Media Stars: Statistical Significance and Research Impact

Abel Brodeur
University of Ottawa

Nikolai M Cook
Wilfrid Laurier University

Anthony Heyes
University of Birmingham

Taylor Wright
Brock University

December 10, 2024

Abstract

How efficiently do scientific results make their way into the wider world? Applying multiple methods to the universe of hypothesis tests reported in the abstracts of articles in the three leading health journals (British Medical Journal, Lancet, New England Journal of Medicine) between 2016 and 2022 we explore the role of statistical significance as a driver of attention. A research finding with significance that places it marginally inside the arbitrary 5% threshold attracts 60 to 110% more real world attention than a finding with significance marginally outside that threshold. We explore separately measures for news, social media, policy and academic attention. The results have important implications for the (in)efficiency of science translation and incentives in research.

KEYWORDS: Hypothesis testing - Statistical significance - Knowledge mobilization - Popular science - News Media - Social Media - p-Hacking - Publication bias - Research credibility

JEL CODES: B41, C12, I10, L82.

Authors: Brodeur: University of Ottawa and Institute for Replication. abrodeur@uottawa.ca Cook: Wilfrid Laurier University. ncook@wlu.ca. Heyes: University of Birmingham a.g.heyes@bham.ac.uk. Wright: Brock University. twright3@brocku.ca Laura Denniston and Abigail Marsh provided excellent research assistance. We thank participants of the 12th Berkeley Initiative for Transparency in the Social Sciences (2024 BITTS) Annual Meeting in Berkeley, the Birmingham Behavioral Workshop, Wilfrid Laurier University, Ted Miguel and James Rockey for helpful advice. Errors are ours.

1 Introduction

For scientific research to have a tangible impact requires that it attract attention in the "real world". Particularly in applied fields, an academic paper that attracts minimal attention beyond the academic community might be regarded as a failed endeavor (Fuoco, 2021). As such it is important to understand what drives attention to one scientific result versus another. Here we provide evidence of the distorting role of statistical significance.

Statistical significance in the context of hypothesis testing is a complex concept (Yaddanapudi (2016)) and summary statistics such as p-values are frequently misinterpreted. For example, it is commonly thought that the p-value tells us something about the likelihood that the null hypothesis is false, which it doesn't (Tanha et al. (2017)). As Siegfried (2010) notes "(A) common misinterpretation of the p-value is that it measures how likely it is that a null (or no effect) hypothesis is true. A p-value of 0.05 means that there is only a 5 percent chance of getting the observed results if the null hypothesis is correct. It is *incorrect* to transpose the finding into a 95 percent probability that the null hypothesis is false" (emphasis added). As Woolston (2015) observes, "(I)t is only by convention that smaller p-values are interpreted as stronger evidence that the null hypothesis is false." Nonetheless, numerous studies have shown that the pattern of statistical significance has a strong influence on how researchers evaluate the importance of research Chopra et al. (2024) and the likelihood of publication Andrews and Kasy (2019). Research published in leading academic journals has been shown to be substantially distorted with respect to statistical significance by the combined phenomena of p-hacking and publication bias, reducing the credibility of published results across numerous disciplines and research fields.[1] Confidence intervals, the range of plausible values for parameter

[1]Publication bias is the phenomenon whereby journals are more likely to publish research containing statistically significant results than insignificant ones, while p-hacking arises if researchers deliberately or inadvertently make modeling choices such as to enhance the statistical significance of results. The extent of each have been shown to be extensive and to substantially compromise the credibility of published research in areas such as economics (Brodeur et al., 2016), psychology (Simmons et al., 2011), neuroscience (Button et al., 2013) and science more generally (Head et al., 2015). We also point the reader to the provocatively titled article "Why Most Published Research

being estimated, convey other information, in particular with respect to estimation precision, but plausibly still encourage a typical reader to focus on whether or not the confidence interval includes zero.

However most people learn about scientific research not through their own reading of peer-reviewed academic journals. Rather they consumer the results of that research second-hand, via summaries provided by news media, social media, and so on. This paper is about the next link in the chain – how statistical significance impacts the attention paid *outside* academia to results found in an (already distorted) corpus of published academic research.

The purest "junk science" view of newsworthiness is depicted in the cartoon of Figure 1. The model of research envisioned here is one in which many spurious experimental studies are conducted, with only a study that delivers a non-null (or positive) result attracting outside attention. This is obviously a caricature, but reflects a common observation that popular science is largely about non-null results.

If the intuition in the cartoon, that statistical significance influences external attention, is a systematic feature of the real world, it has important implications for science mobilization. The green jelly bean "result" in Figure 1 is obviously spurious, but correct inference from the p-value would require the reader knowing about the other trials. Without that the nature of what can be learned from a study is not so obvious. Now consider a case in which the treatments involved were not jelly beans of different colors but a series of *ex ante* plausible medical or lifestyle interventions.

We test the link from statistical significance to external attention in the context of health research. We study the universe of articles published in three leading health journals: British Medical Journal (BMJ), The Lancet, and New England

Findings are False" (Ioannidis, 2005), with 12,714 Google Scholar citations at time of writing, describes how lines of inquiry afflicted by these two phenomena will be characterized by underpowered studies (Ioannidis et al., 2017), disproportionately many false positives, inflated effect size estimates and low replicability. Psychology is one literature frequently cited as an example (see Nuijten et al. (2016) and Open Science Collaboration (2015)). Another is well-published experiments in business disciplines (such as marketing, economics, finance, and the like (see Brodeur et al. (2022)).

Journal of Medicine (NEJM) for the period 2016 through 2022, inclusive. For each article, we extract the statistical significance of all hypothesis tests reported within the results sections of the structured abstract. In 30% of cases, these are presented as exact p-values. In the remainder of cases, statistical significance is reported via confidence interval, which we then convert to a p-value for consistency. Some abstracts report one hypothesis test, and so yield one p-value, others report several (the median number per article is three). We then investigate the relationship between those p-values and measures of the article's impact or, equivalently, the amount and intensity of external attention the article receives.

The primary measure of attention that we use is the widely-recognized "Almetric Attention" score, often reported on university websites and elsewhere via a dynamic many-coloured 'donut' logo. Altmetric is a data science company that applies a proprietary algorithm that converts the intensity of attention that an article receives in newspapers, blogs, policy documents, social media sites, etc. into a single integer index. Weights are used to account, for example, for the readership of the citing newspaper (the New York Times counts for more than a local news outlet) or the reach of those who promote it on social media (for example, the number of followers if research is quoted on Twitter). While our main focus will be on this aggregate score, we will draw insights by examining both the individual components which contribute to the index, and excluded measures such as one for attention from the academic community.

With emphasis on whether a result is statistically significant at better than 5% (in other words $p<0.05$)—the usually focal threshold against which a result is deemed to be statistically significant or not, and the standard against which the editorial guidelines of journals that we study instruct authors to judge results—we investigate the relationship between statistical significance and impact in four different ways.

First, we plot p-values against attention measures, allowing for visual identification of an association between statistical significance and attention, which we then

confirm more formally.

Second, we estimate fixed effects regressions that include a dummy variable that takes the value 1 if a result is associated with a p-value less than 0.05. This approximates the binary way in which a consumer of research may mentally compartmentalize results; either a significant effect of X on Y was found, or it wasn't.[2] However, this is a coarse approach since it turns a continuous measure of statistical significance, the p-value, into a categorical one, which ignores the possible impact of variation within significance ranges.

Third, we apply the caliper method, common in the publication bias literature, comparing attention outcomes for p-values falling within narrow ranges either side of the arbitrary 5% significance threshold. Our preferred variant applies calipers of width 0.01, which compares the attention received by research with p-values between 0.04 and 0.05 (the *just* significant) to the attention received by research with p-values between 0.05 and 0.06 (the *just* insignificant).

Finally, we borrow from the regression discontinuity (RD) literature to test for a discontinuity in the relationship between p-value and attention at the p=0.05 threshold. Acknowledging concerns that our evidence of p-values (the running or forcing variable) in our sample are manipulated, such manipulation would tend to bias our estimates to zero. We also follow RD best practice advice from Calonico et al. (2017), Calonico et al. (2019), Cattaneo and Titiunik (2022), Cattaneo et al. (2023) in considering different polynomial structures and bandwidths.

In brief, across the various methods we apply, we find a consistent influence of statistical significance on attention. Visual analysis suggests double attention when comparing p-values just above versus just below the 5% statistical significance threshold. More formal caliper and RD methods suggest that, other things equal, crossing the threshold into statistical significance increases real world attention paid

[2]Such discretisation is also embedded in, for example, statement like "the result was significant at better than 5%", or the use of asterisks and other eye-catchers in tables of results. The American Economic Association changed their editorial guidelines to prohibit the use of such eye-catchers as a result of the findings of Brodeur et al. (2016), ruling that the use of eye-catchers was distorting research practice.

to a research result by 60 to 110%. Looking into underlying mechanisms we find this composite increase is particularly driven by social media engagement.

Studying the top health journals rather than, say, health research in economics journals, is attractive for at least two reasons. First, with the focus on health many might contend that the subject matter is more important. Certainly there is *much* more non-academic interest in articles published in the top health journals studied here than even those published in the top 5 – at least as measured by Altmetric. For example, an Altmetric score of around 37 places an article in the top 20% of articles published in *The American Economic Review*, while the mean across the health research sample is 857. Second, health journals have a standardized way in which abstracts are written, and authors are required to comply with a journal-mandated style. This not only facilitates and streamlines the process of harvesting and organizing data, but plausibly gives a more consistent correspondence between the 'main result' of research and the content of its abstract.

Why does this matter?

First, consider the layperson who learns about research not from academic journals rather through the platforms captured by Altmetric. It is natural to ask whether such 'popular science' is good science, for example by giving a fair impression of the research. If not, the layperson can be expected to have a distorted, perhaps very distorted, view of the underlying body of knowledge. Insofar as this distorted view guides subsequent actions, for example making health choices based on scientific studies selectively reported in newspapers and magazines, this likely implies a static welfare loss. Moreover, Oster (2020) shows that such inefficiency can be dynamically reinforced. If a new research finding linking a behavior to positive health outcome stimulates take-up of that behavior among individuals who engage in other, perhaps unobserved, positive health behaviors, this mechanism could *confirm* the originally specious finding in later observational analyses if those unobservables cannot be accounted for. They provide empirical evidence of this dynamic effect using US panel data on vitamin and supplement use. Relatedly,

Vivalt and Coville (2023) shows that policy professionals, in a series of incentivized vignette experiments, tend to update their assessment of intervention efficacy more in response to good news than bad news (a hypothetical study delivering results in support of an effect) and are relatively insensitive to the width of confidence intervals.

Second, attention bias can distort researcher incentives and so the future direction of the underlying body of research itself. Among others, University managers and funding agencies increasingly valorize various measures of external engagement (including using, in some cases, Altmetric as a measure of performance) in appointment, promotion, grant award, and other decisions important to researchers. Pressure to generate not only well-published (in the academic sense) but externally attention-worthy research can be expected to generate mutually reinforcing incentives for p-hacking in the research community.

Note that it is not a defense of science mobilization practice to say that end-users "only care about what works, not what doesn't." The lesson from Figure 1 is not that green jelly beans cause acne, so that is what users will be interested to know, and it is good practice to report only that. Good reporting requires context in that the spurious and specious character of the positive result is understood alongside alongside the many parallel null results.

By investigating potential bias in the linkages from statistical significance to academic research to "real world" attention, our study complements the growing and important literature that documents the presence and extent of p-hacking and publication bias in the research corpus of various disciplines (Andrews and Kasy (2019); Bruns et al. (2019); DellaVigna and Linos (2022); Doucouliagos and Stanley (2013); Gerber and Malhotra (2008a); Gerber and Malhotra (2008b); Havránek (2015); Havránek and Sokolova (2020); Simonsohn et al. (2014); Vivalt (2019)). For the health sciences, various forms of bias have also been documented (Boutron and Ravaud (2018); Brown et al. (2018); Easterbrook et al. (1991); Fanelli (2009); Franco et al. (2014); Garattini et al. (2016); Turner et al. (2008); Zarin and Tse

(2008)).[3] Two relevant studies include Adda et al. (2020) and Dumas-Mallet et al. (2017). Adda et al. (2020) systematically examine p-values in phase II and phase III drug trials from ClinicalTrials.gov, revealing an upward jump for phase III results by small industry sponsors. Linking trials across phases, they provide evidence that early favorable results increase the likelihood of continuing into the next phase.

In terms of study of external attention, Dumas-Mallet et al. (2017) provide a case study of the determinants of newspaper articles coverage for meta-analyses investigating biomarkers and risk factors associated with four psychiatric disorders, four neurological pathologies and four somatic diseases. They find that coverage was more likely to attend to early rather than subsequent studies for some topics (*e.g.* those investigating 'lifestyle' effects), that studies reporting null effects were not covered, and that findings from less than half of studies reported in the news were assessed as valid by later meta-analysis.

We provide what we believe to be the first systematic exploration of the relationship between statistical significance and media coverage of health research. We do this using multiple methods, already outlined, and combine a large sample size with methods that pay particular attention to *marginally* significant or insignificant results to reinforce our conclusions. Our sample covers all health fields and we document the relationship across a broad range of venues including traditional media, social media, policy engagement, and attention in the academic community.

1.1 A motivating example

Pereyra-Elías et al. (2023) appearing in the British Medical Journal (Disease in Childhood), reported associations between how long a child was breastfed for and their academic performance at age 16 in a nationally representative sample of 5012 children followed by the Millennium Cohort Study in England. The main results compare children who were breastfed for 12 or more months (treated) against those not breastfed (control) for six outcome measures relating to performance in the

[3] A small literature also documents the effect of media coverage on citations (e.g., Dumas-Mallet et al. (2020)).

standardized GCSE assessments taken by almost all children in England at age 16. Those are (a) probability of failing in English, (b) probability of failing in Mathematics, (c) probability of achieving a High Pass (A or A*) in English, (d) probability of achieving a High Pass (A or A*) in Mathematics, (e) probability of achieving 5 or more GCSE's at Pass level, (f) an "Attainment 8" metric which adds the marks of a student in their best eight GCSEs, including English and Maths (which are double-weighted).

We make no assessment about the merits of the research design or its execution. However, two features of the study are of interest for current purposes.

First, the pattern of statistical significance associated with the main hypothesis tests. The central estimates in the study are relative risk ratios (RRRs) such that the null result of no association between breastfeeding and a particular performance metric is rejected if the confidence interval does not contain any value less than 1. For two of the outcome measures, the associations are insignificant, while four achieve significance at the focal 5% threshold. However, Figure 4 in that paper shows the boundaries of the 5% confidence intervals for those four outcomes to be 1.00, 1.00, 1.01 and 1.02, each significant at the focal 5% threshold, however the results are (very) borderline, particularly for a large-sample exercise.

Second, we observe that this was among the most highly attended to pieces of academic research published in 2023. It was widely reported in media outlets around the world (for example CNN, Fox News, CBS News) and in the main national newspapers in the United Kingdom was written up under headlines such as "Breastfed Children get Best GCSE Exam Marks" (The Times), "Breastfed Children More Likely to Achieve A Grade GCSE's" (The Daily Telegraph), "Breastfeed children to give them better GCSE Results, says Landmark Study" (The Independent) and "Kids Breastfed for at least a Year are 38% More Likely to get A's" (The Daily Mail). At time of writing it has an Altmetric of 1266 which places it in the top 0.03% of research articles cataloged by Altmetric.

The question that interests us relates to the causal relationship between these

two observations. Would an otherwise identical piece of research have attracted similarly widespread attention had the knife-edge significance tests gone the other way - had the boundaries of the confidence intervals been 0.99 rather than 1.00 and 1.01? Of course such a counterfactual does not exist, but the suite of methods that we apply here - categorical regression, the caliper method, regression discontinuity - deliver what we believe are credible estimates.

2 Data

We collect all research articles published in three leading health journals, namely The Lancet, British Medical Journal (BMJ), and the New England Journal of Medicine (NEJM), for the seven year period from 2016 to 2022, consisting of 2796 articles in total.[4] Though there is some variation in topic coverage between the journals (we discuss possible implications later), each publishes high impact research, has global readership, and are frequently cited in traditional media, social media, and are influential in policy discussion. Almost all articles report research findings that is statistical in character.

For each article in the sample we combine three types of data: (1) the statistical significance of any results reported in the article abstract, (2) the extent to which the article received attention outside, and (3) additional article attributes.

First, from the collected abstracts, we use a combination of manual reading and programmatical (regular expression) text extraction to identify the statistical significance of the contained hypothesis tests. To do so, we keep only statistics reported in the 'results' section of the structured abstracts. We then identify all instances of "$p < \#$","$p \leq \#$","$p = \#$","$p > \#$" (where # represents any number). In our later analysis we will keep only the third of this list, since coarse reporting of the sort offered by the others does not allow us to make meaningful statements about

[4]These journals also publish other materials. Specifically we collect; "Research" from the British Medical Journal, "Articles" from The Lancet, and "Original Articles" from the New England Journal of Medicine. Examples of other publications in these three journals include; News, Comments, Education (BMJ), Editorials, Obituaries, World Reports (Lancet), Perspectives, Images in Clinical Medicine, and Correspondence (NEJM).

where the p-value is within the broad intervals indicated. As many results are not expressly reported as p-values, and instead are reported as confidence intervals or risk-ratios, we then extract '#.# to #.#'. From these extractions we are careful to identify if the reported statistic is a ratio test statistic (where if the 95% confidence interval contains 1 would be considered a null) by searching the preceding characters for the string 'ratio.' An extensive manual audit by one of the authors found no errors.

Second, for each article we obtain the Altmetric Attention Score. As noted in the introduction, this is a systematic, popular, and widely-applied indicator of the amount of attention that research receives outside academia. It is embedded in the research pages of many universities, institutes and individual researchers, often in the form of a dynamic colored doughnut with the Altmetric score displayed in the center. The score is derived from an algorithm which combines mentions in newspaper articles, blogs, social media posts, policy reports and so on.[5] As noted, weights are applied both between and within source categories. Details of the algorithms applied can be found under the "About Us" tab at altmetric.com.[6] Weightings are occasionally updated. The Attention Score is not normalized and is scale free. The average score varies across disciplines and journals, but according to Altmetric a rule of thumb is that a score greater than 20 indicates substantial

[5]Altmetric tracks article mentions in the news, social media, policy spaces, and Wikipedia. In order to be tracked, mentions need to link back to the article in question using a DOI or URL (including publisher, arXiv, PubMed, or institutional repository links). However, not all sources use these links and so Altmetric also uses text mining in order to count mentions. This text mining approach requires the name of at least one author, the title of the journal, and a publication date (though Altmetric also fuzzily searches for articles +/− six weeks of the news article's date). Lastly, the text mining results are then compared with metadata from Crossref to attribute the mentions to the correct article.

[6]To flag a few features, news outlets are placed in tiers based on their reach and therefore more popular sources like the New York Times will have a larger contribution to Attention Scores; articles that are linked on Wikipedia receive a static score and do not increment with additional links; inclusion in policy documents receive a score for each separate source; X (formerly known as Twitter) reposts and quote tweets are down-weighted because they are second-hand attention and the combined total of reposts is rounded to the nearest whole number; X mentions also have modifiers based on the number of followers, frequency of research-related tweets, and bias (the diversity of journals in a user's tweets in an attempt to separate promotion from engagement - for example the social media accounts of many journals refer only to papers published in that journal, so can be discounted as advertising rather than true engagement).

attention. We collect both main Altmetric[7] and the constituent measures for news, social media, and policy. In addition the Mendeley measure of interest in the academic community.

Third, the attributes that we collect for each paper are the journal of publication, year of publication, number of authors, whether the research reported a randomized controlled trial, whether the research reported a meta-analysis, whether the abstract of the paper referred to COVID-19 (or variants) and whether the paper reported pre-registration.

Summary statistics are presented in Table 1. Overall, 10,404 test statistics were collected from 2,796 articles; the median article reports three. There is a fairly even split between the three journals (32% from BMJ, 41% The Lancet and 27% NEJM). The majority of statistical results are reported as confidence intervals (70%), while a not insignificant minority are reported as exact p-values. Around half (51%) of test statistics are from a controlled trial, while 14% are from a meta-analysis. 53% of test statistics are from articles that indicated pre-registration (which includes all controlled trials, but also some of the meta-analyses). The average Altmetric score is 857, though this is characterized by sizable variation with a standard deviation of 1741. The distribution of this variable is plotted in Figure 2, which is winsorized from above at 2000 for presentation purposes only. The distribution of the Altmetric Attention Score across articles is strongly positively skewed - however note that a score of 20 is considered influential by Altmetric, a score which 99.6% of the sample's articles achieve.

We present box plots of Altmetric Score by journal in Figure 3; each box is bounded at the 25th and 75th percentile score, while the internal line represents the median. The ends of the whiskers represent the lower and upper adjacent values, while individual dots reaching up to 2000 (we again winsorize for presentation only - the highest score in our sample is 3,744) represent a box plot's outside values.

[7]It is important to note that the Altmetric Attention Score is a measure of attention and not quality of research articles, therefore it is difficult to say whether a particular score is "good" or "bad". Further, it only tracks public and direct attention that research articles receive.

Overall, articles in all three health journals receive a great deal of attention on average.

3 Methods

Our objective is to characterize how the statistical significance of a research result predicts non-academic interest in it. While our paper is not primarily focused on the twin phenomena of publication bias and p-hacking, there is little reason to believe that the two are absent (See Section 4.1). In terms of interpreting our results, the presence of both (or indeed either) phenomena has the following countervailing effect to our estimates' magnitude and statistical significance: If it is the case that a piece of research has a statistically *in*significant result, it is reasonable to suspect that in order for that research result to have been published it must be of 'better quality' or have some positive characteristic that makes it more engaging, more novel, better executed, or superior in some way. We take as supporting evidence the anecdote that publishing a null result is 'harder' than attempting to publish research which rejects the null (recently confirmed in Economics by Chopra et al. (2024)). That some positive research characteristics would be present, holding statistical significance fixed, would lead to *more* post-publication attention. Said differently, those just *in*significant articles have a certain unobserved positive characteristic which raises their average Altmetric score, while those *just* significant articles did not need this characteristic in order to be published, following the presence of publication bias. In this respect, our estimates of a positive effect of significance on attention can be regarded as conservative.

In our main analysis, we first plot p-values against Altmetric scores in a way that allows for visual inspection of patterns. We then estimate three groups of specifications.

Our particular attention is on the effect of a hypothesis test rejecting the null result at the 5% level. This is the threshold usually applied to delineating significant from insignificant results and is also mandated or encouraged within the editorial

guidance to authors provided by the three journals.[8]

First, we regress Altmetric Score on an indicator variable that takes the value 1 if the corresponding p-value is less than 0.05 (indicating statistical significance at the 5% level) and 0 otherwise. More formally, $T_i = \mathbb{1}(p_i \leq 0.05)$. We estimate the following by ordinary least squares:

$$y_{iatj} = \alpha + \beta_1 T_i + \lambda_t + \theta_j + \varepsilon_{iatj} \tag{1}$$

where y_{iatj} is the Altmetric Score associated with p-value i which appears in the abstract of article a published in year t and journal j. The regressor of interest is the indicator variable T_i. Year and journal fixed effects are λ_t, and θ_j, respectively. Standard errors are clustered at the article level.

Second, we modify our regressions to emulate the calipers implemented in Gerber and Malhotra (2008c). This has the effect of looking at Altmetric Scores associated with p-values just above and just below the $p = 0.05$ threshold. While caliper regressions have been widely used to study publication bias, where the difference in density above versus below a threshold is taken as evidence of manipulation or bunching (examples include Gerber and Malhotra (2008a), Gerber and Malhotra (2008b), Brodeur et al. (2016), Vivalt (2019), and Brodeur et al. (2020)), here we apply the method to investigate attention outcomes on either side of the threshold. Specifically, we conduct the previous estimation with the additional restriction of $p_i \in [0.05 - h, 0.05 + h]$ where h may take on different, typically quite small, values. The underlying assumption is that research results on either side of the threshold *within a very small neighborhood* are comparable with the exception of statistical

[8]The Lancet, under "For Authors: Preparing your manuscript" provides separate guides depending on method used, but in each of these requires reporting of results, constriction of confidence intervals, forest plots etc. based on the 5% threshold. The BMJ "Requirements for Authors" does not mandate it, but the worked examples provided are based on 5%. The 2018 changes to editorial practice at NEJM are described in the *New Guidelines for Statistical Reporting in the Journal* (Harrington et al (2019)) which increased the primacy of the 5% threshold but also, interestingly, encouraged authors to be more selective in the set of results for which significance statistics are presented: "The new guidelines discuss many aspects of reporting of studies in the *Journal*, including a requirement to replace p-values with estimates of effects or associations and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity."

significance (while noting the attenuating bias of publication bias discussed earlier).

Third, we estimate the following motivated by the regression discontinuity (RD) literature. We apply the off-the-shelf statistical package described in Calonico et al. (2017) which estimates the RD treatment effect at the cutoff (see also Calonico et al. (2019), whose convention we follow):

$$\tau = \tau(\bar{x}) = \mathbb{E}\{Y_i(1) - Y_i(0)|X_i = \bar{x}\} \tag{2}$$

Where the score, index, or running variable is X_i (in our setting a p-value) and treatment status is determined as $T_i = \mathbb{1}(X_i \geq \bar{x})$ for the cutoff $\bar{x}$ (in our setting $T_i = \mathbb{1}(p_i \leq 0.05)$) under the potential outcomes framework (where $Y_i(0)$ represents a unit's potential outcome under control and $Y_i(1)$ a unit's potential outcome under treatment). Specifically, we use the 'standard' set up where (without covariate-adjustment and using a linear polynomial) the treatment effect τ is estimated by:

$$\hat{Y}_i = \hat{\alpha} + T_i\hat{\tau} + X_i\hat{\beta}_- + T_iX_i\hat{\beta}_+ \tag{3}$$

Where the RD treatment effect estimate $\hat{\tau}$ is obtained by regression of Y_i on a constant, a treatment effect indicator T_i, the running variable X_i, and their interaction T_iX_i using only units where $X_i \in [-h, h]$ (where the bandwidth h is data driven). We apply the 'default' triangular kernel, which down-weights observations linearly away from the cutoff. Our RD estimates later will prove robust to a quadratic specification and adjustment for covariates.

4 Results

4.1 A First Stage: Publication Bias and p-hacking in the Sample

As a preliminary exercise, before turning to Altmetric and research impact, we describe the pattern of statistical significance in our data set.

First, we plot the distribution of p-values and the associated distribution of z-statistics. Each has been used in the recent literature on research credibility (the

interested reader is encouraged to see Elliott et al. (2022) and Brodeur et al. (2016) for economics).

Figure 4 depicts the distribution of p-values, using bins of width 0.01.[9] Overall 77% of the p-values in our sample are statistically significant at the 5% level. There is a noteworthy step down at statistical significance; the step from the (0.04-0.05) bar to the (0.05-0.06) bar is much more pronounced than that between bars on either side.

While the p-curve in Figure 4 is drawn for consistency with the rest of the paper, for the purposes of this section we can see patterns more clearly by looking at the corresponding distribution of z-statistics in Figure 5. Here we see a clear dearth of both insignificant and marginally insignificant results, with bunching to the right of the 5% threshold value of 1.96. The 'camel' shape, with a trough in mass in the range between about 1.5 and 1.96 reappearing immediately above 1.96 is indicative of presence of "marginal" p-hacking (Brodeur et al. (2016)). The right panel of the figure offers the interested reader a comparison with the distribution reported in Brodeur et al. (2020), although with the caveat that the data collection methodology differs in several ways.

4.2 Attention Bias Method 1: Diagrams

Having established that the statistical significance of the underlying body of research drawn from the three journals is itself distorted, with an over-representation of significant and, in particular, marginally significant results, we now turn to our main questions - the relationship from significance to attention.

We begin with a visual analysis, in order to highlight what the 'raw' data suggests. To each p-value we assign the containing article's Altmetric score. We then calculate the mean of those Altmetric scores for all p-values in a series of bins each of width 0.0025. Figure 6 provides a plot of those averages for that series of bins, with the size of each bubble proportionate to the count of p-values in the corresponding

[9] Later analysis will examine results by journal and method, see Figures A1 and A2 respectively.

bin.[10]

There are three things to note from this figure.

First, the bubbles to the right of the vertical $p = 0.05$ line are visibly smaller than those to the left of it, with a sharp change in bubble size at that threshold. This reflects that the underlying body of research, defined by the universe of statistical results published in the three journals in the seven year period, has features consistent with p-hacking and/or publication bias around the 5% significance level.

Second, apart from their size, the bubbles to the left of the vertical $p = 0.05$ line are visibly higher than those to the right. In other words, the Altmetric scores are greater for papers with significant p-values. The intent of the categorical regressions of Section 4.3 will be to examine this more formally.

Third, considering only p-values within a narrow range on either side of the vertical $p = 0.05$ line, it is apparent that the bubble immediately to the left of the $p = 0.05$ line is visibly higher than that immediately to its right. In other words there is a jump at that threshold: the average altmetric score of the left bubble is 1275, while for the right bubble it is 654, nearly double. The intent of the caliper analysis in Section 4.4 will be to examine this more formally. The RD approach in Section 4.5 represents an alternative approach to testing for discontinuities at the significance thresholds but using data points from the full support.

Figure 8 repeats this exercise for the constituent parts of the composite Altmetric score: attention from News Media, Social Media, and Policy Makers as well as an excluded measure of attention in the academic community, Mendeley readership. Qualitatively similar patterns can be seen in each panel, although perhaps most so for News, Social, and Policy attention.

[10] We provide the same data in Figure 7 with the notable exclusion of the many p-values whose value is less than 0.0025. This has the effect of increasing the resolution of the figure - the following descriptions still apply.

4.3 Attention Bias Method 2: Categorical Fixed Effect Regression

We next investigate the relationship between statistical significance and media attention using the full sample and categorical classification, while accounting for characteristics such as article journal or vintage.

Column 1 in Table 2 reports the results of estimating Equation 1. The dependent variable is the composite Altmetric score. The estimated coefficient on the indicator implies that if a research result achieves statistical significance at better than 5%, then the attention it receives as measured by Altmetric is higher by 264 'points.' When compared to the average attention statistically insignificant research receives, significance increases attention by about 44%.

Columns 2 through 4 report the result of re-estimating Equation 1 but replacing the overall Altmetric score as the dependent variable by each of its components - News, Social, Policy - in turn followed by the Mendeley academic attention measure. In each case, the coefficient on the dummy variable is large, positive, and statistically significant. While each of these metrics have different scales, making these columns' coefficients incomparable to one another, when scaled against the average for statistically insignificant research results, we find that statistical significance increases attention (however measured) by between 44% and 52%.

Table 3 explores heterogeneity by journal (columns 1 through 3) and methodology (4 through 6) by re-estimating column 1 of Table 2 on the corresponding sub-samples. We can see that the estimated coefficients are consistently positive. Unsurprising given eroded sample sizes and subsequently larger standard errors, statistical significance is not achieved at conventional levels in the NEJM and RCT sub-samples. The effect sizes, expressed in percentage terms at the bottom of each column, point to particularly pronounced effects of statistical significance for articles published in The Lancet and for meta-analysis articles, with statistical significance increasing attention by 76% in those cases. We do not speak to what might explain such differences, though possible explanations for the between-journal variation include possible differences in the 'type' of research each journal accepts. Even within

topic it may also reflect different writing conventions and reporting standards within the communities of scholars that publish in different journals, or different practices enforced by those journals.

Table 4 presents some robustness results. Column 1 reports the outcome of re-estimating on the whole sample but excluding year and journal fixed effects. The estimated coefficient is somewhat larger with the fixed effect controls removed. In column 2, we recognize the fact that our sampling window includes the COVID-19 period, an exceptional period for both journalistic and public interest in health research. Table 1 notes that 11% of our sample is drawn from abstracts that include the (case insensitive) word "covid." A plausible concern is that anomalous attention to COVID-19 related attention could dominate or at least distort our inference. We investigate this possibility in two ways. First, reported in column 2 of Table 4, we add to the main specification a dummy that takes the value 1 if the abstract of a paper includes the word "covid" and 0 otherwise. Second, reported in column 3, we re-estimate on a pre-COVID-19 sample which comprises all papers published before 2020. The latter approach removes more thoroughly any potential threat to our results from being confounded by COVID-19, but substantially degrades the sample size. In each of columns 2 and 3 the estimate of the coefficient of interest can be seen to remain positive. However, the estimates effect sizes are somewhat smaller, at 35% and 33% respectively.

Another concern relates to the mapping of p-values to outcomes. In an ideal world each article would contain a single 'result' and report a single p-value, that associated with that single result, in its abstract. However that is often not the case. The median abstract contributes three p-values to our sample. This variation in the number of test statistics each article provides means that, absent weights, an article that offers six hypothesis tests counts "twice" as much as an abstract that provides the median number. In column 4, we apply article-level weights that ensure each of the 2,796 articles in the estimation are equally weighted. The results are similar, if not larger, than our main estimates. Further, in each of the articles, the main

outcome of interest may be measured in a number of ways, or the abstract may report additional secondary or otherwise supporting results. This makes it harder to link attention to an article to any particular hypothesis test. This could have several effects. For example, this introduces measurement error into the regressor of interest, and if that measurement error is classical, this attenuates our regressions' estimates. To probe whether this may be confounding results we conduct we re-estimate the main specification but only on those p-values drawn from abstracts containing one, two, three and four p-values - in other words excluding papers that report 'lots' of test results in their abstracts. The results of doing this are reported in column 5 of Table 4. The estimated coefficient value is in line with the rest of the coefficient estimates in the Table.

4.4 Attention Bias Method 3: Caliper Method

The preceding approach converts an inherently continuous variable, a p-value, and renders it binary. While this may correspond with the way in which statistical significance is often verbalized - respectively using phrases such as 'significant at 5%' - the coarseness of the categories means that results associated with quite different p-values are treated equally. In other words, a result associated with p=0.01 would be regarded as the same, in terms of statistical significance, as one associated with p=0.049.

In this section we refine the analysis by applying calipers that focus on p-values within narrow bands or neighborhoods of the statistical significance threshold. This approach also serves to reinforce the causal interpretation of the differences in outcome variables. If the reported level of statistical significance is not *causing* the difference in attention then we would not expect there to be discernible differences in attention paid to research results with p-values marginally above an arbitrary significance threshold, than with p-values marginally below.

Table 5 reports the results of estimation using calipers of three different widths. Our preferred caliper width is 0.02, which remains "narrow enough" to be confident

the underlying research results are comparable while at the same time ensuring a reasonable number of data points on which to estimate the specification reported in column 2. The result there implies that, other things equal, a research result with p-value in the interval 0.03-0.05 attracts an Altmetric score 277 points higher, or garners about 61% more attention, than it would had the result been associated with a p-value in the interval 0.05-0.07.

Columns 1 and 3 report the results of repeating the procedure but with narrower (0.01) and wider (0.03) calipers respectively. As expected the sample sizes vary with caliper width but the results prove consistent in sign and significance across columns. The larger treatment effect for the narrow bands is consistent with the single elevated bubble at 0.0475 in Figure 6.

Table 6 investigates mechanisms by re-estimating the preferred specification but on the separate attention elements. Estimated treatment effects are positive in each column though statistical significance is not achieved at conventional levels for Policy. Recall that the various metrics have different scales so the coefficient estimates are not comparable between columns, though the estimated treatment effect can be seen at the foot of each column.

Table 7 presents some heterogeneity results for the caliper estimation, re-estimating the main specification but on sub-samples defined by journal and methodology respectively. Since the caliper method involves discarding most of our data (all p-values not falling in the interval 0.3 to 0.7) subsequent sub-sampling means that the results in this table are estimated on small numbers of data points, and should be interpreted in that light. Looking across columns we can see that coefficient estimates are in all cases positive and in three columns statistically significant. The BMJ and NEJM sub-samples do not achieve significance at conventional levels though they are not far away, even in these small samples, and we again do not want to over-interpret the difference between journals. On the other hand, the result from the RCT subsample in column 5 offers further evidence that results from controlled trials may not follow the pattern seen in the wider sample. We can see

that for meta-analyses, attention is particularly sensitive to the study delivering a p-statistic the 'right' side of 0.05.

Table 8 reports the results of conducting, separately, the same series of robustness exercises that we reported for the analysis using the categorical regression design in Section 4.3. First, dropping the Year and Journal FEs (column 1). Second, assessing the possibly distorting role of COVID-19 by including an additional control for papers that include the word 'covid' or synonyms in their abstract (column 2) or estimating on that subsample of papers published before 2020 (column 3). Third, estimating using article weights (column 4) or only using test statistics derived from papers that feature 4 or less test statistics in their abstracts (column 5). We are cautious not to over-interpret variations between columns in light of the heavily eroded sample sizes in some cases. However for the purposes of assessing robustness we can see by looking across columns that in all cases the coefficient estimates retain sign and order of magnitude with level of statistical significance varying across columns. Overall the results point to the overall findings of the caliper analysis being robust to these variations.

4.5 Attention Bias Method 4: Regression Discontinuity Design

Our final approach applies a method inspired by the regression discontinuity methodology which attempts to discern whether there is a discontinuity or jump in the Altmetric score (outcome variable) as the p-value (forcing or running variable) traverses an arbitrary significance threshold. It does this by fitting a polynomial on the data points that lie within specified ranges, or bandwidths, of the threshold, with the weight attached to any particular point varying with its proximity to that threshold (following Cattaneo and Titiunik (2022), see also Cattaneo et al. (2023) for a practical guide).

We acknowledge that this is not a true RD exercise since the p-value associated with a particular result is something that a researcher may be able strategically to manipulate. Indeed that is precisely what p-hacking is and what we examined in

brief in Section 4.1. In essence what we seek to do here is look for a distinct break or step at 0.05 and the RD toolkit provides an excellent and familiar tool for doing that, albeit we need to be reflective in interpretation.

As already noted RD requires the researcher specify the degree of polynomial to be fitted, and the bandwidth. Following best practice and to reduce the risk that estimates be idiosyncratic to any particular modeling assumption we estimate both quadratic (Table 9) and linear (Table 10) specifications (i.e., polynomials of degrees 2 and 1) for each of four different intervals. In the first, we apply the data-driven 'optimal' threshold following Calonico et al. (2017). For the remaining three, we use arbitrary and shrinking bandwidths, namely the intervals of +/-0.01, +/-0.02, +/-0.03.

The results of the preferred quadratic RD specification are reported in Table 9 (qualitatively similar results from a polynomial of degree one are presented in Table 10). The RD estimate at the top of each column is the estimated jump in the Altmetric score at p=0.05. Looking across the columns we can see that the estimated discontinuity is roughly stable in size across columns, and in each case statistically significant. Given consistency we remain agnostic here on the preferred modeling choices. The effect sizes in the columns point to a discontinuity in the Altmetric score of 663 to 744 points. This is larger than that derived from our previous analyses and corresponds to the Altmetric score increasing by between 110 to 160% when the p-value moves from being an epsilon outside the 5% threshold to an equivalent result an epsilon inside.

Consistent with the previous analyses, we now apply the RD methodology to attention's components, as well as heterogeneity and robustness exercises.

First, Table 11 investigates the separate attention elements. Estimated treatment effects are positive for each measure of attention, with statistical significance associated with increases in overall attention as well as news (column 2), social media (column 3), and policy (column 4). Academic readership, measured via Mendeley, is not statistically significant but is positive. Of course, the various metrics

components have different scales so the coefficient estimates are not directly comparable between columns, but we have provided a scaled treatment effect against the untreated mean in the foot of each column.

Second, Table 12 presents heterogeneity, re-estimating the main RD specification on journal and methodology sub-samples. Across columns we can see that coefficient estimates are in all cases positive and in three columns statistically significant. The BMJ and NEJM sub-samples do not achieve significance at conventional levels. On the other hand, the result from the RCT subsample in column 5 offers further evidence that attention for randomized controlled trials may not follow the pattern identified in the wider sample. We can see that for meta-analyses and other research, attention is particularly sensitive to the study delivering a p-value on the 'right' side of 0.05.

Finally, Table 13 reports the results of conducting, separately, a series of robustness exercises. In column 2 we add fixed effects for journal and year (following the suggestions of Calonico et al. (2019) for RD). In column 3 we control for the potentially distorting role of COVID-19. In column 4 we examine only articles published during the pre-COVID period. In column 5, we apply abstract weights (that is each abstract has an identical weight in the analysis - to assuage concerns that abstracts which offer more test statistics could potentially be over-weighted). In column 6, we restrict the sample to articles which present less than the median number of test statistics. While we are again cautious not to over-interpret variations between columns, we can see by looking across columns that in most cases the coefficient estimates retain sign, significance, and order of magnitude. Overall the results point to the overall findings of the RD analysis being robust to sensible variation in research decisions.

5 Conclusions

The statistical significance of research results, and in particular the use of null hypothesis testing which classifies scientific findings into 'significant' or 'insignificant',

has gathered notoriety in recent years. There have been widespread calls for reform of practice, and in some cases to discontinue the reporting of statistical significance test statistics altogether (e.g., Frank et al. (2021)). Null hypothesis test results, p-values and other test statistics are frequently misinterpreted in both academic and non-academic writing (Adams et al., 2019). Nonetheless p-values and other measures of statistical significance have been shown to have an important impact on how researchers evaluate research (Chopra et al. (2024)) and how journals decide which research results to publish (Andrews and Kasy (2019)). Unsurprisingly, researchers are found to manipulate their practices to deliver 'better' p-values, in particular p-values that place them under the arbitrary 0.05 threshold, or equivalently to generate confidence intervals that do not include zero, which allows claims of 'statistical significance' (Brodeur et al. (2016)). This compromises the credibility of the research base, and creates a corpus of published research that over-represents false positives, and have with little doubt been a major contributor to the non-replicability crisis now afflicting numerous research fields.

Our focus in this paper has been on how statistical significance affects which research results receive attention outside the academic community. We try to answer: "(T)o what extent does popular science paint a misleading picture of the underlying research?" Schoenfeld and Ioannidis (2013) study 50 common ingredients drawn at random from a popular cookbook, and for most find at least one published study pointing to each of a positive, negative, or statistically insignificant association with cancer. With such non-consensus even in published research the process whereby actors such as journalists select what to mobilize matters. As per Figure 1, news and social media interest in a study finding that *something* has a statistically significant association with cancer/hair loss/longevity/academic success is much greater, anecdotally at least, than that in an otherwise identical study finding that *something* does not. With an application to a large set of papers published in three elite health journals, and using a host of methods, we find that when a research finding is statistically significant, even just inside the 5% threshold, it gathers about 60 to

110% more real world attention compared to an otherwise identical finding.

Why does this matter? While a fully-formulated welfare analysis is outside the scope of this paper it can be expected that the distortion that we identify will misinform behavior and reduce the utility of those who act on it. Those actors may be individuals, for example in the case of health advice, or may be policy practitioners. Furthermore Oster (2020) outlines a mechanism that could allow such a distortion to be exacerbated through time, and provides evidence in support of it from a study of vitamin and supplement use among Americans – a setting adjacent to if not exactly within our topic of study. Outside the personal health sphere we can think of many other behaviors that are 'research informed' albeit potentially not, such as voting on climate issues or making parenting decisions.

It is also plausible that career-motivated researchers, "impact"-motivated journal editors, and others involved in the research production process may or already have adjusted to the incentives that such attention bias implies, such that the future evolution of science itself may be misdirected.

References

Adams, R. C., Challenger, A., Bratton, L., Boivin, J., Bott, L., Powell, G., Williams, A., Chambers, C. D., and Sumner, P. (2019). Claims of Causality in Health News: A Randomised Trial. *BMC Medicine*, 17(1):1–11.

Adda, J., Decker, C., and Ottaviani, M. (2020). P-Hacking in Clinical Trials and How Incentives Shape the Distribution of Results Across Phases. *Proceedings of the National Academy of Sciences*, 117(24):13386–13392.

Andrews, I. and Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–94.

Boutron, I. and Ravaud, P. (2018). Misrepresentation and Distortion of Research in Biomedical Literature. *Proceedings of the National Academy of Sciences*, 115(11):2613–2619.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, 110(11):3634–60.

Brodeur, A., Cook, N., and Heyes, A. (2022). We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell Us about Publication Bias and p-Hacking in Online Experiments. IZA Discussion Paper 15478.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Brown, A. W., Kaiser, K. A., and Allison, D. B. (2018). Issues with Data and Analyses: Errors, Underlying Themes, and Potential Solutions. *Proceedings of the National Academy of Sciences*, 115(11):2563–2570.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M., et al. (2019). Reporting Errors

and Biases in Published Empirical Findings: Evidence from Innovation Research. *Research Policy*, 48(9):103796.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2):372–404.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2023). A Practical Introduction to Regression Discontinuity Designs: Extensions. arXiv preprint arXiv:2301.08958.

Cattaneo, M. D. and Titiunik, R. (2022). Regression Discontinuity Designs. *Annual Review of Economics*, 14:821–851.

Chopra, F., Haaland, I., Roth, C., and Stegmann, A. (2024). The null result penalty. *The Economic Journal*, 134(657):193–219.

DellaVigna, S. and Linos, E. (2022). RCTs to Scale: Comprehensive Evidence from Two Nudge Units. *Econometrica*, 90(1):81–116.

Doucouliagos, C. and Stanley, T. D. (2013). Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity. *Journal of Economic Surveys*, 27(2):316–339.

Dumas-Mallet, E., Garenne, A., Boraud, T., and Gonon, F. (2020). Does Newspapers Coverage Influence the Citations Count of Scientific Publications? An Analysis of Biomedical Studies. *Scientometrics*, 123:413–427.

Dumas-Mallet, E., Smith, A., Boraud, T., and Gonon, F. (2017). Poor Replication Validity of Biomedical Association Studies Reported by Newspapers. *PloS One*, 12(2):e0172650.

Easterbrook, P. J., Gopalan, R., Berlin, J., and Matthews, D. R. (1991). Publication Bias in Clinical Research. *Lancet*, 337(8746):867–872.

Elliott, G., Kudrin, N., and Wüthrich, K. (2022). Detecting p-Hacking. *Econometrica*, 90(2):887–906.

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PloS one*, 4(5):e5738.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science*, 345(6203):1502–1505.

Frank, O., Tam, C. M., and Rhee, J. (2021). Is it Time to Stop Using Statistical Significance? *Australian Prescriber*, 44(1):16.

Fuoco, R. (2021). How to Get Media Coverage and Boost your Science's Impact. *Nature*.

Garattini, S., Jakobsen, J. C., Wetterslev, J., Bertelé, V., Banzi, R., Rath, A., Neugebauer, E. A., Laville, M., Masson, Y., Hivert, V., et al. (2016). Evidence-Based Clinical Practice: Overview of Threats to the Validity of Evidence and How to Minimise Them. *European Journal of Internal Medicine*, 32:13–21.

Gerber, A. and Malhotra, N. (2008a). Do Statistical Reporting Standards Affect what is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*, 3(3):313–326.

Gerber, A. S. and Malhotra, N. (2008b). Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results? *Sociological Methods & Research*, 37(1):3–30.

Gerber, A. S. and Malhotra, N. (2008c). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1):3–30.

Havránek, T. (2015). Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting. *Journal of the European Economic Association*, 13(6):1180–1204.

Havránek, T. and Sokolova, A. (2020). Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say "Probably Not". *Review of Economic Dynamics*, 35:97–122.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The Extent and Consequences of p-Hacking in Ccience. *PLoS Biology*, 13(3):e1002106.

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):e124.

Ioannidis, J. P., Stanley, T. D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *Economic Journal*, 127(605):F236–F265.

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., and Wicherts, J. M. (2016). The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349(6251):aac4716.

Oster, E. (2020). Health Recommendations and Selection in Health Behaviors. *American Economic Review: Insights*, 2(2):143–160.

Pereyra-Elías, R., Carson, C., and Quigley, M. A. (2023). Association between breastfeeding duration and educational achievement in england: results from the millennium cohort study. *Archives of Disease in Childhood*.

Schoenfeld, J. D. and Ioannidis, J. P. (2013). Is everything we eat associated with cancer? a systematic cookbook review. *The American journal of clinical nutrition*, 97(1):127–134.

Siegfried, T. (2010). Odds are, it's wrong. *Science news*, 177(7):26.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-Curve: A Key to the File Drawer. *Journal of Experimental Psychology: General*, 143:534–547.

Tanha, K., Mohammadi, N., and Janani, L. (2017). P-value: What is and what is not. *Medical journal of the Islamic Republic of Iran*, 31:65.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., and Rosenthal, R. (2008). Selective Publication of Antidepressant Trials and its Influence on Apparent Efficacy. *New England Journal of Medicine*, 358(3):252–260.

Vivalt, E. (2019). Specification Searching and Significance Inflation Across Time, Methods and Disciplines. *Oxford Bulletin of Economics and Statistics*, 81(4):797–816.

Vivalt, E. and Coville, A. (2023). How do policymakers update their beliefs? *Journal of Development Economics*, 165:103121.

Woolston, C. (2015). Psychology journal bans p values. *Nature*, 519(7541):9–9.

Yaddanapudi, L. N. (2016). The american statistical association statement on p-values explained.

Zarin, D. A. and Tse, T. (2008). Moving Toward Transparency of Clinical Trials. *Science*, 319(5868):1340–1342.

6 Figures

Figure 1: 'Significant' by XKCD

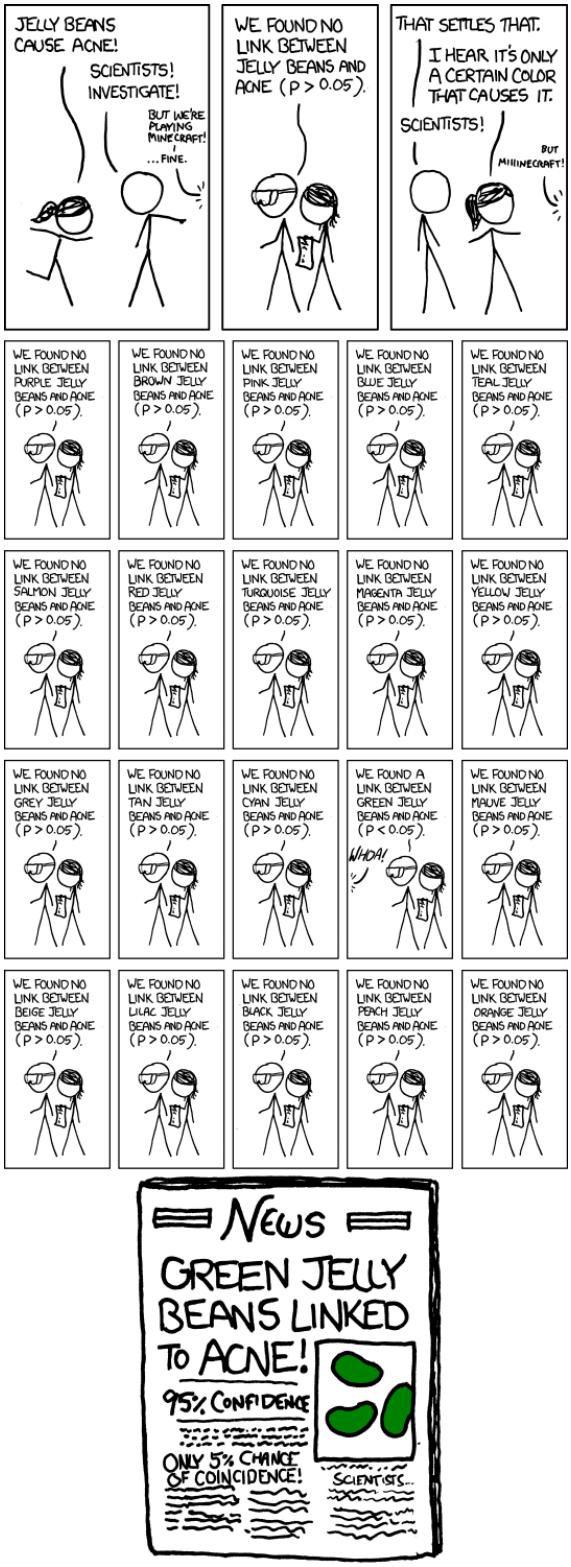
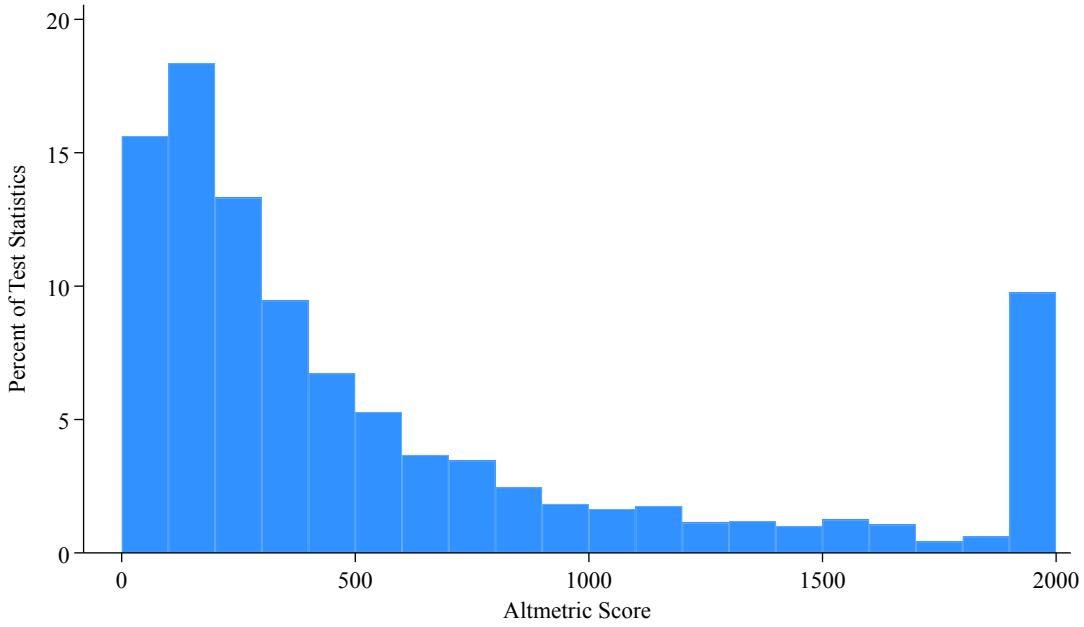
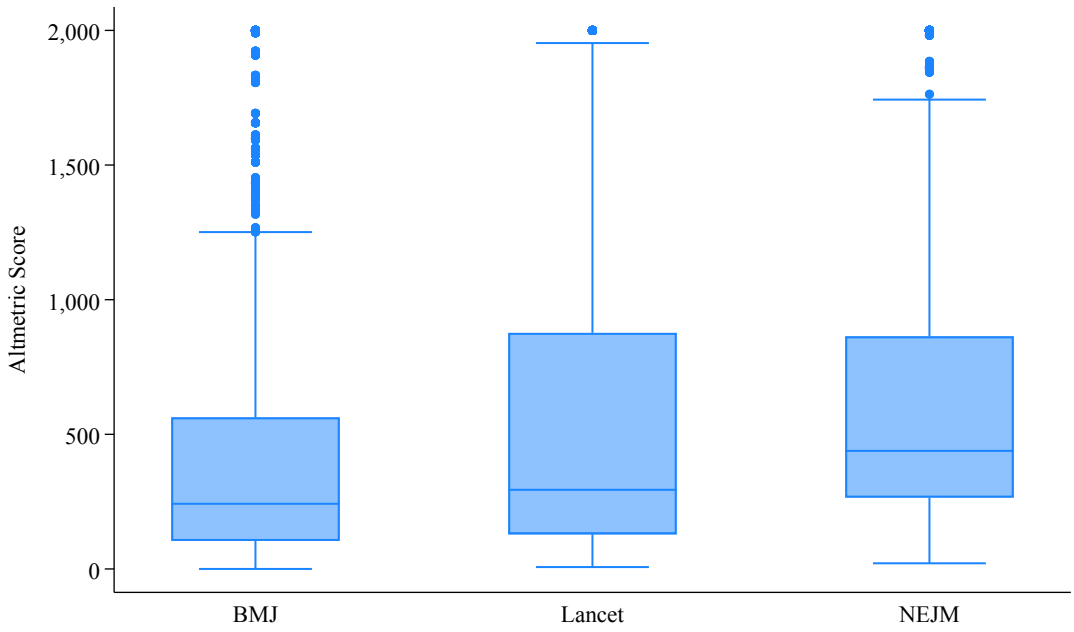


Figure 2: Altmetric Score



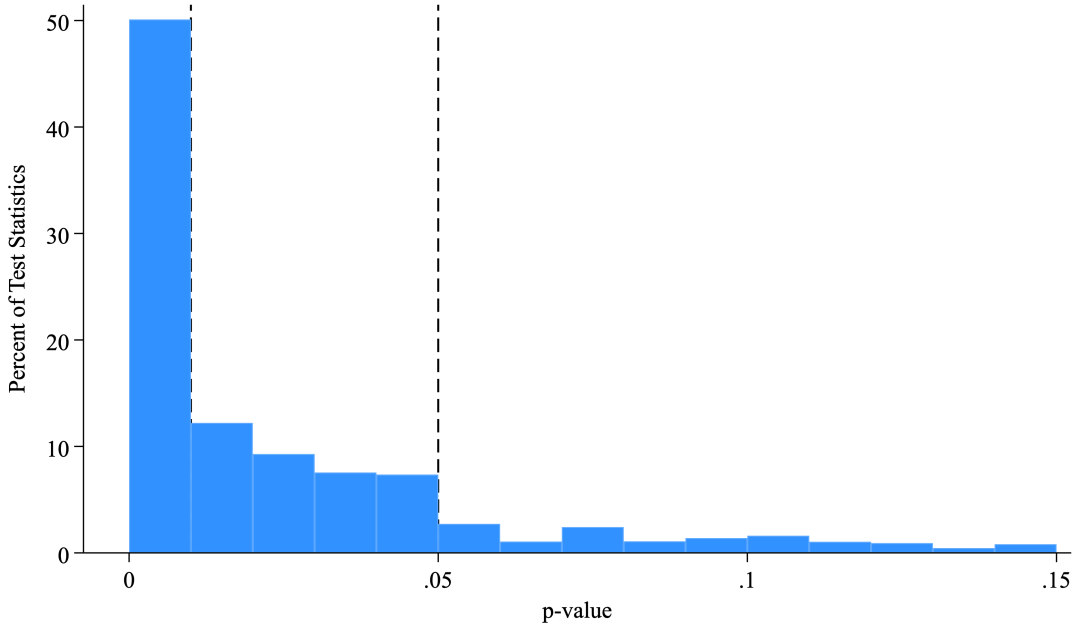
Notes: An observation is a test statistic (N= 10,404). Bins of width 100. Altmetric score winsorized above 2000. Average Altmetric Score 857. Median Altmetric Score 327.

Figure 3: Altmetric Score By Journal



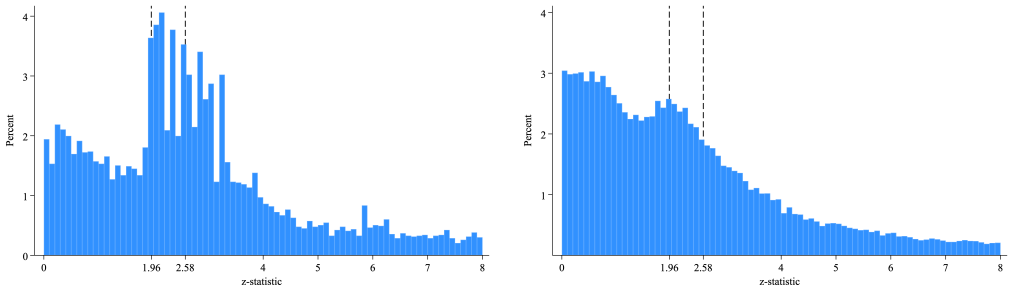
Notes: An observation is a test statistic. Altmetric score winsorized above 2000.

Figure 4: p-Curve: Distribution of p-Values



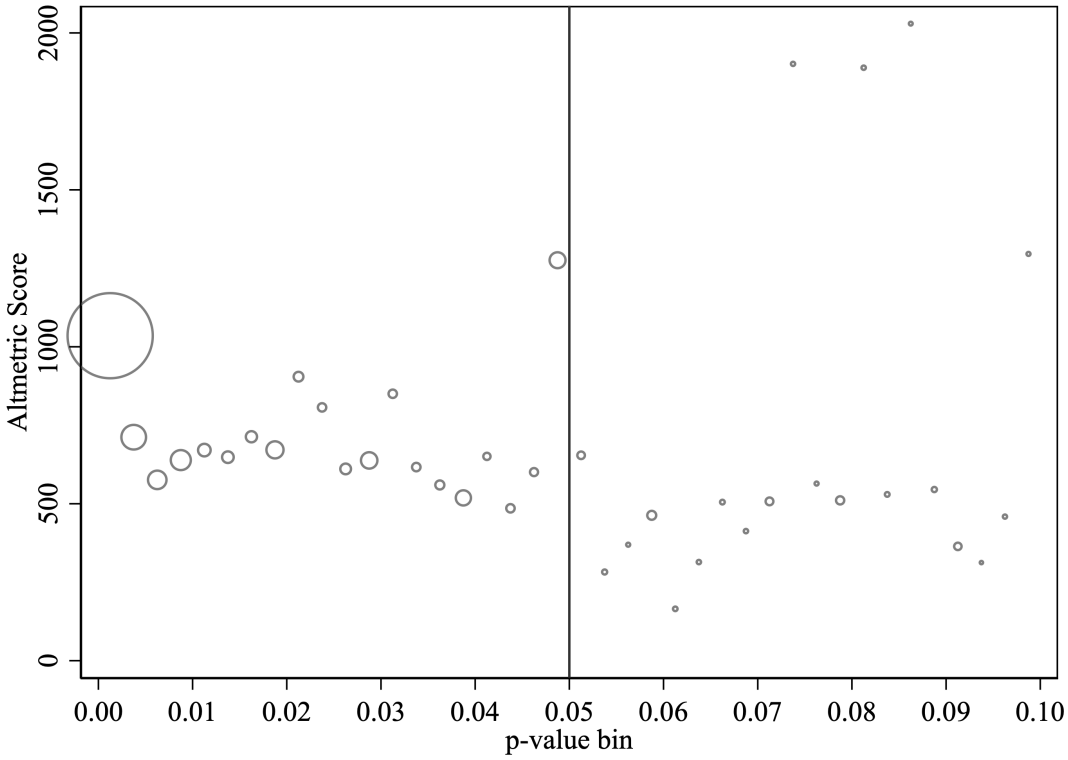
Notes: An observation is a test statistic. The p-curve (histogram of p-values) for full sample. We present [0.0001,0.150] for readability. Bins are 0.01 wide.

Figure 5: t-Curve: Distribution of Test Statistics



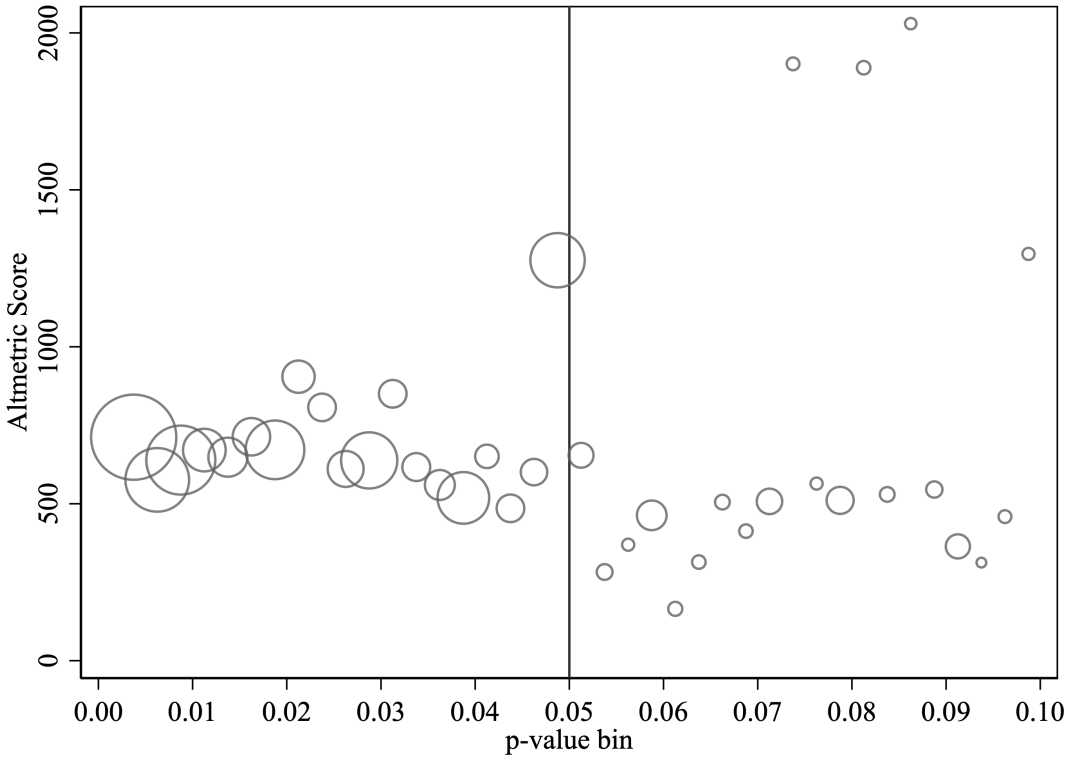
Notes: An observation is a test statistic. Bins of width 0.10. $z > 8$ not displayed for readability. Left panel: Test statistics from BMJ, Lancet, and NEJM articles. Right panel: Test statistics from leading Economics articles (courtesy Brodeur et al. (2016)).

Figure 6: Attention and Statistical Significance [0.00,0.15]



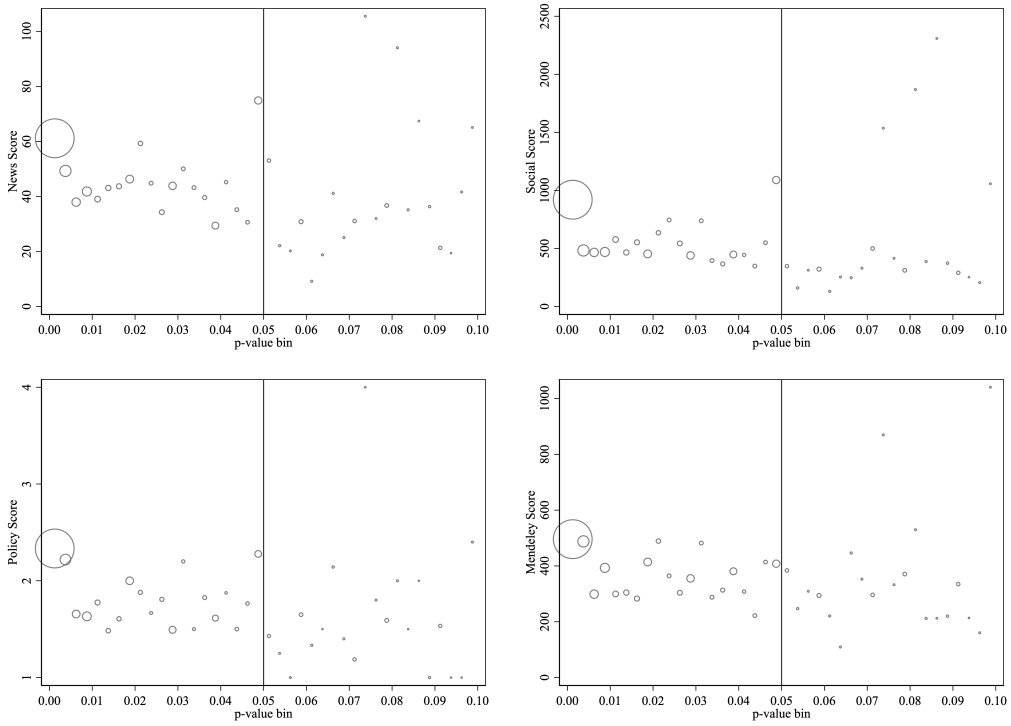
Notes: The vertical value is the average altmetric score per p-value bin. The horizontal variable is the average p-value (in bins 0.0025 wide). Markers sized proportional to number of tests in p-value bin. p-values in the range [0.00,0.15] displayed. Dotted lines are provided at 5% statistical significance threshold.

Figure 7: Attention and Statistical Significance [0.0025,0.15]



Notes: The vertical value is the average altmetric score per p-value bin. The horizontal variable is the average p-value (in bins 0.0025 wide). Markers sized proportional to number of tests in p-value bin. p-values in the range [0.0025,0.15] displayed. Dotted lines are provided at 5% statistical significance threshold.

Figure 8: Attention Categories and Statistical Significance [0.00,0.15]



Notes: Each panel relies on a different media attention variable: News, Social Media and Policy metrics, and the Mendeley measure of attention in the academic community. The horizontal variable is the average p-value (in bins 0.0025 wide). p-values in the range [0.00,0.15] displayed. Dotted line provided at the 5% statistical significance threshold.

7 Tables

Table 1: Summary Statistics

	Mean	Std. Dev.	Min.	Max.
p-value	0.10	0.225	0.00	1.00
Prop. Confidence Intervals	0.70	0.459	0.00	1.00
Prop. Exact p-value	0.30	0.459	0.00	1.00
Year of Publication	2019.05	2.004	2016.00	2022.00
Number of Authors	17.01	13.763	1.00	123.00
Mentions COVID-19	0.11	0.318	0.00	1.00
Meta-Analysis	0.14	0.349	0.00	1.00
Pre-Registered	0.53	0.499	0.00	1.00
Altmetric Score	857.33	1741.454	0.00	30744.00
News	51.65	86.793	0.00	895.00
Social	732.75	2077.664	0.00	42212.00
Policy	0.91	1.559	0.00	18.00
Mendeley	427.84	905.343	0.00	33654.00
Citations	351.63	881.839	0.00	35749.00
Observations	10404			

Notes: Each observation is a test statistic.

Table 2: Attention and Components

	(1) Altmetric	(2) News	(3) Social	(4) Policy	(5) Mendeley
p < 0.05	264.41***	17.07***	223.61***	0.33***	137.78***
	(50.67)	(2.43)	(65.15)	(0.05)	(22.50)
Year FE	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y
Obs.	10,404	10,404	10,404	10,404	10,404
Avg. Sig.	932.68	56.26	800.73	0.99	461.19
Avg. Not. Sig.	603.71	36.15	503.94	0.63	315.59
Scl. Eff. Not Sig.	0.44	0.47	0.44	0.52	0.44

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. Dependent variables indicated in column titles. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 3: Attention by Journal and Method

	(1) BMJ	(2) Lancet	(3) NEJM	(4) Meta	(5) RCT	(6) Other
p < 0.05	166.68***	442.09***	151.74	372.59***	71.25	310.57***
	(53.97)	(92.64)	(105.84)	(126.98)	(59.55)	(88.63)
Year FE	Y	Y	Y	Y	Y	Y
Journal FE				Y	Y	Y
Obs.	3,292	4,256	2,856	1,472	5,283	3,649
Avg. Sig.	584.71	1144.66	991.61	1096.61	728.33	1117.69
Avg. Not. Sig.	420.26	578.14	794.86	488.52	616.06	628.14
Scl. Eff. Not Sig.	0.40	0.76	0.19	0.76	0.12	0.49

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. The dependent variables is Altmetric Score. Column titles refer to subsample restrictions. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 4: Attention (Robustness)

	(1) No FE	(2) C19 Control	(3) Pre-C19	(4) Weighted	(5) Less than 5
p < 0.05	328.97***	213.32***	152.37***	201.40***	170.56***
	(59.53)	(48.12)	(38.26)	(49.51)	(55.09)
COVID-19 FE		Y			
Year FE		Y	Y	Y	Y
Journal FE		Y	Y	Y	Y
Obs.	10,404	10,404	5,813	10,404	5,813
Avg. Sig.	932.68	932.68	622.15	932.68	749.43
Avg. Not. Sig.	603.71	603.71	464.90	603.71	577.00
Scl. Eff. Not Sig.	0.54	0.35	0.33	0.33	0.30

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. The dependent variables is Altmetric Score. Column titles refer to robustness exercises. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 5: Caliper for Attention

	(1)	(2)	(3)
p < 0.05	180.75**	277.54***	304.52**
	(83.08)	(92.84)	(129.29)
Year FE	Y	Y	Y
Journal FE	Y	Y	Y
Obs.	1,228	755	407
Prop. Sig.	0.80	0.80	0.73
p-window	[$\tau \pm .03$]	[$\tau \pm .02$]	[$\tau \pm .01$]
Avg. Sig.	754.18	789.89	995.56
Avg. Not. Sig.	528.25	455.61	495.97
Scl. Eff. Not Sig.	0.34	0.61	0.61

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. The dependent variables is Altmetric Score. Columns differ by p-window or caliper width. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 6: Caliper for Attention and Components

	(1) Altmetric	(2) News	(3) Social	(4) Policy	(5) Mendeley
p < 0.05	277.54***	12.22**	300.79***	0.06	54.67*
	(92.84)	(5.64)	(98.87)	(0.09)	(30.15)
Year FE	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y
Obs.	755	755	755	755	755
Prop. Sig.	0.80	0.80	0.80	0.80	0.80
p-window	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$
Avg. Sig.	789.89	47.81	647.93	0.70	370.05
Avg. Not. Sig.	455.61	32.82	289.22	0.62	310.17
Scl. Eff. Not Sig.	0.61	0.37	1.04	0.10	0.18

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. Columns differ by the titled dependent variable. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 7: Caliper for Research by Journal and Method

	(1) BMJ	(2) Lancet	(3) NEJM	(4) Meta	(5) RCT	(6) Other
p < 0.05	140.10	485.67***	131.76	548.93*	14.93	659.89***
	(128.76)	(185.28)	(99.99)	(281.79)	(72.43)	(248.06)
Year FE	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Obs.	253	275	227	116	439	200
Prop. Sig.	0.77	0.85	0.76	0.82	0.77	0.84
p-window	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$
Avg. Sig.	542.32	1086.01	671.54	1265.48	463.44	1179.67
Avg. Not. Sig.	473.09	382.98	493.65	344.38	450.98	543.06
Scl. Eff. Not Sig.	0.30	1.27	0.27	1.59	0.03	1.22

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. Columns differ by the titled subsample. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 8: Caliper for Attention (Robustness)

	(1) No FE	(2) C19 Control	(3) Pre-C19	(4) Weighted	(5) Less than 5
p < 0.05	334.28***	220.48**	147.80*	248.83***	193.73*
	(113.94)	(91.71)	(75.79)	(80.63)	(105.20)
COVID-19 FE		Y			
Year FE		Y	Y	Y	Y
Journal FE		Y	Y	Y	Y
Obs.	755	755	474	755	461
Prop. Sig.	0.80	0.80	0.79	0.80	0.79
p-window	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$	$[\tau \pm .02]$
Avg. Sig.	789.89	789.89	592.30	789.89	648.87
Avg. Not. Sig.	455.61	455.61	431.32	455.61	458.99
Scl. Eff. Not Sig.	0.73	0.48	0.34	0.55	0.42

An observation is a test statistic. The primary independent variable is an indicator which takes the value 1 if the test statistic's p-value is less than or equal to 0.05. Columns differ by the titled robustness check. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated with ordinary least squares with standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 9: Regression Discontinuity for Attention (Preferred)

	Discontinuity at the 5% Threshold			
	(1)	(2)	(3)	(4)
p < 0.05	663.22***	722.24***	729.88***	743.88***
	(228.78)	(244.90)	(251.95)	(256.59)
Obs.	10,404	1,228	755	407
p-Window	Optimal	$[\tau \pm 0.03]$	$[\tau \pm 0.02]$	$[\tau \pm 0.01]$
Poly. Deg.	2	2	2	2
Prop. Sig.	0.77	0.80	0.80	0.73
Avg. Sig.	932.68	754.18	789.89	995.56
Avg. Not. Sig.	603.71	528.25	455.61	495.97
Scl. Eff. Not Sig.	1.10	1.37	1.60	1.50

An observation is a test statistic. The primary independent variable is the estimate of $\hat{\tau}$ from Equation 3. Columns differ by bandwidth. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated using Calonico et al. (2017), with default settings, a polynomial of order two, and standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 10: Regression Discontinuity for Attention (Linear)

	Discontinuity at the 5% Threshold			
	(1)	(2)	(3)	(4)
p < 0.05	648.07***	696.46***	704.15***	718.93***
	(215.84)	(232.07)	(244.10)	(250.46)
Obs.	10,404	1,228	755	407
p-Window	Optimal	$[\tau \pm 0.03]$	$[\tau \pm 0.02]$	$[\tau \pm 0.01]$
Poly. Deg.	1	1	1	1
Prop. Sig.	0.77	0.80	0.80	0.73
Avg. Sig.	932.68	754.18	789.89	995.56
Avg. Not. Sig.	603.71	528.25	455.61	495.97
Scl. Eff. Not Sig.	1.07	1.32	1.55	1.45

An observation is a test statistic. The primary independent variable is the estimate of $\hat{\tau}$ from Equation 3. Columns differ by bandwidth. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated using Calonico et al. (2017), with default settings, a polynomial of order one, and standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 11: Regression Discontinuity for Attention and Components

	Discontinuity at the 5% Threshold				
	(1)	(2)	(3)	(4)	(5)
	Altmetric	News	Social	Policy	Mendeley
p < 0.05	663.22***	24.26*	775.36***	0.42**	41.86
	(228.78)	(14.54)	(225.36)	(0.17)	(77.53)
Obs.	10,404	10,404	10,404	10,404	10,404
p-Window	Optimal	Optimal	Optimal	Optimal	Optimal
Poly. Deg.	2	2	2	2	2
Prop. Sig.	0.77	0.77	0.77	0.77	0.77
Avg. Sig.	932.68	56.26	800.73	0.99	461.19
Avg. Not. Sig.	603.71	36.15	503.94	0.63	315.59
Scl. Eff. Not Sig.	1.10	0.67	1.54	0.67	0.13

An observation is a test statistic. The primary independent variable is the estimate of $\hat{\tau}$ from Equation 3. Columns differ by dependent variable. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated using Calonico et al. (2017), with default settings, a polynomial of order two, and standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

Table 12: Regression Discontinuity for Attention by Journal and Method

	Discontinuity at the 5% Threshold					
	(1) BMJ	(2) Lancet	(3) NEJM	(4) Meta	(5) RCT	(6) Other
p < 0.05	38.96	1331.15***	316.67	1557.55***	15.55	1098.34***
	(201.06)	(335.77)	(201.64)	(570.41)	(136.95)	(340.38)
Obs.	1,311	1,248	1,294	516	2,385	952
p-Window	Optimal	Optimal	Optimal	Optimal	Optimal	Optimal
Poly. Deg.	2	2	2	2	2	2
Prop. Sig.	0.39	0.42	0.33	0.45	0.36	0.41
Avg. Sig.	616.55	850.15	708.16	993.85	543.02	970.94
Avg. Not. Sig.	420.26	578.14	794.86	488.52	616.06	628.14
Scl. Eff. Not Sig.	0.09	2.30	0.40	3.19	0.03	1.75

An observation is a test statistic. The primary independent variable is the estimate of $\hat{\tau}$ from Equation 3. Columns differ by the titled subsample. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated using Calonico et al. (2017), with default settings, a polynomial of order two, and standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

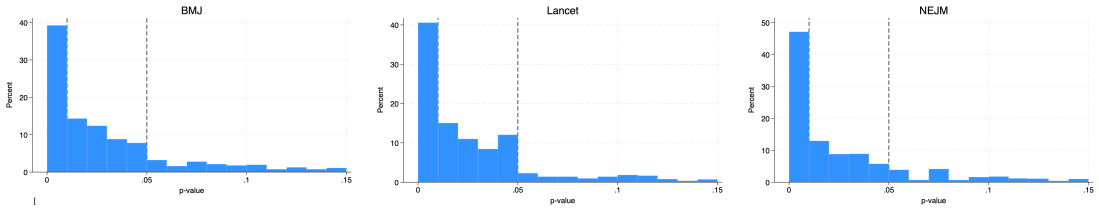
Table 13: Regression Discontinuity for Attention (Robustness)

	Discontinuity at the 5% Threshold					
	(1) No FE	(2) Yes FE	(3) COVID	(4) Pre-COVID	(5) Weighted	(6) Less than 5
p < 0.05	663.22***	423.95*	374.97*	529.33***	409.02**	219.56
	(228.78)	(221.22)	(220.42)	(184.06)	(176.99)	(220.93)
Obs.	10,404	10,404	10,404	5,813	10,404	5,813
p-Window	Optimal	Optimal	Optimal	Optimal	Optimal	Optimal
Poly. Deg.	2	2	2	2	2	2
Prop. Sig.	0.77	0.77	0.77	0.75	0.77	0.73
Avg. Sig.	932.68	932.68	932.68	622.15	932.68	749.43
Avg. Not. Sig.	603.71	603.71	603.71	464.90	603.71	577.00
Scl. Eff. Not Sig.	1.10	0.70	0.62	1.14	0.68	0.38

An observation is a test statistic. The primary independent variable is the estimate of $\hat{\tau}$ from Equation 3. Columns differ by the titled robustness exercise. Avg. Sig. refers to the average dependent variable value for significant test statistics. Scl. Eff. Not Sig. refers to the indicator variable's regression coefficient divided by the average dependent variable value for not significant test statistics; the 'percent' effect on the control group when treated. Estimated using Calonico et al. (2017), with default settings, a polynomial of order two, and standard errors clustered at the article level. * $p < 0.10$ ** $p < 0.05$ and *** $p < 0.01$.

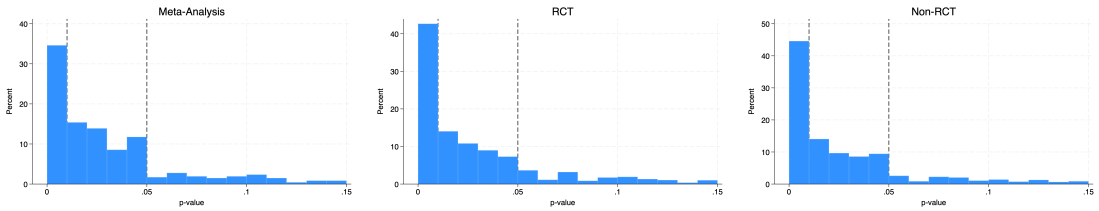
8 Appendix Figures

Figure A1: p-curve by Journal



Notes: Histograms of p-values in the sample of articles (also called a p-curve). Each panel presents p-values in the range [0.001,0.150], separately by journal of publication.

Figure A2: p-curve by Method



Notes: Histograms of p-values in the sample of articles (also called a p-curve). Each panel presents p-values in the range [0.001,0.150], separately by method applied in the article (meta-analysis, RCT, or other).