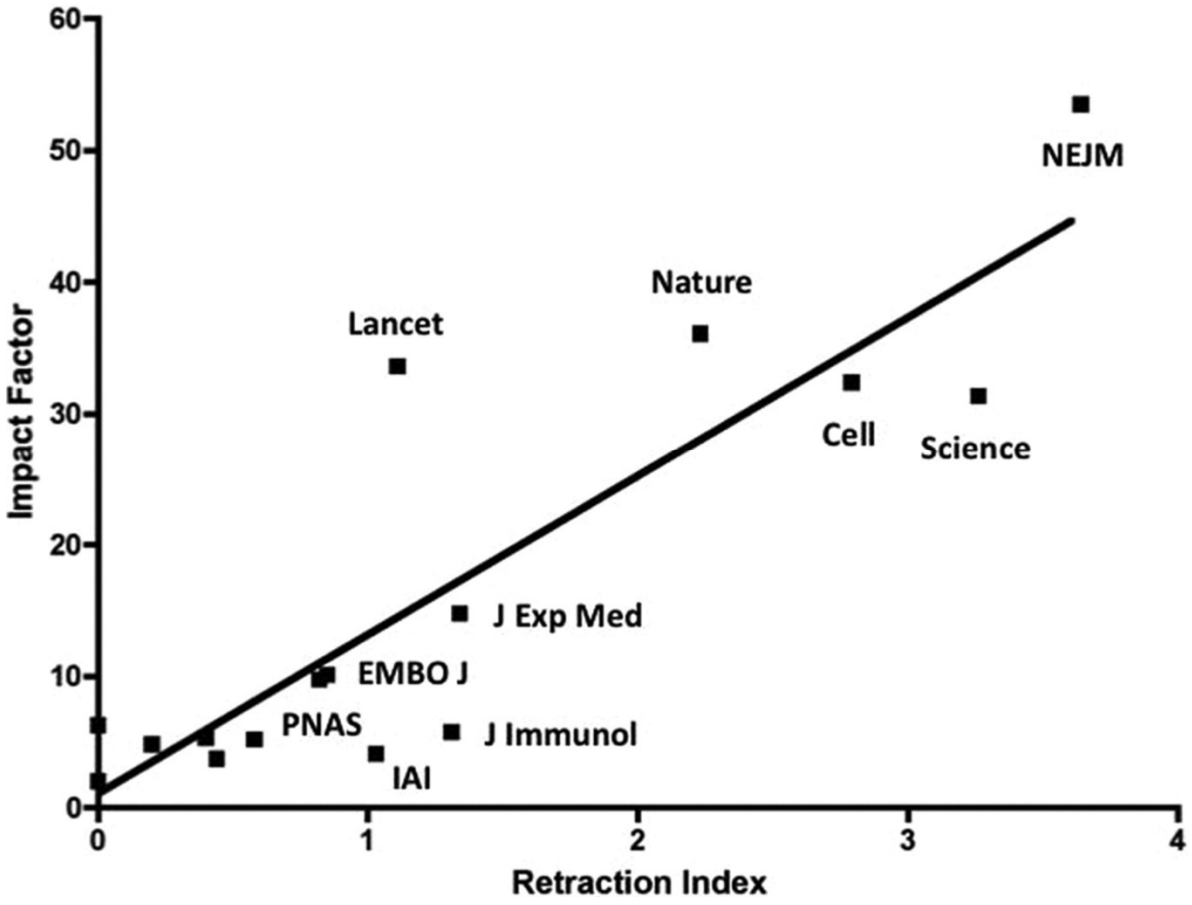

Figure 2: *(Fang & Casadevall, 2011) Journal impact factor on the y-axis and calculated retraction index on the x-axis, showing correlation between the larger the impact factor the greater the retraction index*

Moreover, this report helped to reveal that top-tier journals were not doing enough to protect their readership from reading false reports and were distributing bad

science, which was something that should have been paramount for their strict peer-review processes and would need to be corrected!

Initial adoption of the ARRIVE guidelines.

Following these reports, which questioned the integrity of the peer-review process and reporting methods of some of the most prevalent top-tier journals, it seemed they were obligated to respond. Since these Journals needed away to reinforce their commitment to publishing quality and reproducible works, they found the method of attack with the endorsement of the ARRIVE guidelines. The Journals Nature and PLOSone's responded that it was of utmost importance for the authors submitting works to the journal to include all details of the works and being careful not to omit any necessary detail. They published a release on their author's instructions webpage which included a link to the ARRIVE guideline and further recommended that all submitting authors *should* review and include the suggested reporting key elements of the ARRIVE guideline in their reports (Leung, Rousseau-Blass, Beauchamp, & Pang, 2018). Moreover, Nature's response included an additional correction following these accusations of errors of omission when they removed the word limit previously imposed on their methods section (Collins & Tabak, 2014), which would allow authors excuse that they could not include all necessary information due to this limitation.

This change seemed to be a huge win for the integrity of reporting in the scientific community and its readership. This manuscript production and reporting change hoped to create a more robust reporting environment for future publications and further should

have helped start correcting the portion of irreproducible preclinical works which were being caused or attributed to errors of omission. Disappointingly Nature's inclusion of the word "should" made the reference to the ARRIVE guidelines only a suggestion and not a policy or mandate for manuscript acceptance and publication, which allowed many authors to ignore this additional step before manuscript acceptance and once again push the ARRIVE method out of the spotlight.

Adoption Woes for ARRIVE

Although Nature had included the ARRIVE guidelines suggestion on their home page and instructions for authors website, the adoption of the ARRIVE guidelines was stagnant. This lack of adoption and inclusion of the ARRIVE guideline for proper methods and results reporting was discussed by Baker et al. in their 2014 manuscript entitled; "Two Years Later: Journals Are Not yet Enforcing the Arrive Guidelines on Reporting Standards for Pre-Clinical Animal Studies." Baker et. al completed a review of manuscripts published in PLOS Biology and confirmed that the twenty key areas described in the ARRIVE guidelines still remained largely under-reported (Baker, Lidster, Sottomayor, & Amor, 2014) one year after the ARRIVE guidelines were referenced by Nature and PLOS on their websites.

Further, they stated that in some scientific areas upwards of 80 to 90% of the published works in PLOS failed to include even two of the 20 key points layout by the ARRIVE guidelines. Disappointingly, this report confirmed that authors were still being either lazy or fraudulent in their reporting of research methods, while almost completely

ignoring the ARRIVE reporting guidelines, and the research community didn't seem to care. Furthermore, this lack of adoption of ARRIVE left open the doors to extreme risk for authors to perform irreproducible studies and subsequently for journals to publish these misleading reports.

Pushback to ARRIVE adoption

As the movement to accept the notion of irreproducibility in peer-review publications started to gain momentum, as tends to happen with reform movements, it all at once came to a screeching halt. Authors started one after the other producing works which would break-down the reform movement and try to discredit the author who had previously pointed out an issue they were encountering. These authors would try their best to explain away irreproducibility as we had come to understand and accept it to this point. Arguments would include examples like; the way the scientific community was describing irreproducibility was very confusing it required further dissection, and reproducibility was further confused across scientific areas of study. Additionally, authors would try to say that the negative connotation or association that came with correcting errors in previously published works was the basis for authors not actively correcting their published works. As being part of a retraction notice for such revisions even if the overarching results and conclusion of the work were unchanged could tarnish your name, it was easier to ignore the issue until someone asked you about why they were having trouble reproducing your work.

Expanding on this notion of risk of irreproducibility, the work of Goodman et al in 2016 further inspected reproducibility issues throughout the scientific community. The authors started by again reiterating the problem of reproducibility by summarizing the already reported data on the crisis. (Goodman, Fanelli, & Ioannidis, 2016) Once the irreproducibility crisis, as the authors believe it should be understood and documented, was recognized and confirmed to their audience, the authors then moved toward an argument that the use of the word reproducibility as they just described was in fact not well understood across the science disciplines and more broadly the scientific community.

Furthermore, they inferred that before the research community could truly make any attempt at discussing where investigators are going wrong and taking further corrective action, there was a need to obtain a better understanding of the terminology associated with irreproducible studies. Further, they defined that reproducibility should be broken down into 3 specific areas which include. 1. Methods reproducibility 2. Results reproducibility 3. Inferential reproducibility. (Goodman et al., 2016) The authors went on to conclude that these areas, which were included in discussion by previous authors and all of which are discussed and covered by the ARRIVE guidelines, were in fact not well understood and needed to be discussed at length. Following the extensive breakdown of each of the 3 areas, they concluded that the research community *must* accept and use the author's binning terms before we start discussion into potential solutions to the reproducibility crisis. The vocabulary disarray that exists within the scientific community with regard to reproducibility is vast and each binned term or area of reproducibility had different proposed resolutions techniques with minimal overlap between each.

Following this publication, further reporting related to the defining terminology defense of the failure of the scientific community to take any corrective action up to this point in time was released in 2017. In this report, entitled: "Correcting Honest Pervasive Errors in the Scientific Literature: Retractions without Stigma", the authors discuss the apparent stigma related with the term retraction. Moreover, they try to attribute the lack of self-reporting methodological errors and/or errors of omission observed following publication with the negative connotation of the word retraction. Further, the authors continue to argue that journals should try to create a method for dealing with errors which were self-reported, moving away from the catch-all term of retraction, reiterating that it would promote authors to correct their work without being associated with the negative term "retraction" (Baskin, Mink, & Gross, 2017). Finally, the authors concluded that if *Journals* created a design method for Authors to correct their works, it would act the same as the retraction and correction process yet would be termed differently to have more positive reinforcement. Thus, the scientific community should expect an immediate increase in the amount of self-reported corrections, which in turn would support an effective solution to the internal errors or errors of omission problem.

Although the authors are right that treating each of these areas as specific problems and tackling them individually is correct, their overarching statements and oversimplification may be a bit of a misstep in the defense of the apparent inactivity to this point. While these areas do in fact have solutions which can be approached by differing methods and further they are correct that in some cases these methods truly have nothing in them which overlap but to say that you can't diagnose irreproducibility as a whole through the application of the ARRIVE guidelines as a first step is wrong. All

these simplification and clarification suggestions do is try to reassign blame and protect those who have been exacerbating the issue. Journals and authors should start by instituting and conforming to the ARRIVE guidelines in the publications before over-analyzing and further complicating an already convoluted problem.

ARRIVE has yet to arrive

Shortly after the pushback works were published, Leung et al, completed a comprehensive review of manuscripts published before and 8 years after the adoption of the ARRIVE guidelines by PLOS and Nature. The inclusion criteria for their groupings were that they must have been published in either of these 2 journals and the works were either published before the release of the ARRIVE guidelines suggestion and method word limitation was lifted or after. Once they were a group as either before ARRIVE or after, the authors did their best to understand the inclusion of the key factors discussed in the ARRIVE guidelines. Following this review, the authors concluded that they saw only marginal increases in reporting across the 20 guideline areas, adding they had expected the reporting of these key guideline areas to have increased, yet their results disappointingly show no such correlation (Leung et al., 2018). Furthermore, this report confirmed yet again that authors were still omitting key reporting areas which could result in studies that were irreproducible.

The result was very disappointing and led to further concern. The reason for this concern comes from the understanding that there are two main reasons for an author's work to be irreproducible. The first is attributed to the authors omitting details from their

science unintentional, while the second is attributed to the authors omitting details from their work intentionally. The ARRIVE guidelines were created to help rectify the first issue and expose individuals who may try to exclude details of the work intentionally. Moreover, correcting irreproducibility created through errors of omission and/or poor study design reporting still pose a risk which all investigators need to be aware of when doing a literature review. There still seems to be an inconsistency in adoption, understanding, and use of the ARRIVE guidelines. We hope to investigate this gap in our small research environment and promote all authors reporting from our field endorse the use of the ARRIVE guidelines in their future reports, when applicable.

Financial Impact of the Irreproducibility Crisis

The published literature of the financial impact of the irreproducibility crisis is greatly debated, some reports claiming an impact as large as $28 billion annually wasted (Figure 3) (Freedman, Cockburn, & Simcoe, 2015). There has been great debate on how correct this number really is, but let us ruminate in the fact that even if the problem is 1/28th the 2015 reported amount there is still at least billion dollars being wasted. For more perspective, we considered the annual budget for the NIH in 2019 was $39 billion, with about $23 billion going to research project grants (RPBs). If the evaluation of waste from 2015 of $28 billion is correct it becomes a bit more staggering as this amount outweighs the total budget of the NIH for RPBs. Again it should be stated that this is federal funding only, but private funding is even higher than that for industry.

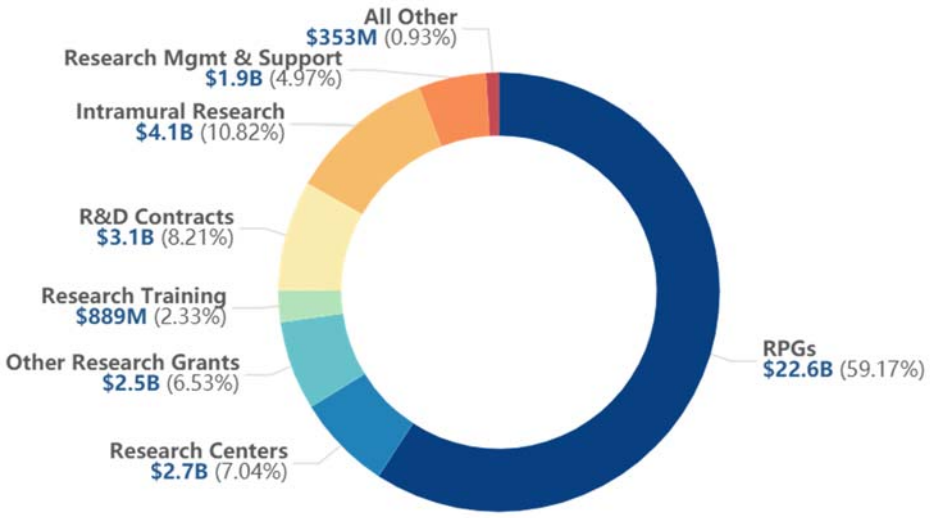

Figure 3: *(NIH, 2019) Total NIH Budget Authority: FY 2019 Operating Plan*

The potential waste reported from Freedman et. al in 2015 is cause for great concern yet, let us reflect on how they reach this reported number and further what they reported to be the 4 essential problem areas in the literature today.

Their first postulation was related to the reported waste amount. The author's estimate started with the 2012 published amount of a life science research in the US being $114.8 billion. It is important to note that this amount included the pharmaceutical industry, government investments, non-profits, and academic investments; as the investigation of data irreproducibility across each of these subcategories varies and in some cases, there has been no analysis completed to date. Further, they concluded that of this $114.8 billion about $56.4 billion was invested into the area of pre-clinical studies.

Their second postulation was related to the percentage of reports that are assumed to be irreproducible. They came to this number based on the works which we have touched on above which report as low as 18% irreproducibility and in some cases as great

as 88% irreproducibility. Following a probability distribution based on these reports, the mean of this distribution was 53%. As such they move forward with this number as their accepted figure for irreproducibility across all scientific areas. I will point out again that these reports did not all come from the same areas of science, so this outcome can be assumed to be a very rough extrapolated estimation.

Their final postulation was to combine the first two assumptions. As such, they supposed that both their $56.4 billion spent on pre-clinical research and their 53% irreproducibility numbers are correct assumptions, and with simple math, they came to a number of $28 billion dollars of annual waste in pre-clinical studies. Freedman et. al also reported that of this calculated amount approximately 31.5% could be attributed to government spending or tax dollars of the American people. Now, this becomes a bit more concerning to the populace if we again do the math. $56.4 billion being spent x 31.5% being spent by the government (taxpayers) totaling $17.76 billion/ year on pre-clinical research. Now again assuming that 53% of this research is irreproducible $17.76 x 53%, gives us $9.4 billion taxpayer dollars being wasted annually on studies that are irreproducible.

Following this big announcement, Freedman et. al then reported on their reasoning behind this described irreproducibility. (Figure 4) As can be seen from their figure they believed the biggest areas of concern came from (1) problems with overall study design, (2) incorrect biological reagents and/or reference materials, (3) poorly thought-out or poorly reported laboratory protocols, and (4) problems with data analysis

and reporting (Lowe article)

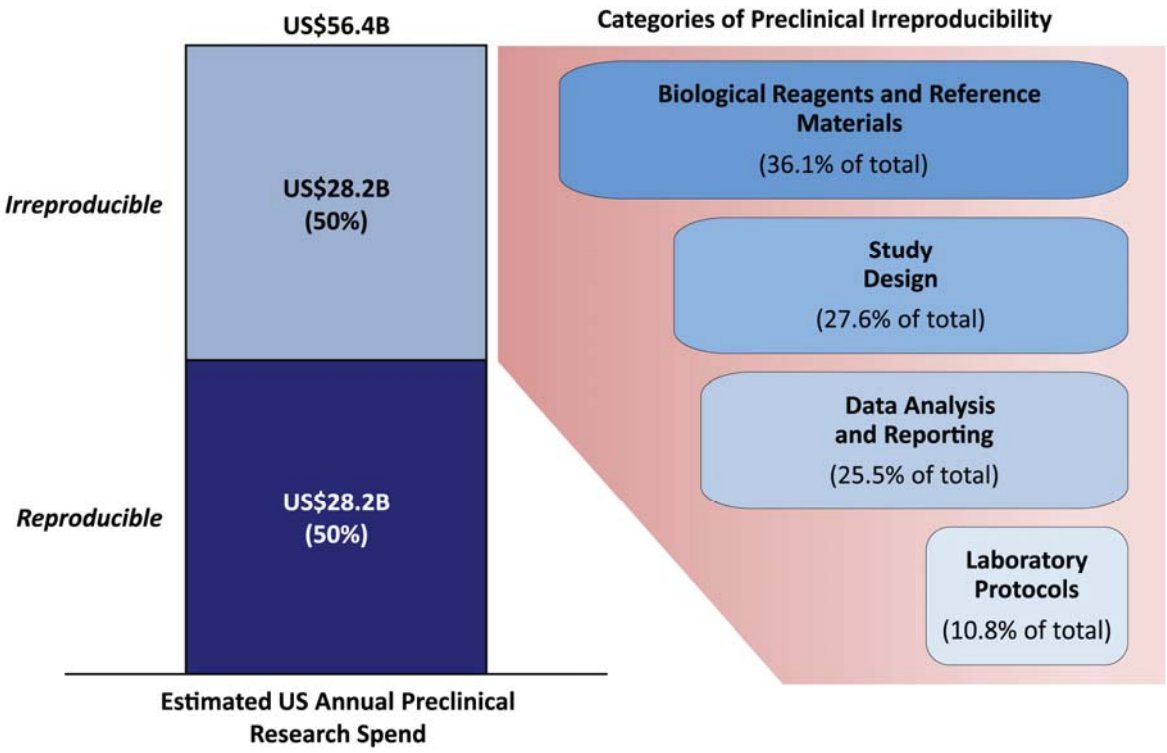

Figure 4:*(Freedman et al., 2015) Chart of the breakdown of irreproducibility causes*

As we have already discussed most of these areas in-depth previously, suffice it to say that these authors concluded similarly that these error areas were for almost exacting reasons as all the other previously discussed reports. Additionally, it is worth repeating that all of these topics and their related errors of omission are all covered by the ARRIVE guideline checklist.

The potential for corrective action, for the segment of irreproducibility created by errors of omission, exists with the ARRIVE guidelines. Yet, there has been no real movement toward their adoption and application. It is our hope through the application of a short 20 question survey to gain a better understanding of all small research environments to help elucidate some of the reasoning why ignorance of the irreproducibility problem and the ARRIVE guidelines still persists.

Chapter II

Materials and Methods

The research method for this project followed a cross-sectional survey study design and evaluated a small population's knowledge of the irreproducibility crisis in science, the ARRIVE guideline, and opinions on the financial significance of reproducible work and application of the ARRIVE guidelines. Although this survey was only a point in time view and opinion-based, by design, its goal was to inform us of potential lack of exposure and/or misunderstandings that were present in our small research community. This research was undertaken following the approval of the Committee on the Use of Human Subjects (CUHS) which serves as the Institutional Review Board for the Cambridge and Allston campuses at Harvard University.

Survey Design- data collection methods

We choose to undertake a cross-sectional survey design as opposed to successive sampling or a longitudinal design due to the expedited nature of the research collection period. Moreover, we did not have the financial support required to allow for multimodal survey collections techniques, which may have had the ability to enhance the study. With proper funding we may have chosen to enhance this study by being able to hire support study staff for manual collection of survey (door-door), costs for mailing service delivery

of survey (broadening the reach), or expenses for the use of survey services which reimburses subjects/individuals for their time following completion of the questionnaire (usually a set amount per survey-based estimated completion time). As such, we implemented an email mailing list for survey completion requests and additionally undertook group messaging social media techniques to allow for a broader outreach to our tailored target participants.

Survey Design- target population

Our target population was a cluster sample of a small research community in the greater Boston area. Our survey was designed to ensure that we would be able to collect enough responses to be able to extrapolate our respondent's answers and trust it to be descriptive of the target populations, a correct design relies heavily on specific inclusive/exclusion criteria or inferences about our target population.

Our first assumption was that the institution had approximately 150 working staff, comprises of researchers, technicians and research support staff. This initial inference was based upon a lab which consists of approximately 30 investigators across 11 research units, and the associated research support staff (ex. post-docs, research technicians) and no-research support staff (ex. department and grant administrators). We estimated the size of the population-based of a current mailing list which was compiled of just under 100 individuals at the time. This number was increased to include additional support staff who were not included on the mailing list, and individuals who were lab collaborators and would be targeted for response through social media solicitation.

Initially our inference placed our target population to around 138 individuals. While we further included a 10% increase for those individuals who were new hires or previous lab staff not currently on the mailing list and finally rounded to the whole number of 150 potential respondents.

Once we reached the size of our target population, we then proceeded to understand how many responses will be required for the study to obtain acceptable power. We performed sample size power calculations such that with a confidence level of 95% and a confidence interval of 10% the minimum number of required responses to ensure our data could be extrapolated as a representative population sample, we would need 59 respondents to our survey. We went further with our power calculations to include a tighter alpha of 5%. Once adjusted, our sample size power calculation with a confidence level of 95% and a confidence interval of 5% the minimum number of required responses to have a representative population sample we would need 108 respondents to our survey.

We initially felt that with the current decrease in response rates to population surveys (Harris-Kojetin & Groves, 2017) and the fact that we were working with a limited time-table and funding support and moreover that this analysis was not trying to prove cause and effect but rather trying to understand a majority opinion for a small sample size that a 10% confidence interval would be acceptable but of course a 5% confidence interval would be preferable. Additionally, we assumed that the 10% confidence interval would be acceptable as to avoid a major potential pitfall of the study was to require a confidence interval of 5% and not being able to collect enough respondents to complete the survey data collection and further analysis.

Survey Design- Survey Development

Demographics considerations

Which demographics we would include, collect, and report in our survey became of significant importance to us as if differed greatly for general study designs. Normally, during the creation of a study survey, you strive to include as much demographic information from your target population as possible, as with each additional demographic response you are able to further segment your target population and infer specifics about this tailored population segment. For our purposes this was true but we also had to consider that our target population was small, reasonably diverse and very intimate, such that with too many demographic questions we would be able to potentially drill down and understand who had made these responses, eliminating the anonymity of the study. Further, we considered that our target population had at least some overlap/contact with similar survey studies and may come to the same anonymity conclusion which may potentially hinder their successful completion of the survey or discourage them to provide their honest opinions about a specific topic. Potentially, rendering our results to be skewed by such bias would be extremely problematic for our research questions. As such we included the minimum required demographic information or what we deemed to be the essential defining characteristics of our target population, to successfully segment our small population while still producing anonymity of the respondents.

Survey Question development

Our survey's goal was to understand our target population's knowledge of the irreproducibility crisis in science, the ARRIVE guideline, and opinions on the financial significance of reproducible work and application of the ARRIVE guidelines. Considering that we were in search of an understanding of multiple related topics the number of study questions could be vast. The initial creation of questions of interest totaled out around the 35-40 mark, understanding that there is a negative correlation between increased total time required to complete a survey and survey response rate. Limiting the number of questions of interest became a top priority during survey creation.

Our goal was to limit the total time to complete the survey to around 5 minutes to make the survey a manageable time commitment for our projected respondents. We made the assumption that it would take on average 30 seconds for individuals to read, comprehend and answer each question, with more time for opinion-based questions and less time dedicated to demographic responses. Realizing that we would have to cull our initial survey from 35-40 questions to 20 questions, we started by grouping questions of similar subtopics, from there we were able to better understand the overlap between questions, narrowing the focus of our questions and making each question response more impactful.

Survey Design- Qualtrics build and beta-testing

Upon completion of question creation and finalizing the subtopic design, we had what we deemed to be a final questionnaire. Once the questions were finalized by the study staff, building the survey in Qualtrics (Provo, Utah) began. The first build of the survey had each question presented one at a time and requiring an answer before moving forward in the survey. Furthermore, the back button of the survey was initially functional so that the participant was able to move both forward and backward through the survey as many times as they like.

In the hopes of eliminating response bias potentially created through leading responses by the question make-up or the order in which questions were presented or answered and then re-answered, we first eliminated the utility of the back button so the question(s) on a specific survey page had to be answered in full once before they moved to the next question or set of questions on the following page. We did this in hopes that participants would not be biased by a question from one sub-area to another and return to change their answers. Secondly, based on our initial subtopic design for each question set, we tried our best to have questions which related on the same page, so that these questions would allow a better understanding of the subject matter and allow the participant to make responses with the maximum amount of data on the topic as we would disclose through the study. Thirdly, we reordered the subtopic questions allowing for good viewing presentation when the survey page was populated on desktop monitors or mobile devices. This was meant to eliminate the use of the scroll to see all questions in a specific sub-area, making the survey more user-friendly. Finally, we designed the subtopics in an order which we assumed would help to eliminate any question leading

bias and one which would match the introduction of topics in the production of our thesis to make the analysis of the results of each question easier for the study staff following data collection.

Once we felt that the order of subtopics, the design of the questions (number and presentation) and the elimination of survey bias were performed to the best of our ability, the next step was to perform a beta-test of our survey, for any potential unforeseen issues. Limitations with our beta-test were the fact that our survey population was so small and with specific details that to find a comparable beta-survey population would prove difficult. Nevertheless, we were able to produce a limit (n=10) number of people to take the beta -survey for question understanding and allow us to modify questions and answers before full release to our population of interest.

Issues which were found in the beta- release included a few grammatical and one spelling error (there instead of their). Additionally, a few of the demographic questions needed to be modified so that we could effectively collect our participant information in the correct segments. For example, we had listed MD, Ph.D., Masters, Bachelors, High School-GED, Other. Yet, as our survey design capability was limited through the Qualtrics programming, due only to user inexperience, we were unable to create successful production of the other free text box while still holding the page layout in a manner in which we deemed acceptable and pleasing for the subject participant. We decided that it would be easier to keep the format of the survey page presentation the most user-friendly as possible by eliminating the free text box associated with the other option. Moreover, we added the JD and other doctorate degrees to the Ph.D. answer, doing our best to allow for the best interruption of the data set following production.

Another change that was made to the demographic responses to allow us to segment the population which would allow for better result analysis was to change the area of a degree from “other degree type” to “other science-related degree” and “other non-science related degree”. This modification would allow us to delineate respondents who are part of the scientific team from the study support staff more effectively. Of course, we understood that some study support staff may have a science-based degree and would be included in our science group in error but this type of error was deemed to be marginal and acceptable for the interpretation of our data set. The inclusion of these changes was paramount for the betterment of the results interpretation and analysis.

Following re-writing the questions based on these suggested edits we were able to have new confidence in our final draft survey and following the approval of the Committee on the Use of Human Subjects (CUHS) at Harvard University, the study was published online and made active for responses. In hind-sight failure to run initial beta-testing could have created much confusion when trying to parse through the data sets and assign a meaningful interpretation to the collected responses.

Survey Design- Data analysis and display

The majority of the design processes for our survey was spent on question design. One aspect of question design considered during the production of this population survey was the answers. The design of respondents answer selections went through numerous iterations. What follows is the thoughtful analysis undertaken for each of the answer sets

and its related question. Answer set selection criterion which was true for almost all questions were the following.

As this was a cross-sectional analysis of a clustered population, we were looking to understand briefly, at this point in time, whether respondents were aware or unaware of specific underlying topics of, or related, to the irreproducibility crisis. As such we designed responses wherever possible to be of a binary nature. A section that was made to help our study population choose for or against a certain query, forcing them to be sided in their response. This method held true from almost all of our question responses and lead to almost all questions being of yes or no nature during the initial design. A slight change was made following the beta-release of the survey what followed was an evolution in answer design, which will be further discussed.

The evolution from the binary response when applicable to adding a tertiary selection happened as follows. In instances where our questions where querying our population on whether or not they had knowledge of something, we were concerned that subjects might unknowingly or unfairly weighting the selections yes or no due to a question bias or bad memories. Naturally, for questions in which this may apply we, therefore, added a “not sure” selection. We selected this choice to allow individuals who were confused by the question or failing to feel comfortable with their own knowledge of the subject in favor of or against, the ability to not choose. Further, the choice would allow them to continue the survey without having biased the responses due to some confusion.

It should be noted for questions that we included this “not sure selection” these answers were eliminated from the results and further the statistical analysis. Such that we

could perform a statistical analysis which was derived from the fact that the outcome measured was of a binary nature. Further, we separated our respondents of our survey into two comparable categories “science base degree” and “non-science based degrees” to create our sub-population. Such that we now had binary data with different users in each group and only two groups. We were able to use the N-1 two proportion test to evaluate statistical significance between groups of our population in their responses to the specific questions which fit these decision tree assumptions. For data reports which did not fit these assumption either summary statistics of means and standard deviations were reported or percentile charts and graphs were reported.

Although the statistical analysis was completed on the outcomes of the survey it should be noted that we only reference the evaluations or the related p-values in the results and discussion sections to show potential relation and never to prove a cause and effect. As our cross-sectional study design is only of a point in time understanding of a small clustered population and to extrapolate this data set to a wide population or conclude that these responses are reasons attributed to the larger scientific community would be unfounded. This survey and its outcomes were designed to increase the exposure of the irreproducibility crisis and the ARRIVE guidelines and promote future investigation therein.

Chapter III

Results

The survey entitled *ARRIVE and Documentation Short-falls*, which had the approval of the Harvard Institutional Review Board for human subject research, ran from the 22nd of November 2019 until the 22nd of December 2019. During this published time our survey collected a total of 130 entrants. Following study closure, the total data set was evaluated and the following inclusion criteria were set.

The first of the inclusion criteria was that all respondents needed to answer the survey between November 22nd and December 22nd, 2019. We had 3 results included in our data set which were erroneously not deleted following the beta-testing sessions, these surveys were thus excluded. Our next inclusion criteria were that all surveys needed to be completed in full for the responses to be included in our study. This excluded 28 surveys that were not 100% complete. Upon further investigation of the 28 surveys which were excluded for incomplete responses; we determined 2 had failed the completion requirement due to the fact that the respondents failed to click the enter survey button before proceeding into the survey. Following this discovery, we modified our completion requirement to exclude the first question. Thus, if respondents had successfully completed all other questions of the survey, excluding number one, we deemed these surveys valid and included them back into a final response set. This brought our number of respondents back to 101.

These 101 respondents were then evaluated. The average completion time for the survey was 3.8 ± 3.1 (expressed in minutes ± STDEV). The spread of our respondent's backgrounds was segmented as could be expected and correlated nicely with the true make-up of our target population. With 47 respondents coming from what we called the Scientific Segment (SS) and 54 coming from the Non-Scientific Segment (NSS). A breakdown of "area of study" can be appreciated in (Figure 5). The figure illustrates the reported "area of study" for our SS and all of the NSS or support staff responses were pooled into the Other- Non-science related field for this chart.

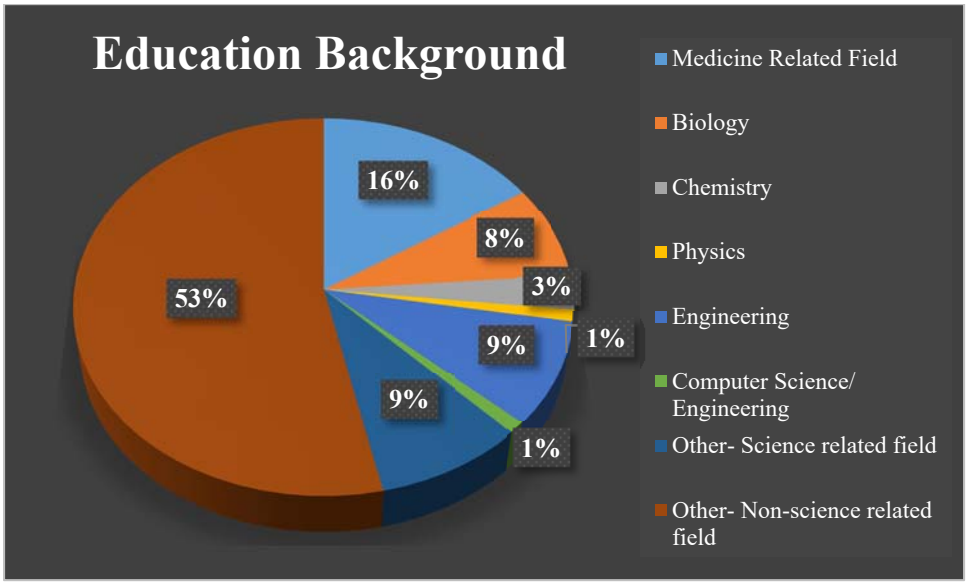

Figure 5: *Breakdown of the educational areas of study from SS and NSS respondents*

As can be observed above, 53% came from our NSS education areas, 16% reported they had studied a medical-related field, 9% was reported for both the fields of engineering and Other-Science related field of study, 8% had backgrounds in biology, 3% chemistry and 1% for both physics and computer science/engineering.

To gain a further understanding of our population we asked respondents to provide the highest degree achieved related to their educational backgrounds. The results of this question can be visualized in Figure 6.

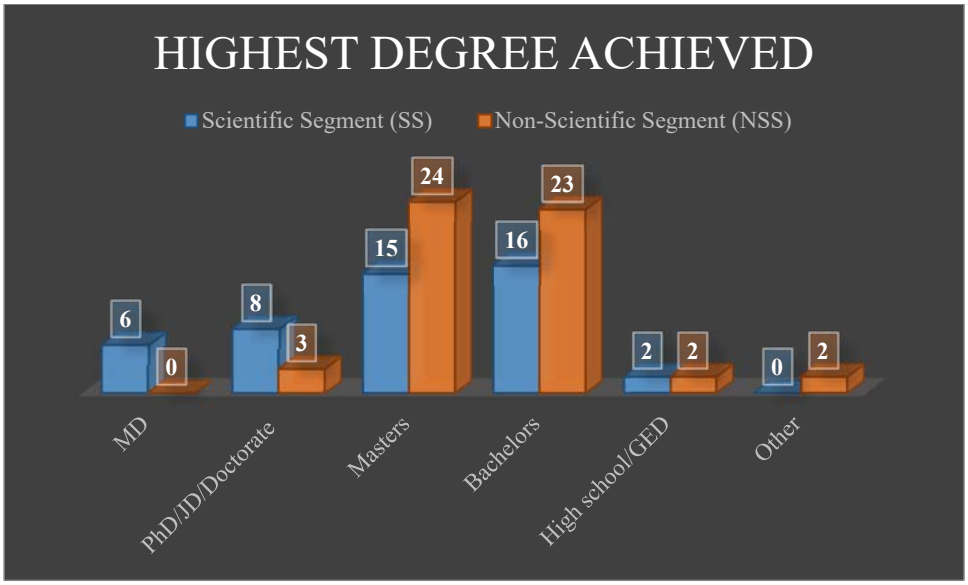

Figure 6: *Highest degree achieved by our SS (blue) and NSS (red) respondents*

Our SS reported 6 MD, 8 Ph.D./ JD/ doctoral degrees, 15 master's degrees, 16 bachelor's degrees, 2 high school or GED and 0 other degrees. Our NSS reported 0 MD, 3 Ph.D./ JD/ doctoral degrees, 24 master's degrees, 23 bachelor's degrees, 2 high school or GED and 2 other degrees.

Other next line questions for our target population aimed to understand their knowledge of the replication crisis in science, and the work currently being done to correct the issue. Our target population reported that less than half of both our SS group and/or NSS group had heard of the replication crisis. Corresponding to 45% of SS and 30% NSS, respectively.

Displayed in Figure 7 you can see that the majority in both groups responded as having no prior knowledge of the replication crisis in science.

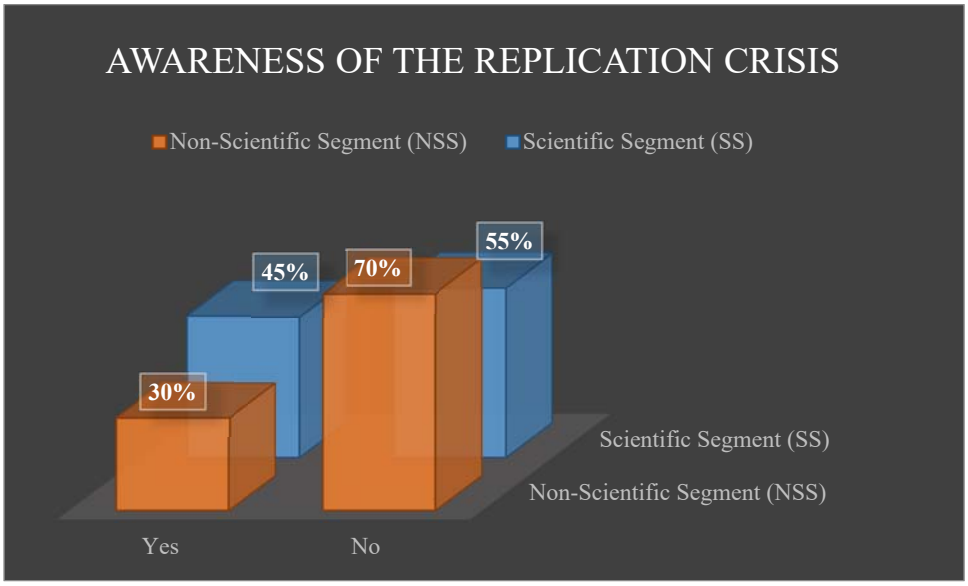

Figure 7:*Awareness of the replication crisis yes or no SS (blue) NSS (red)*

The responses to the next question shown in Figure 8, reported that 53% of our SS had knowledge that scientists have been working over the 2 decades, and are currently working to improve replicability or reproducibility in scientific reports. More surprisingly our NSS reported the 78% of respondents were unaware of any previous or current corrective action to improve replicability or reproducibility in scientific reports. The full results can be visualized in Figure 8 below. As there was a visual trend between groups for this question we decided to further investigate if there was statistical significance between the two subgroups of our population. The results of…revealed a value of 3.4352, with p value of 0.00058, meaning that the result between groups is statistically significant ($p < 0.05$)..

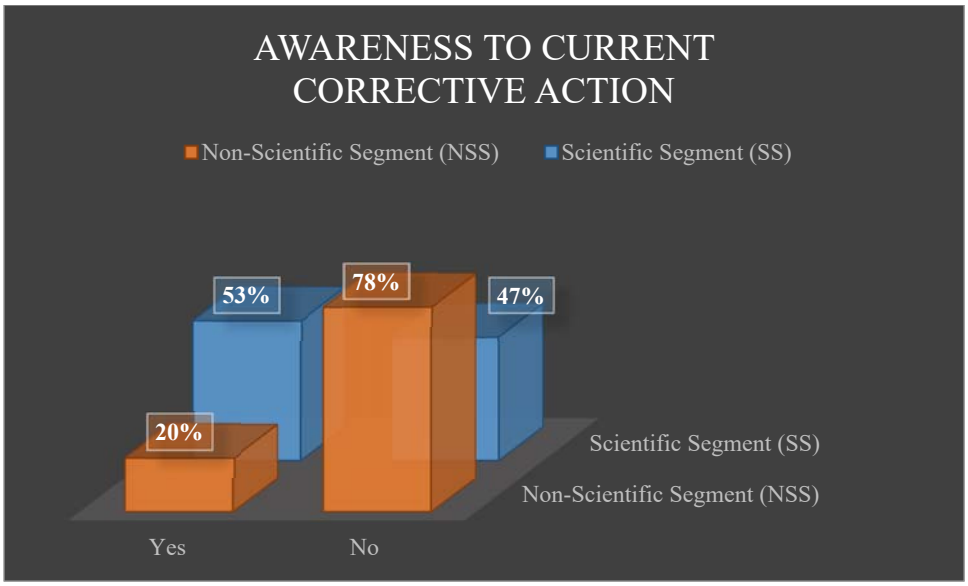

Figure 8: *Respondents awareness to corrective action for the irreproducibility crisis SS (blue) and NSS (red)*

All respondents answered yes to the question "Do you believe proper data collection and reporting methods are an important part of all pre-clinical (animal) research studies?"

Our next question looked to elucidate our population's understanding of the annual financial investment in pre-clinical research. Although reports exist for financial investment across all of science and the world as a whole, we were more focused on the investment of the NIH annually as they are a major supporter of our target population's research efforts. Moreover, this focus would help us in our discussion and implementation of the ARRIVE method. The published value for NIH pre-clinical annual spending is approximately 18 billion dollars a year. The results from our SS and NSS are shown in Figure 9

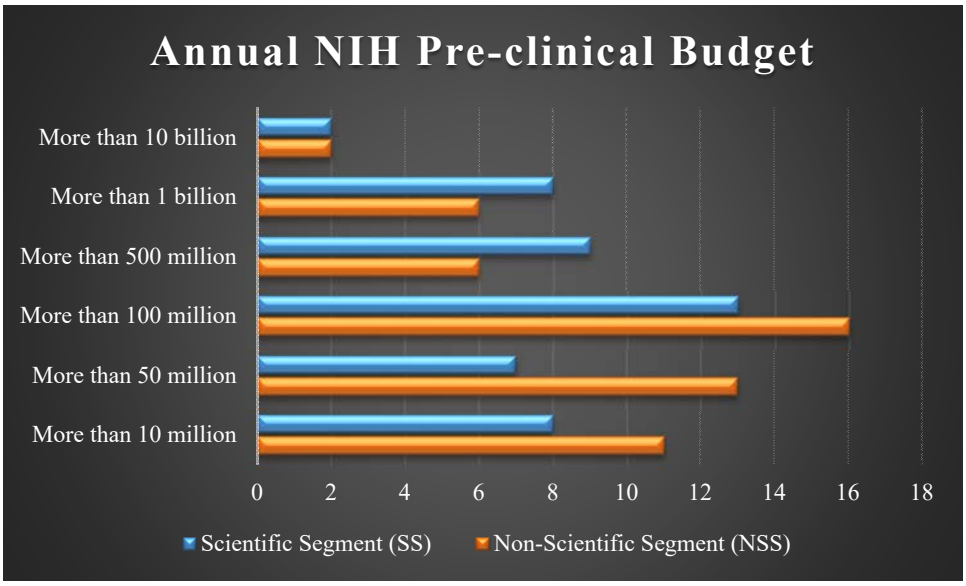

Figure 9: *Annual reported pre-clinical budget of the NIH, y-axis shows US dollars and x-axis in the number of respondent to select each bin, SS (blue) NSS (red)*

Only 2 respondents from our SS and NSS responded with the correct answer of over 10 billion dollars annually. This demonstrated that 95% of our target population did not correctly appreciate the annual financial investment of the NIH in preclinical research. The median answer for both groups was “More than 100 million” with 24% and 34%, SS and NSS respectively. Building on this financial commitment of the NIH we asked respondents if they thought; “Should all studies utilizing taxpayer dollars be required to collect and report study data in a harmonized way to increase reproducibility and reliability by reducing errors of omission?” exactly 94% (Figure 10) of respondents from both groups, answered yes to this question.

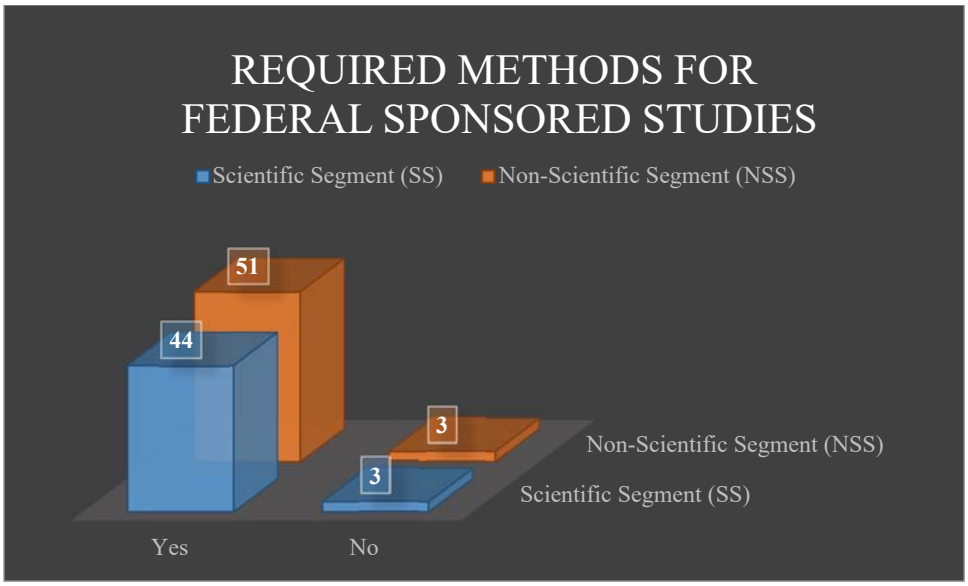

Figure 10: *Respondents answered yes or no to whether or not harmonized methods for NIH studies should be required, SS (blue) NSS (red)*

Our next question built off the required data collection and reporting methods piece of the prior question, by asking our respondents if they have ever heard of the "…ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines for proper preclinical research reporting". Table 1 demonstrates that both our SS and NSS were relatively unaware of the ARRIVE guidelines at the time of this cross-sectional analysis.

	YES	NO
Scientific Segment (SS)	11%	89%
Non-Scientific Segment (NSS)	7%	93%

Although 11% of the SS and 7% of the NSS had reported knowledge of the ARRIVE guidelines, only 3 respondents reported ever referencing the ARRIVE guidelines during either study creation or manuscript preparation (6% SS and 0% NSS).

Following the introduction of all of the 20 ARRIVE guidelines, we next moved to ascertain whether or not our population thought it was reasonable of authors publishing a manuscript to include all 20 of these guidelines in their reports. As can be seen in Figure 11, the majority thought that it was, in fact, reasonable to include a description of all 20 guidelines when reporting on their scientific findings (87% SS and 81% NSS). We further asked, "if authors include a description of these 20 reporting best practice guidelines it could improve the standards of reporting and replicability of pre-clinical studies?"; our population responded overwhelmingly yes (100% SS and 97% NSS).

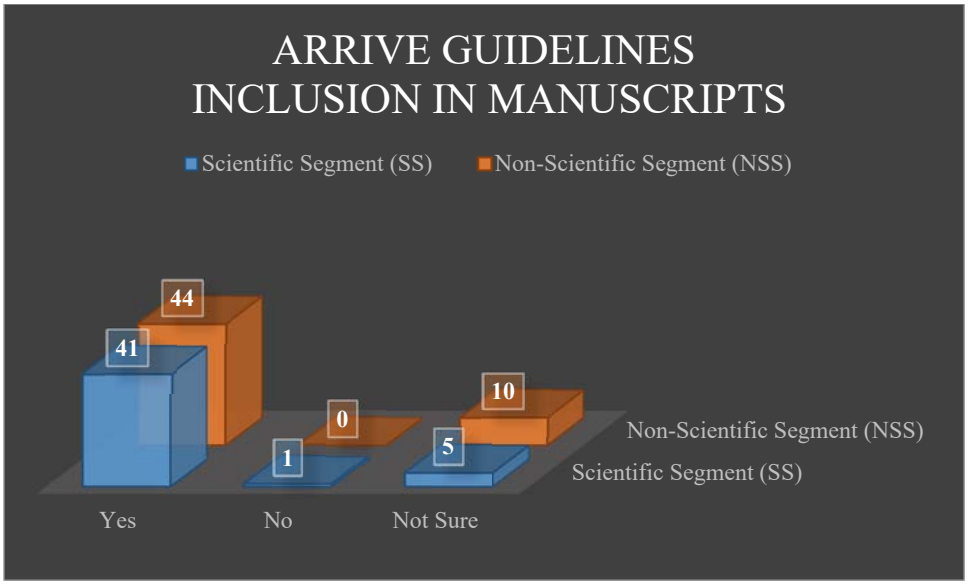

Figure 11: *Respondents answered yes or no related to the inclusion or not of ARRIVE during manuscript preparation, SS (blue) NSS (red)*

Further, we sought to understand the number of respondents in our population who had ever been paid from a research project which had been supported by the NIH. Our NSS population reported only 13% of its respondents had, as opposed to 53% of our SS (Figure 12).

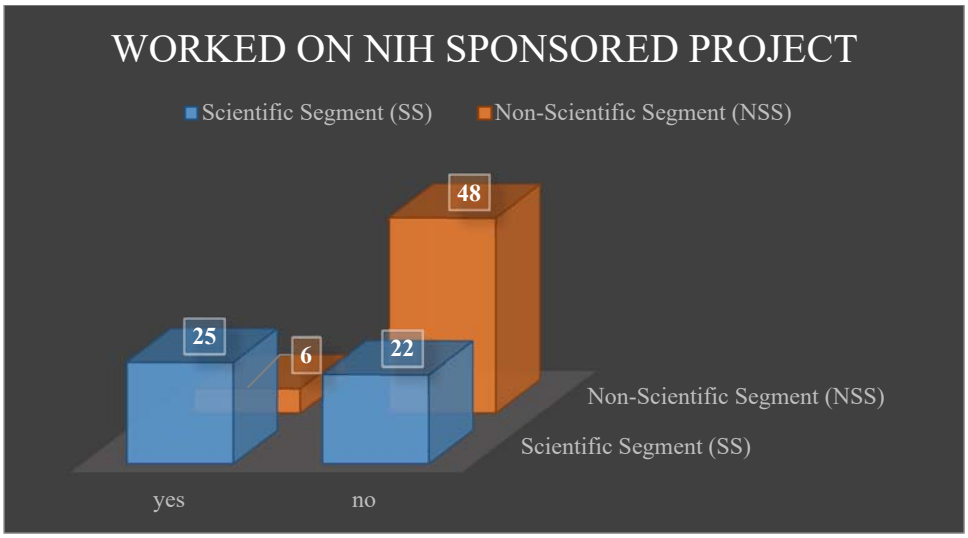

Figure 12: *Report responses of subjects being personnel of federal grants, SS (blue) NSS (red)*

Once we understood our population's relation with the NIH funding and their stance on proper data reporting related to the ARRIVE guidelines. Our next question aimed at understanding of our population's thought on making ARRIVE as a required reporting tool for a research project which was sponsored by the NIH.

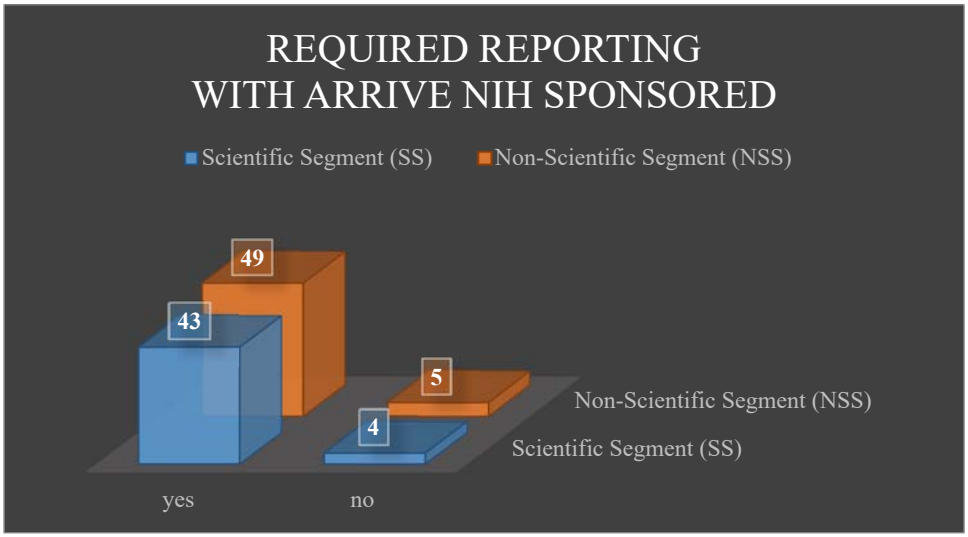

Figure 13: *Respondents answered yes or no to whether or not ARRIVE should be a reporting requirement for all NIH sponsored studies, SS (blue) NSS (red)*

Our population agreed with the idea of required reporting for all 20 ARRIVE guidelines during manuscript preparation when the study was sponsored by the NIH, with 92% of our SS and 90% of our NSS responding yes (Figure 13).

Further, we expanded the prior question by not limiting the sponsor type to just the NIH, asking if ARRIVE reporting should be used no matter whom the research study is sponsored by.

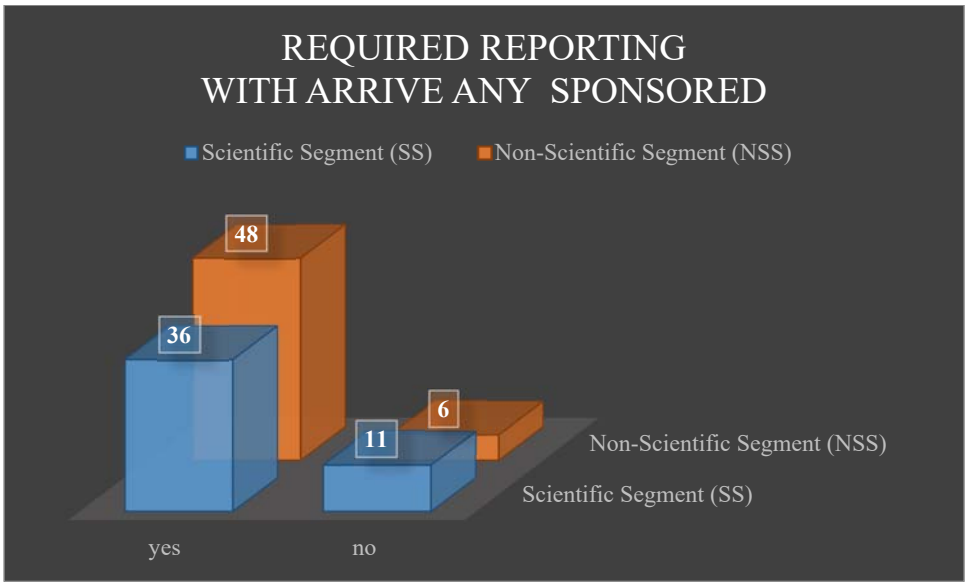

Figure 14: *Respondents answered yes or no to whether or not ARRIVE should be a reporting requirement for all sponsored research studies, SS (blue) NSS (red)*

We saw a slight dip in support but still a yes response from an overwhelming majority, 77% of our SS and 88% of our NSS (Figure 14).

We next focused on the potential financial investment required for the ARRIVE guidelines to be utilized by researchers during the publication phase of their work. With that we asked respondents if they thought; "…initial application of the ARRIVE guidelines could save time or money during publication?"

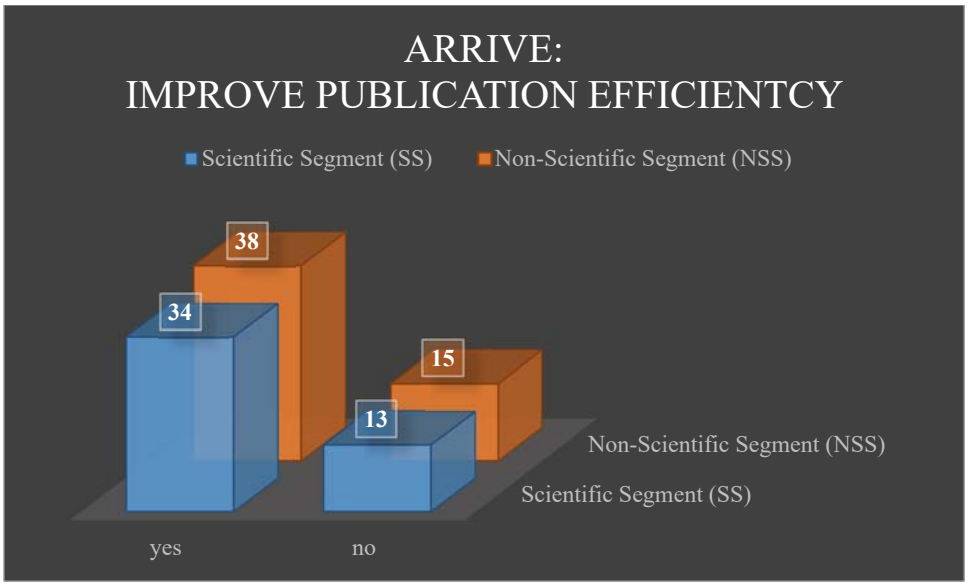

Figure 15: *Respondents answered yes or no to whether a saving of time and money by utilizing arrive during publication, SS(blue) NSS (red)*

Our population agreed that it could save time and money by using ARRIVE method during publication, 72% of our SS and 70% of our NSS (Figure 15).

Our final question looked at whether or not researchers who were trying to reproduce other's work would benefit from the inclusion of ARRIVE. We asked; “Do you think the application of the ARRIVE guidelines could save time or money for investigators looking to reproduce another’s work?”

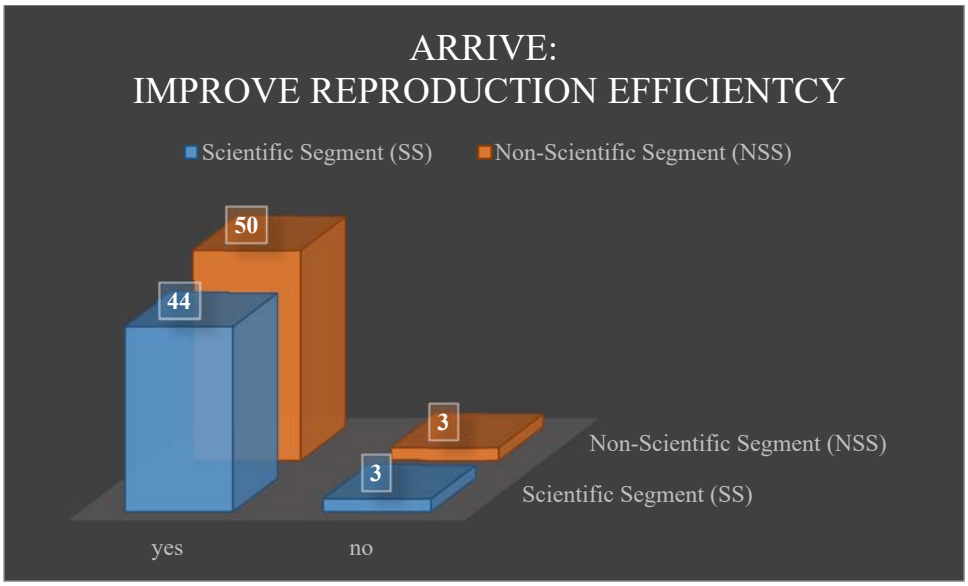

Figure 16: *Respondents answered yes or no to whether researcher trying to reproduce other's work would save time and money if ARRIVE was utilized by the primary researchers, SS (blue) NSS (red)*

Our population agreed that it could save time and money if ARRIVE guidelines had been applied when trying to reproduce other's work, 94% of our SS and 93% of our NSS (Figure 16).

Chapter IV

Discussion

The irreproducibility crisis in the sciences is a perplexing issue, which will require a unified community working together for corrective action to be successful. The first requirement in launching this corrective action movement will be, acceptance that there is, in fact, a problem with the way researchers across science in areas, including but not exclusive to designing their study, carrying out their methods and reporting their findings. Once researchers accept that irreproducibility is a problem, they can then start working toward a resolution to solve some or all of these related issues. My biggest concern coming into this research project was that researchers were unaware of this issue and how vast the potential for irreproducible studies was. As of now, there seems to be a very limited push toward corrective action for this issue. This means that researchers are either unaware of the problem or do not wish to investigate steps toward a resolution and are ok with the status quo.

To reiterate, in order to move forward with the corrective action, each of the scientific community's segment populations must be aware and accepting that there is actually a problem and be actively working toward a resolution. As this irreproducibility issue can be attributed to two large catch-all categories either "errors of omission" or fraudulent science, more needs to be done in regards to specific and detailed reporting. Movement is needed to make this suggestion a requirement in the peer-review process,

due to the fact that at least a moment in time, when an author's works are exposed as irreproducible, they quickly fall back on the acceptable excuse that they must have "omitted in error" key information required to reproduce their studies. Allowing this excuse to remain an option only perpetuates this as acceptable bad behavior and continues the cycles of preparation and publication of irreproducible works. Further, this acceptable excuse permeates to others rapidly and leaves the scientific community unable to discern the etiology of specific irreproducible events. This is significant as the future of the scientific community and sponsor research funding will ultimately be wasted.

To elaborate, if the work of Freedman et. al is correct, the scientific community could be wasting upwards of 10's of billions of dollars a year in the generation of irreproducible science (Freedman et al., 2015). As the competition for proper grant support is extreme, it is striking that researchers are not more concerned about the fall-out of this irreproducible research, mainly because their peers, who are potentially securing funds over them, are not completing work which is reproducible and further this potentially fraudulent and definitely wasteful practice. Furthermore, irreproducible publications are allowing researchers to secure even more funding, over their prospective competition and thus furthering the cycle in the wrong direction.

If scientists are going to start correcting the problem they first must become aware of and accept the issue that is irreproducibility. Second, they must come up with a way to weed out the fraudulent irreproducible studies from those who have mistakenly forgotten to include specific details. Finally, researchers can move toward correcting the areas including but not exclusive to; designing their study, carrying out their methods and reporting their findings. The results from this survey showed us that only half of the

members of our “scientific segment” (SS) and approximately a quarter of our “non-scientific segment” (NSS) of our research community were aware of the irreproducibility crisis in science, which is quite concerning as these individuals either work toward producing or work to support the publications of scientific findings. Moreover, the fact that two-thirds of our population completely unaware that the scientific community has been working toward corrective action for the irreproducibility crisis over the last decade is also concerning.

If our community of researchers is unaware that there is a problem, then we can’t begin to hope that they are actively working toward resolving this issue. It should be noted that all respondents of our survey agreed that proper data collection and reporting methods are important to all research in the preclinical field. Yet, only 9% (11% SS and 7.5% NSS) of our respondents had even heard of the ARRIVE guidelines. This lack of awareness in our population is especially concerning because approximately half of this community has manuscript (s) which are published in the same top-tier journals which include a reference to the ARRIVE guidelines. Not only do these journals include a reference on their homepage but again suggest authors to reference the guidelines during manuscript preparation. Something which our survey shows is not happening in our study population is that 6% of our SS having referenced the ARRIVE guidelines during study design and 0% have ever consulted the guidelines again during manuscript preparation.

Although the literature has not reported on the irreproducibility crisis in our study population’s area of research, it would not be surprising if a future study that focused on this community’s research area potentially found similar irreproducible results. Moreover, as some of our data collection methods are strikingly similar to this

neuroscience report of irreproducible methods (Button et al., 2013) entitled; “Power failure: why small sample size undermines the reliability of neuroscience.”, finding similar irreproducible datasets in our area of study could be a real possibility.

Compounding our population’s unawareness related to the replication crisis in science and their relative lack of knowledge about the ARRIVE guidelines was their under-appreciation for the amount of annual NIH funds earmarked for pre-clinical research. It is estimated that the NIH earmarks and spends approximately $17.76 billion dollars a year on pre-clinical research (Freedman et al., 2015). When exploring our population’s knowledge of this fact, we were surprised to find that only 2 members of each segment, SS and NSS, chose the correct annual bin amount (More than 10 billion). It is eerily concerning that our research community has such an under-appreciation of the investment of tax-payer dollars.

If our research community, a community that is dependent on this type of funding source, is unaware of the federal government’s annual investment in pre-clinical research, it can only be assumed that the general public would also under-appreciate this investment. Further, if approximately 53% of this research could be assumed irreproducible or flawed(Freedman et al., 2015), then we assume they are also unaware that approximately $9.4 billion dollars could be wasted each year. Work needs to be done by our scientific community to correct such an egregious waste, because, as of now, there seems to be limited knowledge to the fact that this waste is being incurred annually. If we fail as a community to correct the cracks in our peer-review process and method reporting systems it is only a matter of time before the public loses faith in these scientific reports,

and demands reform which may include decreased funding support, further straining an already overburden and highly competitive funding-pool.

Our population overwhelmingly believes that our publicly funded research studies should be required to collect and report data in a harmonized way to help decrease irreproducibility. The unified response in support of such reporting demonstrates that our population still believes in ethical science and proper reporting methods, which would lead us to believe they would be in support of using the ARRIVE guidelines during reporting, but to be able to utilize these guidelines they must know that they exist. Our reported responses to our inquiry to their knowledge of the existence of the guidelines, demonstrated that the majority of our population was not aware of them, with only 11% of our SS and 7% of our NSS reporting awareness. This further would make it quite difficult for our researchers to have already referenced the ARRIVE guidelines during study creation or during manuscript production, confirmed by only 6% of our SS and 0% of our NSS reporting they have ever referenced ARRIVE during these crucial research times. Although, the lack of knowledge of the ARRIVE guidelines could be the main culprit for lack of adoptions, it may not be the only motive we must consider.

Research study design, manuscript development and further publication through the peer-review process can be an arduous undertaking. Compounding this with the additional reference to the suggested ARRIVE guidelines may prove to be a formidable opponent to ARRIVE's adoption success. As the ARRIVE guidelines are a complete and all-inclusive guide for all areas of peer-review scientific reporting, they are involved and lengthy. To suggest to authors who have most probably already completed data collection, and interpretation and subsequent manuscript preparation, to add a voluntary

lengthy review and report of these 20 areas on top of this already difficult publication process could be daunting.

Not surprisingly, the potentially daunting nature of a 20 guideline review and report could be the reason why the scientific community has yet to embrace and implement ARRIVE. As such a group of authors moved to understand which of the 20 guidelines might be able to be condensed or culled entirely, these authors created two standard sets from the 20 ARRIVE guidelines: one which described the essential 10 or the minimum required to proper reporting, and another which they entitled the recommended set, that again describes the full 20 guidelines with examples of implementation and interpretation (Percie du Sert et al., 2019). The authors went on to describe how the lack of adoption of ARRIVE cannot be corrected without editors of the peer-review journal getting behind them and at least requiring a review of the essential 10.

This group felt that lessening the burden by half for the initial adoption process could help the research community start supporting the adoption of ARRIVE and hopefully make it a requirement for each new research report. Thus, strengthening the peer-review and reporting process will hopefully improve reproducibility. The implementation and adoption of ARRIVE was a query of our research population. Our population revealed 77% of our SS and 88% of our NSS were in support of making ARRIVE guidelines review and reporting a requirement during study reporting with any sponsor.

A potential reason for why the ARRIVE guidelines have not been adopted could be that the additional effort required from researchers to include the details of each of the

20 guidelines during their publication process is time-consuming and tedious. As the adoption and implementation of the ARRIVE guidelines would require significant initial time investment from researchers when producing these publications, it becomes unlikely that they would undertake this voluntarily. As the implementation of these details would require researchers to carry a significant monetary investment when initially including each of these sections in their future publications, it could easily be seen as a considerable deterrent for the adoption of ARRIVE by these time and money conscious investigators. Further supporting the significant modification by Percie du Sert et al., of the ARRIVE guidelines to the essential 10, which may help to improve the potential adoption and improve publication reproducibility for future investigators.

Although the adoption of ARRIVE has yet to gain momentum, the understanding of the reproducibility crisis in science is growing. Moreover, the push from young investigators to modify methods and reporting procedures is ever-growing. To this end, there have been multiple reports from young investigators and post-docs that they are being pushed by their seniors to follow the old method and reporting procedures (Poldrack, 2019). This is something that will need to be corrected as these young investigators are promoted into the decision making positions, the hope is that they do not lose their prowess to improve the system by the time they get to those posts.

Recently, reports from senior scientists have given credibility to the notion from the young investigators and post-docs that they are being pushed to use the well-established but questionable methods and reporting techniques. Specifically, this author tries to put the onuses of improving methods and reporting procedures to correct reproducibility back on the tenured professors, like himself, suggesting that although

these methods may have been the best available or the most common and allowed tenured professors to reach their high-ranking posts, they must understand that this should not remain the status quo. They should use their high-rank and liquidity to make sure that the science their young investigators are completing is more through, higher statistically powered and better reported. Furthermore, although the responsibility applies to all researchers, a great burden should be shouldered by the senior investigators. Who should be open to and promote improvements in data collection, increasing statistical power and cultivating better reporting procedures. These members of the scientific community are financially and ethically responsible to ensure that the irreproducibility which exists in research at this point is corrected moving forward.

There will never be an all-encompassing, or cure-all method or guideline to correct these attributes of the irreproducibility crisis (Alstrup & Sonne, 2019; Begley & Ioannidis, 2015; Kilkenny, Browne, Cuthill, Emerson, & Altman, 2010; Poldrack, 2019). There are many recommendations available for how to attack and solve the problem of irreproducibility, and further who's responsible for actioning these proposals (Alstrup & Sonne, 2019; Bauer, Bechtold, & Swiontkowski, 2019; Freedman et al., 2015; Percie du Sert et al., 2019; Poldrack, 2019). Yet, there is an agreement at this point, and there continue to be new reports across all scientific areas, that irreproducibility crisis is real. As such, there needs to be a greater response to this issue and it must happen sooner rather than later, to ensure that funding is not being wasted and we are not losing precious time which we can never get back.

After a comprehensive review of the current methods available, we support the adoption and implementation of the ARRIVE guidelines, as they not only serve as a

check-list for investigators but they can also be viewed as a teaching tool for how to develop, produce, and report science across plethora of research areas. Working not only to generate more research studies that are reproducible but also help to refine the researcher's scientific methods for the future, is our goal. It should be noted again that our research community agreed that the production of reporting methods should be harmonized across science which inherently supports a method like the ARRIVE guidelines or a similar framework.

Finally, it should be stated that our survey was slightly underpowered to produce a confidence level of 95% and a confidence interval of 5% with 101 respondents instead of the required 107 for this power calculation. Yet, it was closer to the 107 than it was 59 respondents which relate to the power calculation with a confidence level of 95% and a confidence interval of 10%. It should again be noted that we were not able to find a study on irreproducibility related to scientific publications specific to our target population's area of research. Although, we hesitate to say that none would exist following an attempt at reproduction. Further research is required in this area to confirm or deny whether this is accurate

It is our hope that this survey will help by spreading the knowledge of the irreproducibility crisis in science through-out our research community, suggesting that there is, in fact, an issue across all areas of research in our community, and that the community has to take the responsibility for the quality improvement needs across study design, data collection and reporting methods. Introducing the ARRIVE guidelines as a suggested method for immediate improvement, will get our scientific community

thinking and questioning their current methods, or at the very least “googling” to confirm our notion that there is, in fact, an irreproducibility crisis in science today.

Chapter V

Bibliography

Alstrup, A. K. O., & Sonne, C. (2019). 3Rs as part of preclinical neuropsychiatric translational crisis, and ARRIVE guidelines as part of solution. *Acta Neuropsychiatr, 31*(6), 348-349. doi: 10.1017/neu.2019.40

Baker, David, Lidster, Katie, Sottomayor, Ana, & Amor, Sandra. (2014). Two Years Later: Journals Are Not Yet Enforcing the ARRIVE Guidelines on Reporting Standards for Pre-Clinical Animal Studies. *PLOS Biology, 12*(1), e1001756. doi: 10.1371/journal.pbio.1001756

Baskin, P. K., Mink, J. W., & Gross, R. A. (2017). Correcting honest pervasive errors in the scientific literature: Retractions without stigma. *Neurology, 89*(1), 11-13. doi: 10.1212/WNL.0000000000004106

Bauer, T. W., Bechtold, J. E., & Swiontkowski, M. F. (2019). JBJS Will Require Adherence to ARRIVE Guidelines for Animal Research to Reduce Bias and Improve Quality of Reporting. *J Bone Joint Surg Am, 101*(21), 1891-1893. doi: 10.2106/JBJS.19.01001

Begley, C. G. (2013). Six red flags for suspect work. *Nature, 497*(7450), 433-434. doi: 10.1038/497433a

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533. doi: 10.1038/483531a

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res, 116*(1), 116-126. doi: 10.1161/CIRCRESAHA.114.303819

Bissell, M. (2013). Reproducibility: The risks of the replication drive. *Nature, 503*(7476), 333-334.

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. *Nat Rev Neurosci, 14*(5), 365-376. doi: 10.1038/nrn3475

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505*(7485), 612-613.

Fang, F. C., & Casadevall, A. (2011). Retracted science and the retraction index. *Infect Immun, 79*(10), 3855-3859. doi: 10.1128/IAI.05661-11

Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *PLoS Biol, 13*(6), e1002165. doi: 10.1371/journal.pbio.1002165

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Sci Transl Med, 8*(341), 341ps312. doi: 10.1126/scitranslmed.aaf5027

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *R Soc Open Sci, 5*(1), 171511. doi: 10.1098/rsos.171511

Harris-Kojetin, B. A., & Groves, R. M. (2017). In B. A. Harris-Kojetin & R. M. Groves (Eds.), *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington (DC).

Ioannidis, John P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine, 2*(8), e124. doi: 10.1371/journal.pmed.0020124

Kilkenny, C., Browne, W., Cuthill, I. C., Emerson, M., Altman, D. G., & Group, N. C3Rs Reporting Guidelines Working. (2010). Animal research: reporting in vivo experiments: the ARRIVE guidelines. *J Gene Med, 12*(7), 561-563. doi: 10.1002/jgm.1473

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 Biol, 8*(6), e1000412. doi: 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. doi: 10.1371/journal.pone.0007824

Lawrence, P. A. (2003). The politics of publication. *Nature, 422*(6929), 259-261. doi: 10.1038/422259a

Leung, V., Rousseau-Blass, F., Beauchamp, G., & Pang, D. S. J. (2018). ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One, 13*(5), e0197882. doi: 10.1371/journal.pone.0197882

NIH. (2019, October 2019). FY 2019 Operating Plan. *Total NIH Budget Authority: FY 2019 Operating Plan.* Retrieved Nov 23 2019, 2019, from https://report.nih.gov/nihdatabook/report/5

Osherovich, Lev. (2011). Hedging against academic risk. *Science-Business eXchange, 4*(15), 416-416. doi: 10.1038/scibx.2011.416

Percie du Sert, Nathalie, Hurst, Viki, Ahluwalia, Amrita, Alam, Sabina, Avey, Marc T., Baker, Monya, . . . Würbel, Hanno. (2019). The ARRIVE guidelines 2019: updated guidelines for reporting animal research. *bioRxiv*, 703181. doi: 10.1101/703181

Perrin, S. (2014). Preclinical research: Make mouse studies work. *Nature, 507*(7493), 423-425. doi: 10.1038/507423a

Poldrack, R. A. (2019). The Costs of Reproducibility. *Neuron, 101*(1), 11-14. doi: 10.1016/j.neuron.2018.11.030

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov, 10*(9), 712. doi: 10.1038/nrd3439-c1

Steen, R. G. (2011). Retractions in the scientific literature: is the incidence of research fraud increasing? *J Med Ethics, 37*(4), 249-253. doi: 10.1136/jme.2010.040923

Tsilidis, Konstantinos K., Panagiotou, Orestis A., Sena, Emily S., Aretouli, Eleni, Evangelou, Evangelos, Howells, David W., . . . Ioannidis, John P. A. (2013). Evaluation of Excess Significance Bias in Animal Studies of Neurological Diseases. *PLOS Biology, 11*(7), e1001609. doi: 10.1371/journal.pbio.1001609