Spring 2019

The reproducibility crisis in scientific research

Sarah Eline

The Reproducibility Crisis in Scientific Research

Sarah Eline

Senior Honors Thesis

While evidence-based medicine has its origins well before the 19th century, it was not until the beginning of the 1990s that it began dominating the field of science. Evidence-based practice is defined as "the conscientious and judicious use of current best evidence in conjunction with clinical expertise and patient values to guide health care decisions" (Titler, 2008, para. 3). In 1992, only two journal articles mentioned the phrase evidence-based medicine; however just five years later, that number rose to over 1000. In a very short period of time, evidence-based medicine had evolved to become synonymous with the practices that encompassed the medical field (Sackett, 1996). With evidence-based medicine came a decline in qualitative research and a shift towards quantitative research. This shift changed the focus from primarily exploratory research to a type of research that involves systematic empirical investigation through the use of statistics, mathematics, and computational techniques (DeFranzo, 2011). With the introduction of computers and online databases came an increase in quantitative research and the use of statistics, which allowed for evidence-based medicine to grow exponentially in the early 1990s (Zimerman, 2013).

The push for evidence-based practices (EBP) in all fields - not just specifically in medicine - has directly led to the proliferation of both research journals and journal articles. In 2012, there were approximately 28,100 active scholarly peer-reviewed journals publishing 1.8 to 1.9 million articles a year, and that number continues to grow (Rallison, 2015). The United States National Library of Medicine's premier bibliographic database, MEDLINE, shows statistics for the number of citations by year of publication. The total number of citations in the United States have almost quadrupled since 1970, from 64,161 to 278,341 in 2016 (National Institutes of Health [NIH], 2017). This exponential growth of scientific papers makes it increasingly more difficult for researchers and medical professionals to keep track of all that is

relevant to their fields and has created a growing concern about both the quality and quantity of the research being published.

The aforementioned concerns, which were sparked by the sheer number of published research, led to a multifield investigation regarding the reproducibility of published scientific literature. Initial questions were raised following a few landmark studies that did not hold up to replication by scientists in the field. In 2015, Brian Nosek, a social psychologist at the University of Virginia, and the head of the Center for Open Science set out with a group of researchers to conduct the biggest replication study of its kind (Baker, 2015). Nosek selected 100 original papers from three leading psychology journals to see if he and his research team could reach similar conclusions. The results were concerning: two-thirds of the replication results were so weak that they did not reach statistical significance (Open Science Collaboration, 2015). John Ioannidis, an epidemiologist and highly cited author of *Why Most Published Research Findings are False,* claimed that the true replication-failure rate could likely exceed 80%, since Brian Nosek's Reproducibility Project targeted only highly acclaimed journals (Baker, 2015). The results of this study opened up a conversation that spans much further than the field of psychology and reaches deep into every field of research.

The reproducibility crisis could be responsible, in part, for some of the conflicting headlines appearing in health-related news. Headlines often contain mixed information in the field of nutritional research. Articles have touted that a glass of red wine can help prevent cancer, chocolate can help with concentration, and even that coffee shows signs of cancer-fighting agents. However, the following week, there is yet another headline stating just the opposite. Therefore, it leaves the public wondering what evidence they should and should not accept. A quick Google search about red wine and the following results read: "red wine is bad for you, says

experts," or "an extra glass of wine a day will shorten your life by 30 minutes," followed by, "mounting evidence shows red wine antioxidants kill cancer." Schoenfeld and Ioannidis (2013) found that most food ingredients they studied (72%) were associated with both cancer risks and benefits. These false articles often influence dietary guidelines, are used to help shape public health policy, and influences individuals to change their lifestyle habits. With all of this mixed information available, it is challenging to make definitive conclusions often needed for evidence-based practice.

A further examination of the health field revealed numerous alarming discrepancies in biomedical research, which potentially affects new drug development and clinical treatments. According to Glenn Begley, the former head of oncology research at Amgen, "the ability to translate cancer research into successful clinical application has historically been very low especially in the field of oncology" (Begley & Ellis, 2012, para. 2). The quality of published preclinical data is one of the most significant contributions to the failure in oncology trials (Begley & Ellis, 2012). Over the years, the lack of quality data became more evident and led Begley and Ellis (2012) to conduct a reproduction study on 53 landmark papers. Begley and Ellis (2012) brought in the original researchers to help control for technical differences in research; however, the results showed that only six of the 53 (11%) were able to reproduce their own results. Begley is not the only one to come to this shocking conclusion. The German drug company Bayer reported that only 25% of published preclinical studies could be validated to allow the projects to continue (Begley & Ellis, 2012). This inability of the biomedical industry and clinical trials to validate the majority of results points to the existence of a systematic problem in the field that could potentially affect both the safety and effectiveness of clinical treatments and drugs on the market today.

The aforementioned concerns became even more evident when the present editor-in-chief of one of the most prestigious journals *The Lancet*, publicly stated, “much of the scientific literature, perhaps half, may simply be untrue” (Horton, 2015, para. 2). This quote is troubling given that these studies are being used to develop drugs and vaccines, as well as help to train medical staff and educate medical students (Walia, 2015). Dr. Marcia Angell, a physician and past editor of the prestigious *New England Medical Journal* stated,

> it is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgement of trusted physicians or authoritative medical guidelines. I take no pleasure in this conclusion, which I reached slowly and reluctantly over my two decades as an editor of the New England Medical Journal (Walia, 2015, para. 3)

These bold testimonies from highly respected individuals have brought these issues to the forefront of research journals and solidified the need for vast change in many aspects of scientific research.

The lack of reproducibility hinders progress and threatens the reputation of science as a whole. In biomedical research, small changes in conditions or natural variability in biological systems can cause a study to be irreproducible. However, the scale by which studies are unable to be replicated is concerning and has an effect on the translation of studies into clinical practices, not to mention the wasting of valuable resources and tax dollars. As this issue came to the forefront, it became clear that scientists needed to come together to address this topic and formulate possible solutions. In 2015, the Academy of Medical Science, Wellcome Trust, Medical Research Council and Biotechnology, and Biological Sciences Council (BBSRC) held a symposium surrounding the issue of reproducibility in their fields in hopes to find causes and solutions and work to regain the public's trust in scientific research. The overall causes discussed

in this symposium and other literature surrounding this topic include statistical error, culture surrounding research, funding corruption and bias, and lack of data sharing and transparency (Academy of Medical Sciences, 2015).

Statistical Error

Ioannidis (2005) explained how statistical error and lack of statistical knowledge contributes to much of the non-reproducible scientific literature. The high rate of non-replication in research can be attributed to the convenient strategy of using P-values as the main conclusive measure of statistical significance. Statistical significance is one the most influential metrics in determining if a study is published in a scientific research journal (Ioannidis, 2005). In order for a result to be statistically significant in most fields of science, the p-value must fall below 0.05. If this is the case, by definition, scientists can declare their result statistically significant (Resnick, 2017). The concept of statistical significance and p-values can be a hard to grasp, even for some scientists conducting research. To fully understand the impact that statistical error has on the reproducibility crisis, a brief explanation is necessary.

To begin, on the one hand, the experimental or alternative hypothesis is one that states there is a statistically significant relationship between two variables (Gonzalez, 2019). On the other hand, the null hypothesis is essentially the devil's advocate argument (Resnick, 2017). The null hypothesis states that there is no difference between the two variables, and this is what the researcher is trying to prove wrong in order to accept the experimental hypothesis and draw conclusions (Gonzalez, 2019). For example, rejecting the null helps researchers to understand the rarity of their results, and allows them to draw the conclusion that their alternative hypothesis could be true. In other words, the p-value helps to quantify the rareness of the results. If the researcher has a low p-value, it means that the data would rarely occur just by chance alone

(Resnick, 2017). However, it is important to note that researchers cannot rule out the null hypothesis indefinitely, which is the reason a threshold of 0.05 is set in place (Harris, 2017).

A significant area of confusion lies in the meaning of getting a p-value of less than 0.05. The biggest misconception is that a $P<0.05$ means that there is a 95 percent chance that the finding is correct, and a five percent chance that it is wrong (Harris, 2017). Instead, achieving a $P<0.05$ means that if one assumes the null hypothesis is true and the experiment was conducted 100 times, one would see the same results only five times (Resnick, 2017). There is a vast difference between the two definitions, and that is what often leads to the inflated trust placed in this value. Johnson (2013) proved this misconception wrong through advanced statistical techniques. There is a 25 to 30 percent chance that the null hypothesis is still true when the p-value is 0.05 (Johnson, 2013). This percentage is a very big difference from the five percent that many people, including some researchers, believe to be the truth.

P-values were not originally designed to be the end all be all for statistical significance. Ronald Fisher, who was the first influence on the idea of p-values, emphasized that experiments should be performed many times to see if the results hold up and the p-value remains below the threshold (Harris, 2017). Scientists of today have not held on to this advice and have instead begun to abuse the use of the p-value to prove their data to be true. Most studies just barely reach a p-value of 0.05, and this is often because studies are designed from the start to reach that exact mark. Consequently, p-hacking is being used by many researchers to reach statistical significance and get their findings published.

Uri Simonsohn, an economist at the University of Pennsylvania, became concerned with the growing number of findings that did not seem plausible in his field (Harris, 2017). Simonsohn and his colleagues set out to see how easy it would be to show that a finding was

true, when in fact it was not, and the results were concerning. Interestingly, it became clear how easy it is for scientists to look at their data and pull out bits that support their hypothesis and throw away ones that do not. Of particular concern part is that researchers can watch their data as they are being generated and the moment they reach statistical significance, they can stop the tests and declare their findings. This practice is extremely concerning because researchers are ignoring the fact that more data could result in a different outcome, but they are choosing to ignore that data in order to declare statistical significance (Harris, 2017). This process is called p-hacking, and it is plaguing the research field.

Brian Wansink, a researcher at Cornell University's Food and Brand Lab, is an example of a researcher who found himself in the spotlight due to statistical discrepancies in his research of which he claims to have not been aware of (Bartlett, 2017). Four papers that Wansink was co-author of were found to contain not just one or two statistical discrepancies, but approximately 150 inconsistencies. He was accused of recycling data from past experiments and was quoted telling a post-doctoral student that, "there's got to be something here we can salvage," when discussing their data (Bartlett, 2017, page or paragraph number of direct quote). This situation is a prime example of p-hacking, which shows that if one tampers with the data long enough, a finding may be revealed that looks significant, but in reality, is meaningless. A lot of people believe that researchers are often fooling themselves, and this was exactly the case for Brain Wansink. He said that he was unaware of the reproducibility crisis and the term p-hacking, until he found himself accused of such behavior (Bartlett, 2017).

As if p-hacking was not enough of a problem in the field, another popular term and practice that has been contributing to a lot of the statistical error in research is HARKing. This term edifies the process of creating a hypothesis after the results are known (Harris, 2017).

Clearly, this process is opposite of the scientific method and is extremely problematic. Scientists are taking data and running multiple tests, which they then use to make a hypothesis for which they already see a promising result. This practice is stepping over the line from confirmatory research to exploratory research, which is not what statistical tests were designed to do. While most research of this kind begins with good intentions, it is clear that scientists are abusing the tools of statistical analysis, and the results are misleading and inappropriate (Harris, 2017).

To pose a solution to this problem, Johnson and other scientists, including Ioannidis, believe that more stringent standards for p-values should be met. Johnson suggested that a $p<0.005$ should be the standard in the field. In this case, researchers can be confident that there is, in fact, a 95 percent chance that their findings will remain statistically significant if the study is run again (Harris, 2017). In essence, scientists hope that with a lower p-value, less false positives will be present in the literature, and it will require scientists to develop better research designs, increase sample sizes, and improve statistical techniques. However, this proposal was not met with unanimous support. Some scientists worry that this requirement will slow down the process and impede young doctoral students with limited budgets (Resnick, 2017). Ioannidis admitted that, "statistical significance alone doesn't convey much about the meaning, importance, clinical value, or the utility of research" (Resnick, 2017, page or paragraph number). However, Ioannidis purported that we live in a scientific culture that relies on p-values because of their quick and easy methods (Resnick, 2017).

An alteration in the threshold for p-values is the first line of business before other problems in statistics can be addressed (Resnick, 2017). Supporters of lowering the p-value also agree that findings that reach a p-value of 0.05 should still be considered; however, they should be classified as suggestive. Findings that reach 0.005 should be classified as statistically

significant. Findings that are suggestive are still important in advancing science; however, they should require further testing, and they should elicit some caution among readers when they are listed as breakthroughs in the scientific headlines (Resnick, 2017).

At the root of statistical error in science is the lack of a solid foundation of statistics among scientists and medical professionals. Casadevall believed that the reproducibility crisis begins in the way we train our scientists (Harris, 2017). According to Leonard Freedman's estimate, one in four irreproducible results in biomedicine are a result of analytical error. He believes that a significant reason for analytical error is the lack of statistics training among biomedical researchers. With the move towards evidence-based practice, and developments in technology, big data sets are the reality. Biology, which once was largely a descriptive science, is now dealing with large data sets and complicated analytical techniques (Harris, 2017). Consequently, it is clear that the education system needs to catch up and integrate more math and science courses into their curriculum.

Casadevall and Fang (2016) believed that an obvious step in strengthening rigor in scientific research is through proper training in experimental design, statistics, error analysis, logic, and ethics. However, currently, even statistics is not always a requirement in graduate school curriculums. Furthermore, statistics is not always a prerequisite course for medical school, and once in medical school, statistics education is minimal (Kaplan, 2019). This lack of in-depth statistics training is problematic, particularly considering the fact that medical professionals are expected to understand and educate their patients on the literature in the field. Therefore, it is necessary that medical professionals and researchers have the statistics knowledge that allows them to be critical about what they read in the literature in order to avoid blindly trusting the research.

Culture Surrounding Research

Through adjusting the scope and looking at the big picture of the reproducibility crisis, it becomes clear that there is a more overarching issue at hand, and that is the culture surrounding science as a whole. To begin, job security in the field of research is extremely scarce. A study conducted by the National Institutes of Health, using data from 2008, showed that only 21 percent of post-docs will receive a tenure-track job (Harris, 2017). While there may be variation in this data since 2008, it is clear that this shortage of spots results in a highly competitive environment in the field of research. The large pool of researchers and shortage of job opportunities forces hiring committees to base their decisions on researchers publishing history and, more importantly, on the impact factor of the journals in which researchers were published. An impact factor is essentially a score allotted to a journal based on how often its papers are cited in the literature (Schekman, 2013). The impact factor is critical because scientists rely on getting published in journals with high impact factors in order to receive subsequent grants, promotions, tenure, and to boost their reputation in the field (Resnick, 2017).

The focus on publishing is also the reality in academia. University professors rely on publishing, particularly in high impact factor journals, to boost their chances of receiving tenure, getting promoted, and receiving grant money (Harris, 2017). This practice has caused a dangerous mentality that is referred to as publish or perish. Resnick (2017) recounted a young scientist who shared, "I feel torn between asking questions that I know will lead to statistical significance and asking questions that matter" (p. 12). In order to keep funding and to progress their careers, researchers are stuck with the decision to choose projects that will likely succeed and produce positive results, rather than to ask difficult research questions that may result in greater progress. This duality exemplifies that the true root of the problem is not strictly

statistical error and the reliance on statistical significance, but instead may be the institution of science, which has incentivized the behaviors to allow it to fester in the field of research (Resnick, 2017).

Schekman (2013 believed one of the biggest reasons that the culture of science is to blame for the reproducibility crisis is because "people know what it takes to get their paper into one of these journals, and they will bend the truth to make it fit because their career is on the line" (Harris, 2017, p. 177). This pressure to publish specifically in journals with high impact factors lures scientists to cut corners and tempts them to focus on data that may help their research look better (Schekman, 2013). Rather than incentivizing quality research designs and proper statistical techniques, journals are creating a culture that incentivizes being first, even if that means being sloppy and cutting corners to get published (Harris, 2017). Journals with high impact factors know that highly cited research is usually flashy and eye-catching. Therefore, that is what journals tend to publish, even if it means it is wrong and will not stand the test of time. Therefore, in order to get published, researchers are having to adjust to that standard and choose research that makes bold claims, rather than encouraging replication studies, or meta-analyses, which have an even greater impact on the field (Schekman, 2013).

The problem is that journals put too much weight on single studies rather than incentivizing replication studies, meta-analysis, and even the publication of negative results. A study conducted by *PLOS* assembled a database of studies in biomedical science which included the initial study, follow-up studies, and meta-studies (Resnick, 2017). The *Dow Jones Factiva* newspaper was then searched to see how each type of study was covered. Ultimately, initial studies were five times more likely to be covered than follow-up studies, and meta-analysis reviews were rarely covered at all. The most concerning findings were that out of 1,475

newspaper articles, only 75 of the articles reported null findings (Resnick, 2017). Null findings can be just as impactful in the field of science; however, not only are they rarely reported, they are discouraged. This hesitancy is creating a culture that only incentivizes positive study results, and researchers' careers are dependent upon those findings. If a researcher has worked on a study for years and he/she comes to realize the study is not holding up as originally planned, the researcher is forced to either accept that time and money has been wasted, or it may lure the researcher to tweak his/her data to still get published. This practice is perpetuating the publish or perish mentality, and the institution of science needs to be better at rewarding failure (Resnick, 2017). Scientists should not fear the idea of failure, but instead should welcome it because failed studies can be just as impactful to advancing science. However, this change in mentality can only occur if the incentives for scientists are altered as well.

Scheckman (2013) believed one solution to the problem of incentives is a new breed of open-access journals. Allowing free access eliminates the need for promoting expensive subscriptions. Furthermore, open access journals should accept papers that meet quality standards and can be edited by working scientists who can assess the worth of the papers without regard for the number of citations as a measure of success. A current example of this is *eLife*, an open access journal funded by the Wellcome Trust, the Howard Hughes Medical Institute, and the Max Planck Society (Scheckman, 2013). *eLife* is a real-world example of a journal that is publishing credible, world-class science in a more accessible manner. Along with this change, however, funders and universities must also adjust. Grant funders and academia must begin to change their standards and stop basing positions and grants on the number of published papers, but instead base success on the quality of science (Scheckman, 2013). This shift in mindset

would help to change the incentives and alleviate the pressure to publish just for the sake of publishing.

Funding Corruption and Bias

While fraud in science is not a huge contributing factor to the reproducibility crisis, it does play a role in the culture of research. The federal Office of Research Integrity cites cases of scientific misconduct annually, such as the one at Rowan University School of Osteopathic Medicine, where an associate professor intentionally fabricated data that led to eight published papers and grants from the National Institute of Health. These stories rarely make the news, the researchers are rarely punished, and retraction statements are vague (Harris, 2017). Casadevall and Fang (2012) investigated published retractions and found that 70 percent of the retractions they found were due to misconduct rather than simply error. Retractions were also more common in high profile journals, which may be due, in part, to the need to publish in those journals to advance researchers' careers (Casadevall & Fang, 2012).

Fraud in the context of corporate influence over research has led to a great deal of the bias and mistrust in the field. Research is often done at the wishes of companies that, in turn, have a large financial stake in the results (Ioannidis, 2011). This financial stake is especially true of the pharmaceutical industry, which has essentially bought out the medical profession. A prime example occurred in 1999, when a new leading drug called Avandia came onto the market to help treat diabetes (Whoriskey, 2012). Follow-up studies resulted in some alarming conclusions that showed an increased risk of heart attacks from taking Avandia. As a result, the drug company GlaxoSmithKline was forced to release their original data from the study. Ultimately, it was determined that important data were omitted from the original analysis, and the company was accused of knowing about the increased risk of heart disease all along. Upon further

investigation, the financial connection between the drug maker and the researchers was alarming. The study was funded by GlaxoSmithKline, and the 11 authors received monetary compensation from the drug company. While it is not clear whether the financial association between the drug company and the research led to the report and data being altered, what is clear is that any form of financial connection increases bias (Whoriskey, 2012).

While retractions are common, the public is rarely informed when initial studies are disconfirmed (Resnick, 2017). Allison, a researcher at the University of Alabama, Birmingham, found out how difficult it can be to get retractions published, so that the public can be informed of disconfirmed studies (Harris, 2017). When Allison and his colleagues sent letters to journals pointing out clear mistakes, they were shocked when they were asked to pay up to $2,000 dollars to publish these letters and set the record straight on other researchers' studies (Harris, 2017). One of the reasons for this practice is that scientists hate to admit mistakes, and journals who publish them are the same. This reluctance to admit to mistakes is partly due to pride, but mostly due to the fact that error can put a black-mark on one’s record and impede a scientist's ability to advance his/her career, secure tenure, and receive subsequent funding (Harris, 2017). This lack of accountability is further evidence that a more fault-free system of admitting mistakes is necessary in the field. The culture of science needs to be altered to allow researchers to not feel as though their career is on the line if their study fails. This renewed culture would help foster a better research environment that could potentially reduce sloppy research and questionable practices in the field.

Lack of Data Sharing and Transparency

One overarching and powerful solution to reducing poor quality research designs, p-hacking, HARKING, and outright fraud in the field of research is to increase transparency of

research designs and data. One of the reasons statistical error and fraud are not often caught is because most published methods are either vague or deliberately kept secrete (Resnick, 2017). Good quality research that the public can trust requires disclosure and openness, and the best way to boost the reproducibility of studies is through pre-registration of research. This practice is one of the easiest ways to ensure that scientists are not able to fraudulently hand-pick data, change their hypothesis, or alter any part of their study (Walia, 2015). Nosek, who completed the aforementioned replication study in the field of psychology, created a platform that serves to accomplish this goal of transparency, called the Open Science Framework (Harris, 2017). OSF is acts as a data repository that invites scientists to register their hypothesis in advance to ensure that their results were, indeed, confirmatory (Harris, 2017). OSF provides scientists a platform for organizing their entire experiment through an interface where they can store all aspects of their research in a safe and open manner (Center for Open Science, n.d.). The biggest barrier of convincing scientists to switch over to this program is at the moment, there is not much of an incentive to share data, especially prior to the experiment being completed. With the competitive nature of research, it is difficult to convince scientists to share their data, given the chance that another researcher could take on their research and potentially get published first. This lack of data sharing reiterates the idea of the culture of science and the need to advance one's career over all else. While open data and methods do increase the risk of other scientists using one's data, open data also helps detect error prior to publishing, and allows for quicker progress of science through collaboration. The OSF program, in turn, helps to increase the trust that the researchers, medical professionals, and the public has in published scientific literature.

How to Avoid Poor Research

Following this discussion of the reproducibility crisis, the question remains about what medical professionals can do to best spot poor research and avoid basing clinical procedures or medical advice on false research. Ultimately, it is important that health care professionals are educated on the reproducibility issues facing scientific research, so that they can be more critical when it comes to reading the literature. Begley (2013) explained the best ways to recognize poor data in pre-clinical papers that will likely lead to nonreplicable studies that cannot be fully trusted. Begley (2013) suggested that every reader evaluating a research paper should ask the following six question: "were experiments being performed blind, were basic experiments repeated, were all results presented, were there positive and negative controls, were reagents validated, and were statistical tests appropriate" (Begley, 2013, p. 434).

Studies that are performed unblinded are inherently more biased in nature and should be read with caution. Replication of experiments throughout a study can be expensive and time-consuming; however, they provide evidence for the validity and reliability of a study. A study that does not disclose all data and results is lacking transparency and the ability for future researchers to validate their work is not possible. In addition, it is imperative that clear and detailed methods are presented in the literature, so that the study can be replicated by future researchers. Moreover, it is also important to note whether studies use positive and negative controls in their experiment and whether they disclose data on both (Begley, 2013). Additionally, it is critical that studies have validated their reagents by making sure they test their materials prior to the experiment. For instance, there is a lot of research that shows nonreplicable studies that were due to contaminated cell lines, which were believed to contain certain cancer cells, but turned out to be different cancer cells (Harris, 2017).

The final and potentially most important question to ask is regarding the researcher's statistical analysis. Appropriate statistical techniques should be decided before the onset of a study. Appropriate tests depend on many factors, including the type of research question being asked, the type of data collected, the number of groups in the study, and the number of data sets (Nayak & Hazra, 2011). To ensure that proper analysis and conclusions can be made, it is imperative that appropriate tests are performed.

With the sheer number of new studies published every week, it may be difficult for medical professional to fully dive deep into the methods and research design of every research study they come across. Therefore, there are quicker ways to spot questionable research. To begin, if a study is simplistic, universal, or definitive, it is a good idea to proceed with caution while reading. Any finding in the literature that is supported by only a single study should be further backed up by subsequent research, and until then should be considered questionable. Studies claiming to be groundbreaking or are being pushed by people or organizations that have a financial stake in the studies' success, should be considered biased at best (Freedman, 2010). Finally, medical professionals should be leery of small studies with small effect sizes. Small studies are more likely to present false positives. and having a small effect size can result in data that is statistically significant. However, with an effect size so weak the results could be deemed meaningless (Oxenham, 2015).

Conclusion

With a reproducibility crisis of this scale affecting the field of science, it is often difficult to see clear steps that can be made towards a more trustworthy and rigorous research environment. Knowing the core causes of the reproducibility crisis, which include statistical error, lack of data transparency, fraud and corporate influence, and the culture of science and

research, helps scientists to better understand where changes can be made. To reiterate, adjustments in the p-value can have tremendous effects on the value of statistical significance and will result in far less false positives in the literature. Starting at the foundation of education, statistics courses need to be better implemented into curricula to give all professionals a stronger knowledge base to use when conducting research and evaluating studies. Altering incentives in the field can help to alleviate the culture surrounding research that promotes the publish or perish mentality. With changes in this area, a system can be created that is more forgiving of failure and promotes publishing of null findings, meta analyses, and replication studies. In addition, promoting more open access journals, and platforms such as the Open Science Framework, will allow for greater transparency in data, methods, and research designs. Finally, as a medical professional, researcher, or lay person, it is important to be critical and to address every study with the aforementioned questions to assure that the individual is reading and trusting studies with rigorous study designs, correct statistical analyses, and ethical research practices.

References

Academy of Medical Sciences. (2015). Reproducibility and reliability of biomedical research: Improving research practice. Symposium Report. Retrieved from https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf

Baker, M. (2015). Over half of psychology studies fail reproducibility test. *Nature*. doi:10.1038/nature.2015.18248

Bartlett, T. (2017). Spoiled science: How a seemingly innocent blog post led to serious doubts about Cornell's famous food laboratory. *The Chronicle*. Retrieved from https://www.chronicle.com/article/Spoiled-Science/239529

Begley, G., & Ellis, L. (2012). Raise standard for preclinical cancer research. *Nature, 483,* 531-533.

Casadevall, A., & Fang, C. F. (2016). Rigorous science: A how-to-guide. *mBio*, *7*(6). doi: 10.1128/mBio.01902-16.

Center for Open Science. (n.d.). Title of original webpage. Retrieved from https://cos.io

DeFranzo, S. (2011). *What's the difference between qualitative and quantitative research.* Retrieved from https://www.snapsurveys.com/blog/qualitative-vs-quantitative-research/

Fang, F. C., Steen, R. G., & 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, D. (2010). Wrong: Why Experts* Keep Failing us-And How to Know When Not to Trust Them. New York, NY: Little, Brown and Company.

Gonzalez, K. (2019). *What is a null hypothesis? Definition & examples.* Retrieved from https://study.com/academy/lesson/what-is-a-null-hypothesis-definition-examples.html

Harris, R. (2017). *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions.* New York, NY: Basic Books.

Horton, R. (2015). Offline: What is medicine's 5 sigma? *The Lancet* (385).

Ioannidis, P. J. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124. doi: 10.1371/journal.pmed.0020124

Johnson, V. (2013). Revised standards for statistical evidence. *PNAS*, 110(48), 19313-19317. doi:

https://doi.org/10.1073/pnas.1313476110

Kaplan. (2019). The prerequisites for medical school. Retrieved from https://www.kaptest.com/study/mcat/the-prerequisites-of-medical-school.

Nayak, B. K., & Hazra, A. (2011). How to choose the right statistical test?. *Indian journal of ophthalmology*, 59(2), 85-6.

National Institutes of Health. (2017). Medline citation counts by year of publication (as of mid-December 2017). Retrieved from https://www.nlm.nih.gov/bsd/ medlinecit_ counts_yr_pub.html

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349, aac4716. doi: 10.1126/science.aac4716

Oxenham, S. (2015). Believe it or not, most published research findings are probably false. Retrieved from https://bigthink.com/neurobonkers/believe-it-or-not-most-published-research-findings-are-probably-false.

Resnick, B. (2017a). Study: Half of the studies you read about in the news are wrong. *Vox.* Retrieved from https://www.vox.com/science-and-health/2017/3/3/14792174/half-scientific-studies-news-are-wrong

Resnick, B. (2017b). What a nerdy debate about p-values shows about science and how to fix it. *Vox.* Retrieved from https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005

Schekman, R. (2013). *How journals like nature, cell, and science are damaging science.* Retrieved from http://openscience.ens.fr/ABOUT_OPEN_ACCESS/ARTICLES/2013_12_09_The_Guardian_Nobel_Prize_winner_against_top_journals.pdf

Schoenfeld, J., & Ioannidis, J. (2013). Is everything we eat associated with cancer? A systematic

cookbook review. *American Society for Nutrition, 97*, 127–34.

Titler, G. M. (2018). *Patient Safety and Quality: An Evidence-Based Handbook for Nurses.* Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK2659/

Walia, A. (2015). Editor in chief of world's best-known medical journal: Half of all the literature is false. *Collective Evolution.* Retrieved from https://www.collective-evolution.com/2015/05/16/editor-in-chief-of-worlds-best-known-medical-journal-half-of-all-the-literature-is-false/

Whoriskey, P. (2012). As drug industry's influence over research grows, so does the potential for bias. Washington Post. Retrieved from https://www.washingtonpost.com/business/economy/as-drug-industrys-influence-over-research-grows-so-does-the-potential-for bias/2012/11/24/bb64d596-1264-11e2-be82-c3411b7680a9_story.html?utm_term=.4018f20fb048

Zimerman, A. (2013). Evidence-based medicine: A short history of a modern medical movement. *AMA Journal of Ethics, 15*, 71-76. doi: 10.1001/virtualmentor.2013.15.1.mhst1-1301.