2018

Reproducibility crisis in science: causes and possible solutions

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Thesis

REPRODUCIBILITY CRISIS IN SCIENCE:
CAUSES AND POSSIBLE SOLUTIONS

by

DANIEL A. DRIMER-BATCA

B.A., University of Michigan, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2018

Approved by

First Reader ____
Alan Fine, M.D.
Professor of Medicine

Second Reader ____
Carl Franzblau, Ph.D.
Professor of Biochemistry

REPRODUCIBILITY CRISIS IN SCIENCE:
CAUSES AND POSSIBLE SOLUTIONS

DANIEL A. DRIMER-BATCA

ABSTRACT

Part I. Claims to knowledge require justification. In science, such justification is made possible by the ability to reproduce or replicate experiments, thereby confirming their validity. Additionally, reproducibility serves as a self-correcting tool in science as it weeds out faulty experiments. It is therefore essential that experimental studies be replicated and confirmed. Recently, attempts to reproduce studies in several fields have failed, leading to what has been referred to as "a crisis of reproducibility." This crisis is largely a result of the current culture in the scientific world. Specifically, it is a result of a system that incentivizes individual success in the form of publications in high-impact journals over collaboration and careful conductance of research. This environment contributes to the crisis of reproducibility by increasing biases, incentivizing researchers to engage in manipulative statistics, decreasing quality control and transparency, and increasing the likelihood of researchers engaging in fraudulent behavior.

Possible solutions to the problem of irreproducibility could tackle individual factors. A more prudent approach would be to focus on changing the current culture in the scientific world. Increased transparency had been suggested as a way to solve this problem. There is currently a movement advocating for increased transparency in science through "open science."

Part II. Retraction of scientific papers due to evidence of research misconduct is on the rise, having increased tenfold from 2000 to 2009. Previous work on this topic focused on published retraction notices, using notices to identify the percent of retracted articles that were caused by research misconduct. This study utilized a different approach. Using the Office of Research Integrity database, we first identified publications that resulted from research misconduct. We then searched those articles to determine whether they were indeed retracted. Once retraction notices were identified, they were scored based on scoring elements reflecting guidelines for transparency. Lastly, we investigated whether a correlation exists between the quality of a retraction notice and journal impact factor. Our findings suggest that 21% of papers containing data derived from scientific misconduct are not retracted. Moreover, the quality of retraction notices varies, with some elements more likely to be present than others. No significant correlation between retraction notices and journal impact factor was found.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

COPE	Committee on Publication Ethics
NIH	National Institute of Health
ORI	Office of Research Integrity

PART I

INTRODUCTION

The first part of the thesis will explore the so-called crisis of reproducibility, including the background to the problem, its extent, and significance. In particular, the reasons for the crisis will be investigated, and the remediation strategies that have been proposed or implemented will be described.

Importance and Historical Background of Reproducibility in Science

Science, as an enterprise, is concerned with obtaining knowledge about the universe through the implementation of the scientific method. While the question of what constitutes knowledge—or if we can ever truly know anything—is a contested topic, I believe most would agree with Plato’s definition of knowledge as a “justified true belief.” To claim something to be known requires, at the very least, some degree of justification for that claim. And since science is concerned with obtaining knowledge, it must provide justification for the validity of its findings. Reproducibility, or replicability, allows scientists to weed out false claims or bad science, thereby ensuring that the results of a scientific experiment are correct. In fact, successfully reproducing an experiment is one of the few tools available to scientists in their pursuit of justification—so much so that philosopher Karl Popper once stated: “Non-reproducible single occurrences are of no significance to science” (Popper, 1959). It is important to note that reproducibility does not necessarily provide us with knowledge that the conclusion of an experiment is true. For example, the observation that maggots will eventually appear in unrefrigerated raw meat does not provide justification for the theory of spontaneous generation. Similarly, reproducibility is not an essential requirement for all scientific fields (after all, we cannot replicate the evolution of human life, or the big bang, in a laboratory). So while reproducibility does not guarantee the validity of conclusions derived from experimental results, it confirms the validity of the process that leads to them and is, therefore, a cornerstone of experimental science and the scientific method.

The importance of reproducibility in science had been known for centuries. In their 1985 book, Leviathan and the Air-Pump, Shapin and Schaffer describe a well-documented debate between the 17th-century scientists Robert Boyle and Thomas Hobbes. The debate revolved around the integrity of Boyle's newly invented air-pump, a machined he used to study the properties of air and, specifically, the controversial concept of a vacuum. After learning that the Dutch scientist Christiaan Huygens was able to levitate a drop of water inside a jar using his own version of the air-pump, Boyle attempted to replicate the experiment with no success. Convinced that unless the experiment could be replicated, the value of its findings would be questionable, Boyle invited Huygens to England and together they were able to reproduce the phenomenon. While Hobbes denied the existence of a vacuum, Boyle maintained that by successfully repeating the experiment, its findings could not be denied. At its core, the debate was about the proper way of obtaining knowledge—with Boyle’s notion that it should be through the conduction of experiments and their successful reproduction prevailing.

More recently, however, concerns have been raised regarding the reproducibility of many published experiments. Irreproducibility had been reported to varying degrees in fields such as medicine, chemistry, engineering, environmental science, and psychology, to name a few (Baker, 2016). As attention given to the problem increases so does our understanding of its extent, and today many scientists talk of a “crisis of reproducibility.”

The Reproducibility Crisis

In 2016, the journal *Nature* published the results of a survey completed by 1,576 scientists from various fields (Baker, 2016). Over 70% of polled researchers admitted to having attempted and failed to reproduce another researcher’s published experiments, and more than half reported failing to reproduce their own experiments. When asked whether there is a reproducibility crisis, 90% of surveyed researchers stated that there is either a significant or slight crisis. Only 3% held the view that there is no crisis.

The poll showed that the problem of irreproducibility affects many fields, including biology, physics and engineering, environmental science, and medicine (Baker, 2016). Data suggest that the most affected fields are psychology and biomedicine. In 2012, researchers attempted to reproduce the results of 53 cancer studies, which they had deemed "landmark" studies. They failed to reproduce 47 of the 53 papers (Begley & Ellis, 2012). Other studies have also tried and failed to replicate the results of many preclinical biomedical studies presented in prestigious journals with irreproducibility numbers ranging from 75% to 90% (Yaffe, 2015). A similar problem of irreproducibility exists in observational studies, with Ioannidis reporting that 80% of such studies either fail to be replicated or yield results that are significantly smaller than originally stated (Ioannidis, 2015). Similarly, in one of the largest replication study in psychology, conducted in 2015 by Brian Nosek et al., attempts to replicate 100 original studies published in three psychology journals were successful in only 39 of the 100 papers (Baker, 2015). A separate study from that same year, which attempted to replicate a separate set of 100 psychology papers, achieved statistical significance in 36% of replications, compared with 97% in the original studies (Aarts et al, 2015).

Failure to replicate experiments is especially impactful when it occurs in the basic sciences or in preclinical studies, as those types of studies provide the background and basis for future experiments. Preclinical studies, in particular, serve as the basis for new drug development, an expensive and resource consuming process. Irreproducibility of preclinical studies, which suggests potential problems with the original experiments, has the potential to affect the success of drug development. Indeed, recent studies have blamed the declining success rates of Phase II trials for new drug developments on the irreproducibility of much of the published data which serves as the basis for the drug development (Prinz, Schlange, & Asadullah, 2011).

Defining the Terms—What Constitutes Reproducibility?

As the attention given to—and publications on the topic of—the problem of irreproducibility increases, questions have been raised regarding the terms used to describe the problem, and their meaning (Oransky & Marcus, 2016).

According to a 2015 report by the National Science Foundation Subcommittee on Replicability, reproducibility means “the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator. [For example], a researcher uses the same raw data, builds same analysis files, and same statistical procedures to make sure that same results obtained as in published study.” Replicability, on the other hand, refers to “the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected.” (Bollen, 2015). The same report also acknowledges the lack of a consensus in science on the meaning of these terms, noting that different terms can refer to the same thing, while the same term can be used to describe different things. Given the lack of an agreed upon definition, it is often hard to know what researchers mean when using these terms.

Furthermore, it is not clear from the definitions of these words what constitutes reproducibility or replicability. For example, it is not clear whether successful reproduction entails complete agreement between the results of the original and replication study, or if small deviations that do not alter the conclusion fall under successful reproduction. It is similarly unclear whether studies that yield the same data but are interpreted differently constitute successful reproduction.

A recent paper by Goodman, Fanelli and Ioannidis (2016) suggests that many of the terms used in discussing this issues are proxies for the word “truth,” a fact they criticize. They conclude their paper by pointing out the importance of defining these terms in a universally accepted manner, and of better understanding the relationship between the terms we use and the truth of a scientific claim.

Specific Aims

Specific aims of the first part of this thesis include the following:

1. Review of relevant literature to identify reasons for the crisis in reproducibility.
2. Analysis of possible solutions to the crisis in reproducibility.

WHY DO SCIENTISTS FAIL TO REPRODUCE RESEARCH?

Numerous reasons have been identified as contributing factors to the crisis of reproducibility. While varying, a common factor for those reasons is the “overflow” of science. Siebert et al. offer the following explanation. An increased demand for limited resources in the scientific world, such as jobs, postdoctoral positions, and grants, combined with a perceived decrease in funding, had led to increased pressure to produce results. Given this so-called overflow, publications in high-impact journals have become a measuring stick for success. This, in turn, creates increased pressure on scientists to produce high-impact publications, to the point where the publications themselves, and not the science behind them, had become the goal of scientists (Siebert, Machesky, & Insall, 2015). As will be discussed shortly, the increased pressure to publish and the sheer amount of science being produced has led to biases, decreased quality control and transparency, and other factors that have contributed to the problem of reproducibility.

Biases

Biases have been known to affect thought and reasoning for centuries. Francis Bacon identified four types of biases, which he called idols, as early as 1620, in his philosophical work *Novum Organum.* Today, biases continue to affect human reasoning and the way we conduct science. These biases come in different forms and are exacerbated by increased competition and pressure to publish.

Perhaps the most obvious type of bias, referred to by Bacon as “idols of the cave,” is personal biases. Such biases are often unconscious, and stem from the desire to support one’s theory, to refute an opposing theory, or to publish a new discovery on a topic. These biases affect how investigators conduct experiments or analyze data, and could even lead researchers to overlook calculation mistakes if the results are aligned with their expectations (Nuzzo, 2015; Maccoun & Perlmutter, 2015).

Another, perhaps less obvious, type of bias is publication bias—a term referring to the fact that research findings are either published or not publication depending on the nature of the results (Dickersin, 2005). The idea behind publication bias is that in an ideal scientific world, the valid results of every legitimate, well conducted and honest study should be published. In such a world, the scientific literature would reflect the true and full body of research that had been conducted on a specific topic, with no artificial biases towards certain outcomes. Yet this is not the case. In today’s scientific landscape, the results of a significant proportion of conducted studies are never made public. This fact had been known since at least 1959, when a paper by Sterling noted that an astonishing 97% of studies published in a number of reputable psychology journals reported statistically significant results (Sterling, 1959). In that same paper, Sterling coined the term “publication bias” to refer to the fact that “successful” studies—that is, studies with positive and statistically significant results—are more likely to be published. In 1979, Rosenthal further commented on this issue, terming it “the file-drawer problem,” in reference to the fact that results which do not support the researchers’ hypothesis end up in the file drawer, rather than in a journal (Rosenthal & Hernstein, 1979). More recently, it was suggested that as many as 95.8% of papers on the topic of cancer prognostic markers reported statistical significance (i.e. positive results) (Kyzas, Denaxa-Kyza, & Ioannidis, 2007). Additional studies have taken different approaches to investigating this phenomenon. A 2013 study utilized a database containing all drug-evaluating clinical trials approved by the Ethics Committee of a Spanish hospital over a period of 7 years. The study found that the publication rate in peer-reviewed journals for completed trials with positive results was 84.9%, compared with 68.8% for studies with negative results. The researchers concluded that trials with positive results were more likely to get published than those with negative results (Sune, Sune, & Montoro, 2013). There is also direct evidence for this phenomenon in the form of admission to bias by parties involved in the publication process (Song et al, 2009).

A closely related problem to the non-publication of negative results is that of outcome reporting bias. Outcome reporting bias, which affects published papers, is defined as the selective reporting of some, but not all, outcomes. A 2013 study by Riveros et al. investigated selection bias by utilizing the fact that the US Food and Drug Administration (FDA) Amendments Act requires results from clinical trials of FDA-approved drugs to be made available in an online database. The researchers compared the results of clinical trials as they appeared in the online database with the results of the same studies as they were presented in published journals. They found that the results in the online database were far more complete and included findings that were not presented in journal publications, suggesting the existence of a selective reporting bias (Riveros et al, 2013). Other studies have confirmed that statistically significant results are more likely to be fully reported, whereas non-significant results are more likely to be partially reported (Dwan et al, 2008).

Interestingly, the relative lack of negative results in the literature is not so much a result of journals failing to publish such papers, but of researchers failing to submit them. An analysis of 745 manuscripts submitted for publication in JAMA revealed only a slight, statistically-insignificant increased rate of publication for studies with positive-results over those with negative-results (Olson et al, 2002). A study looking into publication bias in clinical trial reporting similarly concluded that non-publication was mostly a result of failure to submit manuscripts, not of rejection by journals (Dickersin et al, 1987). Such findings would be expected given the overflow in science and the increased pressure on scientists to make new, impactful discoveries.

Fanelli, Costas, and Ioannidis (2017) identify additional bias patterns which negatively impact the scientific literature. The “small-study effect” refers to the observed phenomenon of smaller studies showing larger, more significant treatment effects than large studies (Schwarzer, Carpenter, & Rücker 2015). A particular problem that arises from the presence of small-study effects is their impact on meta-analysis reviews. Meta-analysis reviews are believed to provide better evidence than any single study on a specific topic, which they achieve by pooling the results of many studies on a particular subject. However, meta-analyses that include many small studies have the potential of being slanted towards a particular result, which is not reflective of the truth. It had been suggested that the small-study effect is likely due to a combination of factors. Some of those factors, like publication or reporting biases, have been discussed previously. Other factors include low methodological quality of small studies, such as a bias towards selecting participants who are more likely to produce significant results, or mere-coincidence (Nuesch et al, 2010; Schwarzer et al, 2015).

Another observed phenomenon is the diminishing effects obtained when repeating a study. Stated differently, earlier studies reporting a certain effect are likely to overestimate the effect’s magnitude relative to later studies. This has been termed “the decline effect,” and is attributed to the decrease in publication bias on a specific topic over time (Schooler, 2011). A related phenomenon is the “early-extreme effect,” which describes the increased likelihood of early studies to report extreme effects. This too is attributed to the publication bias, as extreme findings are more likely to be published early (Ioannidis & Trikalinos, 2005).

Citation bias describes the fact that studies with strong magnitudes of effects and large statistical significance are more likely to be cited. Since it is common to look for references in articles when investigating a certain question in the literature, citation bias may lead to a distorted view of the field (Jannot et al, 2013).

Lastly, industry bias is associated with the increased role of industries in the scientific world, and the financial support they provide for research. The increased involvement of companies in scientific research has been linked to increased reporting bias in the literature. A 2003 study found that research funded by drug companies is less likely to be published than research funded from other sources, and that the results of published industry-funded research are more likely to have results favorable to the sponsor (Lexchin et al, 2003).

These bias patterns diminish reproducibility in multiple ways. Personal biases do so by potentially introducing sloppy work or questionable practices to studies.

Attempts to reproduce such studies by scientists who do not share the same motivations as the original researchers might, therefore, alter how the research is conducted or what data is used, and could result in unsuccessful attempts to replicate original studies. Publication and outcome reporting biases affect irreproducibility by creating a scientific literature in which certain results are overrepresented. Since the overrepresentation of those results is not an accurate representation of all sampling done by researchers, it is not a true representation of those phenomena in the world. It thus follows that attempts to reproduce those studies in an unbiased matter are more likely to fail than succeed. Lastly, biases such as small-study and decline effects lead to results that can often be explained by small sample sizes and poor statistics. Attempts to reproduce such studies with larger sample sizes are likely to obtain different conclusions, thus making the original studies irreproducible.

Bad Statistics

An issue closely related to the biases discussed above, bad statistics are another reason some findings cannot be reproduced. The pressure to publish once again plays a role, as it incentivizes researchers to engage in practices that increase the likelihood of finding significant results. Some of those practices include using flexible statistical analyses and conducting small studies with low statistical power. Of particular consequence is the issue of low statistical power. It has been shown that low statistical power (due to small sample size) affects studies in three ways. First, it reduces the likelihood of detecting a true effect. Second, it reduces the likelihood of statistically-significant results being a true representation of an effect. Third, even when describing a true effect, it tends to exaggerate its magnitude (Button et al, 2013). Given these facts, it is disconcerting that many studies have been found to have low statistical power. A study published in *Nature* concluded that the average statistical power of studies published in the field of neuroscience is between 8% and 31% (Button et al, 2013). For comparison, the conventionally accepted minimum value for statistical power is 80%. Similarly, a 2017 study of statistical power in biomedical science found that approximately half the studies had statistical power in the 0% to 20% range (Dumas-Mallet et al, 2017). To illustrate what this means for reproducibility, consider a study with a 20% statistical power. By definition, conducting an exact replication of such a study will, on average, only yield the same results 20% of the time.

An additional practice that had been suggested as contributing to the probability of obtaining false-positive results is flexibility in data collection. Data-collection flexibility refers to the practice of either continuing data collection until the desired result is obtained, or ceasing data collection as soon as one is found (Nosek, Spies, & Motyl, 2012). In a 2011 survey, approximately 70% of polled behavioral scientists admitted to having engaged in data-collection flexibility at least once (John, Loewenstein, & Prelec, 2012). Additionally, Simmons et al. conducted statistical analyses that show how a continuous increase in sample size leads to fluctuations in p-values, further illustrating how data-collection flexibility can impact results (Simmons, Nelson, & Simonsohn, 2011)

Quality Control

Another consequence of the pressure to publish is a willingness to forgo adequate quality control measures when conducting experiments, for the sake of expediency. One major quality control issue, which affected scientists for over 50 years, is the cross-contamination, misidentifying and mislabeling of cell lines (Freedman et al, 2015). Immortal cell lines are cells which, due to mutations that cause loss of cell cycle checkpoint pathways, are able to proliferate indefinitely in vitro (Irfan Maqsood et al, 2013). Such cells are widely used in laboratories as a model to study multicellular organisms, and to develop new drugs. Consequently, when a laboratory purchases a specific cell-line, which they believe to be derived from a specific type of cancer, they might, in fact, be unknowingly working with cells derived from a different type of cancer. They may even be working with cells derived from a different species (McCook, 2015). The fact that cross-contamination and misidentification of immortal cell-lines is a problem has been known since at least 1966, yet as of 2012, an estimated 15% of human cell lines were not derived from the claimed source (Masters, 2012). Despite the prevalence of this problem, efforts to curb the phenomena had only recently gained traction (Nardone, 2008). To that end, in 2012 the International Cell Line

Authentication Committee (ICLAC) was established in order to “make cell line misidentification more visible and to promote awareness and authentication testing as effective ways to combat it” (http://iclac.org/about-iclac/). As of February 2018, the ICLAC database lists 451 cell lines that are misidentified, and 49 cell lines that come from different species.

This issue can have costly consequences. Take for example the case of Radoslaw Stachowiak, who, in October 2014, published a paper in *Current Microbiology* comparing how *Listeria* invades three different cell lines. When the authors found out that all three cell lines investigated were in fact derived from a single source—the HeLa cell-line—they were forced to publish a correction to their paper (Stachowiak et al, 2015). Such a mistake is not only costly due to the time and money spent by the original team who published the findings, but also due to the amount of citations these studies receive, and the lost time and effort to reproduce these sometimes unreproducible findings (Neimark, 2015).

These problems are accentuated by a number of factors. An article by Freeman et al. identifies a number of problems which contribute to the irreproducibility of papers due to cell-line misuse (Freedman et al, 2015). For one, there are data to suggest that most laboratories fail to conduct cell line authentication before publishing a paper; a 2014 *Nature Cell Biology* editorial reporting the results of an audit of papers published between August to December 2013 showed that only 19% of publications authenticated cell-lines used in their studies (An update on data reporting standards, 2014). Sharing of cell-lines among scientists, as opposed to obtaining such cells from a reputable source such as a cell bank, is another contributing factor. In addition, Freeman et al. identified the lack of cheap, fast, and commercially available authentication techniques, as a problem. The article also notes the lack of universally enforced reporting guidelines for reporting cell authentication in publications. Having such guidelines in place, they claim, would “help ensure the credibility, reproducibility and translatability of the data and results” (Freedman et al, 2015).

The issue of quality control is not limited to cell-line contamination. As part of an effort to create a drug screening library, Corsello et al. sought to confirm the identity and purity of 8,584 chemical compounds sourced from different vendors. To their surprise, they found that 2,482 compounds, comprising 29% of the 8,584 testes, failed quality control standards (defined as impurities making up 15% or more of the reagent) (Corsello et al, 2017). Such impurities directly impact the ability of researchers to conduct and replicate experiments. Take for example the case of a Stefan Knapp, a German scientist who obtained three different results when conducting the same experiment on three different occasions, each time using the same compound sourced from a different vendor (Bakes, 2017). A similar incident occurred in 2012, when 18 separate vendors were found to be supplying the wrong compound, which they advertised as the leukemia drug bosutinib (Levinson, Boxer, & Ramchandran, 2012).

Data Interpretation

Shortly after Rafael Silberzahn and Eric Uhlmann published their findings that Germans with noble-sounding surnames, such as *Kaiser* (“emperor”) or *König* (“king”) are more likely to hold managerial positions than Germans with surnames that refer to everyday occupations such as *Bauer* (“farmer”) or *Becker* (“baker”), they received a request for their dataset (Silberzahn & Uhlmann, 2013; Silberzahn & Uhlmann, 2015). Using the same dataset that yielded the original conclusion, Uri Simonsohn showed that there is, in fact, no correlation between noble-sounding last names and holding managerial positions. Together, the three authors published a paper refuting the original conclusion (Silberzahn, Simonsohn, & Uhlmann, 2014).

This experience led Silberzahn and Uhlmann, along with Dan Martic and Brian Nosek, to investigate how different teams may interpret similar datasets. To do so, they conducted an experiment in which 29 teams of analysts received the same dataset, which included a number of variables pertaining to soccer players and referees. The teams were asked whether the data support the claim that dark-skinned soccer players are more likely to receive red cards. An analysis of the results showed that different teams took different approaches to analyzing the data, using different variables from the dataset, with 20 teams finding a statistically significant correlation between skin tone and the likelihood of receiving a red card, while 9 teams found no significant correlation (Silberzahn et al, 2017).

The results of this experiment show the power of subjectivity, in the form of which variables researchers choose to use and what statistical analysis they conduct, on the way data is interpreted and therefore on the final outcome of a study. Had any one of the teams' results been published, it would have been cited and become part of the scientific literature. Yet with multiple different conclusions from the same dataset—one is left wondering which conclusion is the "right" one. Moreover, this study illustrates how data interpretation adds an additional layer to the problem of reproducibility.

Data Sharing and Protocol Availability

In a 1675 letter to Robert Hooke, Isaac Newton famously declared "If I have seen further it is by standing on the shoulders of Giants." As a field where progress is made by building on previous knowledge, science depends on the sharing of information. The practice of transparency in science cannot be limited to the sharing of results, but must also include the necessary tools to allow authentication and replication of results. It is therefore paramount that scientists publish the datasets used to obtain their results, along with protocols describing how their research was conducted. Doing so benefits the scientific community in multiple ways. It allows for datasets to be fully explored and analyzed, while also increasing the likelihood of detecting errors in the original study, thereby reducing the likelihood of scientific fraud, and making science more efficient and overall, more trustworthy (Duke & Porter, 2013; Vanpaemel et al, 2015).

Yet despite the obvious importance of sharing data and protocols, it is not ubiquitously practiced. Take for example the field of psychology, where reproducibility is notoriously difficult (Pashler & Wagenmakers, 2012). Despite a 2015 report in which a team of psychologists failed to reproduce 61 of 100 psychology studies (Baker, 2015 April), researchers remain reluctant to share their data. When Vanpaemel et al. (2015) contacted the authors of 394 papers published in APA journals during 2012 with a request to share their data, only 38% of researchers did so. Other fields suffer from similar reluctance as well. In 2002, 47% of geneticists who had tried to receive additional data on published articles, reported having been denied access at least once in the preceding 3 years (Campbell et al, 2002). More recently, a study looking at the availability of raw data for published biomedical articles found that none of the 268 articles investigated in the study had provided full access to their data (Iqbal 2016).

This prevalent lack of data sharing across disciplines raises the obvious question of why it is the case. A number of factors have been identified in the literature.

While funders like the NIH and NSF require that investigators publish and share the data from projects they funded, other organizations—such as military agencies or private pharmaceutical companies—may, in fact, prohibit researchers from making public any data obtained through their funding (Kim & Stanton, 2016). In cases where the funding institution neither requires nor prohibits the publication of raw data, it is largely up to the researchers to decide whether to share it. The fact that data collection can be a difficult and time-consuming process may make researchers reluctant to share their data before making sure they have gotten the most return for their efforts—that is, the most possible publication out of that dataset. Another factor that increases the likelihood of withholding data is the lack of a standardized method for storing and sharing data. This is especially a problem in fields such as engineering and ecology, where researchers report spending a significant amount of time organizing, uploading, and sharing their data (Kim et al, 2013). With fields such as genetics, where data is oftentimes a physical thing such as a reagent or a re-engineered organism, sharing of data can be an expensive ordeal. Lastly, for medical research, in particular, there is the added complexity of the data including both clinical and personal information, which—combined with the known geographical location of a study—may be sufficient information to unveil the identity of the research participants, adding another layer to the problem of sharing data (Grover, 2010).

When it comes to publishing complete protocols—a document which describes in detail every step of a proposed study, standardizes laboratory methods and therefore assists in successfully replicating the study—the situation is not much better. A 2016 study by Iqbal et al. found that of 268 randomly sampled biomedical journal articles, which contained empirical data and were therefore expected to include study protocols, only a single article had a full protocol. A separate study looking into randomized trials funded by the Canadian Institute of Health Research found that in 40% of the investigated papers, major discrepancies in primary outcomes existed between the protocol and publication. One possible explanation, the researchers suggest, is that formal changes were made to the protocol, but were not updated in the version that was made public (Chan et al, 2004). The lack of a clear and full protocol for many studies, combined with the possibility that even for studies where a protocol is available, it may not be a true description of the study, make replication challenging.

Fraud

Though not as significant a reason for the crisis of reproducibility, scientific misconduct is nonetheless a contributing factor. Fang et al., who in 2012 conducted a review of all biomedical and life-science research articles indexed by PubMed as retracted, claimed that 43.4% of the 2,047 identified papers were retracted due to suspected or confirmed fraud (Fang, Steen, & Casadevall, 2012). Fraudulent behavior includes fabrication of results—that is, reporting results that were not obtained—and falsification—which refers to the manipulation of different aspects of the research, including its results. Engagement in fraudulent behavior by scientists contributes to the problem of irreproducibility on multiple levels. First, the results of such studies cannot be reproduced, as they were either falsified or fabricated. Second, scientific studies which were derived from fraudulent behavior are cited by other studies, leading to propagation of faulty and irreproducible data in the scientific literature. Moreover, results of papers derived from scientific misconduct continue to be cited even after the original article had been retracted (Bornemann-Cimenti, Szilagyi, & Sandner-Kiesling, 2016).

WHAT CAN BE DONE

Perhaps ironically, one way to combat the factors that contribute to the crisis of reproducibility is more reproducibility. In theory, more reproduction attempts would increase the likelihood of identifying irreproducible results, thereby fulfilling its role as a mechanism for self-correction. However, without proper incentives to encourage pursuit and publication of negative results, and a reduced emphasis on novelty over quality, this is unlikely to happen.

Many of the proposed solutions address individual causes of irreproducibility. For example, journals devoted to the publication of negative results (such as Journal of Negative Results) had been suggested as a way to address publication biases that lead to overwhelmingly positive results in the literature. However, such journals are likely to fail for a number of reasons. First, it does not address the current incentives driving researchers to publish in prestigious, high-impact journals, and since negative-result journals are unlikely to become prestigious or high-impact, scientists are unlikely to seek publications in such journals. Second, it had been shown that the relative scarcity of negative results in the literature is likely a function of failure to submit such studies for publication, and not a selection bias by journals. Another proposed strategy, meant to address quality control issues, is that the NIH begin requiring cell-line authentication to be conducted when submitting grant applications that involve the use of cell-lines

(Neimark, 2015) reasonable solution, it may be hindered by the current lack of cheap, fast authentication tools, and does not address the larger issues affecting reproducibility.

A possible strategy to tackle the bigger picture issues is requiring any publication claiming new results to include replication of their findings (Begley & Ellis, 2012). Such a requirement could work in cases where data collecting is relatively easy, yet it would be nearly impossible for large-scale studies where data collection could take years. Furthermore, it has the potential of deterring researchers from taking risks or seeking innovation due to the trouble involved with replicating their findings (Nosek, Spies, & Motyl, 2012).

An effective solution should address not just specific causes of the problem, but the core of it as well. At its core, the problem of irreproducibility—whether it's due to biases, bad statistics, fraud, unavailability of complete data or protocol, data interpretation, or poor quality control—stems from the current system which favors innovation and extreme results over careful analysis. Increased transparency had been suggested as a way to solve this problem. In the short term, transparency would help eliminate the individual causes responsible for the crisis of reproducibility. More importantly, in the long term, it will help shift the focus of science from personal achievements (in the form of groundbreaking publications in high-impact journals) to collaborative knowledge accumulation.

One way to achieve transparency is through “Open Science.” The idea behind open science is that scientists would share their complete methods, data, and results in easily and freely accessible databases (Nuzzo, 2015). Making complete data and methods public would make data-sharing simpler, and allow other researchers to replicate studies and verify the results. It would also allow other scientists to identify potential errors in experimental design or data interpretation, and could serve as a tool to “outsource” data interpretation in order to avoid the issues discussed earlier. Increased transparency would also serve to deter scientists from engaging in fraudulent behavior.

To further eliminate biases, researchers would be required to publish a pre-study document detailing their proposed research, what data—and how much of it—will be collected, what relationships will be investigated, and what statistical methods will be utilized (Dickersin, 2015). Such a pre-study document will prevent publication and outcome reporting bias, as researchers will be expected to publish the results of all conducted investigations, as reported in the pre-study document, regardless of the direction or strength of the results. It will also prevent sample-size manipulation.

CONCLUSION

The crisis of reproducibility in science is widespread across multiple fields. The reasons for the crisis are numerous, yet they all stem from the current scientific culture of promoting competitiveness over collaboration. In this landscape, novelty and positive results are valued more than the pursuit of additional evidence for or against already published studies; researchers are rewarded for their results, not the quality of their research, and are therefore more likely to engage in questionable practices (such as statistical manipulations, foregoing quality control, or fraud); there is no incentive to share data among scientists. This, in turn, leads to decreased trust in scientific claims and wasting of valuable resources.

Attempts to fix the problem of reproducibility, and by extension the problem of “bad science,” should focus on increasing transparency. In fact, a shift towards more open science is already underway. The Center for Open Science was launched in 2013 with a mission to “increase openness, integrity, and reproducibility of research” (https://cos.io/). Projects such as Dataverse (https://dataverse.org) and Dryad (https://datadryad.org) are offering open databases for sharing and analyzing data. Journals are increasingly adopting policies to encourage open science (Munafò et al, 2017). On a legislative scale, attempts to promote open science include a 2016 document drafted by Dutch scientists, The Amsterdam Call for Action on Open Science, which calls for full and open access for all scientific publications, and promotes an environment where “data sharing and stewardship is the default approach for all publically funded research” (Vollmer, 2016).

This current movement towards an open science is promising and has the potential to make science more collaborative, replicable, and transparent.

PART II

INTRODUCTION

The second part of the thesis will focus on instances of research misconduct and their associated retraction notices. Specifically, it will present data on the fate of NIH-funded research publications that have been found to contain data derived from scientific misconduct, and investigate the nature of retraction notices associated with those publications.

Retractions on the Rise

Retractions of scientific papers are on the rise; in the decade spanning 2000 to 2009, there was a tenfold increase in the number of published retractions (Wager & Williams, 2011). According to *Retraction Watch*, there were a total of 684 retractions in the scientific literature in 2015 as compared to 467 in 2013. Notably, in the biomedical sciences, the percent increase in retractions is outpacing the increase in the total number of overall publications (McCook, 2016).

While there are various reasons for retracting a paper, most can be broadly categorized as attributable to unintentional errors or due to research misconduct. The latter category includes articles in which authors volitionally engaged in fabrication or falsification of data, or plagiarism. Just as is the case with retractions in general, retractions due to scientific misconduct are thought to be on the rise (Fang et al, 2012).

Though guidelines for retraction notices underscore the importance of full transparency (Kleinert, 2009), it remains unknown to what extent authors who engage in research misconduct inform readers that research misconduct has occurred. Indeed, individual journals may not have clear policies regarding what is an appropriate composition of a retraction notice, particularly in cases of misconduct. Moreover, there is the basic issue as to whether there is a retraction at all in these circumstances.

It is noteworthy that previous work in this area focused on published retraction notices, evaluating the percent where research misconduct was cited as a cause (Steen, 2011; Fang et al, 2012; Grieneisen & Zhang, 2012; Damineni et al, 2015). Since we cannot be certain that all papers associated with research misconduct are retracted or even whether authors admit to misconduct in published retractions, these types of analyses are limited. To understand the nature of retractions in this context, therefore, requires following the fate of papers documented to involve misconduct.

To begin to address this complex issue, we utilized the Office of Research Integrity database. This approach allowed us to identify a population of publications that resulted from research misconduct and, in turn, to assess their associated retraction notices. Specifically, we determined the presence or absence of a notice, its composition, and whether there were correlations with journal impact factor.

METHODS

Identification of Articles Involved in Research Misconduct

The Office of Research Integrity in the US Department of Health and Human Services maintains a website (https://ori.hhs.gov/) that lists Case Summaries with names of investigators who engaged in documented research misconduct on projects supported by NIH funding. The outcomes of investigations into these cases along with the titles of associated published articles that contain data obtained by research misconduct are detailed. For some investigations, Case Summaries were published in a newsletter also available at the website. All such listed published articles for a 10-year period (2007-2017) formed the basis of this investigation.

These articles were searched on PubMed to determine if there were published retractions and to assess the quality of any published retraction through a scoring system that was developed. In addition, the impact factor (IF) of the journal was recorded. Impact factors were obtained from the Web of Science from the year in which the article was retracted, except in one case where the article was retracted after the original journal ceased publication. In this case, the most recent available IF was used. If there was no retraction, the IF of the year in which the ORI published its misconduct findings was used.

Assessment and Scoring of Retraction Notices

We drew upon the retraction guidelines established by the Committee on Publication Ethics (COPE) and the International Committee of Medical Journal Editors to identify four elements of an optimal retraction notice. These include: (1) identity of the author who initiated the retraction; (2) use of unambiguous language that communicates the reason for retraction is research misconduct; (3) identity of the author responsible for the misconduct; and (4) identification of data/figures that are not valid and led to the retraction. Notices received a point for of each of these elements. In addition, a single point was given for the mere publication of a retraction notice, independent of whether they contained any of these elements. Scores thus ranged from 0 to 5 with 0 accorded if there was no published retraction and 5 for published retractions that contained all 4 elements (table 1). So-called corrections, expressions of concern, or paper withdrawal were scored 0.

Statistical Analysis

All data were analyzed using SAS Studio software, version 3.71 (Cary, NC). We calculated total scores for each incident of research misconduct based on our retraction scoring system, described by mean and standard deviation. We assessed for an association between journal impact factor and mean retraction score using ANOVA. In subanalyses, we assessed only incidents of research misconduct for which no retraction was issued, using Fisher's exact test to determine an association between journal impact factor and whether a retraction was issued. A two-sided alpha &lt;0.05 was consider the threshold for statistical significance

Table 1. Retraction Notice Scoring System. Summary of scoring system used to evaluate and score retraction notices of paper retracted due to scientific misconduct. Retraction notices were awarded either 0 or 1 points based on 5 criteria, for a total score ranging from 0 to 5.

		Points Awarded	
		1	0
Scoring Criteria	Retraction notice present?	Yes	No
	Author initiated retraction?	Yes	No
	Does retraction notice state findings of misconduct?	Yes	No
	Is responsible author identified?	Yes	No
	Does notice specify affected data/figures?	Yes	No

RESULTS

Using this database, we identified 200 papers that contained data derived from documented misconduct. Of those papers, 42/200 (21%) were not retracted as of November 2017. The 42 unretracted papers include 9 instances for which a correction was published, 1 expression of concern, and 2 withdrawals with no retraction notices. The remaining 30 papers, representing 15% of all identified papers, were not retracted and had no associated notices.

The median retraction score of all identified papers was 3 with a mean of 2.62 +/-1.68. When excluding unretracted papers (score of 0), the median and mean scores were 3 and 3.32 +/-1.11, respectively. The score distribution is shown in figure 1.

With regard to scoring elements, 67% (109/158) of published retraction notices specified which data/figures were derived by misconduct; 63.9% (101/158) cited misconduct as the reason for retraction; 52.5% (83/158) were initiated by the authors; and 46.8% (73/158) identified the author responsible for the misconduct (figure 2).

In the 200 papers identified in this study, there were a total of 65 authors identified by the ORI as guilty of research misconduct. Twenty-four of those individuals authored papers which were not retracted. One author was responsible for 10 unretracted papers, while the rest of the 32 unretracted papers were distributed among 23 different authors.

Publication dates for the unretracted papers range from March 1993 to April 2016. In total, 3 papers were published prior to 2000, 25 between 2000-2009, and 14 on or after 2010. The dates on which the ORI concluded its research misconduct investigations that encompassed these 42 papers were distributed as follows: 8 prior to 2010, 28 during 2010-2015, and 6 after 2015. Notably, for 2 of the unretracted papers the ORI published its findings of research misconduct in 2017.

Journal impact factors for papers identified in our study range from 0.38 to 53.3. The average impact factor was 7.46, with a median of 4.73. No significant association was found between impact factor and retraction score. A trend toward journals with higher impact factor being more likely to retract a paper was observed, though the sample size may be too small to confidently determine.

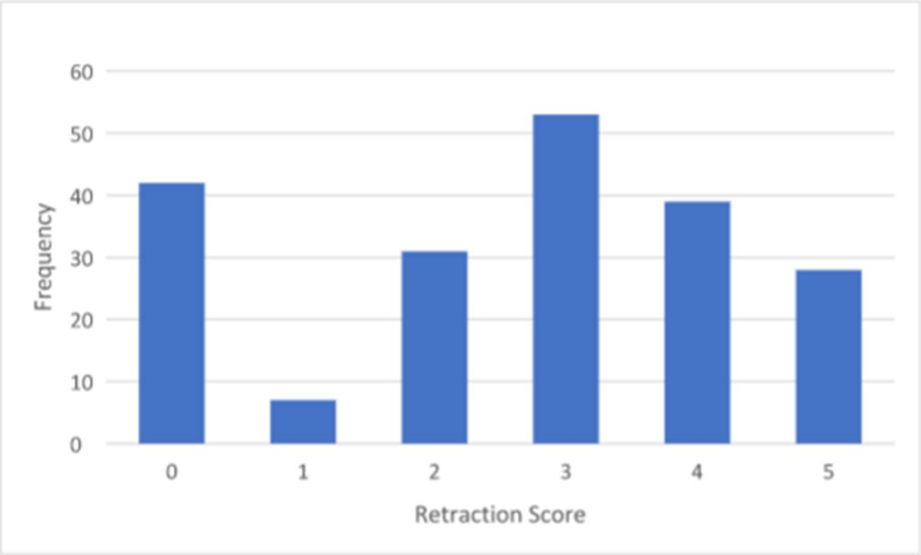

Figure 1. Distribution of Retraction Notice Scores. Total scores for retraction notices were distributed as follows: 42 retraction notices received a score of 0; 7 received score of 1; 31 received score of 2; 53 received score of 3; 39 received score of 4; 28 received score of 5.

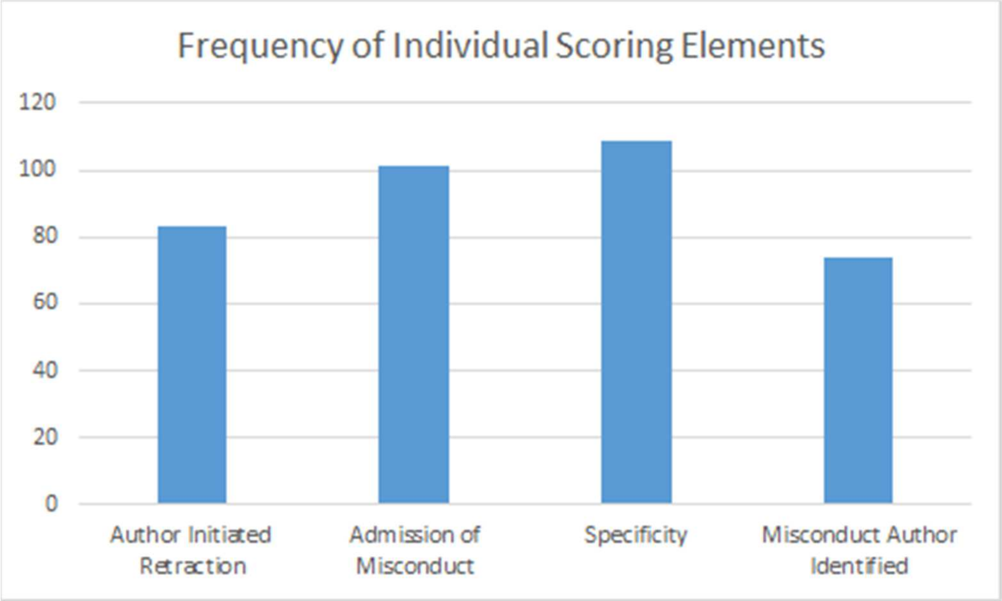

Figure 2. Frequency of Individual Scoring Elements. Some scoring criteria were more likely to be present in retraction notices. For each scoring element, graph shows the total number of retraction notices satisfying that criterion.

DISCUSSION

Summary of Findings

In order to examine the status and quality of retraction notices in papers associated with research misconduct, we took advantage of information available at the ORI website. We found that 21% of such papers have not been retracted with an associated retraction notice at the time this study was conducted. Importantly, 39/42 of unretracted papers identified were published prior to 2015. In this regard, it is estimated that, for articles published after 2002, the average time span between publication to retraction is just under 2 years (Casadevall & Fang, 2013). Based on this, we believe that the lack of retraction notices is not explainable by the timing of publication.

Of the retracted papers that did have an associated retraction notice (158/200), the most frequently present scoring element (67%) was information detailing which figure(s) or data were incorrect. The least common element (46.8%) was information regarding the identification of the author who engaged in research misconduct. Notably, only 63.9% of retraction notices stated in unambiguous language that the reason for retraction was misconduct. In our study, we did not find any statistically significant relationship between overall and individual scoring elements and impact factor.

Limitations

Given that 6/42 of unretracted papers identified were from ORI investigations that concluded on or after 2015, it is possible that in the future more of those papers will be retracted. Another limitation of our study is the relatively small sample size.

Implications

In order to correct the scientific record and restore trust, it is essential that retraction notices be transparent and forthcoming, containing all essential elements. Though the information presented in retraction notices depends, in part, on individual journals' policies, the Committee on Publication Ethics (COPE) provides guidelines for writing such notices. While COPE has over 12,000 member journals, our findings indicate that for 36% of published retractions, there was no mention of research misconduct, demonstrating poor adherence to COPE guidelines. A possible solution, which had been previously proposed (Moylan & Kowalczuk, 2016), is the adoption of a checklist containing key elements identified by COPE that authors have to provide when preparing retraction notices.

To ensure that articles identified as containing data derived from research misconduct are retracted in a timely manner, we suggest that submission of retraction notices be part of the final process that concludes an investigation. If authors refuse, this information should be conveyed and published in the journal. One possibility is for journals to provide links to ORI published investigations.

Retracting papers due to scientific misconduct and publication of an appropriate Collaboration between authors, funding agencies, research institutions, and journals, as well as the establishment of clear adherent guidelines for dealing with the process of retraction and publication of retraction notice are essential. This would help speed the process of retraction, improve transparency, enhance a culture of integrity and rigor, and improve scientific credibility amongst the lay public.

REFERENCES

Aarts, A.A., Anderson, J.E., Anderson, C.J., Attridge, P.R., Attwood, A.,,…(2015). *PSYCHOLOGY*. Estimating the reproducibility of psychological science. Science (New York, N.Y.), 349(6251)

An update on data reporting standards. (2014). *Nature Cell Biology, 16*(5), 385.

Baker, M. (2015, April). First results from psychology's largest reproducibility test. *Nature,* Nature, 4/30/2015.

Baker, M. (2015). Over half of psychology studies fail reproducibility test. *Nature,* Nature, 8/27/2015.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature,533*(7604), 452-4.

Baker, M. (2017). Reproducibility: Check your chemistry. *Nature, 548*(7668), 485-488.

Begley, C.G., & Ellis, L.M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature,483*(7391), 531-3.

Bollen, K. (2015). Reproducibility, Replicability, and Generalization in the Social, Behavioral, and Economic Sciences. Retrieved from https://www.nsf.gov/sbe/SBE_Spring_2015_AC_Meeting_Presentations/Bollen_Report_on_Replicability_SubcommitteeMay_2015.pdf

Bornemann-Cimenti, H., Szilagyi, I., & Sandner-Kiesling, S. (2016). Perpetuation of Retracted Publications Using the Example of the Scott S. Reuben Case: Incidences, Reasons and Possible Improvements. *Science and Engineering Ethics, 22*(4), 1063-1072.

Button, K.S., Ioannidis. J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson E.S.J., Munafò, M.R. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience*, Vol.14(5), p.365

Campbell, E., Clarridge, B., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N., & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. *JAMA, 287*(4), 473-80.

Casadevall, A., & Fang, F. (2013). Why Has the Number of Scientific Retractions Increased? *PLoS One*, 8(7), E68397.

Chan, A., Krleza-Jerić, K., Schmid, I., & Altman, D. (2004). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ : Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne, 171*(7), 735-40.

Collmer, T. (2016, May 3). EU pushing ahead in support of open science. [blog post]. Retrieved from https://creativecommons.org/2016/05/03/europe-moving-right-direction-support-open-science/

Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., Mccarren, P., Hirschman, J.E., . . Golub, R.T. (2017). The Drug Repurposing Hub: A next-generation drug library and information resource. *Nature Medicine, 23*(4), 405-408.

Damineni, R. S., Sardiwal, K. K., Waghle, S. R., & Dakshyani, M. . (2015). A comprehensive comparative analysis of articles retracted in 2012 and 2013 from the scholarly literature. *Journal of International Society of Preventive & Community Dentistry*, 5(1), 19–23

Dickersin, K. (2005) Publication Bias: Recognizing the Problem, Understanding Its Origins and Scope, and Preventing Harm, in Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments (eds H. R. Rothstein, A. J. Sutton and M. Borenstein), John Wiley & Sons, Ltd, Chichester, UK.
doi: 10.1002/0470870168.ch2

Dickersin, K., Chan, S., Chalmersx, T.C., Sacks, H.S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8*(4), 343-353.
Duke, C.S., Porter, J.H. (2013) The Ethics of Data Sharing and Reuse in Biology. *BioScience, 63*(6), 483-489.

Dumas-Mallet, E., Button, K., Boraud, T., Gonon, F., & Munafò, M. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science, 4*(2), 160254.

Dwan, K., Altman, D., Arnaiz, J., Bloom, J., Chan, A., Cronin, E., . . . Siegfried, N. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias (Publication and Reporting Bias). *PLoS ONE, 3*(8), E3081.

Fanelli, D., Costas, R., & Ioannidis, J. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America, 114*(14), 3714-3719.

Fang, F., Steen, R., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), 17028-33.

Freedman, L., Gibson, M., Ethier, S., Soule, H., Neve, R., & Reid, Y. (2015). Reproducibility: Changing the policies and culture of cell line authentication. *Nature Methods, 12*(6), 493-7.

Goodman, S., Fanelli, D., & Ioannidis, J. (2016). What does research reproducibility mean? *Science Translational Medicine, 8*(341), 341ps12.

Grieneisen M. L., Zhang, M. (2012) A Comprehensive Survey of Retracted Articles from the Scholarly Literature. *PLoS ONE,* 7(10): e44118

Groves, T. (2010). The wider concept of data sharing: View from the BMJ. *Biostatistics, 11*(3), 391-392.

Ioannidis, J. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA: The Journal of the American Medical Association, 294*(2), 218-28.

Ioannidis, J., & Trikalinos. T.A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*(6), 543-549.

Iqbal S.A., Wallach J.D., Khoury M.J., Schully S.D., Ioannidis J.P. (2016). Reproducible Research Practices and Transparency across the Biomedical Literature. *PLoS Biology, 14*(1), E1002333.

Irfan Maqsood, M., Matin, M., Bahrami, A., & Ghasroldasht, M. (2013). Immortality of cell lines: Challenges and advantages of establishment. *Cell Biology International, 37*(10), 1038-1045.

Jannot, A., Agoritsas, T., Gayet-Ageron, A., & Perneger, T.V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology, 66*(3), 296-301.

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*(5), 524-532.

Kim, Y., & Stanton, J. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology, 67*(4), 776-799.

Kim, Y., Zhang, Ping, Bellini, James, Crowston, Kevin, Driscoll, Charles, Moraresco, Paul, . . . Stanton, Jeffrey. (2013). *Institutional and Individual Influences on Scientists' Data Sharing Behaviors,* ProQuest Dissertations and Theses.

Kleinert, S. (2009). Committee on Publication Ethics (COPE). COPE's retraction guidelines. *Lancet,* 374: 1876–77.

Kyzas, P., Denaxa-Kyza, D., & Ioannidis, J. (2007). Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer (Oxford, England : 1990), 43*(17), 2559-79.

Levinson, N., Boxer, S., & Ramchandran, R. (2012). Structural and Spectroscopic Analysis of the Kinase Inhibitor Bosutinib and an Isomer of Bosutinib Binding to the Abl Tyrosine Kinase Domain (Structure of Bosutinib Bound to Abl). *PLoS ONE, 7*(4), E29828.

Lexchin, J., Bero, L., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ (Clinical Research Ed.), 326*(7400), 1167-70.

Maccoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature,526*(7572), 187-9.

Masters, J.R. (2012). Cell-line authentication: End the scandal of false cell lines. *Nature,492*(7428), 186.

McCook, A. (2015, December 8). Hundreds of researchers are using the wrong cells. That's a major problem [Blog post]. Retrieved from https://retractionwatch.com/2015/12/08/hela-is-the-tip-of-the-contamination-iceberg-guest-post-from-cell-culture-scientist/

Moylan, E., & Kowalczuk, M. (2016). Why articles are retracted: A retrospective cross-sectional study of retraction notices at BioMed Central. *BMJ Open*, 6(11), E012047.

Munafò M.R., Nosek B.A., Bishop D.V., et al. (2017): A manifesto for reproducible science. *Nature Human Behavior*; 1: 0021. 10.1038/s41562-016-002
Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A., Tschannen, B., Altman, D., . . . Jüni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: Meta-epidemiological study. *BMJ (Clinical Research Ed.), 341*, C3515.

Nardone, R.M. (2008). Curbing rampant cross-contamination and misidentification of cell lines. *Biotechniques* 45, 221–227.

Neimark, J. (2015). Line of attack. *Science (New York, N.Y.), 347*(6225), 938-40.

Nosek, B., Spies, J., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability.

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature, 526*(7572), 182-5.

Olson, C., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J., . . . Pace, B. (2002). Publication bias in editorial decision making. *JAMA: The Journal of the American Medical Association, 287*(21), 2825-8.

Oransky, I., Marcus, A., (2016). What does scientific reproducibility mean, anyway? Retrieved from https://www.statnews.com/2016/06/01/reproducibility-science/

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science. *Perspectives on Psychological Science, 7*(6), 528-530.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*(9), 712.

Riveros, C., Dechartres, A., Perrodeau, E., Haneef, R., Boutron, I., Ravaud, P., & Dickersin, K. (2013). Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals (Results at ClinicalTrials.gov and in Journals). *10*(12), E1001566.

Rosenthal, R., & Hernstein, R .J. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437.

Schwarzer G., Carpenter J.R., Rücker G. (2015) Small-Study Effects in Meta-Analysis. In: Meta-Analysis with R. Use R!. Springer, Cham

Shapin, S., Schaffer, Simon, & Hobbes, Thomas. (1985). Leviathan and the air-pump : Hobbes, Boyle, and the experimental life : Including a translation of Thomas Hobbes, Dialogus physicus de natura aeris by Simon Schaffer. Princeton, N.J.: Princeton University Press.

Siebert, S., Machesky, L., & Insall, R. (2015). Overflow in science and its implications for trust. *ELife, 4*, ELife, 14 September 2015, Vol.4.

Silberzahn, R., & Uhlmann, E. (2013). It pays to be Herr Kaiser: Germans with noble-sounding surnames more often work as managers than as employees. *Psychological Science, 24*(12), 2437-44.

Silberzahn, R., Simonsohn, U., Uhlmann, E. L. 2014. Matched names analysis reveals no evidence of name meaning effects: A collaborative commentary on Silberzahn and Uhlmann (2013). *Psychological Science*, 25: 1504-1505

Silberzahn, R., & Uhlmann, E. (2015). Crowdsourced research: Many hands make tight work. *Nature, 526*(7572), 189-91.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F.,... Nosek, B. A. (2017, September 21). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. Retrieved from psyarxiv.com/qkwst

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-66.

Song, F., Parekh-Bhurke, S., Hooper, L., Loke, Y., Ryder, J., Sutton, A., . . . Harvey, I. (2009). Extent of publication bias in different categories of research cohorts: A meta-analysis of empirical studies. *BMC Medical Research Methodology*, 9, 79.

Stachowiak, R., Jagielski, T., Roeske, K., Osińska, O., Gunerka, P., Wiśniewski, J., & Bielecki, J. (2015). Lmo0171, a Novel Internalin-Like Protein, Determines Cell Morphology of Listeria monocytogenes and Its Ability to Invade Human Cell Lines. *Current Microbiology*, 70(2), 267-274.

Steen, R. G. (2011) Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of Medical Ethics,* 2011;37:249-253.

Sterling, T. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa. *Journal of the American Statistical Association, 54*(285), 30-34.

Suñé, P., Suñé, J., P., & Montoro, J., B. (2013) Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. *PLoS ONE,8*(1), E54583.

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra, 1*(1), Collabra, 01 October 2015, Vol.1(1).

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. *Journal of Medical Ethics*, 37(9), 567-70.

Yaffe, M. (2015). Reproducibility in science. *Science Signaling, 8*(371), Eg5.

CURRICULUM VITAE

[redacted]