A Thorough Analysis on the Misuse of Statistics: How Not to Be an A**hole

David Pham, Liza Bolton

Abstract

This educative piece serves to discuss the misuse and abuse of statistics in the scientific community from multiple perspectives and disciplinary fields. Upon observing some meta-analysis studies on past published research papers, it was found that the three most common errors made in academia are: null hypothesis testing and p-values, inflation of Type I error, and transparency with data. Furthermore, we explore retracted research papers that committed one or more of the above errors and discuss solutions to implement moving forward. These recommendations are made primarily towards researchers and journals, but all statisticians can learn from these mistakes and find ways to spread statistical awareness to their communities.

Keywords: p-value, type I error, transparency, statistical literacy, ethics, significance

Introduction

As a general, yet philosophical introduction, the main goal of a statistician is to measure/quantify uncertainty and try to help make the world a better place. Although science is an objective field, it is also impossible for it to be correct all the time. Statistics is not a field that can predict the future, but merely a practice that can help us infer hypotheses based on existing data. Despite all this, it is still our utmost responsability to ensure that we do science cautiously and accurately. Since the scientific method and existing research build upon themselves, it is crucial that the literature, peer-reviewing phase, and every step of the pyramid are done right to ensure progress. Upon closer inspection, this seems easier said than done because of many factors that impede scientists from producing good statistics. Hurdles such as greed, publication output, funding, fame, and influence block out any warnings of unforeseen consequences, allowing researchers to engage in statistical misconduct. After looking at meta-analysis studies from the late 1900's, as well as retracted papers from this century, we can summarize the three most common culprits in statistical abuse: inappropriate use of null hypothesis testing and p-values, inflation of type I error, and transparency in data. In the next few sections, we shall discuss each misuse, provide some examples and cautionary tales regarding each topic, and some possible remedies.

Inappropriate Use of Null Hypothesis Testing and p-values

P-values are one of the most fundamental but misconstrued measures in statistics. As a reminder, a **p-value** is the probability of seeing a test statistic as extreme as the one you just witnessed in your analysis under the assumption of the null hypothesis. It is *not* the prob-

ability that the null hypothesis is true, nor is it the probability of witnessing a false positive. This is quite a textbook definition, but a gentle reminder is always helpful. There are many papers and resources dedicated to explaining these misconceptions (for example, a paper from Goodman (2008)), but the most common errors in academia persist of very simple mistakes. For example, when you fail to have your p-value below your predetermined significance level, it does *not* mean that you accept your null hypothesis and conclude that there is no effect. The only thing this shows is that your experiment/analysis failed to detect a significant effect and more work needs to be done. Another reckless mistake is having multiple significance levels throughout your experiment. It is imperative that before you conduct an analysis, you decide on **one** hypothesis and **one** a-priori significance level (α) that you will be comparing your p-values to for the rest of the analysis. Dar, Serlin, and Omer (1994) analyzed three decades worth of psychotherapy research and discovered that an overwhelming amount of papers had multiple significance thresholds throughout their research (Figure 1).

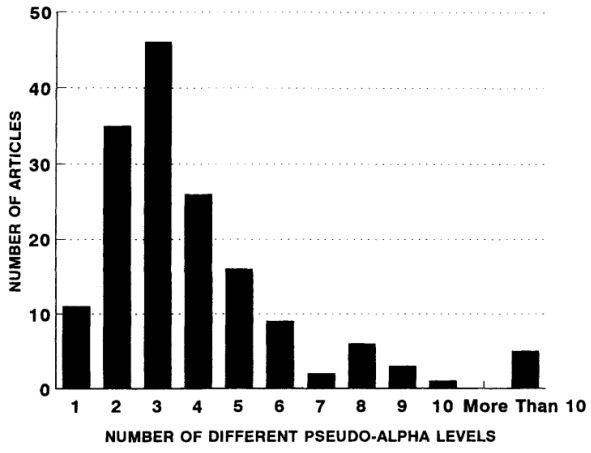

Figure 1: A bar graph from Dar, Serlin, and Omer (1994) showcasing the distribution of pseudo-alpha levels per given paper.

It does not make much sense to just change your choice of α because you can modify it after-the-fact if the significance level does not coincide with a p-value that is larger than you had expected. Furthermore, Dar, Serlin, and Omer (1994) found that some researchers reported p-values that were very close to their significance level (but not quite below it) as "borderline significant" or "a strong trend towards a difference". However, all that p-value really means is just the probability of observing that result by chance under the assumption that the null hypothesis is true. Sadly, the only trend that was found here was a trend of ignorance.

Inflation of Type I Error Rate

A Type I error, also known as a false positive, is when you claim statistical significance on a result and reject the null hypothesis; but in reality, there was no effect. Multiple meta-analyses on various multidisciplinary fields, including papers from Schatz et al. (2005) and Dar, Serlin, and Omer (1994), had revealed that many articles from neuropsychology and psychotherapy respectively had performed multiple statistical tests "in the absence of corrections to the P-value or compensatory use of multivariate analyses". When running multiple experiments to test a hypothesis, it is important to adjust for your p-values to reduce the chance of establishing false significance.

Transparency

Although "transparency" seems like a very loose and simple definition, it is often violated the most and in a variety of different ways. Data exploration is a section that can be exploited behind the scenes very easily, especially if researchers do not allow the public to access this data. Modifying, omitting,

and misinterpreting data are all common occurrences in research and can be often hard to detect since this makes the work less reproducible. As an example, Walach, Klement, and Aukema (2021) had released a very controversial paper arguing that the COVID-19 vaccine should be used more "sparingly", and to reconsider the resources needed to develop the vaccine because it is causally related to adverse effects. From Office (2021), the data used in the retracted paper (i.e., data from the Lareb report in the Netherlands) were used to "calculate the number of severe and fatal side effects per 100,000 vaccinations". The problem from Walach, Klement, and Aukema (2021) is that they assumed these side effects were directly caused by the vaccine. However, this was not the case; healthcare professionals and patients are "invited" to report suspicions of adverse side effects that may or may not be related to getting vaccinated. This led to erroneous conclusions made by the researchers and is a prime example of the correlation $\neq$ causation phenomenon. Some other shady practices done behind the scenes also include p-hacking and HARKing (Hypothesis After Results are Known), which will be covered in the next example.

A Cautionary Tale: Brian Wansink

Finally, there is one last story that needs to be told when discussing statistical misconduct. While it is perfectly acceptable to make a mistake, acknowledge it and correct it, it is beyond unethical to deceive and make *hundreds* of errors to publish statistically significant results at the expense of the credibility of science.

Brian Wansink was a former research at Cornell University who specialized in nutrition psychology and consumer behavior. He has been cited thousands of times and is praised for many of his contributions in food science, as well as in diet and nutrition. Notably, he was very well-received by mainstream media. Wansink was admired for findings that included: people eating less when the serving size/plate is smaller despite the amount of food being kept constant, branding affecting taste and perception, and the amount you eat at a buffet varies depending on a wide variety of factors (including the amount of money you pay). These discoveries have even led to reduced serving sizes and to Wansink helping develop the U.S dietary guidelines in 2010. From this description, it seems like Wansink was a hero in the world of science. What happened? Well, it turns out that a lot of statistical abuse and misconduct were performed behind the scenes to generate these results.

Many of Wansink's research papers had been retracted due to many erroneous conclusions that were supported by irreproducible data and many inconsistencies in the analyses. More specifically, after some heavy skepticism by the scientific community regarding his papers on various relationships from eating at a buffet, an investigation had been conducted by the university to indeed discover that Wansink and the Food and Brand Lab had conducted statistical fraud. At face value, Wansink was guilty of: selection bias, HARKing, p-hacking, publishing claims about children ages 8-11 when the collected data was actually about toddlers, and manipulating data in order to get desirable results (Lee (2017)).

Stephanie Lee, a journalist for the Center for Health Journalism and Buzzfeed, had filed public record requests to take a look at exchanged emails between Wansink's research team. The most appalling email was a message sent from Wansink to his research student at the time, Özge Siğirci from Turkey. S. M. Lee (n.d.) shows an email about Wansink being certain that there was a relationship

in the data and that Özge's initial look was not sufficient. He recommended her to "cut the data and analyze subsets of it" to see if any results come out as significant, and even *provides* her the different groups to partition. Finally, Wansink concludes that it is fundamental to come with results as it will help her "stand out a bit" and increases the "likelihood of (...) getting something publishable" out of her visit.

It is quite baffling that an established researcher would tell this to a student, yet alone condone/perform these practices. As a first start, this is a blatant example of **p-hacking** where you run multiple experiments and make multiple measurements on different subsets of data to generate a statistically significant result. This leads to a much greater chance of committing a Type I error, since the odds of running into a false positive increase due to every experiment being run (where each independent experiment can be a false positive). Furthermore, the report from Zee, Anaya, and Brown (2017) that dissects four of Wansink's articles using the same buffet data (which Wansink also failed to be transparent about since he had claimed they were independent studies) finds **150** combined errors in the papers. Some of these errors include inconsistent sample sizes throughout the papers, vague and misleading wordings, and even trivial calculations in most of Wansink's tables and figures. As an example, the table below (Figure 2) summarizes very basic descriptive measures of the age, height, and weight of all participants.

We have no reason to believe that both groups have equal variances since the data collection process was not even mentioned; hence, we use a simple Welch's t-test to calculate these test statistics (Gelman (2017)):

Table 1

Demographics	\$4 (n = 43)	\$8 (n = 52)	*t*
Age (years)	43.67 (18.50)	44.55 (14.30)	0.25
Height (inches)	68.65 (3.67)	66.51 (9.44)	1.38
Weight (pounds)	184.83 (63.70)	178.38 (45.71)	0.52

Figure 2: Zee, Anaya, and Brown (2017) highlighting a table from Sığirci and Wansink (2015) regarding miscalculations of a few t-statistics.

$$t_{age} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{43.67 - 44.55}{\sqrt{\frac{18.5^2}{43} + \frac{14.3^2}{52}}} = -0.2552$$

$$t_{height} = \frac{68.65 - 66.51}{\sqrt{\frac{3.67^2}{43} + \frac{9.44^2}{52}}} = 1.503114$$

$$t_{weight} = \frac{184.83 - 178.38}{\sqrt{\frac{63.70^2}{43} + \frac{45.71^2}{52}}} = 0.556064$$

Despite being small mistakes, even things that can be computed automatically are not calculated correctly... and there are seven more tables with similar inconsistencies, as highlighted in the appendix of Zee, Anaya, and Brown (2017). Wansink's reckless behaviour and irreproducibility crisis had tarnished his reputation and academic career. But is there a deeper reason on why Wansink (and many other researchers around the world) fall into this pithole of statistical abuse and misconduct?

Why People are A**holes

In science, it seems as if scientists are in a *publish or perish* situation in such a competitive environment. Getting published unlocks the key to many achievements and resources

in life such as a credible reputation, funding, influence, and to even stay in academia as the bare minimum. To get to that step however, you must have an interesting hypothesis with interesting data and *statistically significant* results. Sadly, publishers are very hesitant and hence less likely to review/publish research that did not find anything interesting. This directly contributes to the file drawer problem, where results that do not coincide with the researcher's original hypothesis ends up in a stash, nowhere to be seen ever again. Therefore, discussing our failed work and thought process in our research are things we should not be against in academia. To elaborate, Wansink publicly admitted to selection bias as he had claimed that there were *multiple* attempts at finding a statistically significant result (Gelman (2017)). There was an initial "Plan A" that did not work during the data analysis phase, so Wansink had to move on to plans B, C, D, and so on. However, nowhere in his sequence of papers did he mention this "first plan" and what it was about. He was dire to publish *something*, even it just happened to be all noise.

Hence, greed and publication output are two real factors that lead researchers to publishing nonsense and to claim significance just for the sake of claiming significance. Using Wansink again as an example, his avarice led his research student to treading the same path as he did and producing faulty statistics. It is quite unfortunate that someone flew all the way from the other side of the planet just to engage in statistical misconduct. In general, when it comes to data collection, the process can be quite tedious and expensive; therefore, people might be inclined to obtain *something* from their data. But are researchers and scientists the only ones at fault?

In order for a paper to get published, it must go through an "extensive" peer-reviewing process by a journal. People (i.e., other researchers that work in the same field) will analyze your research and determine if it meets high quality standards set by the institution, as well as science. This entails building upon other work in the field, having a strong theory with well-gathered data, and having claims that are logically backed up with *evidence*, reproducible, and replicable (Gelman (n.d.a)). Unfortunately, we cannot get science right all the time. We are not expected to come up with a solution or discovery that will predict the future, but we *are* expected to conduct science at the highest caliber and ensure that we do it right. This prestige and trust is also built off of the journals and peer-reviewers; nevertheless, it seems all too common that a lot of bad papers slip through the cracks and end up getting published.

To illustrate, Hussain et al. (2020) had written a report concluding that patients with obesity had a higher risk of mortality from COVID-19. However, the article was later withdrawn by the *authors*, not the journal, due to incorrect calculations of the odds ratio for age groups and other inaccuracies in their plots about different patient groups. According to the retraction notice from the journal Hussain et al. (2021), these errors were "unfortunately passed unnoticed during the extremely rapid review and publication process at the peak of the COVID-19 pandemic". Does a busy time period necessarily excuse the editorial board from letting any claims go through the peer-reviewing process? In fact, it is arguable that these peaks should be the prime time to *not* let misinformation and outlandish results without proper statistical analysis slide. These are exactly the moments where we should be careful, because the credibility of science and the outlook of the world depend on us.

Another interesting point to note is that most

of the retracted papers (and more specifically, a good chunk of Wansink's work) from this educative piece were from JAMA, the Journal of the American Medical Association. You can argue that this observation was obtained purely by chance, but it is quite fishy how all of Wansink's work had passed the peer-review phase without any problems brought up by the journal, despite the emporium of errors brought up by Gelman (n.d.b), Zee, Anaya, and Brown (2017), S. Lee (n.d.), and many other critics reviewing his other dozen papers. Nevertheless, journals, peer-reviewers, researchers, and everybody across the board should do better to ensure that science is being done in good faith.

How Not to be an A**hole

There appear to be a lot more problems than we had initially thought of. What can scientists, journals, as well as ordinary statisticians such as ourselves do to alleviate this problem?

Despite the title and overall theme of this paper, we should not bash any of these researchers and journals for their mistakes (no matter how reckless some of them were). Sure, their actions are beyond deplorable; however, we will not be able to move forward if all we do is continue to chastise them. Bad statistics can be produced by good people with initially good intentions. What we can hope is that these people and institutions learn something from these mistakes and continue to use their energy and creativity to make the world a better place (Gelman (2017)).

Before that however, learning to admit when you are wrong is a fundamental human virtue, but not many researchers have yet acquired it. Walach and Wansink are prime examples of this, as they are not quite satisfied with the retractions of their articles. Office (2021) had reached out to Walach and the other authors to respond to the claims of misinterpreting the data as causal, but "were not able to do so satisfactorily". Furthermore, they also did not agree to the retraction. Lastly (but certainly not least), Wansink does not believe that he has done anything inappropriate in his statistical work. He stands by his work and sends a statement to S. M. Lee (n.d.) saying that: "I stand by and am immensely proud of the work done here at the Lab. (...) The Food and Brand Lab does not use 'low-quality data', nor does it seek to publish 'subpar studies'." When there are dozens of researchers around the world criticizing your work (this educative piece included), and there is a 20 page paper discovering over *150* errors in only *four* of your studies, it takes a lot of ignorance to not even acknowledge the feedback and turn a blind eye to the whole situation. In summary, it is perfectly okay to mess up; but after that, *acknowledge* that you made a mistake, and take the correct steps to fix it. But even as Gelman (2017) puts it, **honesty and transparency are not enough**. In academia, it is imperative that you have a sound hypothesis with high-quality data; otherwise, nothing can save you.

Also, depending on the line of work and field, committing a Type I or Type II error could be extremely devastating and put peoples' lives at risk. Indrayan (2018) illustrates an example of experimenting the efficacy of a new drug or regimen. If statistical and clinical significance are both established, but we later find that there was faulty design, data analysis, or incorrect interpretation of the results, this could lead to "far-reaching implications" on the health of many. Hence, specifically for this problem, Indrayan (2018) recommends starting off by choosing a design process that is appropriate for the experi-

ment at hand and meets all the necessary requirements. Next, selecting an adequate sample size that meets your standard error and power threshold is just as crucial. This helps determine the smallest amount of participants required in order to save resources, and listing this procedure in your paper can promote transparency. Then, when actually conducting the experiment or analysis, if you plan to run it multiple times, you *must* adjust for it. To clarify, Brereton (2019) advocates for the **false discovery rate**, which is a method that automatically inflates your observed p-values based on the number of tests you perform. This in turn will greatly reduce the chance of you committing a Type I error. Furthermore, multiple-comparison procedures such as the Bonferroni Correction (dividing the significance level by the number of tests you decide to run) and Tukey's test (a test statistic that corrects family-wise error rate by dividing the difference of means over the standard error) are both great to control the overall chance of obtaining a false positive (Indrayan (2018)). Another piece of advice by Dar, Serlin, and Omer (1994) is to use p-values descriptively instead of inferentially. This entails using test statistics to summarize results *collected from the data* instead of trying to generalize conclusions for the larger population. Ultimately, this strategy entirely removes the "inferential consequence of statistical significance", but also takes away a lot of power from the p-value (which may or may not be a good thing depending on the eyes of the beholder).

To conclude this section, something that we can all do as statisticians is not to be a**holes. What does that mean exactly? Adhering to the code of ethical statistical practice and avoiding the mistakes discussed in this educative piece are some basic tips; but, the most useful piece of advice (and perhaps the most fun/simple to follow) is to continue spreading awareness on statistical literacy and ethics. Not enough people care about the misuse of statistics in the scientific community and in our daily lives, nor does the general population know enough about statistics to join in on the discussion. It was quite shocking that S. Lee (n.d.) was one of the only journalists that initially cared enough to press Wansink on the matter and brought it to the attention of mainstream media. Without her remarkable journalism, the general public would not have even heard about the news. We should continue to find innovative and creative ways to promote statistics education to the world, especially in a technologically-reliant generation. Examples such as this atrocity bring good fun and are accessible to a large audience.

Conclusion

In all, statistics is a field and an art that is hard to get right. Although science cannot be correct 100% of the time, it is expected to be conducted carefully and accurately. When blinded by the wrong intentions, bad and good people will be more likely to produce bad statistics. Time and time again, we have witnessed the misuse and abuse of null hypothesis testing and p-values, claiming significance just for the sake of claiming significance, inflation of Type I errors, and transparency with one's data. Luckily, researchers have plenty of methods to correct and adjust for these mistakes to greatly diminish the chance of any false positives. Furthermore, we call on the scientific community (journals, peer-reviewers, and scientists alike) to more carefully analyze research papers and scrutinize results wherever possible. We must always raise our doubts about anything unclear or unsound, especially with the data and hypothesis. Finally, *all* statisticians can play a role in educating others about statistical

literacy. Figure 3 highlights that we are all learning together, and that we should uplift others into doing the right thing.

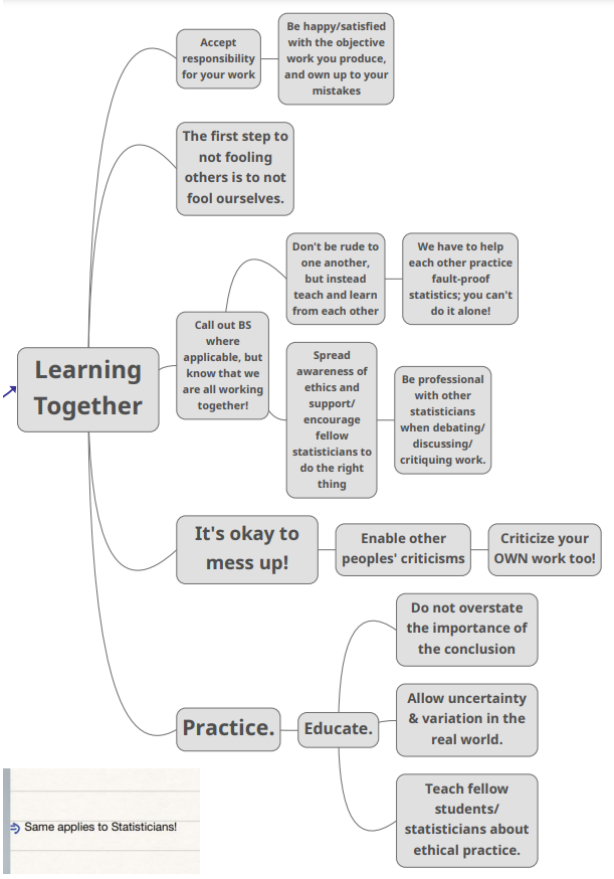

Figure 3: A section of a mind map created by the STA497 team.

Hence, with all this in mind, implementing these considerations will bring us and science a few more steps towards progress in making the world a better place.

Declaration of competiting interest

I declare that I have no conflict of interest.

References

Brereton, Richard G. 2019. "The Use and Misuse of P Values and Related Concepts." *Chemometrics and Intelligent Laboratory Systems* 195 (December): 103884. https://doi.org/10.1016/j.chemolab.2019.103884.

Dar, Reuven, Ronald C. Serlin, and Haim Omer. 1994. "Misuse of Statistical Tests in Three Decades of Psychotherapy Research." *Journal of Consulting and Clinical Psychology* 62 (1): 75–82. https://doi.org/http://dx.doi.org/10.1037/0022-006X.62.1.75.

Gelman, Andrew. n.d.a. "I Fear That Many People Are Drawing the Wrong Lessons from the Wansink Saga, Focusing on Procedural Issues Such as 'P-Hacking' Rather Than Scientifically More Important Concerns About Empty Theory and Hopelessly Noisy Data. If Your Theory Is Weak and Your Data Are Noisy, All the Preregistration in the World Won't Save You. | Statistical Modeling, Causal Inference, and Social Science." https://statmodeling.stat.columbia.edu/2018/03/13/fear-many-people-drawing-wrong-lessons-wansink-saga-focusing-procedural-issues-p-hacking-rather-scientifically-important-concerns-2/.

———. n.d.b. "'Statistical Heartburn: An Attempt to Digest Four Pizza Publications from the Cornell Food and Brand Lab' | Statistical Modeling, Causal Inference, and Social Science." https://statmodeling.stat.columbia.edu/2017/01/25/statist heartburn-attempt-digest-four-pizza-publications-cornell-food-brand-lab/.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve P-Value Misconceptions." *Seminars in Hematology* 45 (3): 135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.

Hussain, Abdulzahra, Kamal Mahawar, Zefeng Xia, Wah Yang, and Shamsi EL-Hasani. 2020. "RETRACTED: Obesity and Mortal-

ity of COVID-19. Meta-analysis." *Obesity Research & Clinical Practice* 14 (4): 295–300. https://doi.org/10.1016/j.orcp.2020.07.002.

———. 2021. "Retraction Notice to Obesity a Nd Mortality of COVID-19.Meta-analysis [Obesity Research & Clinical Practice 14/4 (2020) 295-300]." *Obesity Research & Clinical Practice* 15 (1): 100. https://doi.org/10.1016/j.orcp.2020.12.008.

Indrayan, Abhaya. 2018. "Statistical Fallacies & Errors Can Also Jeopardize Life & Health of Many." *Indian Journal of Medical Research* 148 (6): 677. https://doi.org/10.4103/ijmr.IJMR_853_18.

Lee, Stephanie. n.d. "Emails Showed How a Famous Ivy League Food Lab Was Cooking up Shoddy Data." *Center for Health Journalism*. https://centerforhealthjournalism.org/resources/lessons/emails-showed-how-famous-ivy-league-food-lab-was-cooking-shoddy-data.

Lee, Stephanie M. n.d. "Sliced and Diced: The Inside Story of How an Ivy League Food Scientist Turned Shoddy Data into Viral Studies." *BuzzFeed News*. https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking.

Office, Vaccines Editorial. 2021. "Retraction: Walach et Al. The Safety of COVID-19 VaccinationsWe Should Rethink the Policy. Vaccines 2021, 9, 693." *Vaccines* 9 (7): 729. https://doi.org/10.3390/vaccines9070729.

Schatz, P, K Jay, J Mccomb, and J Mclaughlin. 2005. "Misuse of Statistical Tests in Publications." *Archives of Clinical Neuropsychology* 20 (8): 1053–9. https://doi.org/10.1016/j.acn.2005.06.006.

Siğirci, Özge, and Brian Wansink. 2015. "RETRACTED ARTICLE: Low Prices and High Regret: How Pricing Influences Regret at All-You-Can-Eat Buffets." *BMC Nutrition* 1 (1): 36. https://doi.org/10.1186/s40795-015-0030-x.

Walach, Harald, Rainer J. Klement, and Wouter Aukema. 2021. "The Safety of COVID-19 VaccinationsWe Should Rethink the Policy." *Vaccines* 9 (7): 693. https://doi.org/10.3390/vaccines9070693.

Zee, Tim, Jordan Anaya, and Nicholas J L Brown. 2017. "Statistical Heartburn: An Attempt to Digest Four Pizza Publications from the Cornell Food and Brand Lab." Preprint. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2748v1.