Questions of P-hacking and Data-Replicability in Oncological Therapeutics

Scott Shepetin
May 1, 2021

Professor Gary Smith
Pomona College Department of Economics
Senior Seminar: Final Paper

Section 1: Introduction

Over the past few decades, there has been increasing concern that many published research findings are difficult or impossible to replicate. This phenomenon known as the *Data-Replicability Crisis*--- has been observed across a variety of fields in the social sciences and medicine, among other fields. While a lack of replicability doesn't *entirely* demonstrate all critical findings are false--- it does undermine reliability and credibility—key cornerstones of the scientific method. This poses a particularly pernicious effect in the medical field, because physicians, patients, and health care providers rely on accurate and up-to-date information to best serve their patients. Indeed, it seems a lack of reliability hurts patients in two critical ways:

1) A large percentage of findings in medical research may contain no underlying truth (This hurts patients in the short-term as they may be given faulty advice, such as to take drugs that provide no real benefit.)
2) Stunting further medical innovation: False positives may hurt the medical field by creating a "cloud-of-uncertainty." This may undermine innovation as even *one-faulty-study,* can lead to a rabbit-hole, where further R+D money is wasted. There is some evidence to support the claim that false-positives serve as a growth drag on innovation as (Macleod *et al.* 2014) shows waste across biomedical research accounts for $85 billion annually.[1] This effect is particularly pronounced if false-positives garner widespread media coverage such as Andrew Wakefield's retracted study on measles and autism.[2]

While false medical research is always discouraging, the stakes in the field of oncological therapeutics are particularly high. This is because innovation at the clinical level has the potential to lead to the development of new cancer therapeutics which can both improve survival rates and quality of life for millions of cancer patients around the world.[3] Furthermore, the development of oncological therapeutics is one of the most expensive processes in drug development, so there is a real financial incentive to make sure money is well spent. It is precisely because these stakes are so high that researchers have gone to great lengths to estimate the false-positive rate (sometimes referred to as the "failure-rate") in oncology and to propose strategies to make it lower. Some

[1] Malcolm Macleod, "Biomedical Research: increasing value, reducing waste," *The Lancet,* 2014. A paper which showed that showed the cumulative effective was that about 85% of research investment—equating to $200 billion of the investment in 2010—is wasted.

[2] Andrew Wakefield's retracted study on measles and autism

[3] Begley, C., Ellis, L. Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012). https://doi.org/10.1038/483531a

researchers have argued that because of the intrinsic difficulty of studying cancer, we should not be discouraged by a low-success rate. However, it would nevertheless be disappointing in a) suggesting a poor ability to transfer cancer research into clinical success and outcomes for patients, b) serving as growth drag on true developments in cancer therapeutics, as false-positives drain precious money and time from other initiatives, and c) creating a rabbit-hole of further research tangentially related to the originally false research.

While the larger issue of data irreproducibility has been buzzing around the oncological community for decades, it reached a fever pitch when the biotechnology firm Amgen, reviewed fifty-three "landmark" cancer papers in 2012. These fifty-three papers were selected by researchers Ellis and Begley, because they were supposed to represent some of the most promising and innovative develops in the field of oncology therapeutics --- such as fresh approaches to target cancer or alternative clinical uses for existing therapeutics.[4] Nevertheless, scientific findings were only confirmed in 6 studies, representing a discouraging 11% reproducibility rate.[5] This high failure rate served as an almost existential threat to the field of clinical oncology, highlighting both an existing and historic inability to translate cancer research into successful clinical trials and ultimately more effective drugs. While this 11% number certainly paints a grim picture, it is not totally out of the ballpark of other empirical studies including a Bayer Health team in Germany which found that only 25% of published preclinical studies could be validated "to the point where projects could continue."[6] Taken together, these two studies suggest that oncological research is characterized by *extremely-poor* reproducibility which serves as a meaningful roadblock *early-on* in the pipeline of cancer drug development. Perhaps this is why cancer has such high drug attrition rates. As shown by (*Hutchinson and Kirk, et al.* 2011) only 5% of agents that have anticancer activity in preclinical development are licensed after demonstrating sufficient efficacy in phase three testing.[7]

While this issue of erroneous papers is particularly troubling in oncological research, it effects all of medicine more broadly. It also seems that irreproducibility and *false-positives* are a natural byproduct of hypothesis testing. Hypothesis testing, otherwise known as significance testing, is one of the most common statistical tests utilized in a variety of fields including biology, medicine,

[4] Begley, C., Ellis, L. Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012). https://doi.org/10.1038/483531a

[5] Ibid.

[6] Hutchinson, L., Kirk, R. High drug attrition rates—where are we going wrong?. *Nat Rev Clin Oncol* **8,** 189–190 (2011). https://doi.org/10.1038/nrclinonc.2011.34

[7] Ibid.

phycology, and the physical sciences.[8] In seeking to prove a causal scientific hypothesis, such as that a drug will lower blood pressure, researchers start with a null hypothesis—that nothing is going on—and the scientific hypothesis doesn't hold (β_1=0). This is essentially a default position of skepticism, much like assuming a defendant is innocent until proven guilty.[9] Under this paradigm, scientist compare their results to what they would expect if the null hypothesis were true. They then calculate the probability of observing a specific result, given the null hypothesis is true. This value is known as the P-Value. Prior to an experiment, researchers select a "significance" level so that if the P-value is below the threshold (typically 0.05), it is considered "statistically significant," and the null hypothesis is rejected. In the case of a clinical trial of medication claiming to lower blood pressure, the rejection of the null hypothesis (β_1=0) means researchers conclude (something) does affect blood pressure—a big step into eventually bringing the drug to market.

Although hypothesis-testing serves an important role in the scientific process, the approach is not without faults. Indeed, by utilizing a P-value of 0.05, researchers accept a 5% error rate. This means that if researchers test twenty garbage theories, we would expect one to end up being statistically-significant creating a false-positive result that gets published. Therefore, hypothesis testing will *always* generate some-level of false positives, even without P-hacking, data-mining, or overt manipulation. Considering this unavoidable fact, what becomes critically important is the ratio of false studies to true studies researchers analyze. This is because, throughout the peer-review and selection process, medical journals tend to publish *only* results from researchers who achieve statistical significance causing a survivorship bias. As shown in figure 1, this distortion can lead to an alarmingly high percentage of false paper if a large percentage of theories, researchers analyze are garbage (containing no objective truth). Indeed, if researchers analyze twenty false theories, for every single hypothesis that is true (not an unreasonable ratio considering the difficulty in scientific research) they reach the break-even point, where half of the published medical research is false!

[8] Hutchinson, L., Kirk, R. High drug attrition rates—where are we going wrong?. *Nat Rev Clin Oncol* **8,** 189–190 (2011). https://doi.org/10.1038/nrclinonc.2011.34

[9] Tim Dean, "How do we Edit Science part 2, Significance Testing, P-hacking and Peer Review," *The Conversation.*

Figure 1: Ratio of true/false theories tested vs Accuracy of published research

Ratio of true to false theories tested by researchers	P-Value	Percentage of Published Research which is True[10]
1:1	P= 0.05	95.24%
1:2	P= 0.05	90.91%
1:5	P= 0.05	80%
1:10	P= 0.05	66.66%
1:20	P= 0.05	50%
1:50	P= 0.05	28.57%
1:100	P= 0.05	16.66%
1:1000	P= 0.05	1.97%

*Note the full mathematical equation which estimates the predictive power of positive research findings (PPV) is as follows:

$$\text{PPV} = (1- \text{ß}) *R/ (R- \text{ß}R + a)$$

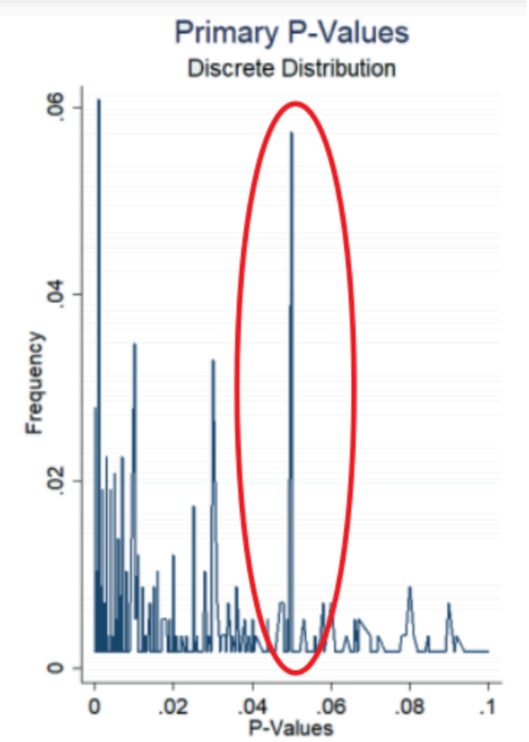
where ß is the Type II error rate, a is the Type I error rate, and R is the ratio of "true relationship" to "no relationship" among the specific sub-field.

As figure 1 shows, even the most well-meaning, ethical researchers will publish some level of false-positives research. This is because finding new medical research is difficult, and there is a publication selection bias (sometimes known as the "file drawer effect") where studies with nonsignificant results have much lower publication rates.[11] This phenomenon has been known for years, but the critical insight, as shown by figure 1, is that as the ratio of true-to-false theories tested *decreases*, the accuracy rate of published research falls. This phenomenon is also displayed in figure two, as published medical research tends to bunch around the 5% significance-level. This suggests not only that published research is not *representative*, but also that there may be subtle and unconscious manipulation by researchers to achieve the all-important 5% statistical significance.

[10] This figure is not totally correct as it doesn't include an assumption about P[reject if null hypothesis is false]. For more full analysis see the equation for PPV: positive predictive value as derived by John. P.A. Ioannidis:
Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2: e124. pmid:16060722

[11] Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychol Bull 86: 638–641.

Figure 2: Understanding the File-Drawer Effect



The emergence of big data and modern computing has drastically lowered the cost of testing new theories. This has given researchers much more flexibility in data-design, including the ability to mine the data without *a prior* hypothesis, only to claim that a statistically significant result has been originally predicted.[12] This practice, known as "HARKing" (Hypothesizing After the Result are Known), is particularly dangerous is fields such as nutritional science, where research produces huge data-sets with many variables. Utilizing statistical software, it becomes easy to mine the "cornucopia of possible variations" to find spurious correlations, creating a slew of false results.[13] Indeed, proponents of sophisticated analytics software even advertise the ability to "mine large data sets for insights as to the solution to many of our society's problems."[14] This logically contributes to the replication crisis because as the ratio of true to false theories tested increases (Figure 1, Column 1) the accuracy rate of published papers (Figure 1, Column 3) decreases.

In addition to data mining and "HARKing," researchers have a variety of tools to increase the likelihood of finding statistical significance. One of the most common is known as P-hacking or

[12] Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. doi:10.1207/s15327957pspr0203_4
[13] Christie Aschwanden, "You can't trust what you read about nutrition," *Fivethirtyeight,* 2016.
[14] Richards, Neil M. and King, Jonathan, Three Paradoxes of Big Data (September 3, 2013). 66 Stanford Law Review Online 41 (2013). Available at SSRN: https://ssrn.com/abstract=2325537

"selective reporting" where researchers misreport true effects size in published studies. As argued by Tim Dean, HARKing is best conceived as a subsection of P-hacking—a more all-encompassing term. Unlike HARKing, which specifically refers to hypothesizing after results, P-hacking reflects that bias that can occur throughout the publication process.[15] Common practices of p-hacking include: conducting analysis midway through an experiment to decide whether to collect more data or recruit additional participants,[16] (Head *et al. 2015)* recording many response variables and deciding which to post-analysis, [17] (Gadbur*y et al.* 2014), the selective deletion of outliers (to artificially inflate statistical significance of tested variables), excluding, or splitting treatment groups in post-analysis,[18] (Ioanndis *et al.* 2005), and stopping data exploration if analysis yields a significant P-value[19] (Bastardi *et al.* 2015). In addition, researchers can always engage in outright fabrication of data and fraud. Perhaps in its less pernicious form, this could involve poor research practices, such as failure to control for bias, low statistical power, and poor-quality control.

From a theoretical standpoint, it is easy to see how P-hacking, data-mining, and overt manipulation by researchers accelerates the rate of false-positives in oncological therapeutics and other areas of medical research. But a key insight from figure 1, is that there will always some level of inaccuracies in peer-reviewed research papers. In a sense, this is the million-dollar question of the replication crisis: How can we begin to untangle the effect of P-hacking from the naturally occurring rate of false-positives that occurs as a result of testing? This is a problem that is extraordinarily difficult to measure as it involves quantifying a few problems.

1. How big is the current replication crisis? (What is the rate of false-positives?)
2. What percentage of false-positives are due to P-hacking and data-manipulation?
3. What is the underlying ratio of false-positives without P-hacking and data-manipulation? This question might involve deciphering what the rate of false-positives was *before* the technological revolution, and whether there are better alternatives that claiming conclusive findings solely on the basis of a single study assessed by a statistical significance of a P-value of 0.05.

As figure 1 shows, it is possible to build a theoretical framework to estimate the likelihood a published research paper is true. However, from a mathematical modeling perspective, it becomes

[15] Tim Dean, "How do we Edit Science part 2, Significance Testing, P-hacking and Peer Review," *The Conversation.*

[16] Head M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biol, 13*(3): e1002106. doi:10.1371/journal.pbio.1002106

[17] Gadbury GL, Allison DB (2014) Inappropriate fiddling with statistical analyses to obtain a desirable p-value: Tests to detect its presence in published literature. PLoS ONE 7: e46363.

[18] Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2: e124. pmid:16060722

[19] Bastardi A, Uhlmann EL, Ross L (2011) Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. Psychol Sci 22: 731–732. pmid:21515736

difficult in that no one really knows (Column 1) the ratio of true to false theories tested. Another way of phrasing this is that it's difficult to put a precise number on the probability that an individual research topic is true *prior-to* issuing a statistical test. Indeed, at the individual-level this task is arduous enough to make even the most fervent Bayesian blink: As not only do we have the issue of evaluating the *prior-probabilities* that research is true, but these predictions must be robust enough to adapt to individual circumstances. Logistically this seems impossible, as the odds of research being true *prior-to* a statistical test could be interrelated with a variety of other factors, including the field of study, the flexibility in design, the tenacity of the researcher, a researcher statistically tools & propensity to P-hack and even whether they are up for tenure. It is here where (Ioannidis *et al. 2006)* had a critical insight.[20] In order, to calculate the (positive predictive value), otherwise known as the accuracy rate of publicized findings, instead of seeking to find, the *prior-probability* of research is true, he just assumed that prior probability would be equal to the average of that specific subfield.

Ioannidis's unique solution allowed for the creation of a basic mathematical model to estimate the odds a recent paper is true (PPV). In particular, he showed that the odds of a being true are interrelated with R: the ratio of the number of "true relationships" to "no relationship" among researchers tested in a field, the statistical power of the study, and level of statistical significance.[21]

- Basic Model: $PPV = (1-\beta)*R/(R-\beta R+\alpha)$ where β is the Type II error rate, α is the Type I error rate, and R is the ratio of "true relationship" to "no relationship" among the specific sub-field.

Note, this general form will be discussed in greater detail throughout the paper, but a critical insight is often quite difficult to reach a $PPV = 0.5$. In an intuitive sense, this means it is difficult for most research to be true, and mathematically assuming $\alpha = 0.05$, a paper is only most likely true if and only if: $(1-\beta)/R > 0.05$:

[20] Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2: e124. pmid:16060722

[21] Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2: e124. pmid:16060722

Section 2: Literature Review: How big is the Data Replicability Crisis in Medicine?

In seeking to discuss the impact of P-hacking and data manipulation in the medical field there is perhaps no more important question then the following: What percentage of published medical findings are replicable? This question is critically important, because it is suggestive of the larger dynamics surrounding the reliability of medical research. Indeed, reproducibility is often considered to be "the defining features of science" because, without it, theories lack the empirical evidence to become accepted as scientific truth.[22] Put bluntly, the size of the reproducibility crisis is important because it suggests what percentage of medical research may be false.

Over the past decade, researchers have developed a variety of methodological approaches to estimate the size of the data-replicability crisis. With that being said, the task of estimating the overall false-positive rate in medicine is still extraordinarily difficult because a) The volume of medical literature is extremely large: (Based on analysis by the National Library of Medicine, there may be around 2.5 million papers in 30,000 medical journals published each year.[23] b) Medical research is very expensive to produce. (This is particularly true for RCTs and clinical trials --- as an example, the cost of developing an oncology therapeutic is approximately $78.6 million,)[24] and c) False-positive rates tend to vary widely by subfield.

When combined in aggregate, these factors make it virtually impossible to accurately test the *overall* reproducibility of findings in the medical field, as doing so would involve constructing a representative "subsample" of all medical findings, evaluating the percentage of the subsampled findings which are reproducible, and then extrapolating overall reproducibility from that subsample. This process is not realistic with financial and time constraints, and difficult to replicate even in fields like phycology---which by best indications may have only a 40% reproducibility rate.[25] With that being said researchers have been able to more effectively estimate replicability in specific medical subfields, as well as establishing a lower bound by assessing rates of fraud and data manipulation.

One medical subfield that has garnered a lot of academic and social interest has been the field of oncology. This is because innovation at the clinical level has the potential to lead to the

[22] Tim Dean, "How do we Edit Science part 2, Significance Testing, P-hacking and Peer Review," *The Conversation.*

[23] It should also be noted, that only a tiny minority of research is published in "Major General Medical Journals." For example, out of the 730,447 articles labeled as "clinical trial" in PubMed as of May 26, 2016, only 18,231 were published in the major medical journals.

[24] "3.1 Costs by Therapeutic Area." *ASPE,* February 2017, aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development/31-costs-therapeutic-area, for more information see figure 1.

[25] Gilbert, Daniel T., et al. "Estimating the Reproducibility of Psychological Science." *Open Science Collaboration, American* Association for the Advancement of Science, 2015.

development of new cancer therapeutics which can both improve survival rates and quality of life for millions of cancer patients around the world.[26] It is precisely because these stakes are so high that researchers have gone to great lengths to estimate the false-positive rate (sometimes referred to as the "failure-rate") in oncology and propose strategies to make it lower. While some researchers have argued because of the intrinsic difficulty of studying cancer, we should not be discouraged by a low-success rate, it would nevertheless be disappointing in a) Suggesting a poor ability to transfer cancer research into clinical success and outcomes for patients, b) Serving as growth drag on true developments in cancer therapeutics: as false-positives drain precious money and time from other initiatives, and c) Creating a rabbit-hole of further research tangentially related to the originally false research.

While the larger issue of data irreproducibility has been buzzing around the medical community for decades, it reached a fever pitch when the biotechnology firm Amgen reviewed fifty-three "landmark" cancer papers in 2012. These fifty-three papers were selected by researchers Ellis and Begley because they were supposed to represent some of the most promising and innovative develops in the field of oncology therapeutics --- such as fresh approaches to target cancer or alternative clinical uses for existing therapeutics.[27] Nevertheless, scientific findings were only confirmed in 6 studies, representing a discouraging 11% reproducibility rate.[28] This high failure rate served as an almost existential threat to the field of clinical oncology: highlighting both an existing and historic inability to translate cancer research into successful clinical trials and ultimately more effective drugs. While this 11% number certainly paints a grim picture, it is not totally out of the ballpark of other empirical studies including a Bayer Health team in Germany which found that only 25% of published preclinical studies could be validated "to the point where projects could continue."[29] Taken together, these two studies suggest that oncological research is characterized by *extremely-poor* reproducibility which serves as a meaningful roadblock *early-on* in the pipeline of cancer drug development. Indeed, as aptly stated by Ellis and Begley, erroneous research papers "spawned an almost entire field, with hundreds of secondary observations, without actually seeking to confirm

[26] Begley, C., Ellis, L. Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012). https://doi.org/10.1038/483531a

[27] Ibid.

[28] Ibid.

[29] Hutchinson, L., Kirk, R. High drug attrition rates—where are we going wrong?. *Nat Rev Clin Oncol* **8,** 189–190 (2011). https://doi.org/10.1038/nrclinonc.2011.34

or falsify its fundamental basis."[30] Perhaps this is why cancer has such high drug attrition rates - as shown by (Hutchinson and Kirk*, et al.* 2011) only 5% of agents that have anticancer activity in preclinical development are licensed after demonstrating sufficient efficacy in phase three testing.[31] While to a certain extent it is not surprising that cutting edge, oncological research has a high false-positive rate (likely over 75%), researchers have proposed a variety of solutions to raise the standards of preclinical research including improving understanding of pharmacokinetics and pharmacodynamics, appreciating the limits of tumor cell lines and animal models, and allowing researches to publish stories that fill in "gaps" rather than publishing perfect stories.[32] While such strategies are unlikely to equalize oncological rates of replicability with other therapeutic areas, as sadly clinical oncology has the highest failure rate, it would serve as a helpful first step.

Recognizing that there is a wide variation in false-positive rates and replicability between medical subfields, researchers have approximated a lower-bound for the false-positive rate of *all* medical research by analyzing rates of fraud and data manipulation amongst researchers. Perhaps most famously, an FDA data-auditing of medical research from 1977-1990 found that approximately 10-20% of R&D funds are estimated to be spent on questionable studies characterized by a "misrepresentation of data, inaccurate reporting, and fabrication of experiment results."[33] Within this broad categorization of data-misrepresentation, (Glick *et al.*) found that 2% of clinical investigators were guilty of serious scientific misconduct.[34] These results are similar to a National Institute of Health (NIH) survey of early-to-mid level career scientists (n=3247) which found that within the previous 3 years, "0.3% admitted to falsification of data, 6% to failure to present conflicting evidence, and 15.5% to changing the study design, methodology or results in response to funding pressure."[35] While these findings are troubling they suggest that classic 'fraud' falsification, fabrication, and plagiarism (FFP) may be less important than more subtle research practices, including P-hacking, data mining, selective reporting of dependent variables, and tinkering of data-

[30] Begley, C., Ellis, L. "Raise standards for preclinical cancer research," *Nature* **483,** 531–533 (2012). https://doi.org/10.1038/483531

[31] Hutchinson, L., Kirk, R. High drug attrition rates—where are we going wrong?. *Nat Rev Clin Oncol* **8,** 189–190 (2011). https://doi.org/10.1038/nrclinonc.2011.34

[32] Ibid, direct quote.

[33] J, Leslie Glick (1992) "Scientific data audit—A key management tool, Accountability in Research" 2:3, 153-168, DOI: 10.1080/08989629208573811

[34] Ibid.

[35] Martinson BC, Anderson MS, de Vries R. 2005. Scientists behaving badly. *Nature* 435, 737–738. (doi:10.1038/435737a)

sets.[36] These findings were echoed by meta-analysis research of survey data by (Fanelli, *et al.* 2009), which found that despite only a "1-3% rate of fraud *per se,* approximately 33.7% of scientist admit to engaging in questionable research practices," with admission rates skyrocketing to 72% when discussing the behavior of colleagues.[37] While these data-points paint a discouraging picture about the current state of ethics in medical research, they may also be suggestive of a toxic 'publish or perish' culture where scientists are incentivized to create publishable results at all costs, thereby increasing bias.

While this is still a developing field preliminary research from (Fanelli, *et al.* 2010) indicates that across all-disciplines, papers are more likely to support a tested hypothesis if "their corresponding authors work in a state that, according to NSF data, produces more academic paper per capita."[38] The size of this effect only *increased* after controlling for the state's per capita R&D expenditure, and although Fanelli was unable to control for the confounding effect of institutional prestige, these results support the hypothesis that competitive academic environments increase scientific bias.[39] This provides some evidence to the traditional "publish or perish," hypothesis which suggests that if scientists are incentivized to create publishable results at all costs, publication bias will increase. There is also some longitudinal data to support this claim, as research from (Fang *et al.* 2013) found that the percentage of scientific articles retracted because of fraud has increased 10-fold since 1975, which has coincided with both an explosion in the number of medical research published and increased competition for scientific funding.[40] While this correlation at the cohort-level, between rates of competition for research funding, and retraction due to fraud does not necessarily imply a causal relationship, it is suggestive that further research needs to be done to determine what is causing increased data-manipulation in medical research.

Finally, despite most research focusing on the replicability rates within a specific medical subfield, or efforts to establish a lower-bound for all medical research by focusing on data-manipulation and scientific misconduct, some researchers have sought to quantify the *overall* false-positive rate of medical research. This work has been most notably completed by (Ioannidis *et al.*

[36] Grimes David Robert, Brauch Chris T. and Ioannidis John P.A, "Modeling Science Trustworthiness under Publish or Perish," *R. Soc. Open Sci: 5:* 171511 https://doi.org/10.1098/rsos.171511

[37] Fanelli D (2009) How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLoS ONE 4(5): e5738. https://doi.org/10.1371/journal.pone.0005738, direct quote.

[38] Ibid.

[39] Ibid.

[40] Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications [published correction appears in Proc Natl Acad Sci U S A. 2013 Jan 15;110(3):1137]. *Proc Natl Acad Sci U S A.* 2012;109(42):17028-17033. doi:10.1073/pnas.1212247109

2005), who analyzed the 49 medical studies from 1990-2003 with more than 1000 signatures.[41] While it is important to acknowledge that Ioannidis paper "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," is not a representative sample of all medical research --- it is a powerful indication of the likely error rate in clinical research studies published in major journals that become widely cited. In terms of methodology, Ioannidis compared the results of the initial highly cited articles against subsequent studies of larger sample size and better-controlled designs, to determine replicability and effect size. Out of these 49 studies, "16% were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged."[42] Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials ($P = .008$).[43] These results suggest that clinical research on the efficacy of medical interventions is sometimes followed by subsequent studies that either reach the opposite conclusion or suggest the magnitude of the original claims were too large. This point about the relative size of effects is quite important, because, particularly with clinical trials, they often show much larger effects in trials than in the real-world because participants are in a perfectly-modulated setting under the appropriate guidance of a health professional. However, it is important to note, that this is still an ongoing debate in medical field and that some evidence from (Benson *et al.* 2000) suggests that at least as it relates to observational studies after 1984, there is no major evidence that they find "consistently larger or qualitatively different effects, than those in randomized, controlled trials."[44] However, this is less likely to be the case for randomized control trials. Indeed, not to digress further, but the typically larger impact medicine has in controlled trials vs. the typical effect it has on patients in real life, may not be only suggestive of the importance of taking medicine as directed, but also of the Placebo effect. While further research is needed to validate these claims and explore whether specific types of studies are more likely to be contradicted than others, it is important to also consider limits to generalizability. This is specifically because highly cited-studies may have either a) a lower false-positive rate than medical research as a whole if it reflects top-quality research or contrarily b) have a higher false-positive rate if widely cited research tends to

[41] Ibid.

[42] Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294(2):218-228. doi:10.1001/jama.294.2.218

[43] Ibid.

[44] Benson, Kjell and Hartz, J. Arthur. "A comparison of Observational Studies and Randomized Controlled Trials N Engl J Med 2000; 342:1878-1886
DOI: 10.1056/NEJM200006223422506

become famous for the 'wrong' reasons (such as wide exposure to the media because of a splashy or trendy result.)

Summary Estimated False-Positive Rates of Medical Research:

- Upper bound: Oncological therapeutic research: 75-88%
- Lower bound: Overt data manipulation 1-3%, studies of poor methodological quality 10-20%, and number of scientists admitting to questionably research practices 36-72%
- Ioannidis's Mid-level estimate: "High-impact studies:" 16% show opposite effect + 16% show initial studies exaggerated effects = 32%

Section 3: Methodological Framework, (How I gathered Data)

To investigate the gap between expectations and the reality of good clinical practice, I am going to engage in the meta-analytical P-curve approach. This was perhaps most famously utilized by a team of German scientists (Belas *et al.* 2017). In their groundbreaking paper, "P-hacking in Clinical Trials: A Meta-Analytical Approach," they were able to compile a dataset consisting of thousands of clinical trials, that contained both primary and secondary endpoints. Primary endpoints are defined as covering the "main effect," of the experiment, thus *directly* influencing the drug approval process --- whereas secondary outcomes detail further information unrelated to the approval process. From a P-hacking perspective there is therefore an incentive for researchers to "tinker" with the P-value of the primary endpoint but not the secondary endpoint. This is what the research from (Belas *et al.* 2017) showed and indeed they observed an "abnormal increase in the P-value frequency at common significance thresholds, while the secondary P-value contained no such anomaly."[45]

The P-curve approach has been used as a strategy to distinguish between selective reporting of results on one hand and truth on the other (Simonsohn *et al.,* 2014). The p-curve approach is just an observation of the frequency of distribution of p-values. The logic that underpins the p-curve is that if a group of studies has no effect, the p-curve should have a uniform distribution, whereas, if the studies have some effect, then the p-curve should be right-skewed. The more statistically impactful the studies measured, the larger the right-skew should become.[46] Therefore, a "right-skewed p-curve, which encompasses a set of independent findings with continuously decreasing p-values from low to high, is an indicator of evidential value."[47] When p-curves differ substantially from this ideal shape --- this is suggestive that studies may be of poor methodological quality or that data-drugging or p-hacking may be occurring.

For my P-curve analysis, I have chosen to analyze published studies from the New England Journal of Medicine. I have made this choice because the NEJM is the most widely read, cited, and influential medical journal in the world.[48] Indeed, NEJM has the highest Journal Impact Factor (74.699) of all general medical journals (Source Clarivate, 2020). NEJM also has broad public access (meaning I was able to find the studies) --- and contains high-quality peer-reviewed research and interactive clinical content. Given, my analysis on P-curves is of studies published in NEJM, this

[45] Direct quote, Belas *et al.* 2017.

[46] Ibid.

[47] Ibid, direct quote.

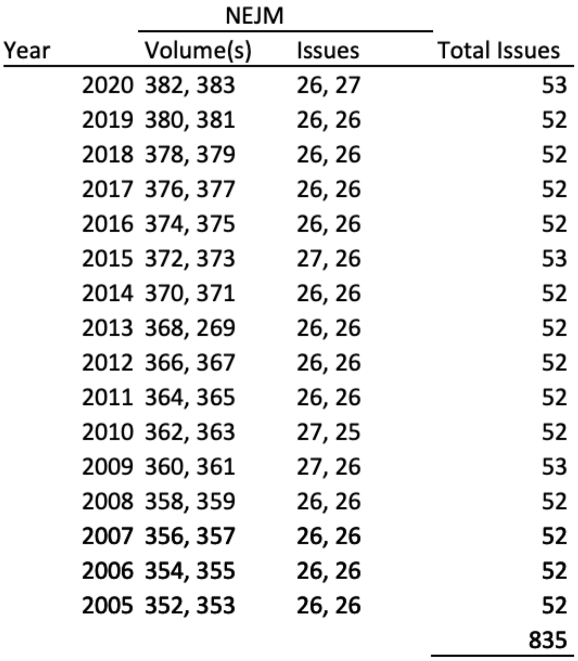
[48] Clarivate, 2020.

suggests that my findings may be generalizable to other high-impact journals, but not to medical journals or medical research more broadly.

In its selection process, each year NEJM receives over 16,000 submissions. Given only 5% of original research submissions achieve publication in the U.S. this suggests NEJM publishes approximately 800 studies each year. In my process, of gathering data I was able to easily obtain NEJM publication from the years 2005-2020. (Please note: for research in the future, I think it would be more interesting to go *further* back in time, particularly as a longer-time range may be more helpful in determining whether rates of p-hacking have increased over time). With that being said, due to the difficulties of collecting data, I have chosen to focus solely on the years from 2005-2020. This further limits generalizability because it is suggestive that results from my p-curve analysis may not be generalizable to other time-periods.

Considering 16 years of analysis, and approximately 800 published NEJM studies/year, my overall dataset consists of approximately **12,800** NEJM papers. Given that obtaining p-values is relatively labor-intensive, I have chosen to go with the approach of sub-sampling within defined time periods. This process was made easier by the relative uniformity of publishing rates. Within each year there are two volumes, and within each volume, there were between 25-27 issues. However, the vast majority of volumes contained exactly 26 issues (84.3%) see Figure 3.

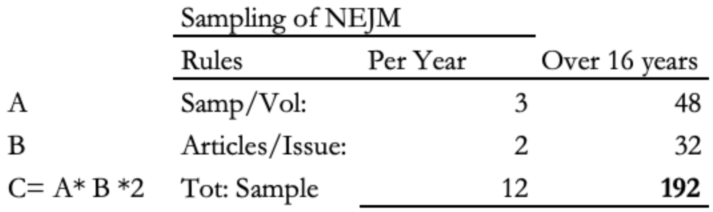
Figure 3: Data-set Overview



	NEJM		
Year	Volume(s)	Issues	Total Issues
2020	382, 383	26, 27	53
2019	380, 381	26, 26	52
2018	378, 379	26, 26	52
2017	376, 377	26, 26	52
2016	374, 375	26, 26	52
2015	372, 373	27, 26	53
2014	370, 371	26, 26	52
2013	368, 269	26, 26	52
2012	366, 367	26, 26	52
2011	364, 365	26, 26	52
2010	362, 363	27, 25	52
2009	360, 361	27, 26	53
2008	358, 359	26, 26	52
2007	356, 357	26, 26	52
2006	354, 355	26, 26	52
2005	352, 353	26, 26	52
			835

As previously mentioned, I have chosen to go with the approach of sub-sampling within defined time periods. Specifically, within each year, I have chosen to sample, 3 issues per volume, and then within each issue 2 papers. This means, that I will analyze a total of 12 studies/year. I made these decisions because I wanted my papers to be relatively uniformly distributed across year, volume number, and issue, to not be over-concentrated in one specific period. This is because during rare times --- such as in public health emergencies like Covid --- the journal can become hyper-focused on one particular issue which may limit generalizability.

Figure 4: My Sub-Sampling Approach



	Sampling of NEJM		
	Rules	Per Year	Over 16 years
A	Samp/Vol:	3	48
B	Articles/Issue:	2	32
C= A* B *2	Tot: Sample	12	**192**

[49]

In terms of my methodological framework, within each volume, I used a random number generator to determine which three issues to analyze. This suggests that it is likely the issues I analyzed are representative of the volumes they are a part of. However, within each issue, I had to pick two studies to gather data on. This process was more difficult because not all of the studies published in NEJM had at least one P-values, let alone primary and secondary variables of interest. While I did not measure this directly, I would estimate less than one-half of all studies had at least one published P-value which is in line with (Belas *et al.* 2017) analysis that 46.71% of 6,081 phase III clinical trials, had at least one primary P-value. Given that I needed at least one P-value and variable of interest, to gather data I had to get a little creative.

To determine which two papers, to analyze within each issue, I developed my own internal protocol. I would spotlight search for key terms and then go through each article until I found one with enough data my process was as follows:

Figure 5: Data Collection Protocol: Articles/Issues

1. Spotlight search for "Randomized,"
2. Spotlight search for "Clinical,"
3. Spotlight search for "Trial,"
4. Spotlight search for "Cancer,"
5. → If no spotlight search yield results proceed chronologically

[49] Unfortunately, this chart does not reflect the most up-to-date information. Due to time and space constraints, instead, of sub-sampling every year I have chosen to subsample 6-years. (2020 & 2019), (2015 & 2014), (2010 & 2009), and (2006 & 2005). However, to remain with a sample size of 192 I have chosen to utilize 6 samples/volume instead of 3. This means I have 192 but only clustered over the designated 8 years instead of 16

To be clear, with regards to my protocol of selecting articles within a given issue, I would proceed with my five-step process. If any one-step returned multiple results, I would proceed chronologically until I found a suitable article.[50] However, if the spotlight search returned no articles or the articles flagged didn't have sufficient results (no P-value or 95% CI) I would proceed to the next step. On the rare, occasion, that there were no suitable articles within an entire issue, and if that was the case, I would have to proceed to the next issue. However, this only happened 6 times (6.25%).

While the process of spotlight searching for "randomized," "clinical," "trial," and "cancer," means that the articles I sampled are not representative of the overall issue they are a part of, I developed this protocol for a very specific reason. This is because clinical trials and randomized studies are the far most likely to be associated with clearly defined primary and secondary outcomes. This is very important because comparing the distribution of P-values for primary and secondary outcomes, is the way I will determine whether there is evidence of p-hacking or data mining. While another strategy involves comparing p-curves over time, to see if there is evidence of more clustering around significant values such as 0.001, 0.025, and 0.05 over time, given the limitations of a small sample size (96 data points) and a timeframe of only 16 years such changes may be hard to observe.

[50] By "suitable" article, I mean one suitable for P-curve analysis. This means for an article to be suitable it needed to at the base-level to have at least one P-value for the primary variable of interest or to have a 95% CI and SS that made calculating a P-value possible.

Section 4: Summary of Data

I generated my dataset in excel. For each of the 96 articles, I generated variables that identify the sample and my variables of interest.

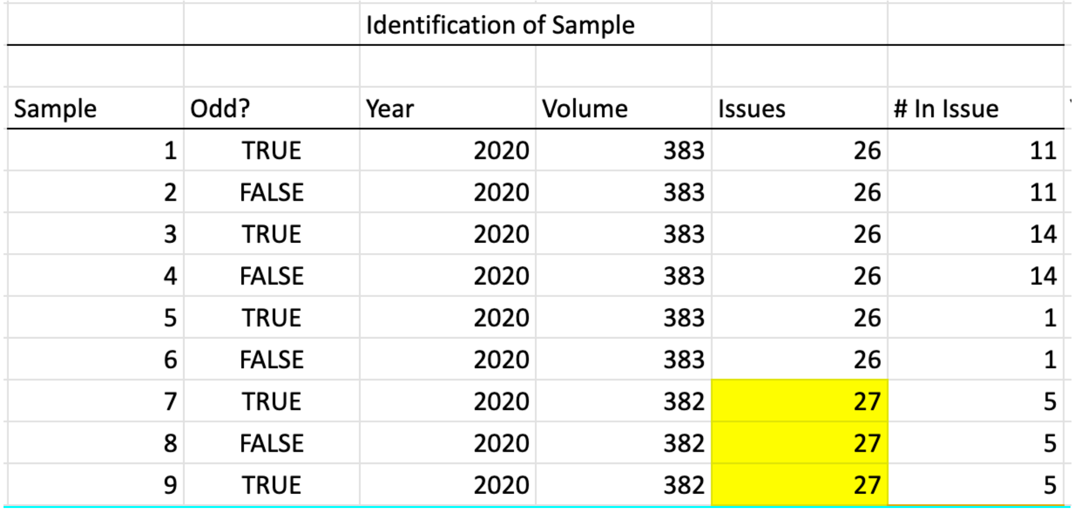
- Identification Variables: {Year, Volume, Issues in Volume, Number in Issue, Author, DOI, Title}
- Variables of Interest
    - Name of Primary Variable of Interest (PVOI)
        - {Exact P-value (PVOI), dummy P<0.005, dummy P<0.025, dummy P<0.001}[51]
    - Name of Secondary Variable of Interest (SVOI)
        - {Exact P-value (SVOI), dummy P<0.005, dummy P<0.025, dummy P<0.001}[52]

There were some articles included in my analysis that didn't list p-values directly. If that was the case, I collected data on the 95% confidence interval on the difference between the treatment and control group, as well as sample sizes, for the treatment and control group. Using this information, I was able to calculate the standard deviation, and Z-score, which ultimately allowed me to solve for P-values using an online calculator. I only collected this information on 11 of the 96 studies (11.4%). This is because the other studies I analyzed directly showed the p-values. This also only occurred in the year 2020, because after calculating this information in the year 2020, I filtered out all data which didn't directly state the p-value for primary and secondary variables of interest. Just to give the reader a clear picture of what my data, looked like I have attached the following figures for samples 1-9.

[51] The variable "dummy P<0.005," is a dummy variable as to whether statistical significance was achieved at the 5% level. (i.e. a one is given if P<0.05, and if not a zero is given.) Same goes for dummy P<0.0025, and dummy P<0.001, except at the 2.5% and 1% significance level.

[52] Ibid. It is also worth noting that I thought the dummy variables would be useful in my results sections, but I ended up not utilizing the data much. This makes it essentially redundant with my P-value, so there is really only one variable of interest.

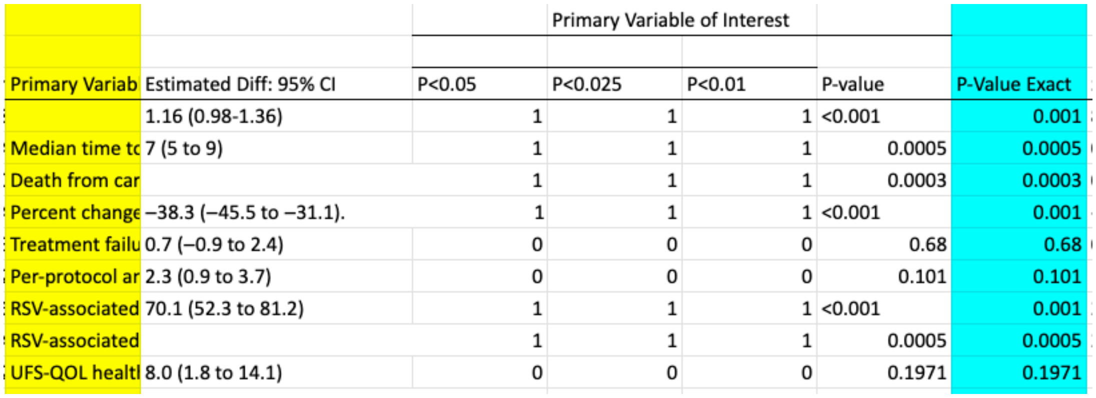
Figure 5: Identification of Sample



		Identification of Sample			
Sample	Odd?	Year	Volume	Issues	# In Issue
1	TRUE	2020	383	26	11
2	FALSE	2020	383	26	11
3	TRUE	2020	383	26	14
4	FALSE	2020	383	26	14
5	TRUE	2020	383	26	1
6	FALSE	2020	383	26	1
7	TRUE	2020	382	27	5
8	FALSE	2020	382	27	5
9	TRUE	2020	382	27	5

*Note: Identification also included Author, DOI, and title, but for spacing purposes I left it out of Figure 5.

Figure 6: Data of Sample (For Primary Variable of Interest)



			Primary Variable of Interest			
Primary Variab	Estimated Diff: 95% CI	P<0.05	P<0.025	P<0.01	P-value	P-Value Exact
	1.16 (0.98-1.36)	1	1	1	<0.001	0.001
Median time tc	7 (5 to 9)	1	1	1	0.0005	0.0005
Death from car		1	1	1	0.0003	0.0003
Percent change	−38.3 (−45.5 to −31.1).	1	1	1	<0.001	0.001
Treatment failu	0.7 (−0.9 to 2.4)	0	0	0	0.68	0.68
Per-protocol ar	2.3 (0.9 to 3.7)	0	0	0	0.101	0.101
RSV-associated	70.1 (52.3 to 81.2)	1	1	1	<0.001	0.001
RSV-associated		1	1	1	0.0005	0.0005
UFS-QOL healtl	8.0 (1.8 to 14.1)	0	0	0	0.1971	0.1971

*Please note, I collected the same data for secondary variable of interest. To see the full dataset, feel free to reach out to smsd2017@mymail.pomona.edu to ask for viewing permission.[53]

In summary, I was able to generate 190 p-values (data points) for the primary variable of interest. However, I was only able to generate 143 p-values for the secondary variables of interest. This suggests, that of the 190 studies I analyzed with a primary variable of interest and p-value **75.26%** had a secondary variable of interest. This makes intuitive sense, as most studies with a primary variable of interest (especially in clinical trials, or randomized experiments) have secondary

[53] The P-value column in my chart, reflects the p-values listed in the studies, whereas the p-value exact changes P<0.001 to P=0.001 to allow for analysis with bar-charts in excel.

variables of interest. In some cases, studies had one primary variable of interest (PVOI) but more than one secondary variable of interest (SVOI). If that was the case, for purposes of continuity I only listed one (SVOI) but chose to list the first-one listed on summary statistics, to avoid any subtle data-manipulation or p-hacking on my part.

Figure 7: Summary Statistics

	Number of Observations	Number of Studies P<0.01	Number of Studies P<0.025	Number of Studies P<0.05
Primary Variable of Interest	190	99	108	135
Secondary Variable of Interest	143	41	55	69

Section 5: Results

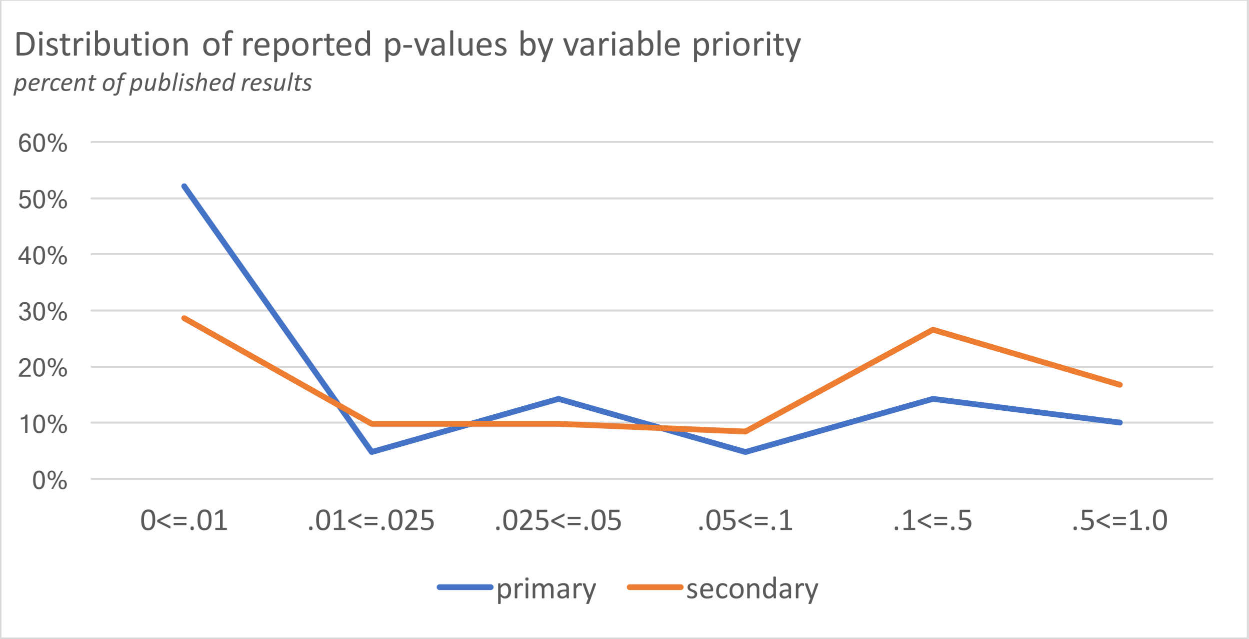
In analyzing results, with my P-curve approach. it is important to keep in mind the two critical findings I suggested in my previous sections may be suggestive of p-hacking, data mining, or broader tinkering.

1) A significant difference between the primary and secondary p-values distributions: particularly if the primary "p-value," clumps around statistically significant values such as 0.01, 0.025, and 0.05.
2) Whether the rate of "clumping," around statistically significance thresholds increase over time. (Is more common in more recent years)

The first point, is particularly important because in clinical trials, drug development, and randomized trials, primary variables of interest play a critical role in approval or acceptance into the journal. Whereas secondary variables of interest, have no such effect but given a tendency to measure relatively similar phenomena should have an approximately similar P-curve. Clumping around "statistically significant variables" would therefore suggest, a likelihood of foul play, data-mining, or p-hacking. Overall, the evidence in my paper, suggests that there is both a *meaningful* difference in the overall shape of the p-curve of primary and secondary variables of interest and that primary variables of interest are *slightly* more likely to "clump," around thresholds of significance. Such data is suggestive of potential p-hacking and data-mining but not conclusively so.

The second point is interesting because there is a fair amount of evidence to suggest that the phenomenon of p-hacking and data-mining has accelerated in recent years. However, given the limitations of my data (i.e. only comparing four time periods that are relatively close), 2020 & 2019 vs. 2015 & 2014, vs. 2010 & 2009, vs. 2006 & 2005, and that every two years only contain a potential of 48 data-points, I went into the experiment with a great deal of caution. This is because with a relatively small sample size of 190, it would be hard to determine that any p-curves over time are representative of broader shifts and not random noise. I have included a few graphs here in case the reader is interested, but I would caution that I think this analysis is preliminary and a relatively small sample size to garner any definitive conclusions about whether the phenomena of p-hacking have increased over time. However, I think there is enough robust evidence to suggest that primary variables of interest have a larger *right-skew* then secondary variables of interest, suggesting higher statistical power.

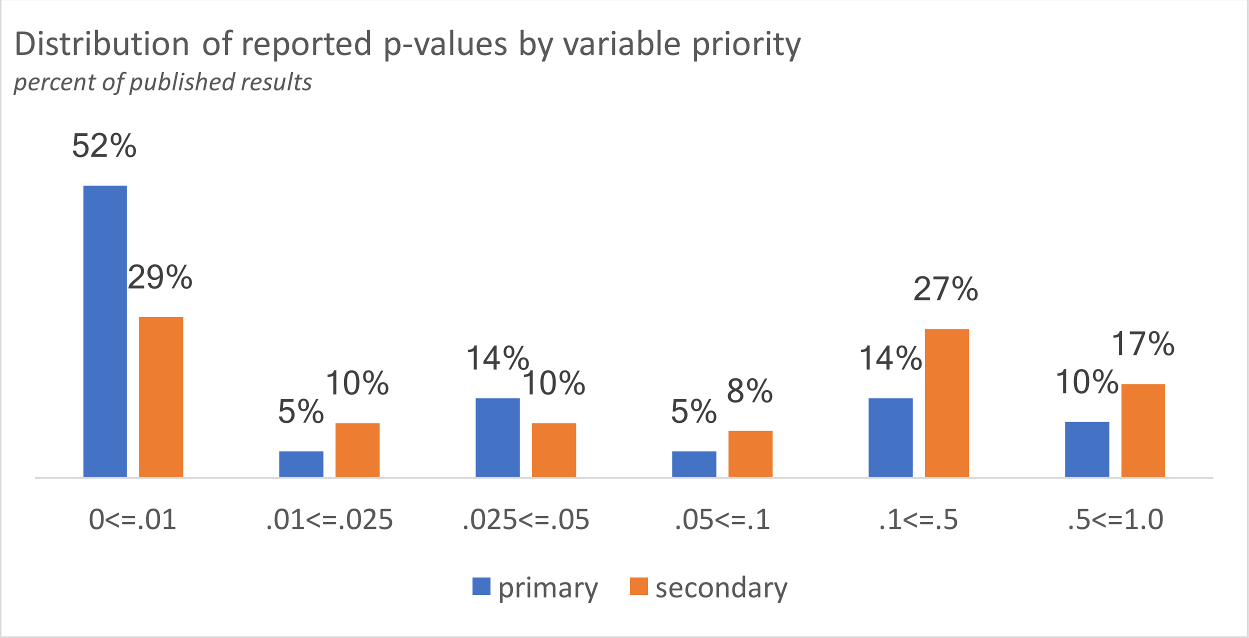
Chart 1: Distribution of reported p-values (line-chart)



* Distribution is such that 0.025< = p < 0.05 (same for every chart)

*It's also worth noting that by almost doubling the number of observations between drafts, the two-sample t test with the null hypothesis that the primary and secondary distributions should be the same has increased to 3.03. This is higher than previously, suggesting a real difference between PVOI and SVOI. It's also interesting that the percentage of PVOI between 0.025 and 0.05 is almost 3 times that between .01 and 0.025 whereas such a phenomenon doesn't exist with the SVOI. This is suggestive that there may be some P-hacking around the 5% threshold for the PVOI.

Chart 2: Distribution of reported p-values (bar-chart)



*Charts 1-2 show the p-values of primary and secondary variables of interest. I have classified p-values for both primary and secondary variables of interest, into the six categories shown above. At the highest level, it's worth noting that primary variables of interest, are associated with higher statistical significance than secondary variables of interest. Indeed, 71% of primary variables of interest have a P-value <0.05 as compared to 57% for secondary variables of interest. This may be suggestive of a) P-hacking/data mining or b) High statistical power of primary variables of interest. It's worth noting that it's slightly difficult to separate out these effects, and simply having primary variables of interest have higher statistical significance, does not in-itself necessarily prove foul play. Instead, what we would need to show would be abnormalities in the larger P-curve.

Chart 3: Clustering of P-values

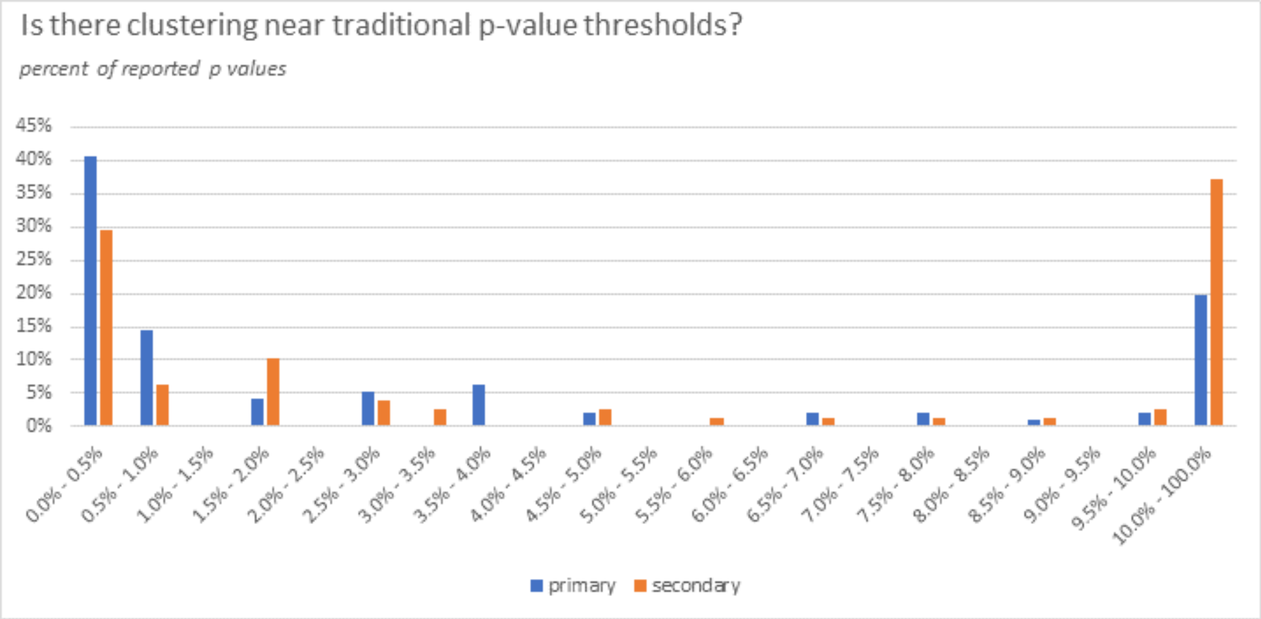
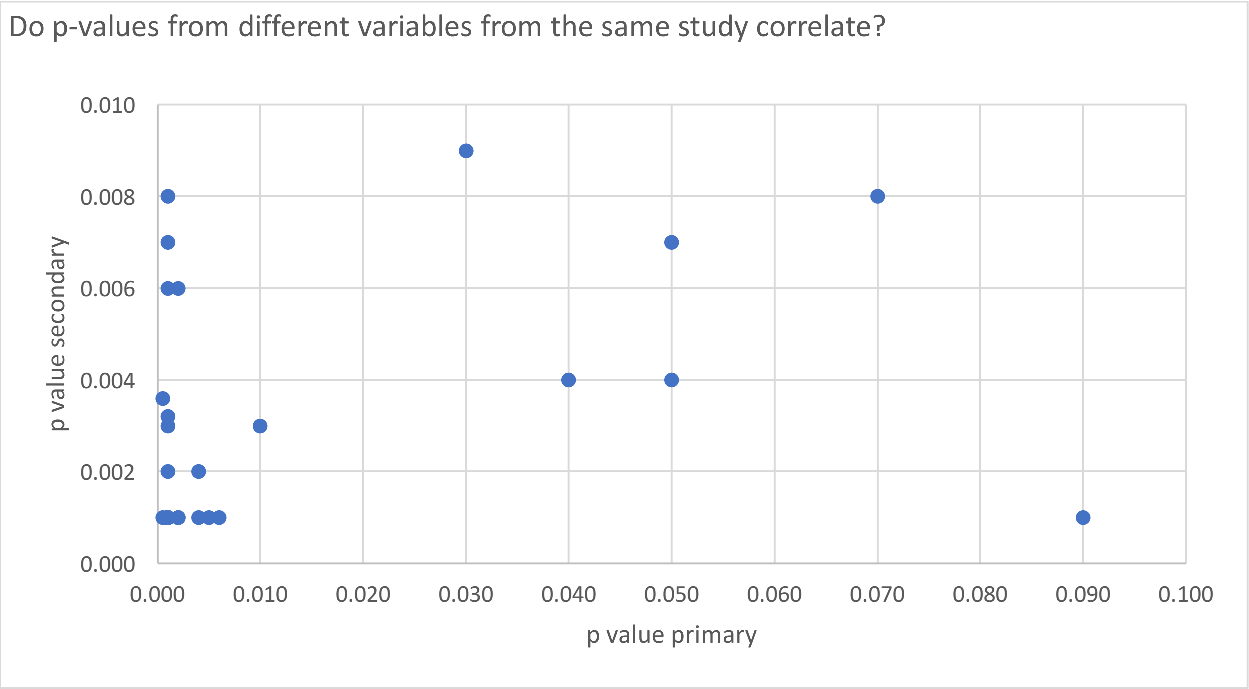


*Chart 3 attempts to better isolate whether the discrepancy between the primary and secondary variables of interest are caused by a) Differences in observational power and statistical significance of two groups or b) P-hacking and data-mining. I will note that at the significant threshold of P=0.05 and P=0.025, there seems to be a slight difference between the primary and secondary variables of interest. This suggests that at two of the pre-determined critical P values there is no conclusively data of clumping in primary values. However, it is worth noting that over 40% of the primary variable of interest, has a P-value of less than 0.005, as opposed to approximately 30% for the secondary variable of interest. This relative increase of almost a third-could be suggestive of data-mining or tinkering but generally speaking this effect, also coincides with increased statistical power of the primary variables of interest. Overall, this suggests very little evidence of data-tinkering and p-hacking.

Chart 4: Are primary and secondary p-values from the study different?



This chart is suggestive of many of the phenomena previously mentioned. Mainly that within the same studies, the P-values of the secondary variable of interest are higher than that of the primary variable. Indeed, the equation for the line of best fit (LOB) is as follows:

$$Y = 0.4102x + 0.1542$$

Chart 5: Another version of distribution of P-values

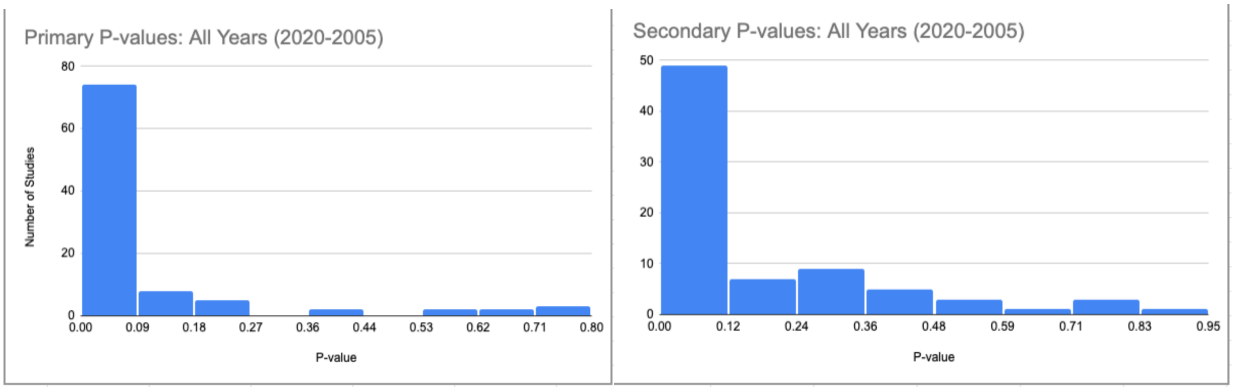


Chart five, is suggestive of the same data as chart 1-2 however, without putting the primary and secondary p-values into uneven bucket-sizes. However, they still need to be edited so both the P-value buckets are the same size, and small enough so that any abnormalities around 0.01, and 0.025 and 0.05 could be observed.

Chart 6: A tale of Time

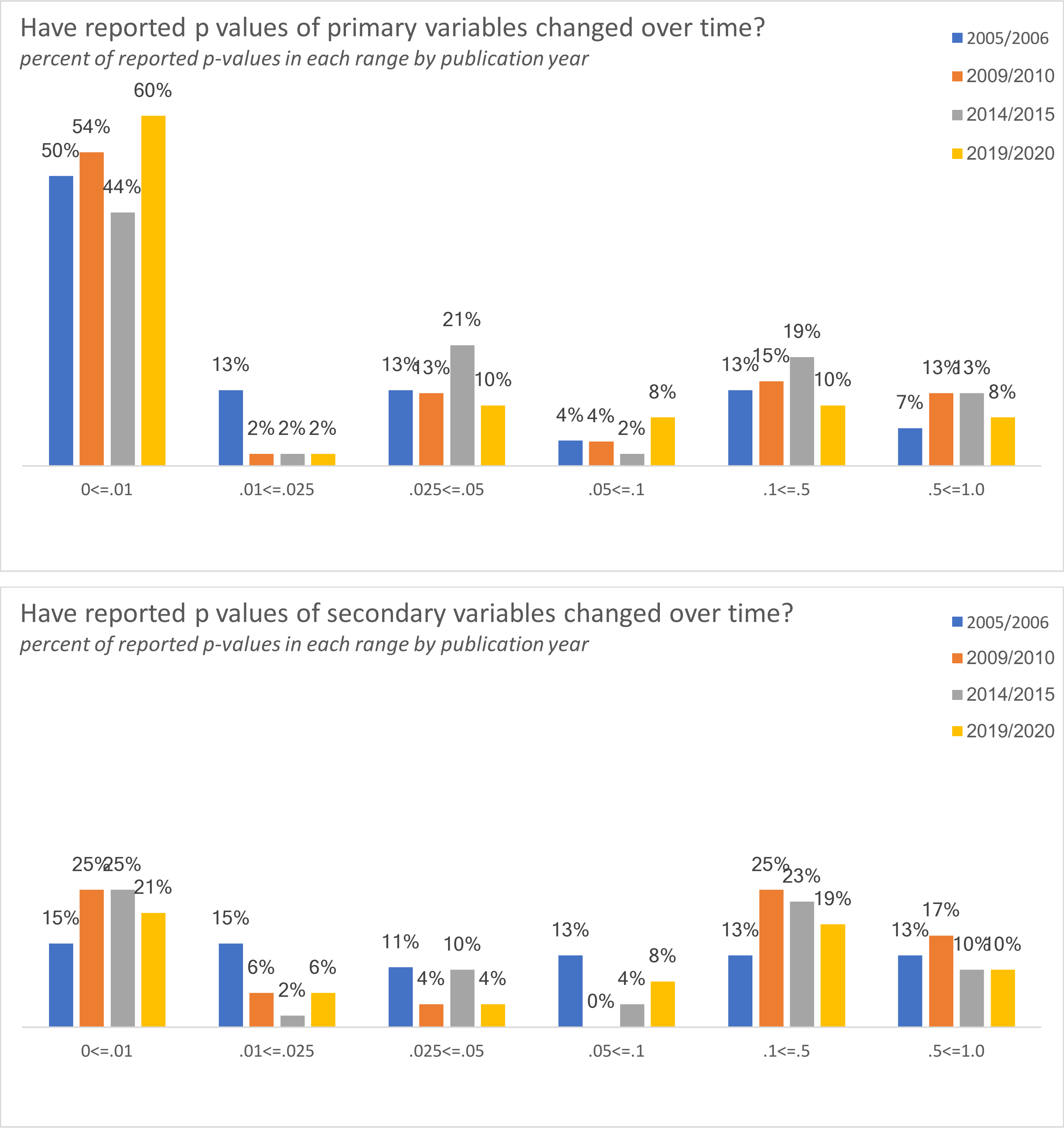


Chart 6 is a messy chart and as previously mentioned I think the analysis of determining whether the prevalence of p-hacking or clustering of variables *increases* over time may be a bit of a fool's errand. With that being said at least in the primary variables of interest there seems to be more

P-values towards the higher in 2006 & 2005, as well as opposed to the other years. Indeed, the percentage of P-values <0.001 is 60% in 2019 & 2020, the highest percentage for PVOI observed in the sample, whereas such a phenomenon does not occur SVOI. So perhaps this is a noisy way of suggesting, that it may be the case that average p-value *decreases* over time, but that is by no means conclusively demonstrated by this analysis. Furthermore, even if that were the case, this could be for a variety of reasons besides for p-hacking, and data mining.

Section 6: Conclusion

Utilizing a P-curve approach I hope to determine whether there is evidence of data-mining and p-hacking. I attempted to do this by comparing primary variables of interest to secondary variables of interest, as statistical significance of PVOIs is essential to publication but the same is not true for SVOIs. In conclusion, I found that PVOI tends to have lower p-values than SVOI. However, it is not clear if this is due to the influence of P-hacking and data-mining as opposed, to simply increased power of primary variables of interest. Based on a sample of 190 data points, it seems that the first effect is certainty true but evidence surrounding the second claim is mixed.

In terms of the question of clumping, I find some evidence that at the significant thresholds of 0.025 and 0.05 the primary variable of interest is more likely than the secondary variables of interest. Indeed, as Chart 2 shows in the PVOI there is a “kink” between 0.025 and 0.05, with 14% of the distribution occurring there as opposed to 5% between 0.01 and 0.025. It’s worth noting that in the SVOI there is no such kink exists, with each segment containing 10% of the distribution. Given that any deviation from the ideal p-curve may be suggestive of p-hacking this suggests there may be some evidence for p-hacking.

While I did find evidence that a significantly higher percentage of PVOI has p-values of <0.01 as opposed to SVOI, this may be suggestive of potential p-hacking and data-bias. With that being said, I am skeptical of this claim as such a result would coincide with the natural increased right-skewing nature of the PVOI given increased statistical impact as opposed to SVOI. The question of why for the same studies PVOI tends to be more statistically significant than SVOI is still open to question and debate. While it may be due to fraud or manipulation, I am skeptical of this case, and instead, think it may be for other benign reasons. For example, it could be that researchers are more confident in the PVOI and test a variety of SVOI to understand side-effects or related causes to a treatment. With that being said, there were a few samples in my observation, were papers would test multiple PVOI and post-analytical only reference the significant finding in the introduction or results section. This was very rare, but could be suggestive of data-tinkering.

In conclusion, by doubling my data-set to 190 data points, I was able to conclude that the PVOI are more right-skewed than the SVOI, suggesting higher empirical power. The question of P-hacking is slightly more complicated, but there is some preliminary evidence to suggests that there is some tinkering around the 2.5% to 5% given the differing distribution of the PVOI and SVOI in that threshold. It's worth noting that my data-set is too small to determine whether the effect of P-hacking and data-mining increases over time.

Works Cited

Aschwanden, Christie. "You can't trust what you read about nutrition," *Fivethirtyeight,* 2016.

Bastardi, A, Uhlmann. EL, and Ross, L. "Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence." *National Library of Medicine,* 2011

Begley, C. Glenn, and Lee M. Ellis. "Raise Standards for Preclinical Cancer Research." *Nature News*, Nature Publishing Group, 28 Mar. 2012. www.nature.com/articles/483531a.

Benson, Kjell and Hartz, J. Arthur. "A comparison of Observational Studies and Randomized Controlled Trials N Engl J Med 2000; 342:1878-1886
DOI: 10.1056/NEJM200006223422506

Costs by Therapeutic Area." *ASPE,* February 2017, aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development/31-costs-therapeutic-area.

Dean, Tim. "How do we Edit Science part 2, Significance Testing, P-hacking and Peer Review," *The Conversation.*

Eggertson, Laura. "Lancet Retracts 12-Year-Old Article Linking Autism to MMR Vaccines." *CMAJ : Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne*, Canadian Medical Association, 9 Mar. 2010, www.ncbi.nlm.nih.gov/pmc/articles/PMC2831678/.

Fanelli D (2009) How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLoS ONE 4(5): e5738.
https://doi.org/10.1371/journal.pone.0005738, direct quote.

Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications [published correction appears in Proc Natl Acad Sci U S A. 2013 Jan 15;110(3):1137]. *Proc Natl Acad Sci U S A.* 2012;109(42):17028-17033.
doi:10.1073/pnas.1212247109

Gadbury GL, Allison DB. "Inappropriate fiddling with statistical analyses to obtain a desirable p-value: Tests to detect its presence in published literature." *PLOS ONE*, 2014.

Gilbert, Daniel T., et al. "Estimating the Reproducibility of Psychological Science." *Open Science Collaboration, American* Association for the Advancement of Science, 2015.

Glick, J. Leslie. "Scientific data audit—A key management tool, Accountability in Research."
1992. 2:3, 153-168, DOI: 10.1080/08989629208573811

Grimes David Robert, Brauch Chris T. and Ioannidis John P.A, "Modeling Science Trustworthiness under Publish or Perish," *R. Soc. Open Sci: 5:* 171511 https://doi.org/10.1098/rsos.171511

Head M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. “The extent and consequences of *p*-hacking in science.” *PLOS Biology,* 2015. *13*(3): e1002106. doi:10.1371/journal.pbio.1002106

Hutchinson, L., Kirk, R. “High drug attrition rates—where are we going wrong?.” *Nature Reviews Clinical Oncology.* **8,** 189–190 (2011). https://doi.org/10.1038/nrclinonc.2011.34

Ioannidis JPA. “Why most published research findings are false.” *PLOS Med,* 2015

Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294(2):218-228. doi:10.1001/jama.294.2.218

Kerr, N. L. “HARKing: Hypothesizing after the results are known.” *Personality and Social Psychology Review, 2*(3), 1998. 196-217. doi:10.1207/s15327957pspr0203_4

Macleod, Malcolm. “Biomedical Research: increasing value, reducing waste,” *Lancet (London, England)*, U.S. National Library of Medicine, 2014. pubmed.ncbi.nlm.nih.gov/24411643/.

Martinson BC, Anderson MS, de Vries R. 2005. Scientists behaving badly. *Nature* 435, 737–738. (doi:10.1038/435737a)

Rosenthal, R. “The file drawer problem and tolerance for null results. *Psychological Bulletin,* 86: 638–641.

Richards, Neil M. and King, Jonathan, Three Paradoxes of Big Data (September 3, 2013). 66 Stanford Law Review Online 41 (2013). Available at SSRN: https://ssrn.com/abstract=2325537