Bad Science: Retractions and Media Coverage.

Eleonora Alabrese*

June 15, 2023

Click here for the latest version.

Abstract

Flawed research can be harmful both within and outside of academia. Even when published research has been retracted and refuted by the scientific community, it may continue to be a source of misinformation. The media can play an important role in drawing broader attention to research, but may also ensure that research, once retracted, ceases to feature in popular discourse. Yet, there is little evidence on whether media reporting influences the retraction process and authors' careers. Using a conditional difference-in-differences strategy, this paper shows that articles that gained popularity in the media at publication and were later retracted face heavy citation losses, while subsequent citations become more *accurate*. Further, authors of such papers see a permanent decline in research output. Lastly, the paper provides evidence that media can influence both the likelihood of retraction and its timing, highlighting that the media can play an important role in contributing to the integrity of the research process.

Keywords: Science; Retractions; Media Coverage; Misinformation; Altmetric; Citations; Career impact.

*University of Warwick. Email address: E.Alabrese@warwick.ac.uk. Acknowledgements: I thank Thiemo Fetzer, Sascha Becker and Andreas Stegmann for the precious mentoring and guidance. I want to thank Arun Advani, Wiji Arulampalam, Manuel Bagues, Gabriella Conti, Mirko Draca, Ludovica Gazze, Vardges Levonyan, Andrew Oswald, Kirill Pogorelskiy, Carlo Schwarz, Paula Stephan, Andrea Tesei, Natalia Zinovyeva and participants of NSEF Workshop 2022 (Napoli Federico II); 2022 European ESA Meeting (Bologna); EEA-ESEM Congress 2022 (Bocconi); CAGE Summer School 2022 (Warwick); PaCSS Politics and Computational Social Science; PolMeth Europe meeting (Hamburg); Warwick Economics PhD Conference 2022 (Warwick); QMUL Economics and Finance Workshop (Queen Mary); Applied Young Economist Webinar; Workshop on the Organisation, Economics, and Policy of Scientific Research 2022 (KU Leuven); Monash-Warwick-Zurich Text-As-Data Workshop; RES 2022 Annual Conference; RGS Doctoral Conference in Economics; International Conference on Social Media and Society 2022; General Online Research conference 2022 (Berlin); Workshop on Political Economy (UniBZ-Brunek), the Political Economy and Public Economics Research Away Day (Warwick) and Warwick internal seminars for insightful comments and suggestions. All errors remain my own.

1 Introduction

Science is not static, and the number of active journals and published articles have grown significantly over time, a proliferation that could influence research quality. The surge in the absolute number of retractions across all disciplines has alarmed many in the scientific community. Nonetheless, retractions remain relatively rare, involving 4 in every $10,000$ published papers. Sometimes the reason is an honest error while 60% is due to some deliberate misconduct (Brainard, 2018). Retracted research often persists and perpetuates misinformation among academic and non-academic audiences (Lewandowsky et al., 2012; The Economist, 2021; Peng, Romero and Horvát, 2022). A recent case study illustrates how in 96% of direct citations post-retraction, research continues to be cited positively and uncritically in support of a medical intervention, without mentioning its retraction for falsifying data, which could harm further research (Schneider et al., 2020). Inattention (Woo and Walsh, 2021) or failure to update beliefs (Goncalves, Libgober and Willis, 2021) are two possible channels that could explain the ongoing use of retracted work. For a flawed study to be quashed, the literature on retractions suggests the visibility and the accessibility of retractions and their associated retraction notices are decisive (Bar-Ilan and Halevi, 2017; Teixeira da Silva and Bornemann-Cimenti, 2017; Cox, Craig and Tourish, 2018; Bordignon, 2020). Yet, there is limited evidence on the potential role of media reporting of science.[1] The media can play an important role in drawing broader attention to research, but may as well ensure that research, once retracted, ceases to feature in the academic discourse.

This work studies whether and how media attention at the time of publication impacts the survival of retracted work within the academic community (i.e. future citations) and how it affects the future career outcomes of its authors (i.e. future publication rate). Two recent examples of well-published fraudulent Covid-19 research illustrate the role of media in drawing attention to research, while potentially interacting with the retraction process (Mehra, Desai, Kuy, Henry and Patel, 2020; Mehra, Desai, Ruschitzka and Patel, 2020). These two studies attracted wide media coverage right after publication, and their findings led to a global halt of hydroxychloroquine trials. However, high scrutiny from the scientific community and an investigation from the Guardian[2] led to a prompt retraction, as the underlining data were fabri-

[1] See Hesselmann et al. (2017) for a review on scientific retractions.

[2] Guardian article.

cated. Nonetheless, it is unclear whether this interaction with the media helps reduce the amount of misinformation (proxied by citations) or increases the cost of retraction (proxied by authors' productivity). This paper sheds light on the empirical relevance of this media attention channel.

Studying the media attention channel is not straightforward. There are concerns related to the timing and the content of the coverage. Media attention may increase the salience of a study, leading to higher future citations (Phillips et al., 1991; Fanelli, 2013), or the media could exert a monitoring role, criticizing a study and updating the public about the shortcomings of a paper (Peng, Romero and Horvát, 2022), potentially reducing future citations (see Whitely, 1994, for an example).[3] Therefore, to test whether media attention leads to differential scrutiny and punishment, I contrast research articles that appeared in the media in a tight window around publication, to those that did not. These *early mentions* (i.e. appearances in either newspapers or blogs within two weeks from publication) are assumed to advertise the original research findings and are less likely attributable to the retraction per se. This approach is suitable given that early media coverage is informative about later coverage (see Serghiou, Marton and Ioannidis, 2021; Peng, Romero and Horvát, 2022).

Furthermore, the endogeneity of media creates potential issues of selection into retraction. Some authors may focus on research questions which are more likely to attract media attention or manipulate their findings to attract more coverage. However, due to their relevance, these questions may receive differential scrutiny that, in turn, could impact their likelihood of retraction and their retraction timing. In addition, popular studies featuring in the media often involve eminent authors (e.g. Ivanova et al., 2013) who might be less impacted by retractions (Azoulay, Bonatti and Krieger, 2017; Jin et al., 2019). I use a refined conditional difference-in-differences estimation strategy to address these issues (Heckman, Ichimura and Todd, 1997; Heckman et al., 1998; Buscha et al., 2012). This approach compares the evolution of citations of retracted papers to that of closely matched control group papers, before and after retraction — while contrasting papers with or without *early mentions*. Control group papers are selected to mimic multiple ex-ante characteristics of retracted papers, those that best simulate the citation path of retracted papers absent the retraction. Papers in the control group are (a) published in the same journal and year, (b) have similar citation trends in the years prior to the retraction, and (c) have attracted compara-

[3]However, Serghiou, Marton and Ioannidis (2021) finds that retracted articles may receive high coverage, but pre-retraction coverage far outweighs post-retraction coverage.

ble media coverage as their retracted counterpart at publication.[4] Matching on *early mentions* is further justified by recent findings that retracted papers experience more coverage than non-retracted papers published in the same journal, similar publication year and author characteristics (see Peng, Romero and Horvát, 2022). This strategy allows to control for (i) selection on matched observables and (ii) selection on time-invariant unobservables, such as authors and teams' prior reputation. In addition, this approach enables a test of the common-trends assumption inherent to such a research design.

Using the same matching strategy, I conduct a difference-in-differences estimation at the author level. This estimation compares the publication record of authors of retracted papers to that of authors of matched control papers before and after their first observed retraction — distinguishing cases where the original publication had *early mentions* or not. Crucially, this analysis allows to capture heterogeneous effects by authors' order of appearance in the coauthorship roster and by authors' seniority (i.e. based on H-index pre-retraction).

There could be other challenges to the validity of the estimates. For example, there could be unobservable paper- and author-specific factors that interact with time to confound these estimates. To allay such concerns, both paper- and author-level exercises incorporate a large set of additional control variables. All paper-level specifications include age, calendar-year, and paper-specific effects. All author-level specifications include career length, calendar-year and author-specific effects.

Despite this rich econometric framework, there may be residual concerns that the endogeneity of media attention — at the publication stage — may not be adequately accounted for. I employ two more strategies to tackle these concerns. First, using methods from computational linguistics, I build a predictive model of media coverage based on words in paper titles. I then use this prediction to separate (i) the impact of the arguably exogenous *excess* coverage on the likelihood of retraction, from (ii) the media coverage that a study may receive due to its authors self-selecting into a general interest topic. Second, I study the relationship between the average coverage of non-retracted articles within a journal-year (*journal visibility*) and the timing of detection of the retracted articles that appeared in those journals.

This paper draws on a comprehensive data collection effort to measure (i) the rep-

[4] From a methodological point of view, Furman, Jensen and Murray (2012); Lu et al. (2013); Azoulay et al. (2015); Mongeon and Larivière (2016); Azoulay, Bonatti and Krieger (2017); Jin et al. (2019) all adopt similar difference-in-differences strategies to estimate the effect of a retraction shock, with the exception of the matching on early mentions.

resentation of research in broader media through *Altmetrics*, along with (ii) papers' citation history and (iii) authors' publication history. Media coverage data are recent, I therefore select papers that were published after 2010 and later retracted. Furthermore, I balance relevance and manageability of sample size by focusing on retractions published in highly ranked journals of each available discipline[5] and listed in the *RetractionWatch* database. Notice that publications in top journals receive more media coverage (Yin et al., 2022), which means that a considerable number of papers receive some coverage. This could introduce a bias as Woo and Walsh (2021) find that top journal citations to retracted articles are more likely to occur outside one's field and are potentially more prone to misinformation. I argue that investigating the relationship between media and punishment after retraction is more compelling for high-impact research. Yet, the average citation penalty for the retracted papers I select ($\simeq$65% of forward citations) is aligned with previous literature that does not specifically focus on high rank journals (Furman, Jensen and Murray, 2012; Azoulay et al., 2015).

Using this sample, I first show descriptive evidence that retracted papers attract some media coverage, although publication and retraction events feature differently across outlets. Newspapers are more likely to cover the publication of a paper, while blogs are more likely to cover its retraction.[6] However, blogs have a more scientific target audience.[7]

Moreover, difference-in-differences estimates at the paper level show that retractions harm citations of retracted papers, and media coverage amplifies this effect (on average, media contributes to a $\simeq$20-28% further reduction in forward citations). This aggravating effect is present only in hard sciences.[8] and the more severe cases of misconduct drive the media effect.[9] To prove that this exacerbating effect of media on future citations is robust, I first show that media is not a proxy for other paper features. Indeed, papers with high cumulative citations (ex-ante) do not drive nor

[5] I retain retractions that either appear in *Scimago* top ten journals or in *Google scholar* top journals for each available category.

[6] Visual differences emerge when looking at the leaning of (US) news coverage, based on the Gentzkow and Shapiro (2010) index of media slant, which deserves further investigation in future work.

[7] RetractionWatch alone represents about one third of all blog mentions for the selected retractions.

[8] The heterogeneity for hard versus social sciences suggests distinct publication practices may impact the visibility of a retraction. The difference across disciplines may also reflect the public perceptions that social science research is less "absolute" as the object of study is more volatile.

[9] The separation between severe and non-severe cases of misconduct comes from the classification developed by Woo and Walsh (2021).

confound media estimates. Estimates are not sensitive to including or excluding the year of retraction in the *Post* indicator. Excluding non-actively cited papers or excluding more recent publications does not impact estimates. Finally, alternative strategies that look at *citation statements*[10] or study the relationship between citation penalty and *journal visibility*[11] confirm that media intensifies the citation penalty after retraction.

I propose two mechanisms that might explain the effect of media on citations: (a) higher scrutiny by the scientific community of a paper that gained popularity; (b) additional information provided to some part of the scientific community which would have otherwise remained unaware of the retraction. For (b), I check whether the textual content of post-retraction citations significantly differs with media coverage. The presumption is that absent the information mechanism, retracted papers with media coverage would experience fewer citations after retraction but with no relative change in their level of *accuracy*. In contrast, I find that with media coverage, new citations mention more often that the paper is retracted. This finding suggests that, with media coverage, scientists become more aware of a retraction and correctly acknowledge it when citing the original paper, reducing potential misinformation.

Consistently, difference-in-differences estimates at the author level show that retractions have a negative impact on authors' future *productivity*. However, this negative impact is large and permanent only if the original publication had some media exposure ($\simeq$10% larger reduction in future publication rate relative to a case with no media). This differential effect of media is significantly larger for first authors, the most visible and main contributors for medical research,[12] which aligns with rational updating from the scientific community. Effects are also evident among senior authors and in cases of severe misconduct.

Furthermore, I show evidence that suggests there is some selection into retraction. On the one hand, papers that attract predictable (endogenous) media coverage are less likely to be retracted.[13] On the other hand, papers with exogenous *excess* coverage get retracted more often. However, both effects are modest. Finally, I show that journals that publish articles that are popular in the media are journals that retract faster (one standard deviation increase in journal visibility implies a reduction in the

[10] The *citation statements* are textual paragraphs where a citation appears.

[11] The *journal visibility* is calculated by averaging the media coverage of non-retracted articles published in a specific journal and year.

[12] Retracted research appear predominantly among life sciences and health.

[13] Words in titles such as climate, stem, meta-analysis, and trial are predictive of media around publication.

timing for detection of 15%).[14]

Overall, this paper investigates whether media coverage of scientific articles influences the auto-correcting process of science by observing media coverage of research articles across outlets and time. Media coverage — at the publication stage — amplifies the penalty for flawed research in terms of both future citations and authors' future publication rate. Although media coverage seems to help the auto-correcting process of science, this implies that (a) plenty of *wrong* science remains unnoticed and (b) that academia needs better strategies to raise the level of scrutiny and reduce incentives for poor-quality research.

The paper contributes to the large body of research on scientific retractions. It is closest to studies that estimate the causal effect of retraction on citations of retracted papers (Furman, Jensen and Murray, 2012), on authors' previous publications (Lu et al., 2013; Azoulay, Bonatti and Krieger, 2017; Jin et al., 2019), or their future research output (Mongeon and Larivière, 2016) and potential spillover to the related field (Azoulay et al., 2015). The main contribution is to show that media coverage — at publication — amplifies the causal effect of retraction on citations of retracted papers, and substantially explains the negative and persistent impact on the future research output of retracted authors. Further, I show that media attention may impact the likelihood of retraction and its timing. To address the risk of selection into treatment in this result, I improve the methodological approach by matching on early media exposure of papers. Only a few studies on retractions mention the role of media (Sugawara et al., 2017; Sarathchandra and McCright, 2017; Serghiou, Marton and Ioannidis, 2021). Among these, this paper is closest to Peng, Romero and Horvát (2022) who use the similar data to show that retractions are ineffective at reducing online attention. They find that retracted papers receive more coverage after publication than non-retracted control papers from the same journals with similar publication years, number of coauthors, and authors' impact. The current paper addresses a complementary question of whether media attention intensifies the effect of retraction on papers' citations and authors' careers, thus reducing misinformation and increasing authors' cost of retraction. As the question differs, the matching strategy also differs. In fact, I compare retracted papers to never-retracted control papers from the same journal and year, with similar pre-trends in citations, and with similar

[14] In these journals, citation penalties for retracted papers are also sizable, which corroborates the main media effect on yearly citations.

salience at publication (i.e. similar early mentions).[15]

The paper also relates to the literature investigating the relationship between science and the media (Weingart, 1998; Phillips et al., 1991; Fanelli, 2013; Ivanova et al., 2013; Sumner et al., 2014; Dumas-Mallet et al., 2020; Ziegler, 2021) to which I contribute by showing that media coverage of subsequently retracted papers can influence the reputation of papers and authors, within science. This work further contributes to the literature on factors influencing citation rates (for example see Card and Dellavigna, 2020; Card et al., 2020; see also Tahamtan, Safipour Afshar and Ahamdzadeh, 2016, for a review of the literature), to which I add that (ex-ante) salience impacts the citations of a paper and its authors' careers in case of a negative event (such as a retraction). Finally, this work relates to the broad literature on misinformation and how media channels influence politics and public policies (for an example see Allcott and Gentzkow, 2017; Lazer et al., 2018; see also Prat and Strömberg, 2013, for a review of the literature). I contribute to this literature by showing that media coverage attenuates misinformation within academia. At the same time, I illustrate that newspapers give more coverage to the publication of a paper than its retraction, which creates the potential for disseminating misinformation to a larger audience.

In the remainder of the paper, Section 2 illustrates the institutional context of retractions. Section 3 describes the data, the sample selection and the main empirical strategy. Section 4 presents descriptive results on the media coverage of retraction. Section 5 and Section 6 provide a detailed presentation of results at the paper- and author-level respectively. Section 7 concludes.

2 Background

Understanding the incentives and governance regulating scientific knowledge production, dissemination and accumulation is crucial to this work. In what follows I discuss relevant aspects of the publication system.

One of the most discussed institutional setting is the peer review system. Articles are submitted and reviewed by independent experts before being accepted for publication. This system is used to maintain high quality standards while allowing a suitable publication timing, even though practices vary greatly across disciplines and

[15]This choice is motivated by the fact that media may impact selection into retraction. Therefore, control papers should be equally likely to be detected (if wrong).

journals. Peer review provides only limited guarantee against bad science.

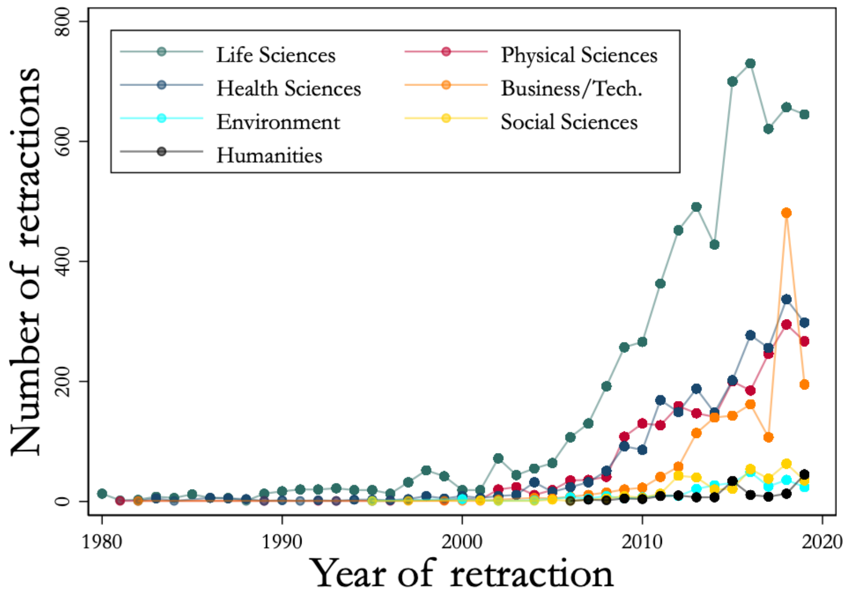
Another aspect which is crucial for the scientific communication is the practice of citing related literature. Citations allows authors to effectively contextualise their research article with respect to pre-existing literature while acknowledging original contributions from previous researchers. Citations are regarded as an indicator of the importance of scientific findings and of their creators and can be negatively impacted by a retraction (Furman, Jensen and Murray, 2012; Lu et al., 2013; Azoulay et al., 2015; Mongeon and Larivière, 2016; Azoulay, Bonatti and Krieger, 2017; Jin et al., 2019).

In academic publishing, a retraction is the result of a procedure used by journals to alert readers that a published article should be removed from the literature. A retraction may occur when a major error (e.g. in the analysis or methods) invalidates the conclusions of the article, or in the presence of misconduct (e.g. fabricated data, manipulated images, plagiarism, duplicate publication, research without required ethical approvals etc). It differs from a correction issued in case of an error or omission which can impact the interpretation of the article, but where the scholarly integrity remains intact. The surge in the absolute number of retractions across all disciplines has alarmed many in the scientific community (see Figure 1). Nonetheless, retractions remain relatively rare, involving 4 in every $10,000$ published papers, of which 60% due to some type of misconduct. Both rates have been rising steadily over time (Brainard, 2018).

A retraction can be initiated by the editors of a journal, by some or all of the authors, or their institution. Retractions are typically complemented by a notice meant to clarify the reason for the decision. But the information contained in notices varies significantly. Some explain the details which lead to the retraction outcome and inform whether an article's results and conclusion should be disregarded entirely or in part. Others are rather brief and vague.

A further element of discussion is therefore the visibility and accessibility of both retractions and notices. "Authors are responsible for checking that none of the references cite retracted articles except in the context of referring to the retraction" (International Committee of Medical Journal Editors 2019). Retractions are usually published and linked to the original publication and can be often identified via different sources (e.g. libraries, databases and search engines) but inaccurate citations still remain. Whether the scientific community is fully aware of the retraction is decisive, yet the current institutional setting does not suffice. Schneider et al. (2020) found

Figure 1: Retractions over time and across subjects.



Note: Numbers reflect the full *RetractionWatch* database as of July 2020, for visual purposes one outlier publisher (e.g. IEEE) was excluded.

that in the case of an infamous clinical trial (Matsuyama et al., 2005), in which data were falsified leading to a retraction in 2008, the retraction is not mentioned by 96% of citations post-retraction. 41% of these inaccurate citations describe the paper in detail, leading to possible disinformation. On the other hand, Piller (2021) looked at the recent case of two high-profile Covid-19 retractions (Mehra, Desai, Kuy, Henry and Patel, 2020; Mehra, Desai, Ruschitzka and Patel, 2020) and finds that 52.5% of the citations do not correctly mention the paper's status. In what follows, I will illustrate that the broad media attention attracted by the latter case could be a relevant factor behind the different survival of the two examples just discussed.

In this respect, recent efforts make use of online media platforms to alert scientists of retractions, as in the case of the specialised blog *RetractionWatch*. This reports on retractions and gathers information surronuding specific retraction events, such as which of the authors is responsible for the article's ultimate fate — information which is usually hard to acquire based on the notice alone. New tools are also emerging as in the case of *Scite.ai*, a recently launched platform which categorises references, monitors retracted papers by searching through Crossref, PubMed, and the RetractionWatch database, and flags both citing and retracted papers on Twitter.

3 Data and Method

3.1 Data

This study combines multiple data sources on scientific publications and their authors which I hereby list in detail.

Retractions. The treatment sample is extracted from the *RetractionWatch* database[16] which Brainard (2018) defines as "the largest-ever database of retracted articles". The dataset contains a list of retracted research articles[17] together with the following information: title, doi, date of publication, date of retraction, journal, name of authors and their institutions, list of reasons for retraction, and when available, a link to the associated blog post reporting on the paper background story.

Journal ranking. I further select papers featuring in either *Scimago* or *Google scholar* rankings. The selected journals appear either as one of the ten highest ranked in Scimago in any of the available subjects or among those listed in Google scholar top publications in any of the existing categories.

Citations and authors' publications. Yearly citations and authors publications are the main outcome of this study and are collected for each article and author using *Scopus*, one of the two largest abstract and citation database of peer-reviewed literature.[18]

Media coverage. Data on online mentions are gathered thanks to *Altmetric*, a company that since 2011 tracks the attention that research outputs receive online.[19] For each paper I retrieve the *Altscore*, an aggregate measure of online mentions (i.e. it combines all mentions across outlets giving a higher weight to outlets such as newspapers, see appendix Table A.1), and details about single mentions (e.g. date, url, author, title, summary).

[16] Version obtained in July 2020.

[17] Dense since the '80s.

[18] For the period considered in the analysis, there exists little difference between Scoups and WoS in terms of coverage (see: Scopus vs. WoS). Scopus though has the advantage of having an API easily accessible via *rscopus*, a library by John Muschelli available on R.

[19] I here focus on sources with the highest number of mentions (i.e. newspapers, blogs and Twitter) though Altmetric collects mentions from numerous additional outlets (e.g. Pubpeer, Wikipedia).

Citation textual content. I obtained data on the content of citation statements quoting the research articles in the sample with the support of *Scite.ai*, a recently launched start-up that uses text analysis to categorize reference statements. For each pair of citing and cited study, statements are categorised as "mentioning", "contrasting" and "supportive".[20] In addition, access is gathered for any statement containing the words "*etract*" or "*ithdraw*".[21]

3.2 Empirical strategy

This work investigates the possibility that media coverage influences scientists' awareness and assessment of research findings (looking at citations of retracted papers) and authors' careers (looking at the publication rate of retracted authors, see Section 6 for author-level analysis). Holding other factors constant, a loss in citations and lower authors' rate of publication, reflects an erosion of trust in the authors' work by the scientific community.

To understand the interplay between the retraction of a paper and the information available online one needs to consider how scientific publications feature in the media and what challenges this poses in terms of identification.

A research article that is accepted for publication may endogenously attract media coverage. Online attention may depend on factors such as the salience of a topic, the importance of the findings, the prestige of authors and publishers, the presence of a press release (Sumner et al., 2014, 2016). Media coverage can therefore bring publicity to a paper increasing future citations (Phillips et al., 1991; Fanelli, 2013) as well as prompt higher scrutiny from the scientific community making any fault more likely to emerge. Online attention can finally inform about the fate of an article as in the case of an expression of concern or a retraction (Serghiou, Marton and Ioannidis, 2021), information that could reach unaware scientists that would otherwise incorrectly cite a flawed article.

Therefore media endogeneity, together with observables and unobservable characteristics of papers and authors can create issues of selection into retraction (treatment). I tackle this challenge using a conditional difference-in-differences strategy (Heckman, Ichimura and Todd, 1997; Heckman et al., 1998; Buscha et al., 2012) which compares retracted papers to matched (never-retracted) controls, before and after the

[20] According to Rosati (2021)

[21] Manually checked to exclude any false positive.

retraction, while contrasting papers with or without *early mentions* (i.e. apperances in either newspapers or blogs within two weeks from publication). Smith and Todd (2005) shows that the difference-in-differences matching estimator performs the best among nonexperimental matching based estimators.

Early mentions are the preferred measure of media coverage as this facilitates identification. These mentions are assumed to broadly advertise the original research findings and are therefore virtually independent from the retraction.

Control papers are choosen to mimic multiple (ex-ante) characteristics of retracted papers, such that they could best simulate the citation path of retracted papers absent the retraction. Specifically, controls are (a) published in the same journal and year, (b) have similar citation trends in the years prior the retraction, and (c) attracted comparable media coverage at publication, as their retracted counterpart.

The rest of this session explains in details the process determining the sample of treated papers, the matching strategy employed to choose control papers, and the main regression model.

3.3 Treatment group

The full RetractionWatch database counts $N_r = 21,968$ retractions starting from 1980. Provided that data availability on online mentions is only relatively recent, I select retracted papers both published and eventually retracted after 2010 ($N_r = 11,258$). Only research articles[22] with non missing paper DOI and retraction notice DOI are maintained.[23] To balance relevance and manageability of sample size, I focus on articles published in journals featuring in either Google scholar top journals by field or among the ten highest ranked journals in Scimago per subject category.[24] Remaining papers are certainly relevant for the scientific community, hence, it is important to study whether in this case disinformation is halted or fostered by media coverage.[25] In addition, publications in reputable journals may be more likely to attract media coverage (Yin et al., 2022), thus helping identification.[26] Next, I exclude articles for which I cannot find any author with at least one publication in the 9 years before the

[22] Excluding for examples: conference abstracts and clinical studies, $N_r = 6,676$.

[23] $N_r = 6,189$.

[24] $N_r = 1,163$.

[25] Woo and Walsh (2021) find that top journal citations to retracted articles are more likely to occur outside one's field and are potentially more prone to disinformation.

[26] This could potentially introduce a bias. Yet, the reader may be reassured that I observe an average penalty for selected papers which is alligned with previous literature (Furman, Jensen and Murray, 2012; Azoulay et al., 2015).

retraction event.[27] Of these I retain cases for which I can find an appropriate control, leading to a final sample of $N_r = 990$.

3.4 Control group

Trends in citations vary across disciplines, age and media coverage, hence, control publications were selected to mimic pre-retraction characteristics of the treated. This strategy draws from the approach first used in the literature on retractions by Furman, Jensen and Murray (2012) and further developed by Lu et al. (2013) and Jin et al. (2019). The main assumption is that treated papers would continue to perform similarly to control ones in absence of a retraction event.

The selection of the control group proceeds in steps. For each retracted paper I search in Scopus for studies[28] published in the same journal and year of the treated.[29] For each retracted i and potential control pair j I compute the measures listed below.

- *Absolute arithmetic distance* in citations.

$$|AD| = \left|\sum_{t=pub}^{retr-1} (c_{it} - c_{jt})\right|;$$

- *Euclidean distance* in citations.

$$ED = \left[\sum_{t=pub}^{retr-1} (c_{it} - c_{jt})^2\right]^{1/2};$$

where c_i indicate the citations paper i receives in year t in the time span between the year of publication pub and the year of retraction $retr$. These measures capture the disparity in citation trends in different ways. AD allows for positive and negative yearly differences to balance over time while any discrepancy is accumulated over time in the case of ED.

- *Early mentions absolute distance* (MD) of blog b and newspaper n mentions whitin two weeks from publication.

$$MD_b = \left|(b_{i,2w} - b_{j,2w})\right| \quad \& \quad MD_n = \left|(n_{i,2w} - n_{j,2w})\right|$$

[27] This is important to study the career impact for retracted authors', hence I need at least some authors with a minimum reputation ex-ante ($N_r = 1,008$).

[28] Articles or reviews.

[29] $N_c = 586.281$ overall results.

The reason for choosing a cutoff close to the day of publication draws from observing that notable studies attract most online publicity around the publication date as suggested by Figure 2 for treated papers and more evidently by Figure A.4 for control papers. Matching media mentions becomes more and more challenging the further away from publication as flawed articles may later prompt additional critical mentions.[30] To capture whether a paper is newsworthy without including mentions related to its misfortune, I focus on a two weeks cutoff from publication date. This threshold is also less sensitive to imprecisions in the publication date compared to a shorter cutoff.

I then retain for each i all j with $|AD| \leq 10$; $MD_b \leq 10$; and $MD_n \leq 10$. These cut offs allow to maximise the number of matches while limiting the maximum conceded distance in either citations or media mentions. These thresholds lie at the extremes of the distribution of distances and improve the quality of matches without affecting results. I rank the remaining j in terms of smallest $MD_b + MD_n$ and select two controls (or one depending on availability) with the minimum ED among those. This final selection leads to a sample of $N_c = 1,969$ control articles.

The quality of selected controls is assessed in Figure A.1 and A.2 of the appendix. The Euclidean distance between the selected controls and the treated paper is dense around zero (in over 68% of the cases this selection yields a perfect match), and the arithmetic distance is fairly centred around zero. No significant difference emerges when comparing treated and control distributions of cumulative citations pre-retraction. Similarly, there is no significant difference in the distribution of early mentions across treated and control groups for either newspaper articles or blog posts.[31] Control papers are marginally more likely to have little citations and no media mentions pre-retraction. In general, the vast majority of published articles have little citations and no mentions in either media outlets at publication.

3.5 Selected summary statistics

Table 1 illustrates a set of distinct summary statistics for treatment and control group. The top of the table looks at variables which should be similar across the two groups for the identification strategy to be successful. ED and AD are on average somewhat close to zero (0.93 and 0.17 respectively) and both groups of papers attracted

[30]Note that for either news or blogs the bulk of mentions appears in week one, grows at a progressively smaller rate in week two and three, and flattens out afterwards.

[31]This remains true when removing observation with no mentions, as shown in Figure A.3.

an average of about 7 citations in the pre-retraction period, substantially confirming the finding reported in Figures A.1 and A.2. Within two weeks from publication papers experience comparable online mentions on newspapers and blogs, even though eventually retracted papers have on average moderately higher coverage (1.04 vs. 0.79 news articles, and 0.24 vs 0.15 blog posts). The age for the two groups of papers is almost identical by construction. Moving to the bottom of the table one can observe that papers take on average two years to be retracted. Furthermore, yearly citations have a distribution that is very skewed, with 32.2% observations actually equal to 0, a Poisson model would therefore better approximate the distribution of the dependent variable. Unsurprisingly, treated papers cumulate substantially less citations over the years as compared to controls (16.8 vs. 33.9 respectively), but attract generally higher online attention with an *Altscore* of 37.5 for retracted papers and 19.1 for controls. In general, a non negligible share of articles experiences some online coverage, most articles are mentioned on social media (60% of retracted papers and 44% of controls) while only a limited fraction appears in newspaper articles (13% and 12% respectively), in addition blogs actively mention over one third of retracted papers while significantly less attention is devoted to controls. Finally, about one tenth of papers in either group appears in either newspapers or blogs around the publication date.

3.6 Estimating Specification

The study employs a difference-in-differences strategy that allows me to compare the evolution of citations of retracted papers before and after retraction relative to citations of a control group of non-retracted studies published in the same journal and year and with a comparable trend in yearly citations before retraction. Treatment and control papers also have similar number of online mentions (on blogs and newspapers) within two weeks from the day of publication (i.e. *early mentions*) to account for unobservable characteristics which make a study newsworthy and could therefore create a problem of selection into retraction.

Therefore, the regression model is the following:

$$\begin{aligned} E[Y_{igt}|X_{igt}] = exp[\alpha + \gamma_1 Post_{igt} + \beta_1 R_i * Post_{igt} + \beta_2 Post_{igt} * Media_i + \\ + \beta_3 R_i * Post_{igt} * Media_i + \delta_i + f(age_{it}) + \delta_\tau] \end{aligned} \quad (1)$$

where i is the treatment (or control) paper, g is the case-level group and includes the retracted paper and its respective controls, t are years relative to the retraction. The dependent variable Y represents a paper yearly citation count and excludes self-citations, as the estimation wants to capture the reaction of the scientific community other than that of the authors involved. *Post* is an indicator variable equal to one for all years after retraction, R is an indicator for retracted articles, and *Media* captures the exposure of an article to online coverage. Different media dummies will be used to indicate articles with or without media mentions. Due to previously discussed issues related to the timing and the content of coverage, the media indicator which is best identified is equal to one if a paper receives at least one online mention within two weeks from publication in either newspapers or blogs and zero otherwise (i.e. $\mathbb{1}[Early\ Mentions > 0]$). Early mentions are assumed to broadly advertise the original research findings and are generally balanced across treatment and control papers. Other media indicators equal one for research papers that receive at least one overall mention in any of the media outlets analysed (i.e. social media, newspaper articles or blog posts). In order to look at different levels of media exposure of each paper, indicators are also derived from the distribution of *Altscore*, an aggregate measure of weighted online mentions.[32] The coefficient β_1 captures the effect of a retraction shock on citations of retracted papers as compared to similar control papers. The coefficient β_3 captures any difference in the effect of the shock for papers that received online attention. Fixed effects are included for each paper δ_i and each calendar year δ_τ while $f(age_{it})$ represents a full set of dummies for years since publication (age) and is meant to flexibly control for the age of the articles.[33] To look at the dynamics of the differential effect of *Media*, estimates will be presented for a model that replaces the indicator *Post* with a full set of dummies for each year relative to the year of retraction.[34] Given the skewed nature of the dependent variable, I follow a long-standing tradition in bibliometric studies, hence I use a pseudo Poisson regression model developed by Correia, Guimarães and Zylkin (2020)[35] where consistency is achieved under the only assumption that the conditional mean of the dependent variable is correctly specified (Gourieroux, Monfort and Trognon, 1984). Finally, standard errors are clustered at the case g level.

[32] See Table A.1 for details about *Altscore* weights across outlets.

[33] Note that the interaction term $R_i * Media_i$ is absorbed by the paper fixed effect.

[34] $E[Y_{igt}|Xigt] = exp[\sum_{t=r-4}^{r-2} \gamma_{1t} * d_t + \sum_{t=r}^{r+6} \gamma_{1t} * d_t + \sum_{t=r-4}^{r-2} \beta_{1t} * d_t * R_i + \sum_{t=r}^{r+6} \beta_{1t} * d_t * R_i + \sum_{t=r-4}^{r-2} \beta_{2t} * d_t * Media_i + \sum_{t=r}^{r+6} \beta_{2t} * d_t * Media_i + \sum_{t=r-4}^{r-2} \beta_{3t} * d_t * R_i * Media_i + \sum_{t=r}^{r+6} \beta_{3t} * d_t * R_i * Media_i + \delta_i + f(age_{it}) + \delta_\tau]$

[35] http://scorreia.com/software/ppmlhdfe/

4 Descriptive results: coverage of retractions

Popular online media like newspapers, blogs and social media, whose target audience is often beyond the scientific community, have been recently active in reporting retracted articles (see Figure 1).[36] In general, media platforms seem to cover both original publications and retractions, but the two events feature to a different extent across outlets, giving raise to potential disinformation. Indeed, Figure 2 shows that mentions in newspaper articles appear predominantly close to the publication date of a study and generally inform the public about its discovery. Rarely is this information updated with a new mention at the time of retraction. On the other hand, blog posts occur mostly around the retraction event. These blogs are often specialized and directly target academics[37] while the wider audience remains exposed to information which is not always complete. This could lead to unintended consequences that deserve further work.

To shed some light into factors that could shape an outlet decision to acover a retraction event or not, I look at US news coverage, classified based on Gentzkow and Shapiro (2010) measure of media slant.[38] Figures A.6 to A.9 contrast the observed mentions for relatively left- or right- leaning outlets. Limited differences seem to emerge as left-leaning news show a somewhat more balanced reporting which deserves to be further studied.

In essence, the rise of the internet and the appereance of new platforms has the potential to direct scientists' (and non-scientists') attention towards "interesting" contributions which in some cases prove to be less reliable (Serra-Garcia and Gneezy, 2021). It is therefore important to investigate whether positive remaining citations, and the retraction process more in general, relate to the media visibility of a research paper and its retraction.

[36] Notice that recent years are likely underreported given retractions take some time to arise and hence feature in the database.

[37] Around a third of blog coverage is from RetractionWatch, the single outlet most committed to inform about scientific retractions (Figure A.5 exclude RetractionWatch mentions).

[38] To maximise the number of observable US press mentions, I take the sample of research articles published and retracted after 2010 with non-missing DOI and whose DOI is different from that of its retraction notice ($N_r = 4,763$) and retain only mentions matching the list of outlets classified by Gentzkow and Shapiro (2010). I remain with 53 retracted papers for which I observe at least one news mention with measurable slant.

5 Paper-level results

Table A.3 shows results for a simple difference-in-differences analysis for the pooled sample of retracted papers and selected controls. Estimates imply that relative to controls, retracted papers experience a 65% (i.e. $1 - exp(-1.06) = 0.65$) loss in yearly citations after the shock and the magnitude is comparable to previous studies (Furman et al., 2012; and Azoulay et al., 2015) which rely on different samples, disciplines and time periods.[39] Figure A.10 illustrates the dynamic of the effect of a retraction. The post-retraction loss in citations increases over time and there is no evidence of pre-trends.[40]

5.1 Main results

Table 5 to A.5 report results from the main specification. The tables differ by measures of media coverage, using indicators for papers with at least one mention within two weeks from publication (early mentions), papers with at least one mention overall in a certain online outlet (any news, blog or social media) or papers that fall in some part of the Altmetric score (Altscore) distribution. Tables highlight the difference-in-differences coefficient $Post * Treatment$, according to which the average citation penalty of a paper after its retraction amount to 55-62% across all specifications. The relative effect for papers that experienced some media coverage is estimated by the coefficient of the triple interaction $Post * Treatment * Media$.[41] Retracted papers with media coverage experience a penalty in post-citations of about 75% (i.e. $1 - exp(-0.96 - 0.45) = 0.76$). Across specifications in Table 5 and Table A.4 the loss in forward citations for retracted papers with media varies between 68-76%, corresponding to a difference of 12.3-15.8 p.p. (or 19.7-28.7%) with respect to retracted papers without media exposure. Furthermore, the effect seems monotonically increasing in the amount of coverage received (see Table A.5). The almost entirety of these estimates is highly significant. Figure 3 represents the dynamics of the additional penalty in presence of (alternative measures of) media coverage. The loss in yearly citations becomes progressively more evident over time without any sign of

[39] Estimates are similar when using an IHS (Inverse hyperbolic sine) transformation of the dependent variable.

[40] Note that effects in the year of retraction are also small due to the fact that papers in the sample get retracted at different points within the year.

[41] Where the *Media* variable is defined in alternative ways across specifications as described at the top of this paragraph.

recovery, and I find no evidence of pre-trends.

5.2 Robustness checks

5.2.1 Highly cited papers differencial

In what follows I intend to increase confidence that the exacerbating effect I showed is solely due to the presence of online attention. It could be that media exposure is actually capturing some alternative paper features, related to media presence, and confounding my estimates. To address this concern I repeat the main excersise looking at whether the effect of retraction on forward citations differs for papers which are highly cited *ex-ante*. The rational being that influencial papers may face higher scrutiny as well as higher chances of featuring in the media (Yin et al., 2022). Reassuringly, Tables A.6 to A.8 show that papers with high cumulative citations before the year of retraction do not drive nor confound media estimates.

5.2.2 Including retraction year into *Post* indicator

Previous estimates illustrate effects on citations for all years strictly after the one of retraction (i.e. excluding the year of retraction). The rationale behind this choice is the fact that papers can get retracted at any point during the year and this can therefore act as a confounder.[42] Nonetheless, Tables A.9 to A.11 show that the main results are not sensitive to this decision. If anything, the additional effect of early mentions is smaller in case of blog mentions (see Table A.9 column (2)). This difference may relate to the fact that most blog mentions appear later when the paper gets discredited. In addition, the fact that effect of early mentions are less significant, may speak to a possible information effect of media which emerges more clearly at a later stage, as captured by overall measures of online coverge in A.10 and A.11.

5.2.3 Actively cited papers

The algorithm for selecting controls attempts to choose papers that could likely mimic the citation path of retracted papers absent the retraction shock. Finding good controls for retracted papers that are not actively cited soon after publication may be

[42] Figure A.10 and Figure 3 show smaller or insignifican effects in the year of retraction relative to the previous year.

challenging and could bias estimates. For this reason I here exclude all retracted papers (and respective controls) with zero citations in any year before retraction. This exercise halves the original sample.[43] Even so, Tables A.12 to A.14 confirm the results all remain robust.

5.2.4 Excluding late published papers

One concern is that for more recently published papers there may be an insufficient time frame to display changes in citation patterns. To this respect I repeat the exercise retaining only older publications. Specifically, I retain only retracted papers (and associated controls) that were published between 2011 and 2017. Tables A.15 to A.17 confirm the results remain virtually unchanged.[44]

5.2.5 Using random control papers

Longitudinal study designs using matched samples are known to be vulnerable to the problem of regression to the mean. Therefore, I created a sample associating two random non-retracted papers published in the same journal and year with each retracted paper. I then follow (Daw and Hatfield, 2018) to understand whether matching on pre-period citations and early media mentions could introduce a bias.

In details, I compare retracted and random control papers to verify that there is no significant group difference in pre-period citations (p-value 0.5943), and no significant group difference in the share of papers with any early media mention (p-value 0.1297). Some differences emerge in the fraction of papers with early blog mentions as well as in their respective level of blog mentions. In addition, early mentions are positively correlated with forward citations (i.e. $\rho \in (0.12, 0.21)$). However, early mentions are, by definition, time-invariant while yearly-citations are highly serially correlated. Taken together, these observations suggest both variables could be used as matching features.

To exclude further doubts, I estimate the main result using this *unmatched* sample (retracted papers and two random controls from the same journal and year) and confirm previous evidence with a few differences. Specifically, Figure A.11 shows the presence of some pre-trends in citations across treatment and control groups, while the main effect is less precisely estimated. Hence, the primary strategy does not seem

[43] 48% observations left in either treated or control group.

[44] Unreported tables show that results hold by removing late published paper, one year at the time.

to introduce a bias while possibly increasing precision (see also Table A.18).

5.2.6 Citation textual content

One additional exercise is that of looking directly at the textual content of citations. *Scite.ai* (a newly launched platform featuring in Nature)[45] scans article PDFs for references to papers and categorises these citations as "mentioning", "contrasting" or "supporting" the research findings.[46] With the platform's support, I built a dataset of yearly citation statements for each classification, paper and year and performed an exercise equivalent to that of Section 5.1. Tables A.30 to A.32 substantially corroborate the main findings. Retracted papers experience a penalty in all type of citation statements after the retraction shock, and for citation statements that only mention the study, this penalty is aggravated in presence of media coverage. No additional change is detected for either contrasting or supporting references. One caveat is that almost the entirety of the citation statements is classified as simply mentioning papers' findings.

5.2.7 Journal visibility and loss in citation

In a final robustness exercise I relate the individual loss in citations obtained comparing each retracted paper to its selected controls for different pre- and post- time windows ($DiD = [E(cit_1^T) - E(cit_1^C)] - [E(cit_0^T) - E(cit_0^C)]$) to the average media coverage of (non-retracted) articles published in the same journal and year (*journal visibility*).[47] Figure A.23 shows that a retracted paper experiences a significantly larger loss in citations if published in a journal with higher average visibility. Notice this negative relationship becomes stronger when looking at wider time windows around the year of retraction. These same conclusions are evident in Table A.29 where alternative measures of media exposure are also used. These findings substantially confirm the main results presented in section 5.1.

[45] Nature article on Scite.ai.

[46] The classification is according to Rosati (2021)

[47] Controlling for year of publication effects, age of paper at retraction effects, number of non-retracted articles within same journal and year, the average Euclidean distance of those non-retracted articles, and the level of (non-self) cumulative citations of the retracted paper before retraction. See Section 5.6 for further details.

5.3 Heterogeneity

5.3.1 Hard vs. Social sciences

Individual disciplines have distinctive publication practices which could create different incentives at publication and therefore lead to heterogeneous effects. Table 6 and A.19 together with Figure 4 illustrate that this may indeed be the case. What consistently emerges across specifications is that, in the case of social sciences,[48] there is no additional penalty associated to retracted papers with media attention. Perhaps one interpretation is that the timing of publications in hard sciences is generally fast; while working papers in social sciences may circulate for longer inside the scientific community. In the latter case, media coverage may therefore offer little room for update on the validity of the study as compared to the former case where online attention may further stimulate the academic discussion around a paper.[49] One caveat is that the subsamples of disciplines are quite small, in particular in the case of social sciences.[50]

5.3.2 Cause of retraction

One aspect I investigate is whether the aggravating effect of online attention differs depending on the reason behind the retraction. Retractions can occur because of honest mistakes or actual misconduct of the authors. Distinguishing the two is relevant as original findings should be entirely discarded in cases of severe misconduct, leading to higher concerns over the spread of disinformation. On the other hand, cases of misconduct may be newsworthy and online discussion may play a special role by circulating detailed information on the case. I therefore divide reasons for retraction in minor, moderate and severe cases of misconduct using the classification developped by Woo and Walsh (2021) (see Table A.21).[51] Table 7 (and A.20) suggests media attention plays a big role in the presence of severe cases of misconduct and

[48]Disciplines are identified using Scopus journal classification. Social sciences are: business and technology, humanities and other; while hard sciences are: life sciences, environment, health and physical sciences.

[49]Related to this, Wohlrabe and Bürgi (2021) suggests that in the case of economics, the practice of releasing working papers before their publication in a journal has a positive impact on citations.

[50]Over 80% of retraction appears in hard sciences (809) of which 12% (95) with early visibility and 59% (475) with Altscore above median. Of the 179 retraction in social sciences 8% (15) have early visibility and 53% (95) have Altscore above median.

[51]The selected sample is divided in 30% (301) cases of minor misconduct, 26% (261) of moderate misconduct and 43% (428) of severe misconduct.

there is little additional penalty associated to minor cases with media attention.

5.4 Information mechanism

This works has so far shown that citations of retracted papers in the literature decline at a faster pace in presence of media coverage. This additional effect of media may be derived by two different mechanisms: (a) higher scrutiny by the scientific community to a paper that gained publicity; (b) additional information provided to some part of the scientific community which would have otherwise remained unaware of the retraction. Although difficult to distinguish, one way to corroborate the information mechanism is to check whether the content of remaining ex-post citations is more "accurate" in presence of media coverage. With the help of *Scite.ai*, I collected for each retracted paper all yearly citation statements that mentioned the retraction. Citation statements were searched for the terms "*etract*" or "*ithdraw*", manually excluding false positives. I then estimate the following regression model:

$$E[Y_{it}|X_{it}] = exp[\alpha + \beta_1 Post_{it} * Media_i + \delta_i + \delta_t + f(age_{it}) + \delta_\tau] \qquad (2)$$

where for each retracted paper i and year relative to retraction t, Y represents the number of citation statements mentioning the paper is retracted, *Post* is an indicator for year strictly after retraction, *Media* is an indicator for whether a paper gained some kind of online coverage. Estimates of β_1 capture the differential change in number of citations "correctly" mentioning the retraction (after the shock) in presence of media coverage. Fixed effects are included for each paper δ_i, each year relative to retraction δ_t, each year since publication $f(age_{it})$ and each calendar year δ_τ. Standard errors are clustered at the retraction level.

Table 9 confirms that the number of references correctly mentioning the cited paper is retracted increases significantly in presence of media coverage. This result support the hypothesis that media coverage provides additional *information* on retractions, hence favouring the belief update of part of the scientific community which would have otherwise remained unaware. One caveat to consider is the small sample of retractions for which an "accurate" yearly-citation is indeed observed (slightly less than 10% of the treated sample).[52]

[52] This is consistent with previous work by Schneider et al. (2020) which finds that, for the case considered, the retraction is not mentioned in 96% of direct post-retraction citations.

5.5 Media and likelihood of retraction

In section 3.2 I argued that a challenge one faces when trying to understand the interaction between the retraction process and media coverage arises from the endogeneity of the latter. To circumvent this issue to some extent and study the relationship between media coverage and the likelihood of retraction, I turn to the text analysis of titles of research articles. This in turn allows me to use the presence of specific words to control for papers' endogenous coverage.

More specifically, I start with the full sample of eventually-retracted articles published (and retracted) after 2010 and for each of these articles I add to the sample twenty randomly selected articles that appear in the same journal and year but were never retracted.[53] I then use the titles of these papers as my corpus of analysis.[54] After cleaning the text according to the Porter (1980) algorithm, Figure 5 shows the most frequent words present in the titles of papers that experience some (Panel A) or no (Panel B) online coverage (in newspapers or blogs) within two weeks from publication. On the one hand, popular papers mention more often words shuch as "cancer", "patient" and "disease". On the other, articles that did not feature in the media often quote different words such as "model" or "system". In what follows I try using this difference to predict an article's coverage.

After building the document-term matrix of words (unigrams and bigrams) that appear in at least 100 titles I randomly split the observations into 90% training and 10% testing subsample. The training sample is used to select words with some predictive power for papers' media coverage based on the lasso selection procedure. The testing sample is then used to compute the out-of-sample performance of the predicted media coverage based on the selection.[55]

The lasso estimates and the set of selected variables (words) depends on the penalty level λ. I obtained alternative lists of selected words using different pro-

[53]This selection facilitate a speedy computation without restricting the corpus of titles. Among the 1008 retracted papers in the sample, 44 have less than 20 associated random controls due to the respective scarsity of potential controls found in Scopus.

[54]$N = 20755$

[55]The lasso estimation minimizes the mean squared error subject to a penalty on the absolute size of coefficient estimates and where λ controls the overall penalty level.

$$\hat{\beta}_{lasso} = argmin \frac{1}{n} \sum_{i=1}^{n} (Media_i - \sum_{j=1}^{p} \beta_j Word_{ij})^2 + \frac{\lambda}{n} \sum_{j=1}^{p} \psi_j \mid \beta_j \mid$$

Due to the nature of the penalty, the lasso sets some coefficients exactly to zero and in doing so removers some predictors from the model.

cedures that choose the optimal penalty level using: (a) EBIC information criteria; (b) AICC information criteria; (c) K-fold cross-validation and (d) Rigorous (theory-driven) penalty levels. These procedures are then repeated including a full set of subject fixed effects, publication year fixed effects and excluding retracted articles from the sample. This strategy allows to estimate the following model:

$$Retraction_{ijp} = \beta_1 Media_{ijp} + \beta_2 \widehat{Media}_{ijp} + \delta_j + \delta_p + \epsilon_{ijp} \quad (3)$$

where for each article i published in year p and journal j, *Retraction* is an indicator for whether the article was retracted, *Media* is a dummy taking value one if the article gained any online coverage (in either newspapers or blogs) within the first two weeks from publication, while δ_j and δ_p absorb journal fixed effects and publication year fixed effects respectively. Estimating the *Media* impact on the likelihood of a retraction (β_1) is challenging as it is difficult to exclude that researchers may choose to investigate salient topics that, given their relevance, are scrutinized differently from the scientific community (see for example Serra-Garcia and Gneezy, 2021) leading to different retractions rates, despite the fact that these topics may be of interest to the general public and hence attract media coverage. The inclusion of $\widehat{Media} = \sum_s \hat{\beta}_{s,lasso} SelectedWord_s$ as predicted from the lasso procedure, where *SelectedWord* represents the number of times a selected n-gram appears in the title of a paper i, allows to control for endogenous topic selection that could otherwise lead to bias. Given that $\widehat{Media}$ is derived from separate estimates, standard errors are bootstrapped and clustered around retraction cases.[56]

Table A.23 shows the correlation between some of the most powerful lasso selected predictors and the *Media* indicator variable. The n-grams with the largest coefficients provide insights into which articles receive media coverage. For example, the word "climate" appears. Similarly, the n-grams "brain", "graphen", "genom" and "stem" all represent research topics of large interest. Also, some research methodologies seem popular as suggested from the n-gram "meta analysis" and "trial". Accuracy ranges between 60 and 76% across procedures and more parsimonious lasso (and logit lasso) seem to provide better-performing selections. The fraction of correctly classified observations reaches up to 86% when a full set of subject and year fixed effects are included and when retracted papers are excluded.[57] Accuracy is calculated

[56]Summary statistics of main variables and a selection of n-grams are displayed in Table A.22.

[57]The most powerful predictors selected with these alternative strategies remain fairly similar (not

after estimating the optimal positive cutoff threshold using the Matthews Correlation Coefficient (MCC). In the area of machine learning with binary classification the MCC is the preferred single metric, especially for imbalanced data (Chicco and Jurman, 2020). The metric ranges $[-1, 1]$ and takes on the value of zero if the prediction is the same as a random guess. Table A.23 shows MCC ranging between 0.37 and 0.45 across different selection procedure.

Equation (3) estimates are reported in Table 8 (Panel A) where despite the differences in n-gram selection and predictive accuracy across models, very similar results emerge across specifications. Evidence suggests that articles with higher *predicted* media coverage are less likely to experience a retraction. The interpretation of this result is twofold. On the one hand, the fact that popular articles are retracted less often seems reassuring and could be due to experienced academics answering salient research questions.[58] On the other, it could indicate that "interesting" research articles may be reviewed with a laxer standard (as suggested in Serra-Garcia and Gneezy, 2021). Under the assumption that predicted media coverage effectively controls for endogenous topic selection, the remaining variation in media coverage is arguably exogenous and therefore allows to estimate the impact of additional attention on the likelihood of being retracted. Estimates show that wider media coverage at publication leads to higher chances of retraction, but the magnitude of this effect remains small. Note that the media variables (observed or predicted) capture very limited variation in the outcome variable. Equivalent results are displayed in Table A.24 for logit estimations, in Table A.25 for direct estimates of residual coverage, in Table A.26 for lasso procedures trained with TF-IDF word scores, and in Table A.27 for lasso procedures trained within subjects and years and excluding retracted articles.[59]

These findings justify selecting controls with early media presence similar to that of their retracted counterparts as allowing the selection into treatment of more popular articles, however small, could otherwise bias the main results reported in section 5.1. Finally, one could be concerned about the common inclusion of both the media indicator and its text-based prediction due to their positive correlation ($\rho \approx 0.3$). To this respect, Table 8 additionally reports the impact of the two regressors separately (see Panel A column (1-2) and Panel B respectively), the magnitudes of coefficients varies only slightly in this case, ressuring us against a collinearity issue.

shown and available upon request).

[58]Notice that predicted coverage is endogenous.

[59]Non reported estimates reveal equivalent results when selecting 50 or 100 random controls per retracted paper.

5.6 Journal visibility and retraction timing

In the following section I offer one way to circumvent media endogeneity and study the relationship between coverage and timing of retractions. In what follows I argue that non-retracted articles, published in the same journal and year as a retracted one, attract online coverage which is arguably *exogenous* to the retracted article's own coverage. Based on this, a good proxy for online *visibility* of a specific journal and year is the average coverage of all non-retracted papers published in there.[60]

$$JVisibility_{jp} = \frac{1}{n}\sum_{k\neq i} Altscore_{kjp} \tag{4}$$

where k are non-retracted papers published in same journal j year p as the retracted paper i. Alternatively I use the average share of $k \neq i$ published in j and p with some media mentions. Hence, I can study the following relationship using an OLS regression in a cross-sectional context:

$$Y_{ijp} = \beta JVisibility_{jp} + \delta_p + \nu X_{ijp} + \epsilon_{ijp} \tag{5}$$

where for each retracted paper i published in j in year p, Y represents either one of the dependent variables: $Time\ to\ retract = (Retraction\ date - Publication\ date) \times \frac{12}{365}$ or $DiD = [E(cit_1^T) - E(cit_1^C)] - [E(cit_0^T) - E(cit_0^C)]$ the individual loss in citations obtained comparing each retracted paper to its selected controls for different pre- and post- time windows (see Section 5.2.7 for results on loss in citations). In addition, δ_p indicates publication year fixed effects, while X_{ijp} controls for $N_{jpk\neq i}$ number of non-retracted papers in same journal-year as retrieved from Scopus, $\frac{1}{n}\sum_{k\neq i} ED_{kjp}$ their average Euclidean distance in citation from the retracted one, and $\sum_{p\leq t<r} cit_{ijpt}$ cumulative citation of i before retraction year r. Standard errors are clustered at the journal level.

Figure 6 shows that papers are retracted faster when published in journals where the average article attracts higher online coverage. Table A.28 (column (1)) illustrates that one standard deviation increase in journal visibility (measured as the average Altscore of non-retracted articles in a journal-year) reduces time to retraction by approximately 15% of its average. Looking across the remaining columns, the relationship is robust to different measures of visibility. The negative association between

[60] The measure is based on the entire pool of papers published in same year and journal as the retracted ones (and excluding the retracted ones).

journal visibility and retraction timing could be driven by higher scrutiny from the scientific community for salient research, together with better detection technologies for visible outlets.

6 Author-level analysis

Previous sections have shown how media coverage impacts the scientific recognition of retracted papers (i.e. citations) as well as the chances and pace of discovery of faulty research. I now turn to investigate the potential impact of online attention on the subsequent research output of the authors of retracted papers. Low authors output after retraction may reflect a combination of factors: (a) erosion of trust in the authors' work by the scientific community, (b) loss of individual resources for research, or (c) any other direct consequence in terms of academic employment. Studying retracted authors' output is important to measure the individual overall cost of "bad science" and allows me to discern the potential role of media on authors' future careers.

6.1 Selected sample

Using the same selection of papers illustrated in Section 3.3 and 3.4, I search all authors of retracted and control papers available in the Scopus library, the retrieved list contains $N_a = 17,991$ distinct authors with any prior publication.[61] For each of these authors, I retrieve their (non-retracted) publications and their corresponding yearly citations and I compute individual output measures focusing on a window of 5 years around the retraction.

Before proceding to the analysis, I select my sample as follows. First, I observe that some retracted authors appear multiple times, therefore I retain only observations relative to their first (in-sample) retraction.[62] Second, I keep cases for which I observe at least one retracted and one control author.[63] Finally, when available, I retain a maximum of three authors per paper (first, mid and last)[64] which provide a final sample of $N_a = 6,718$ authors of which $N_{a,r} = 2,047$ retracted authors, $N_{a,c} = 4,671$

[61] $N_{a,r} = 4,105$ retracted authors and $N_{a,c} = 13,886$ control authors, corresponding to the original sample of retractions $N_r = 900$.
[62] Reducing the underlining sample of retractions, $N_r = 987$.
[63] $N_a = 17,148$, of which $N_{a,r} = 4,077$ and $N_{a,c} = 13,071$ corresponding to $N_r = 874$.
[64] Note these three categories are mutually exclusive and sigle authors are confidered as first authors.

control authors, corresponding to $N_r = 874$ undelining retractions.

In this analysis I intend to study hetherogeneous effects by ranking of appereace in the authorship list, by seniority (based on H-index prior retraction), and by severity of misconduct. Table 2 and Table 3 provide all relative sub-sample sizes.

6.2 Summary statistics

To measure authors' research output I look at: (a) the number of papers published per author per year, (b) the number of papers published per author per year mentioning any source of funding, and (c) the average number of authors across all papers published within the same year. The first is a measure of output *productivity*, the second is an imperfect measure of access to funding, while the last is an indicator for individual collaboration practices. Despite being impefect, these measures allow me to assess, in a within-author analysis, for the presence of career effects of retractions due to media exposure.

Table 4 illustrates the average outputs for this selected sample of academics. In general, authors publish 5.5 papers per year, of which 2.8 have some declared funding support, and with 5.8 authors per paper. First authors seem relatively more junior while last authors are generally more senior (with 3.3 vs. 8.3 publications per year). In medical research, last authors are usually senior researchers with stable careers, whereas first and middle authors can be transient authors who may not pursue a scientific career. Large differences also emerge when comparing authors with high and low (ex-ante) H-index. The former indeed publish more papers (8.3 vs 2 per year) and with a larger set of authors (7 vs. 4 per paper). Output measures of authors associated to different causes of misconduct are generally balanced. Finally, Table A.2 compares average outputs across all authors whose original publication gained initial online attention (or not). Across all categories, authors of newsworthy research have higher publication rates, higher funding support and a larger set of coauthors.

6.3 Estimating Specification

This section investigates whether media coverage influences authors' careers after the reputational shock of a retraction. The worsening of authors' output may reflect a combination of destruction in access to resources for research as well as an erosion of trust in authors' work by the scientific community. To study this I employ a difference-in-differences strategy that compares output measures of retracted authors to that of control authors of similar never-retracted studies,[65] before and after their first observed retraction [66]. Crucially, I further contrast authors whose original publication obtained any media exposure at publication (i.e. *early mentions > 0*) to those who did not.

Therefore, the regression model is the following:

$$E[Y_{aigt}|X_{aigt}] = exp[\alpha + \gamma_1 Post_{aigt} + \beta_1 R_{aigt} * Post_{aigt} + \beta_2 Post_{aigt} * Media_{ai} + \\ + \beta_3 R_{ai} * Post_{aigt} * Media_{ai} + \delta_a + f(CareerLenght_{at}) + \delta_\tau] \tag{6}$$

where a are authors of treatment (or control) paper i of case-level group g[67] in the years t relative to the retraction. The dependent variable Y is either one of the measures of author outputs: (a) the number of publications in each year, (b) the number of yearly publications with grant support, and (c) the average number of authors across all publications of each year. R is an indicator for retracted authors, *Post* is an indicator variable equal to one for all years after retraction, and *Media* captures the exposure of the original (retracted or control) paper to online coverage at the time of publication. Specifically, the indicator is equal to one if the paper received at least one online mention within two weeks from publication in either newspapers or blogs and zero otherwise (i.e. $\mathbb{1}[Early\ Mentions > 0]$). Notice that early mentions are assumed to broadly advertise the original research findings and are generally balanced across treatment and control papers. The coefficient β_1 captures the effect of a retraction shock on retracted authors careers as compared to control authors. The coefficient β_3 captures any difference in the effect of the shock for authors whose papers received online attention. Fixed effects are included for each author δ_a and each calendar year δ_τ while $f(CareerLenght_{at})$ represents a full set of dummies for years since the author's first publication (ever observed in the Scopus library) and is meant to flexibly control for the academic experience of authors.[68] To look at the dynamics of the differential effect of *Media*, estimates will be presented splitting the sub-samples of authors exposed or not to media and replacing the *Post* indicator with a full set of dummies for each year relative to the year of retraction. [69] Given the skewed nature

[65] See Section 3.4 for details on the selection of control papers.
[66] See details on sample selection in Section 6.1
[67] Notice that a group is composed by all authors of a retracted paper and its paired control papers.
[68] Note that the interaction term $R * Media$ is absorbed by author fixed effect.
[69] $E[Y_{aigt\tau}|X_{aigt}] = exp[\sum_{t=r-5}^{r-2} \gamma_{1t} * d_t + \sum_{t=r}^{r+5} \gamma_{1t} * d_t + \sum_{t=r-5}^{r-2} \beta_{1t} * d_t * R_{ai} + \sum_{t=r}^{r+5} \beta_{1t} * d_t * R_{ai} +$

of the dependent variables (e.g. almost 20% of yearly observations see 0 published papers), I use a pseudo Poisson regression model developed by Correia, Guimarães and Zylkin (2020)[70] where consistency is achieved under the only assumption that the conditional mean of the dependent variable is correctly specified (Gourieroux, Monfort and Trognon, 1984). Finally, standard errors are clustered at the case g level.

6.4 Results

Author-level estimates are presented in Table 10. On average, authors' future outputs are negatively impacted by a retraction (about 8.6% loss in forward yearly publications and number of collaborators, corresponding to about half publication per year and half author per paper per year.) and more so for senior authors and severe cases of misconduct. Media exposure adds a further loss in output which is never significant.

Looking more closely into sub-groups, estimates suggests that authors appearing first in the co-authorship list are the ones whose *productivity* is differentially and significantly impacted by media ($\simeq$45% loss in yearly publications and grant supported publications, corresponding to 1.5 less papers published per year, half of which with grant support.). Senior authors and severe cases of misconduct also display a further loss with media which is never significant. This could explain previous findings by Mongeon and Larivière (2016) that first authors are most punished after retractions.[71]

However studying the dynamics of this effect, Figure 7 (together with Figures A.15 and A.19) shows that the negative impact of retraction is large and permanent only if the original publication had some media exposure ($\simeq$10% larger reduction in future publication rate relative to a case with no media). Absent media coverage author outputs are only moderately impacted and may even fully recover by the end of the 5 year window. The differential impact corresponds to 1 less publication per year for authors with media exposure, against half publication less per year, compared to their respective averages. This differential effect of media is evident for first authors (see Figures A.12, A.16, A.20), for authors with high ex-ante H-index (see Figures A.13, A.17, A.21) and authors whose paper was retracted for severe misconduct (see

$\sum_{t=r-5}^{r-2} \beta_{2t} * d_t * Media_{ai} + \sum_{t=r}^{r+5} \beta_{2t} * d_t * Media_{ai} + \sum_{t=r-5}^{r-2} \beta_{3t} * d_t * R_{ai} * Media_{ai} + \sum_{t=r}^{r+5} \beta_{3t} * d_t * R_{ai} * Media_{ai} + \delta_i + f(age_{it}) + \delta_\tau]$ separately for the subsample of original papers with $Media = 1$ (or $Media = 0$).

[70] http://scorreia.com/software/ppmlhdfe/

[71] Excluding few single authors cases does not change the results (not shown).

A.14, A.18, A.22). These figures are all based on split regressions illustrated in Table A.33.

7 Conclusion

Flawed research can be harmful both within and outside of academia. The literature documents that scientific publications lose significant citations after a retraction. Worryingly though, studies also show that retracted publications still get cited long after they are removed from the literature, potentially disseminating misinformation. In the context of scientific retractions, their visibility is a crucial factor, yet there is little evidence on how media reporting may influence the retraction process and authors' careers. This work captures the impact that coverage may have on inducing retraction (however slight), correcting the scientific perception around the research paper, and penalizing later production of authors; however, the information to the general public is not rectified.

I use a conditional difference-in-differences strategy to show that articles that gained popularity in the media — at the time of publication — face heavy citation losses after their retraction while remaining citations become more *accurate* in acknowledging the retraction. This differential effect is considerable for cases of severe misconduct, and it is present only for publications in hard sciences, suggesting distinct publication practices or different topic salience may impact the visibility of a retraction. In addition, retracted authors' future research output is permanently reduced, but only with media coverage (specifically for first authors). I also produce evidence that media can influence the likelihood of retraction and its timing.

Overall, the media seems to help the auto-correcting process of science reducing misinformation within academia. At the same time, this implies that plenty of wrong science goes often unnoticed. The scientific community, thus, needs better strategies to increase the level of scrutiny and lower incentives for *bad* science. For example, journals could increase transparency at submission and systematically check references of newly accepted papers before publication. This research also proves that media platforms can be a useful communication tool, as in the case of *Retraction-Watch* and, more recently, the Twitter bot from *Scite.ai*.[72]

Nonetheless, the scientific information that appears in the media spreads beyond

[72]See: Sciete.ai Twitter bot.

the scientific community. Indeed, while media helps scientists to update beliefs about the credibility of a study and its authors, one question remains about whether this could generate unintended consequences for the main audience of mainstream media: the general public. I show that newspapers, as opposed to blogs, are more likely to report the publication of a paper rather than inform about its later retraction. This possible misinformation can impact public perceptions and behaviour, and therefore deserves further research.

References

Allcott, Hunt and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *J. Econ. Perspect.* 31(2):211–236.

Azoulay, Pierre, Alessandro Bonatti and Joshua L. Krieger. 2017. "The career effects of scandal: Evidence from scientific retractions." *Res. Policy* 46(9):1552–1569.

Azoulay, Pierre, Jeffrey L Furman, Krieger and Fiona Murray. 2015. "Retractions." *Rev. Econ. Stat.* 97(5):1118–1136.
URL: *https://direct.mit.edu/rest/article/97/5/1118-1136/58260*

Bar-Ilan, Judit and Gali Halevi. 2017. "Post retraction citations in context: a case study." *Scientometrics* 113(1):547–565.

Bordignon, Frederique. 2020. "Self-correction of science: a comparative study of negative citations and post-publication peer review." *Scientometrics* 124(2):1225–1239.
URL: *https://doi.org/10.1007/s11192-020-03536-z*

Brainard, Jeffrey. 2018. "Rethinking retractions." *Science (80-.).* 362(6413):390–393.
URL: *https://www.science.org/doi/10.1126/science.362.6413.390*

Buscha, Franz, Arnaud Maurel, Lionel Page and Stefan Speckesser. 2012. "The Effect of Employment while in High School on Educational Attainment: A Conditional Difference-in-Differences Approach." *Oxf. Bull. Econ. Stat.* 74(3):380–396.

Card, David and Stefano Dellavigna. 2020. "What do editors maximize? Evidence from four economics journals." *Rev. Econ. Stat.* 102(1):195–217.

Card, David, Stefano Dellavigna, Patricia Funk and Nagore Iriberri. 2020. "Are Referees and Editors in Economics Gender Neutral?" *Q. J. Econ.* 135(1):269–327.

Chicco, Davide and Giuseppe Jurman. 2020. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC Genomics* 21(1):1–13.

Correia, Sergio, Paulo Guimarães and Tom Z. Zylkin. 2020. "Fast Poisson estimation with high-dimensional fixed effects." *Stata J.* 20(1):95–115.

Cox, Adam, Russell Craig and Dennis Tourish. 2018. "Retraction statements and research malpractice in economics." *Res. Policyh* 47(5):924–935.
URL: *https://doi.org/10.1016/j.respol.2018.02.016*

Daw, Jamie R. and Laura A. Hatfield. 2018. "Matching and Regression to the Mean in Difference-in-Differences Analysis." *Health Serv. Res.* 53(6):4138–4156.

Dumas-Mallet, Estelle, André Garenne, Thomas Boraud and François Gonon. 2020. "Does newspapers coverage influence the citations count of scientific publications? An analysis of biomedical studies." *Scientometrics* 123(1):413–427.
URL: *https://doi.org/10.1007/s11192-020-03380-1*

Fanelli, Daniele. 2013. "Any publicity is better than none: newspaper coverage increases citations, in the UK more than in Italy." *Scientometrics* 95(3):1167–1177.
URL: *http://link.springer.com/10.1007/s11192-012-0925-0*

Furman, Jeffrey L., Kyle Jensen and Fiona Murray. 2012. "Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine." *Res. Policy* 41(2):276–290.

Gentzkow, Matthew; and Jesse Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78(1):35–71.
URL: *http://doi.wiley.com/10.3982/ECTA7195*

Goncalves, Duarte;, Jonathan; Libgober and Jack Willis. 2021. "Learning versus Unlearning: An Experiment on Retractions.".

Gourieroux, Author C, A Monfort and A Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52(3):681–700.

Heckman, J. J., H. Ichimura and P. E. Todd. 1997. "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Rev. Econ. Stud.* 64(4):605–654.
URL: *https://academic.oup.com/restud/article-lookup/doi/10.2307/2971733*

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5):1017.
URL: *https://www.jstor.org/stable/2999630?origin=crossref*

Hesselmann, Felicitas, Verena Graf, Marion Schmidt and Martin Reinhart. 2017. "The visibility of scientific misconduct: A review of the literature on retracted journal articles." *Curr. Sociol. Rev.* 65(6):814–845.

Ivanova, Ana, Mike S. Schäfer, Inga Schlichting and Andreas Schmidt. 2013. "Is There a Medialization of Climate Science? Results From a Survey of German Climate Scientists." *Sci. Commun.* 35(5):626–653.

Jin, Ginger Zhe, Benjamin Jones, Susan Feng Lu and Brian Uzzi. 2019. "The reverse matthew effect: Consequences of retraction in scientific teams." *Rev. Econ. Stat.* 101(3):492–506.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts and Jonathan L. Zittrain. 2018. "The science of fake news." *Science (80-.).* 359(6380):1094–1096.
URL: *https://www.science.org/doi/10.1126/science.aao2998*

Lewandowsky, Stephan, Ullrich K.H. Ecker, Colleen M. Seifert, Norbert Schwarz and John Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychol. Sci. Public Interes. Suppl.* 13(3):106–131.

Lu, Susan Feng, Ginger Zhe Jin, Brian Uzzi and Benjamin Jones. 2013. "The Retraction Penalty: Evidence from the Web of Science." *Sci. Rep.* 3:1–5.

Matsuyama, Wataru, Hideo Mitsuyama, Masaki Watanabe, Ken-ichi Oonakahara, Ikkou Higashimoto, Mitsuhiro Osame and Kimiyoshi Arimura. 2005. "RETRACTED: Effects of Omega-3 Polyunsaturated Fatty Acids on Inflammatory Markers in COPD." *Chest* 128(6):3817–3827.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0012369215496230*

Mehra, Mandeep R., Sapan S. Desai, Frank Ruschitzka and Amit N. Patel. 2020. "RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis." *Lancet* .
URL: *http://dx.doi.org/10.1016/S0140-6736(20)31180-6*

Mehra, Mandeep R., Sapan S. Desai, SreyRam Kuy, Timothy D. Henry and Amit N. Patel. 2020. "RETRACTED: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19." *N. Engl. J. Med.* 382(25):e102.

Mongeon, Philippe and Vincent Larivière. 2016. "Costly collaborations: The impact of scientific fraud on co-authors' careers." *J. Assoc. Inf. Sci. Technol.* 67(3):535–542.

URL: *http://onlinelibrary.wiley.com/doi/10.1002/asi.22883/abstract https://onlinelibrary.wiley.com/doi/10.1002/asi.23421*

Peng, Hao, Daniel M. Romero and Emoke-Ágnes Horvát. 2022. "Dynamics of cross-platform attention to retracted papers." *Proc. Natl. Acad. Sci.* 119(25).
URL: *https://pnas.org/doi/full/10.1073/pnas.2119086119*

Phillips, David P., Elliot J. Kanter, Bridget Bednarczyk and Patricia L. Tastad. 1991. "Importance of the Lay Press in the Transmission of Medical Knowledge to the Scientific Community." *N. Engl. J. Med.* 325(16):1180–1183.
URL: *http://www.nejm.org/doi/abs/10.1056/NEJM199110173251620*

Piller, Charles. 2021. "Disgraced COVID-19 studies are still routinely cited." *Science (80-.).* 371(6527):331–332.
URL: *https://www.sciencemag.org/lookup/doi/10.1126/science.371.6527.331*

Porter, M. F. 1980. "An algorithm for suffix stripping." *Program* 14(3):130–137.

Prat, Andrea and David Strömberg. 2013. The Political Economy of Mass Media. In *Adv. Econ. Econom. Tenth World Congr.*, ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel. Cambridge: Cambridge University Press pp. 135–187.
URL: *https://www.cambridge.org/core/product/identifier/CBO9781139060028A013/type/book_part*

Rosati, Domenic. 2021. "Citations are not opinions: a corpus linguistics approach to understanding how citations are made.".
URL: *http://arxiv.org/abs/2104.08087*

Sarathchandra, Dilshani and Aaron M. McCright. 2017. "The Effects of Media Coverage of Scientific Retractions on Risk Perceptions." *SAGE Open* 7(2).

Schneider, Jodi, Di Ye, Alison M. Hill and Ashley S. Whitehorn. 2020. *Continued post-retraction citation of a fraudulent clinical trial report, 11Â years after it was retracted for falsifying data*. Vol. 125.

Serghiou, Stylianos, Rebecca M. Marton and John P.A. Ioannidis. 2021. "Media and social media attention to retracted articles according to Altmetric." *PLoS One* 16(5 May 2021):1–19.
URL: *http://dx.doi.org/10.1371/journal.pone.0248625*

Serra-Garcia, Marta and Uri Gneezy. 2021. "Nonreplicable publications are cited more than replicable ones." *Sci. Adv.* 7(21):1–23.
URL: *https://www.science.org/doi/10.1126/sciadv.abd1705*

Smith, Jeffrey and Petra Todd. 2005. "Does matching overcome LaLonde's critique of

nonexperimental estimators?" *J. Econom.* 125(1-2):305–353.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S030440760400082X*

Sugawara, Yuya, Tetsuya Tanimoto, Shoko Miyagawa, Masayasu Murakami, Atsushi Tsuya, Atsushi Tanaka, Masahiro Kami and Hiroto Narimatsu. 2017. "Scientific misconduct and social media: Role of twitter in the stimulus triggered acquisition of pluripotency cells scandal." *J. Med. Internet Res.* 19(2):1–10.

Sumner, Petroc, Solveiga Vivian-Griffiths, Jacky Boivin, Andrew Williams, Lewis Bott, Rachel Adams, Christos A. Venetis, Leanne Whelan, Bethan Hughes and Christopher D. Chambers. 2016. "Exaggerations and caveats in press releases and health-related science news." *PLoS One* 11(12):1–15.

Sumner, Petroc, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy and Christopher D. Chambers. 2014. "The association between exaggeration in health related science news and academic press releases: Retrospective observational study." *BMJ* 349(December):1–8.
URL: *http://dx.doi.org/doi:10.1136/bmj.g7015*

Tahamtan, Iman, Askar Safipour Afshar and Khadijeh Ahamdzadeh. 2016. "Factors affecting number of citations: a comprehensive review of the literature." *Scientometrics* 107(3):1195–1225.

Teixeira da Silva, Jaime A. and Helmar Bornemann-Cimenti. 2017. "Why do some retracted papers continue to be cited?" *Scientometrics* 110(1):365–370.

The Economist. 2021. "Zombie research haunts academic literature long after its supposed demise.".
URL: *https://www.economist.com/graphic-detail/2021/06/26/zombie-research-haunts-academic-literature-long-after-its-supposed-demise*

Weingart, Peter. 1998. "Science and the media." *Res. Policy* 27(8):869–879.
URL: *https://doi.org/10.1016/S0048-7333(98)00096-1 https://linkinghub.elsevier.com/retrieve/pii/S004873*

Whitely, W. P. 1994. "The scientific community's response to evidence of fraudulent publication. The Robert Slutsky case." *JAMA J. Am. Med. Assoc.* 272(2):170–173.

Wohlrabe, Klaus and Constantin Bürgi. 2021. "Do working papers increase journal citations? Evidence from the top 5 journals in economics." *Appl. Econ. Lett.* 28(17):1531–1535.
URL: *https://doi.org/10.1080/13504851.2020.1855303*

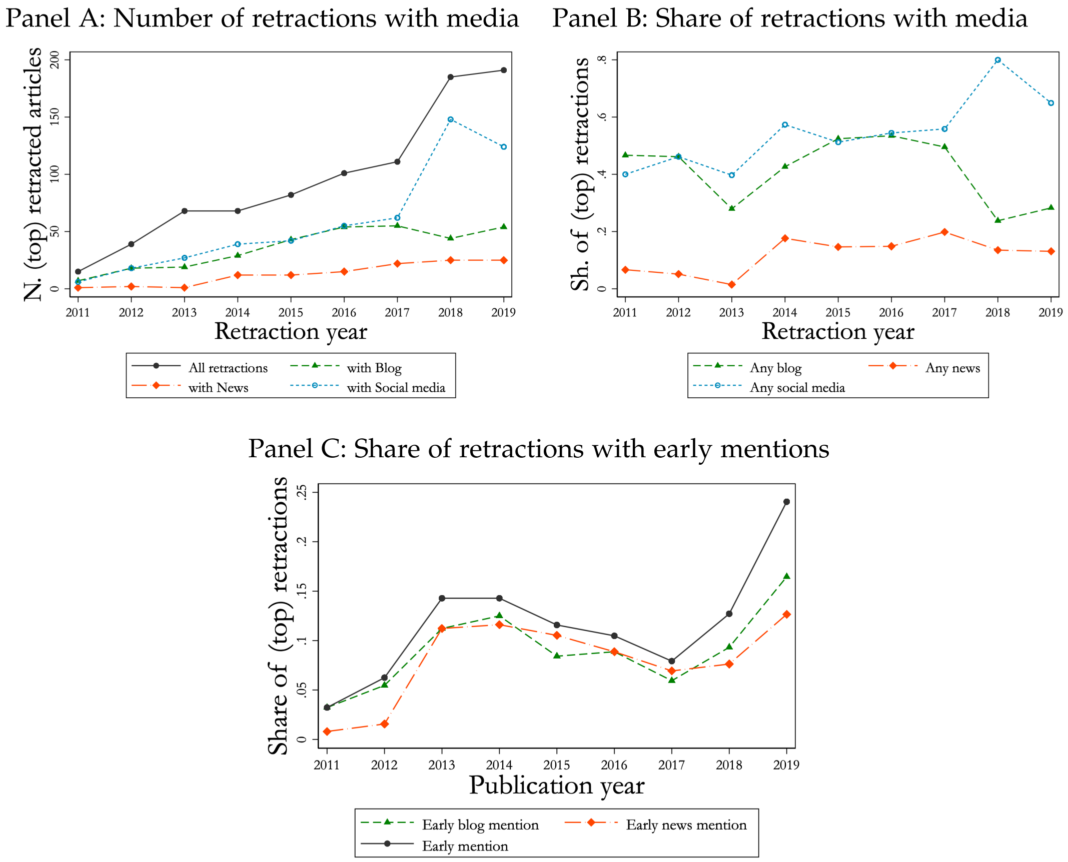
Woo, Seokkyun Joshua and John P Walsh. 2021. "On the Shoulders of Fallen Giants What do references to retracted research tell us about citation behaviors ?".

Yin, Yian, Yuxiao Dong, Kuansan Wang, Dashun Wang and Benjamin F. Jones. 2022. "Public use and public funding of science." *Nat. Hum. Behav.* .
URL: *https://www.nature.com/articles/s41562-022-01397-5*

Ziegler, Lennart. 2021. "What is the Media Impact of Research in Economics ?".

Figures

Figure 1: Media coverage of sample of retracted papers.

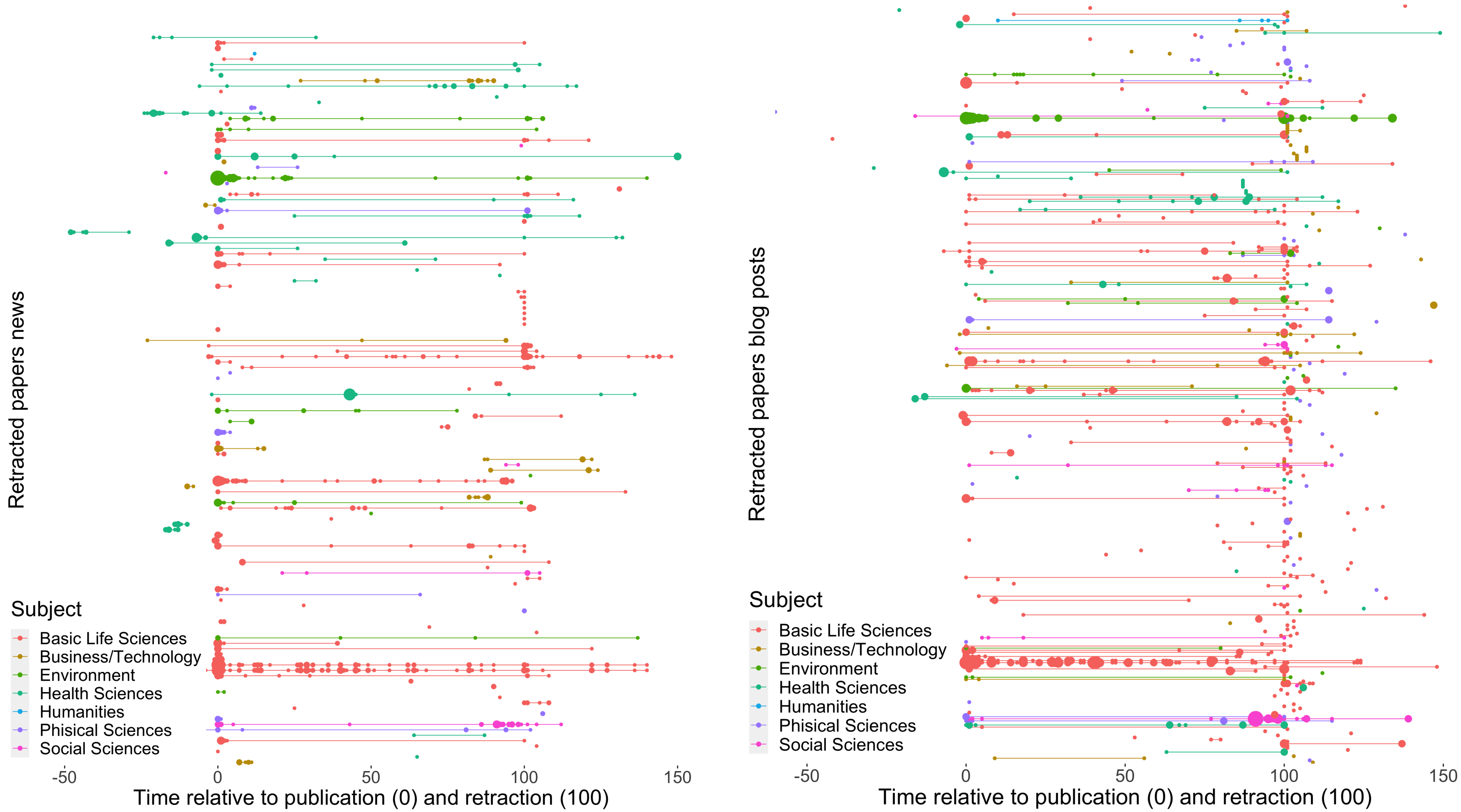


Note: Panel A shows the absolute number of retracted articles in the sample (green) which ever featured in blogs (orange), newspapers (blue), or social media (red), ordered by the year when the retraction occurred. Panel B shows the share of retracted papers that ever appeared in blogs (green), newspapers (orange), or social media (blues), again ordered by year of retraction. Panel C represents the share of retracted articles that were ever mentioned in blogs (green), newspapers (orange) or at least one of the two (blue) within two weeks from publication (i.e. *early mentions*), ordered by year of publication.

Figure 2: Newspaper and blog mentions of retracted articles.

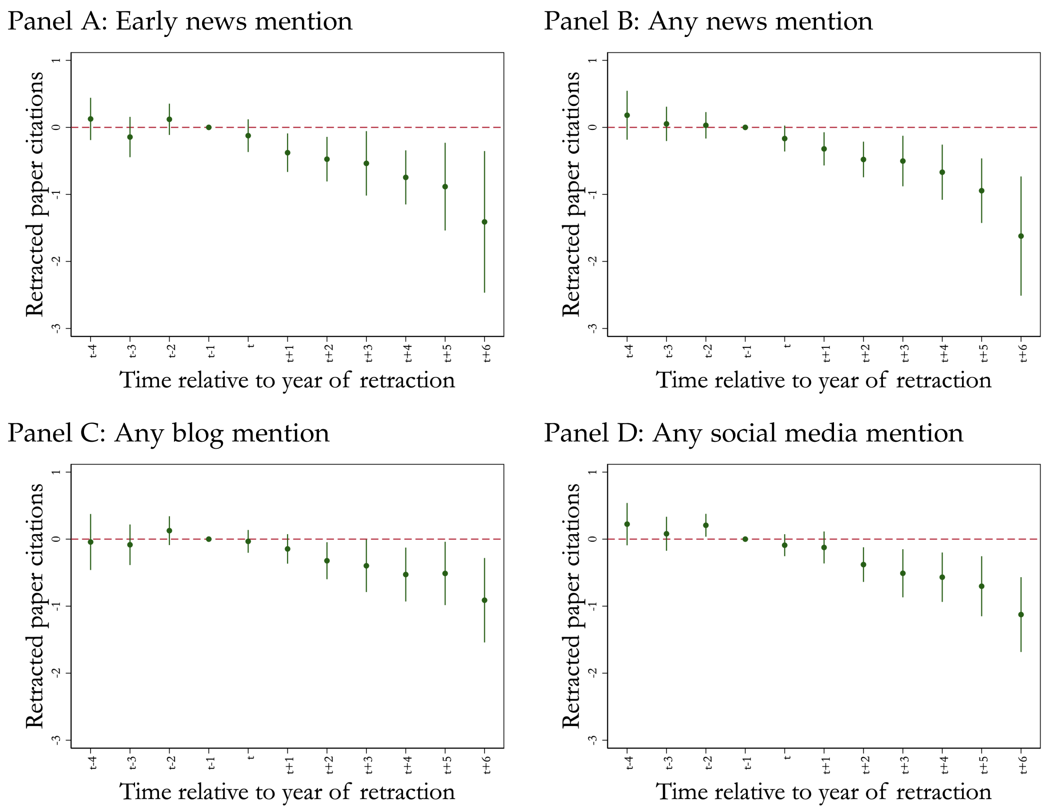
Panel A: News mentions (N=135 retractions)

Panel B: Blog mentions (N=365 retractions)



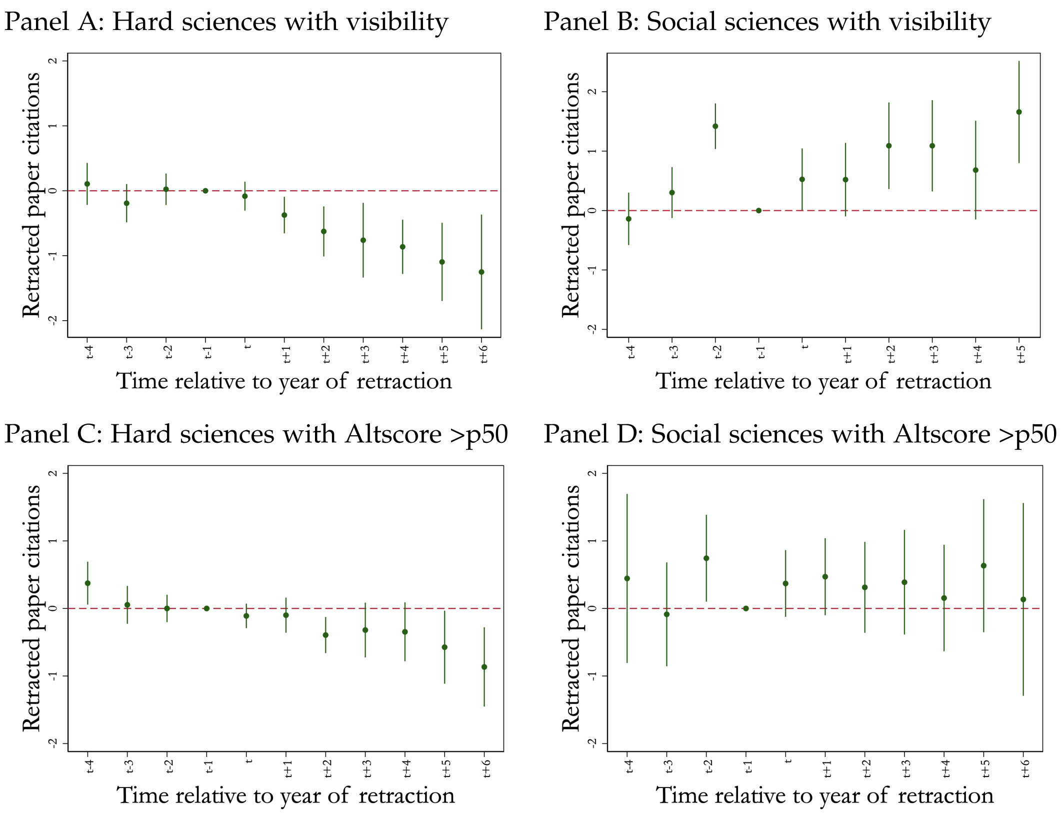
Note: Each line connects the first to the last mention of a single research article on either newspapers (Panel A) or blogs (Panel B) within the considered time window. Dots represent the number of mentions at a certain point in time. The window of analysis focuses on two events: the paper publication date (indexed with 0) and the paper retraction date (indexed with 100). The time score is allocated following the formula $\frac{(t_{mentionposted}-t_{publication})}{(t_{publication}-t_{retraction})} * 100$. The sources of publication date and retraction date are *Altmetric* and *RetractionWatch* respectively.

Figure 3: Dynamics of retracted papers penalty with media coverage



Note: Estimates replicate the following models: Table 5 column (3) for Panel A; Table A.4 column (3)-(4)-(1) respectively for Panel B, Panel C and Panel D. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). The coefficients displayed are that of the interaction between time dummies, a treatment indicator and a media indicator while vertical lines represent 95% CI.

Figure 4: Dynamics of retracted papers penalty with media coverage by discipline



Note: Hard sciences: life sciences, environment, health and physical sciences. Social sciences: business and technology, humanities, other social sciences. Estimates replicate the following models: Table 6 column (3)-(4) for Panel A and Panel B; Table A.19 column (3)-(4) respectively for Panel C and Panel D. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). The coefficients displayed are that of the interaction between time dummies, a treatment indicator and a media indicator, for different subsamples of discipline, while vertical lines represent 95% CI.

Figure 5: Papers' titles wordclouds

Panel A: Titles with media (N=1961) Panel B: Titles without media (N= 18794)

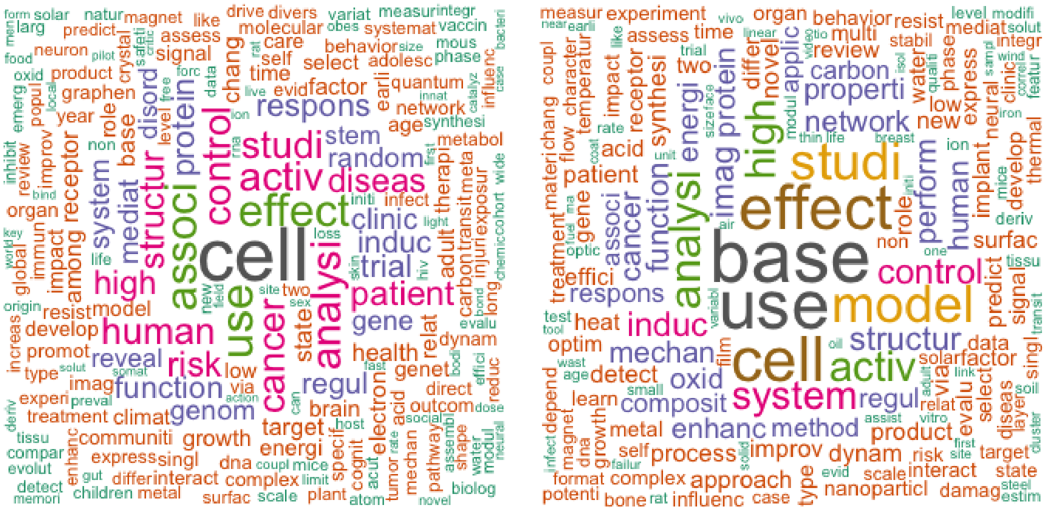
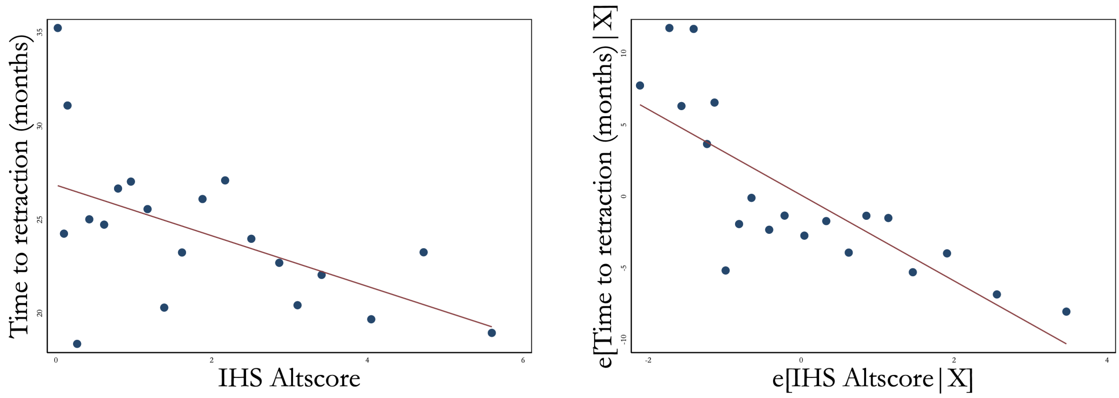


Figure 6: Months to retraction and Journal-year average visibility

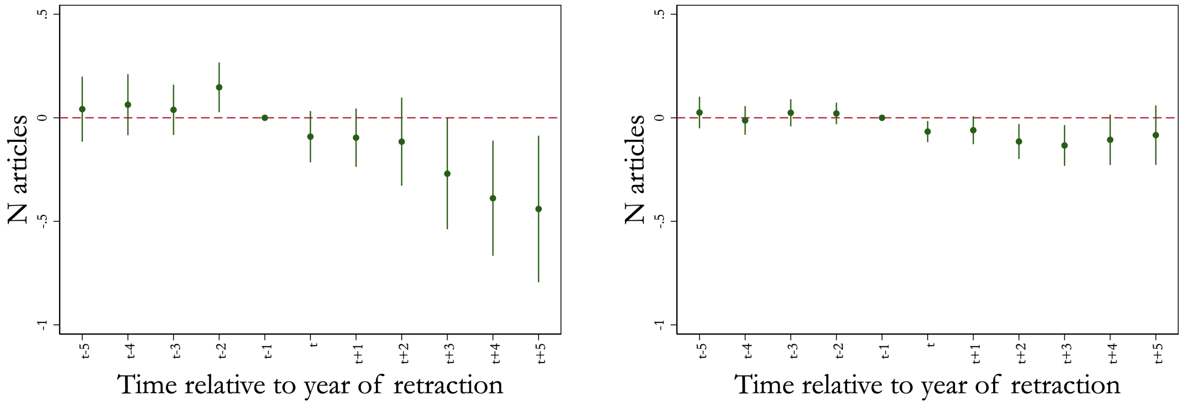
Panel A: Raw data Panel B: Absorb Controls + FE



Note: The vertical axis represents the time intercurring between an article publication and its retraction, expressed in months. The orixontal axis represents the inverse hyperbolic sine transformation of journal visibility, measured as the average Altscore of non-retracted papers that appear in the same yournal and year of the retracted one. Controls include the number of non-retracted articles within same journal and year of the treated, the average Euclidean distance of those from the retracted paper, and the level of (non-self) cumulative citations of the retracted paper before retraction. Publication year fixed effects are included. * p<0.1, ** p<0.05, *** p<0.01.

Figure 7: Dynamics of Author "productivity" with media coverage

Panel A: With Media Panel B: Without Media



Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33 Panel A column (1) and (9) respectively. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Tables

Table 1: Selected summary statistics

	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Diff.
	TREATMENT (N=990)				CONTROL papers (N=1969)				P-value
BALANCING VARIABLES									
Euclidean distance					0.937	2.661	0	45.51	
Arithmetic distance					0.171	1.699	-10	10	
Cum. (no self) citations $(t-1)$	7.103	19.11	0	254	6.807	18.65	0	258	0.6862
Early news mentions	1.037	7.844	0	134	0.787	5.790	0	127	0.3259
Early blog mentions	0.240	1.511	0	28	0.152	0.862	0	21	0.0429
Early mentions	1.278	9.194	0	155	0.939	6.493	0	139	0.2461
Age	5.138	2.612	0	9	5.140	2.612	0	9	0.9860
ADDITIONAL VARIABLES									
Time to retract	2.067	2.021	-0.504	9.353					
Yearly citations (no self)	2.628	4.442	0	56	5.259	12.27	0	354.8	
Cum. (no self) citations	16.83	30.50	0	418	33.91	75.29	0	1,774	
Altscore	37.54	274.5	0	7,128	19.09	130.2	0	3,728	
Tweeters count	32.12	349.9	0	10,105	13.55	165.4	0	5,100	
News count	1.459	8.589	0	122	1.068	6.158	0	113	
Blog count	0.871	3.283	0	65	0.287	1.333	0	27	
Any social media mention	0.597	0.491	0	1	0.443	0.497	0	1	
Any news mention	0.136	0.343	0	1	0.119	0.324	0	1	
Any blog mention	0.369	0.483	0	1	0.110	0.313	0	1	
Any early mentions	0.111	0.314	0	1	0.0945	0.293	0	1	

Note: Self-citations are excluded from citation count. *Early mentions* include all news and/or blog posts published within 2 weeks from publication. *Altscore* is a weighted average of all online mentions across outlets. Media *counts* are the number of outlets/accounts referring to a paper at any point in time. All papers are published/retracted between 2011 and 2020.

Table 2: Author level sample size

	First author	Mid author	Last author	H-index >p50	H-index <=p50	Media	Not media	Severe misconduct	Non-Severe misconduct	Total
Treatment	708 (35%)	650 (32%)	689 (34%)	922 (45%)	1125 (55%)	265 (13%)	1782 (87%)	851 (42%)	1196 (58%)	2047
Control	1639 (35%)	1437 (31%)	1595 (34%)	2412 (52%)	2259 (48%)	500 (11%)	4171 (89%)	1974 (42%)	2697 (58%)	4671
Total	2347	2087	2284	3334	3384	765	5953	2825	3893	**6718**

Note: The sample includes authors of retracted (treatment) papers and authors of matched control papers after the first observed retraction. It includes a maximum of 3 authors per paper (ranked as first, mid or last as per order of appereance) which have at least one publication in the 5 years before the first observed retraction. The H-index is calculated based on pre-retraction publications. Media is an indicator for whether the original publication gained any early popularity in the media. Causes of retractions are classified as Severe based on Woo and Walsh (2021).

Table 3: Author level sub-sample size

	First author	Mid author	Last author	H-index >p50	H-index <=p50	Severe misconduct	Non-Severe misconduct	Sub total
Panel A: With (early) media								
Treatment	97 (37%)	82 (31%)	86 (32%)	154 (58%)	111 (42%)	138 (52%)	127 (48%)	265
Control	172 (34%)	161 (32%)	167 (33%)	300 (60%)	200 (40%)	279 (56%)	221 (44%)	500
Sub-total	269	243	253	454	311	417	348	765
Panel B: Without (early) media								
	First author	Mid author	Last author	H-index >p50	H-index <=p50	Severe misconduct	Non-Severe misconduct	Sub total
Treatment	611 (34%)	568 (32%)	603 (34%)	768 (43%)	1014 (57%)	713 (40%)	1069 (60%)	1782
Control	1467 (35%)	1276 (31%)	1428 (34%)	2112 (51%)	2059 (49%)	1695 (41%)	2476 (59%)	4171
Sub-total	2078	1844	2031	2880	3073	2408	3545	5953
Total	2347	2087	2284	3334	3384	2825	3893	**6718**

Note: The sample includes authors of retracted (treatment) papers and authors of matched control papers after the first observed retraction. It includes a maximum of 3 authors per paper (ranked as first, mid or last as per order of appereance) which have at least one publication in the 5 years before the first observed retraction. The H-index is calculated based on pre-retraction publications. Media is an indicator for whether the original publication gained any early popularity in the media. Causes of retractions are classified as Severe based on Woo and Walsh (2021).

Table 4: Author level summary statistics

	(1) N	(2) mean	(3) sd	(4) min	(5) max	(6) N	(7) mean	(8) sd	(9) min	(10) max
	Panel A: All					Panel B: First author				
N articles	55,116	5.465	8.413	0	170	18,792	3.366	5.689	0	149
N articles with grant	55,116	2.781	5.617	0	156	18,792	1.651	4.069	0	149
N of coauthors	55,116	5.818	7.460	0	100	18,792	5.166	7.271	0	100
	Panel C: Mid author					Panel D: Last author				
N articles	16,798	4.501	7.169	0	129	19,526	8.315	10.55	0	170
N articles with grant	16,798	2.307	4.817	0	129	19,526	4.277	7.036	0	156
N of coauthors	16,798	5.945	8.090	0	100	19,526	6.336	7.015	0	100
	Panel E: H-index above median					Panel F: H-index below median				
N articles	28,822	8.316	10.18	0	170	26,294	2.341	4.023	0	149
N articles with grant	28,822	4.386	6.988	0	156	26,294	1.021	2.585	0	149
N of coauthors	28,822	7.184	8.196	0	100	26,294	4.320	6.224	0	100
	Panel G: Severe cases of misconduct					Panel H: Non-severe cases of misconduct				
N articles	23,614	5.227	8.040	0	129	31,502	5.644	8.677	0	170
N articles with grant	23,614	2.843	5.784	0	129	31,502	2.735	5.488	0	156
N of coauthors	23,614	6.085	7.676	0	100	31,502	5.618	7.287	0	100

Note: All statistics are reported by year. *N of articles* are yearly publications per author. *N of articles with grant* are yearly publications per author that mention any source of funding. *N of coauthors* is the average number of authors across papers published by an author within a year. The sample includes authors of retracted (treatment) papers and authors of matched control papers after the first observed retraction. It includes a maximum of 3 authors per paper (ranked as first, mid or last as per order of appereance) which have at least one publication in the 5 years before the first observed retraction. The H-index is calculated based on pre-retraction publications. Media is an indicator for whether the original publication gained any early popularity in the media. Causes of retractions are classified as Severe based on Woo and Walsh (2021).

Table 5: Retracted papers penalty with early mentions

	(1) Citations	(2) Citations	(3) Citations
Post * Treatment	-0.959***	-0.983***	-0.977***
	(0.059)	(0.058)	(0.058)
Post * Treatment * Early mentions	-0.449***		
	(0.158)		
Post * Treatment * Early blog mentions		-0.396**	
		(0.185)	
Post * Treatment * Early news mentions			-0.418***
			(0.149)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	Y	Y	Y
Pseudo R2	0.709	0.709	0.709
N	15438	15438	15438
N clusters	966	966	966
N full	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table 6: Retracted papers penalty and early mentions by discipline

	(1)	(2)	(3) Hard sciences	(4) Social sciences
	Citations	Citations	Citations	Citations
Post * Treatment * Early mentions	0.592	0.364	-0.520***	0.366
	(0.396)	(0.262)	(0.163)	(0.264)
Post * Treatment * Early mentions * Business/Technology	-0.155			
	(0.522)			
Post * Treatment * Early mentions * Life sciences	-1.169**			
	(0.458)			
Post * Treatment * Early mentions * Environment	-0.917**			
	(0.457)			
Post * Treatment * Early mentions * Health	-1.195**			
	(0.529)			
Post * Treatment * Early mentions * Physics	-0.938**			
	(0.455)			
Post * Treatment * Early mentions * Hard sciences		-0.887***		
		(0.309)		
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.711	0.710	0.718	0.595
N	15399	15438	12980	2419
N clusters	964	966	798	166
N full	16672	16711	13837	2835

Note: Estimates derive from pseudo Poisson specifications. Hard sciences: life sciences, environment, health and physical sciences. Social sciences: business and technology, humanities, other social sciences. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table 7: Retracted papers penalty and early mentions by severity of misconduct

	All	Minor	Moderate	Severe
	Citations	Citations	Citations	Citations
Post * Treatment * Early mentions	-0.158	-0.142	0.096	-0.566***
	(0.335)	(0.360)	(0.196)	(0.193)
P * T * Early mentions * Moderate misconduct	0.239			
	(0.389)			
P * T * Early mentions * Severe misconduct	-0.418			
	(0.389)			
Article, Age & Year FE	Y	Y	Y	Y
Pseudo R2	0.710	0.593	0.694	0.739
N (N clusters)	15438 (966)	4312 (295)	3857 (256)	7269 (415)
N full	16711	4859	4157	7695

Note: Estimates derive from pseudo Poisson specifications. Causes of retractions are classified based on Woo and Walsh (2021). The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table 8: Likelihood of retraction and media coverage

	Panel A: Retraction									
	OLS		EBIC		AICC		CV		Rigourous	
Media coverage	0.009**	0.019***	0.014***	0.021***	0.016***	0.021***	0.016***	0.022***	0.013***	0.020***
	(0.004)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)
Predicted media			-0.078***	-0.074***	-0.067***	-0.068***	-0.066***	-0.068***	-0.135***	-0.140***
			(0.027)	(0.031)	(0.018)	(0.023)	(0.018)	(0.022)	(0.045)	(0.058)
R-squared	0.000	0.001	0.001	0.002	0.001	0.002	0.001	0.002	0.001	0.002
	Panel B: Retraction									
			EBIC		AICC		CV		Rigourous	
Predicted media			-0.059***	-0.065**	-0.048***	-0.060***	-0.048***	-0.068***	-0.104***	-0.127**
			(0.024)	(0.031)	(0.016)	(0.022)	(0.015)	(0.022)	(0.041)	(0.058)
R-squared			0.000	0.001	0.000	0.001	0.000	0.002	0.000	0.001
Pub. year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	N	Y	N	Y	N	Y	N	Y	N	Y
N	20755	20755	20755	20755	20755	20755	20755	20755	20755	20755
N clusters	1008	1008	1008	1008	1008	1008	1008	1008	1008	1008

Note: Estimates from OLS regression. The dependent variable *Retraction* is an indicator for whether a paper was retracted. *Media coverage* is an indicator for whether a paper attracted online coverage at publication. *Predicted media* is media coverage as predicted from the respective *lasso* procedures. Boostrap standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table 9: Citation statements mentioning paper is retracted

VARIABLES	(1)	(2) Statements mentioning paper is retracted	(3)
Post * Early mentions	2.077***		
	(0.457)		
Post * Altscore >p50		1.562**	
		(0.670)	
Post * Altscore 3rd quintile			-0.101
			(1.344)
Post * Altscore 4th quintile			1.137
			(1.223)
Post * Altscore 5th quintile			2.305**
			(1.155)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	Y	Y	Y
Relative yr FE	Y	Y	Y
Pseudo R2	0.361	0.341	0.355
N	531	531	531
N clusters	95	95	95
N full	5591	5591	5591

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations statements received by each paper in a particular year which explicitly mention the retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects, article age indicator variables and dummies for each year relative to the retraction. Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table 10: Impact on authors' careers (interaction)

	(1) All	(2) First author	(3) Mid author	(4) Last author	(5) H-index >p50	(6) H-index <=p50	(7) Severe misconduct	(8) Non-Severe misconduct
	Panel A: N articles							
Post * Treatment	-0.093***	-0.090	-0.071	-0.099**	-0.116***	-0.070	-0.225***	-0.019
	(0.035)	(0.059)	(0.055)	(0.046)	(0.041)	(0.054)	(0.051)	(0.046)
Post * Treatment * Media	-0.121	-0.513***	0.174	-0.110	-0.136	0.032	-0.153	-0.030
	(0.095)	(0.174)	(0.128)	(0.124)	(0.102)	(0.154)	(0.131)	(0.125)
Pseudo R2	0.595	0.495	0.578	0.592	0.566	0.411	0.593	0.598
N	54732	18633	16645	19448	28761	25966	23392	31336
N clusters	874	872	851	870	848	856	367	507
N authors	6666	2327	2065	2273	3326	3340	2795	3870
	Panel B: N articles with grant							
Post * Treatment	-0.061	-0.019	-0.022	-0.089	-0.105**	0.007	-0.165**	-0.014
	(0.045)	(0.075)	(0.068)	(0.054)	(0.049)	(0.071)	(0.069)	(0.054)
Post * Treatment * Media	-0.152	-0.563***	0.036	-0.099	-0.160	0.146	-0.181	-0.020
	(0.110)	(0.194)	(0.149)	(0.156)	(0.117)	(0.181)	(0.160)	(0.130)
Pseudo R2	0.562	0.479	0.540	0.567	0.541	0.386	0.566	0.561
N	50243	16642	15120	18472	28134	22108	21753	28486
N clusters	871	859	832	859	845	833	367	504
N authors	6070	2066	1857	2146	3252	2818	2578	3491
	Panel C: Avg. n collaborators							
Post * Treatment	-0.089***	-0.113**	-0.083*	-0.078**	-0.089***	-0.081*	-0.168***	-0.028
	(0.028)	(0.056)	(0.046)	(0.034)	(0.033)	(0.044)	(0.040)	(0.038)
Post * Treatment * Media	-0.002	-0.118	0.158	-0.024	-0.027	0.041	-0.040	0.080
	(0.085)	(0.154)	(0.148)	(0.090)	(0.089)	(0.181)	(0.122)	(0.104)
Pseudo R2	0.383	0.373	0.386	0.385	0.398	0.312	0.371	0.393
N	54732	18633	16645	19448	28761	25966	23392	31336
N clusters	874	872	851	870	848	856	367	507
N authors	6666	2327	2065	2273	3326	3340	2795	3870
Author FE	Y	Y	Y	Y	Y	Y	Y	Y
Career lenght FE	Y	Y	Y	Y	Y	Y	Y	Y
Calendar Year FE	Y	Y	Y	Y	Y	Y	Y	Y
N full	55116	18792	16798	19526	28822	26294	23614	31502

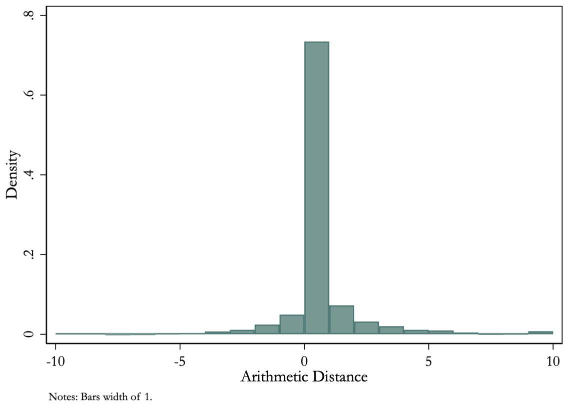
Note: Estimates derive from pseudo Poisson specifications. The dependent variables are: N. published articles x author x year; N. published articles with grant support x author x year; or Avg. n collaborators across all author's publications x year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Media is an indicator for cases where the original publication (either retracted or control papers) had at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate author fixed effects, a full suite of calendar-year effects and carreer lenght indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (e.g. x% loss in publication rate). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Appendix

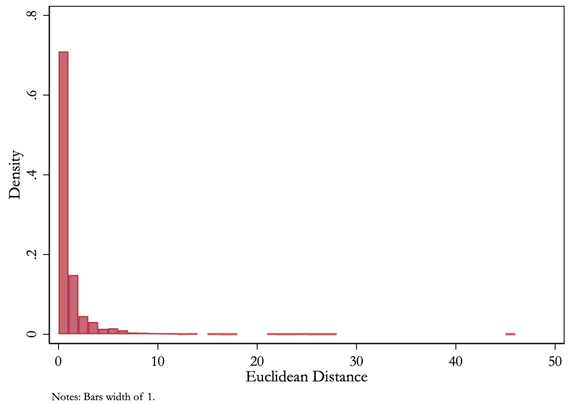
Figures

Figure A.1: Control quality: citations

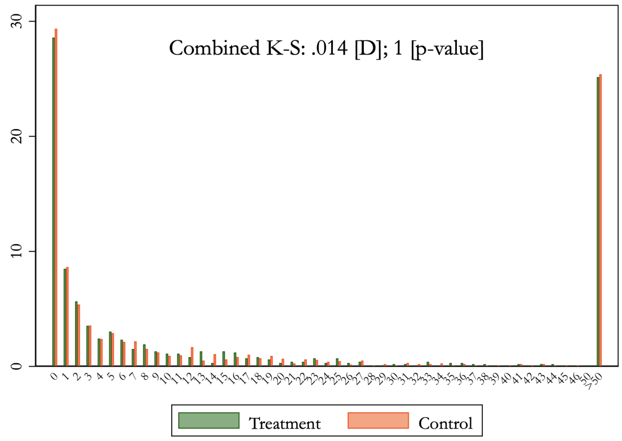
Panel A: AD



Panel B: ED



Panel C: Cumulative citations

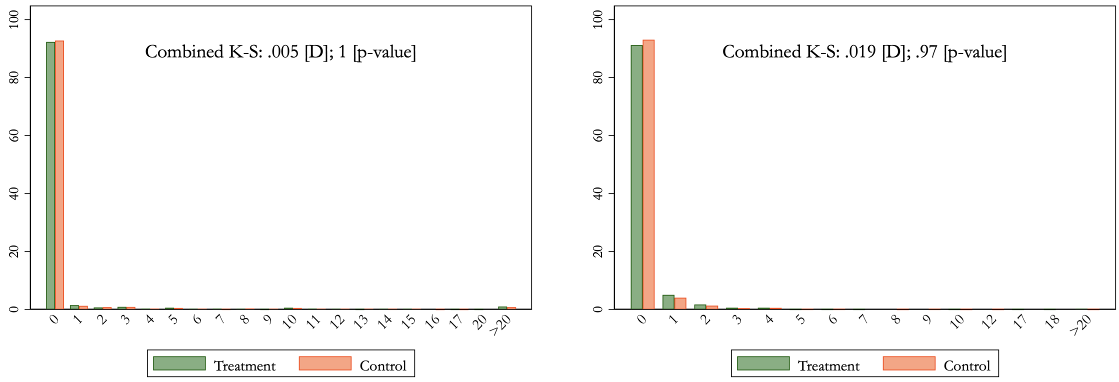


Note: All panels refer to pre-retraction measures. The year of retraction in that of the corresponding treated paper. Panel A shows the distribution of arithmetic distance (AD), panel B shows the distribution of Euclidean distance (ED), and panel C shows the distribution of cumulative citations from publication to the year before retraction and display the result of the Kolmogorov-Smirnov test of the equality of distributions between treatment and control group.

Figure A.2: Control quality: early mentions

Panel A: Newspaper early mentions

Panel B: Blog early mentions

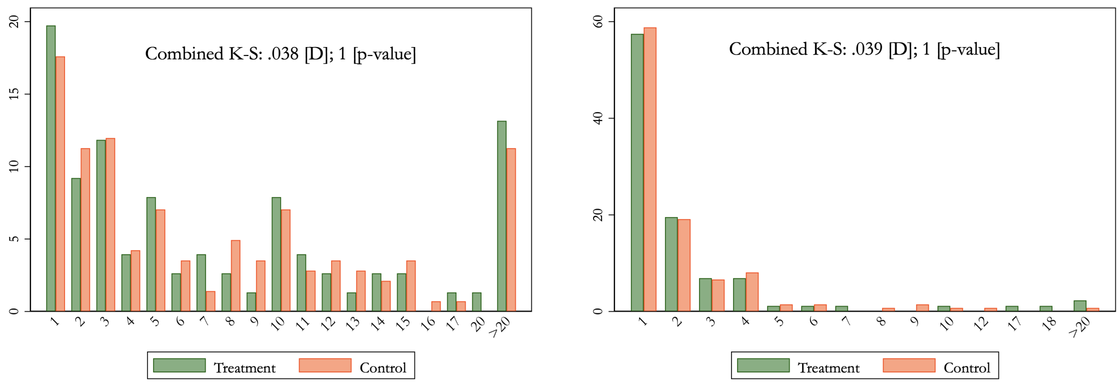


Note: Panels display the distribution of online mentions within two weeks from publication in newspapers (panel A) and blogs (panel B) across treated (green) and control (orange) papers. Both graph report the result of the Kolmogorov-Smirnov test of the equality of distributions across groups.

Figure A.3: Control quality: early mentions (Mentions > 0)

Panel A: Newspaper early mentions

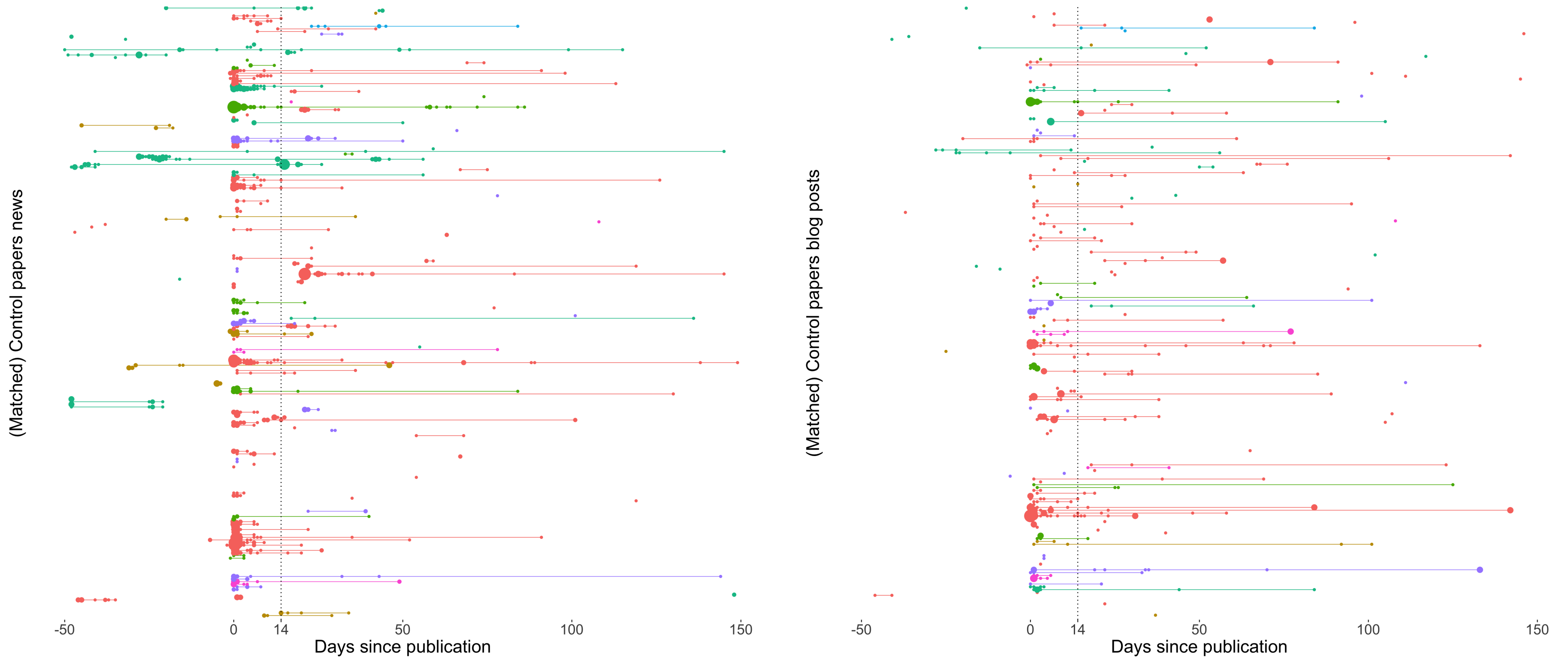
Panel B: Blog early mentions



Note: Panels display the distribution of online mentions within two weeks from publication in newspapers (panel A) and blogs (panel B) across treated (green) and control (orange) papers. Publications with no mentions are excluded. Both graph report the result of the Kolmogorov-Smirnov test of the equality of distributions across groups.

Figure A.4: Newspaper and blog mentions of selected control articles.

Panel A: News mentions (N=235 controls) Panel B: Blog mentions (N=216 controls)

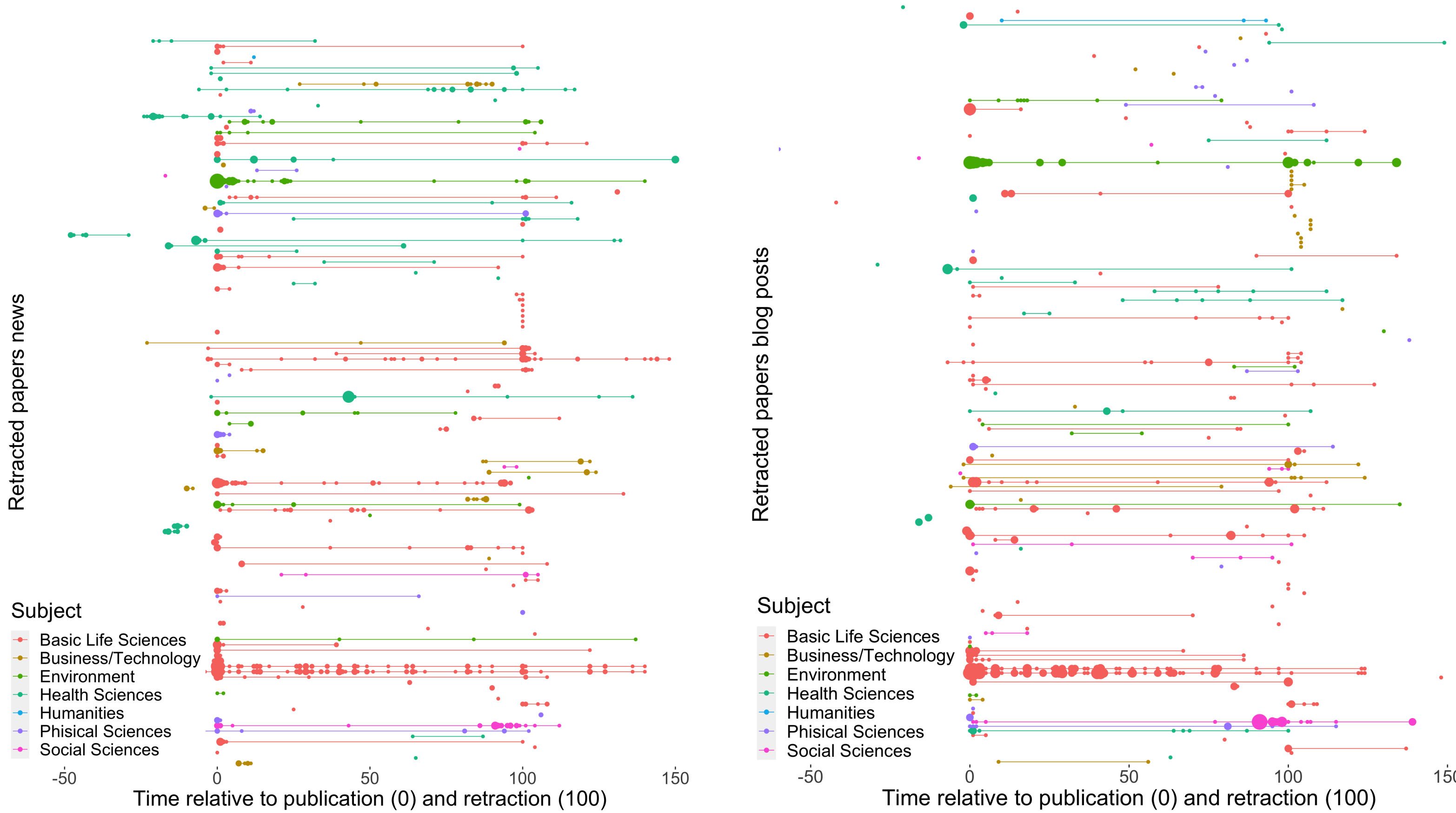


Note: Each line connects the first to the last mention of a single research article on either newspapers (Panel A) or blogs (Panel B) within the considered time window. Dots represent the number of mentions at a certain point in time. The source of publication date is *Altmetric*.

Figure A.5: Newspaper and blog mentions of retracted articles (excluding Retraction Watch from blogs)

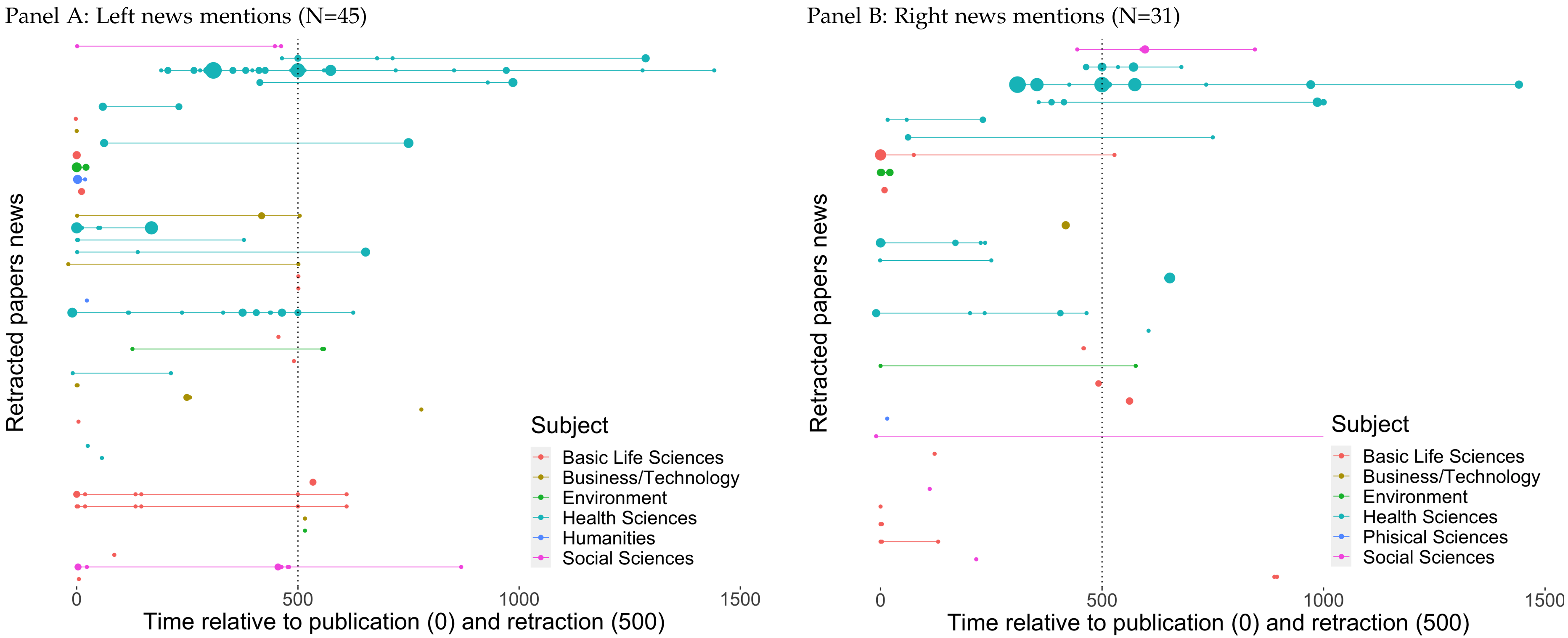
Panel A: News mentions (N=135)

Panel B: Blog mentions (N=171)



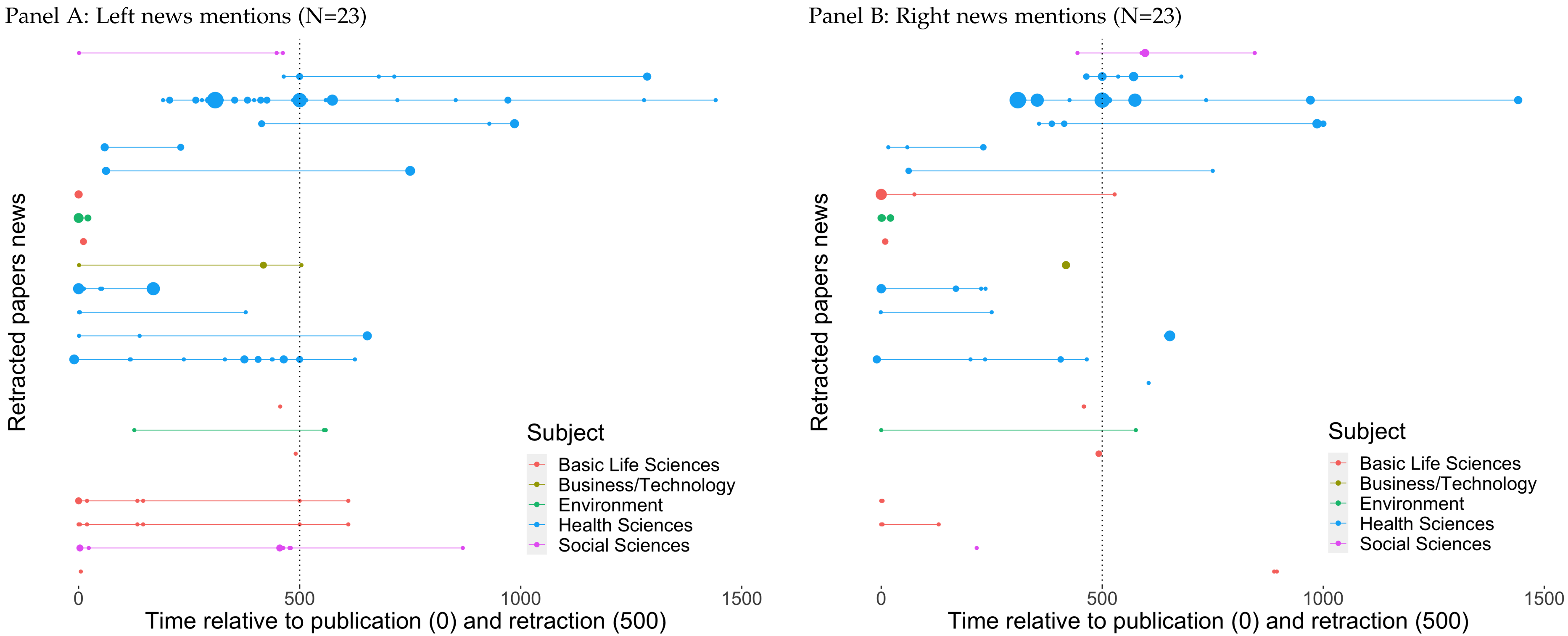
Note: Each line connects the first to the last mention of a single research article on either newspapers (Panel A) or blogs (Panel B) within the considered time window. Dots represent the number of mentions at a certain point in time. Blog mention from *RetractionWatch* are excluded. Source of publication date and retraction date: *Altmetric* and *RetractionWatch* respectively.

Figure A.6: US News media coverage of retracted papers by slant (within sample median).



Note: Each line connects the first to the last mention of a single research article on left leaning newspapers (Panel A) or right leaning (Panel B) within the considered time window. Right (left) leaning newspapers have a slam index (GS10) above median. The sample includes retractions in lower ranked journals. Dots represent the number of mentions at a certain point in time. Source of publication date and retraction date: *Altmetric* and *RetractionWatch* respectively.

Figure A.7: US News media coverage of retracted papers by slant (within sample median and balanced sample).

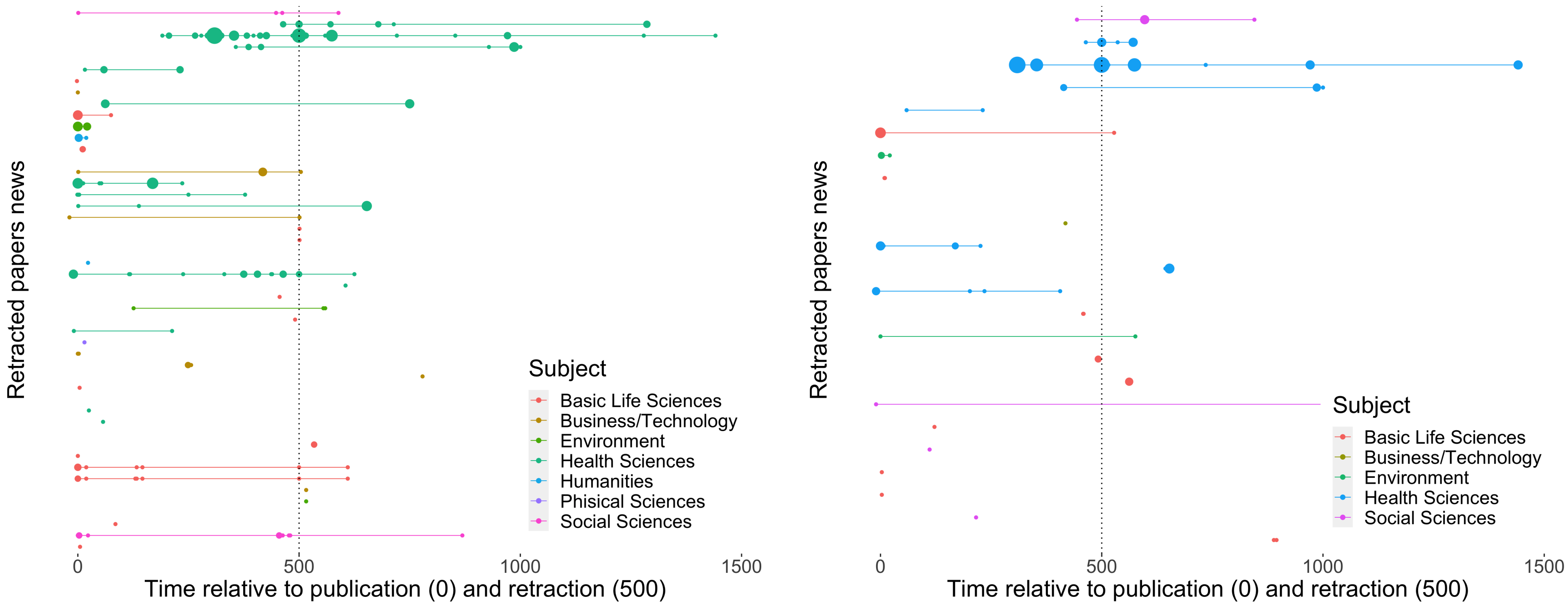


Note: Each line connects the first to the last mention of a single research article on left leaning newspapers (Panel A) or right leaning (Panel B) within the considered time window. Right (left) leaning newspapers have a slam index (GS10) above median. The sample includes retractions in lower ranked journals. Dots represent the number of mentions at a certain point in time. Source of publication date and retraction date: *Altmetric* and *RetractionWatch* respectively.

Figure A.8: US News media coverage of retracted papers by slant (GS10 median).

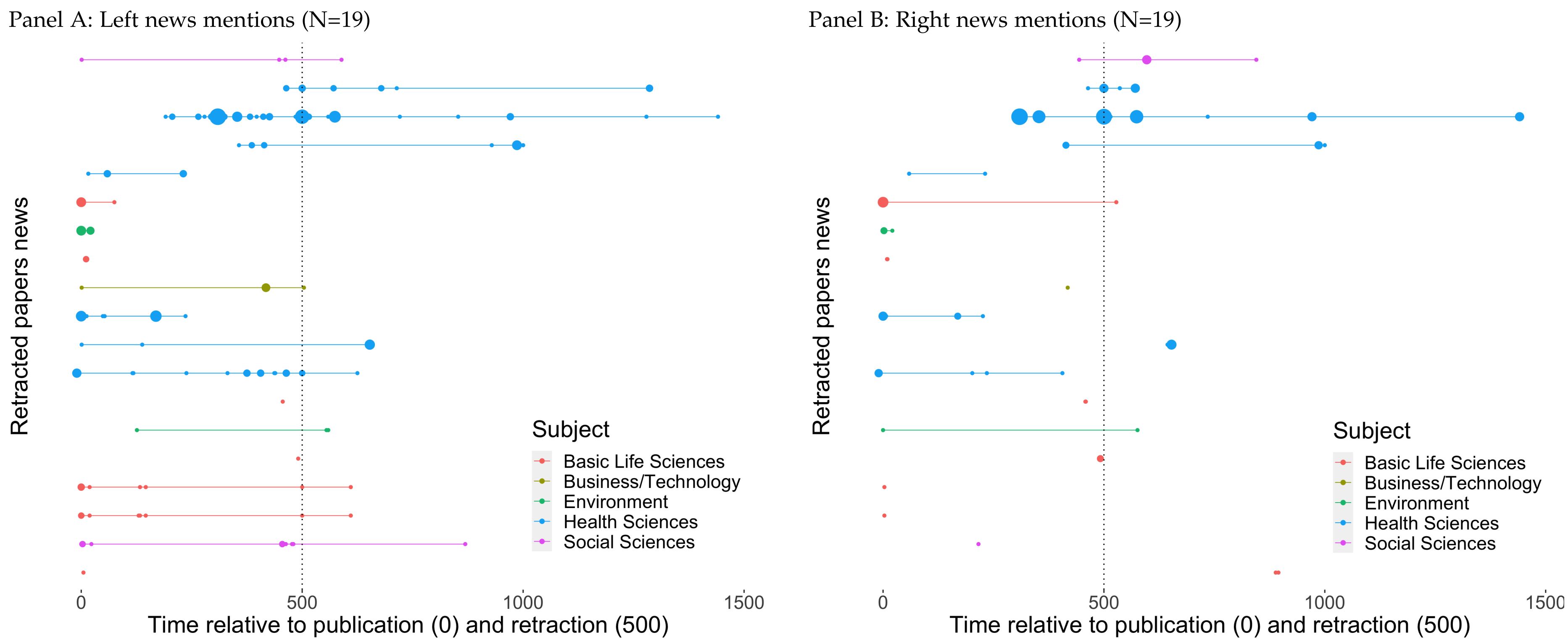
Panel A: Left news mentions (N=48)

Panel B: Right news mentions (N=24)



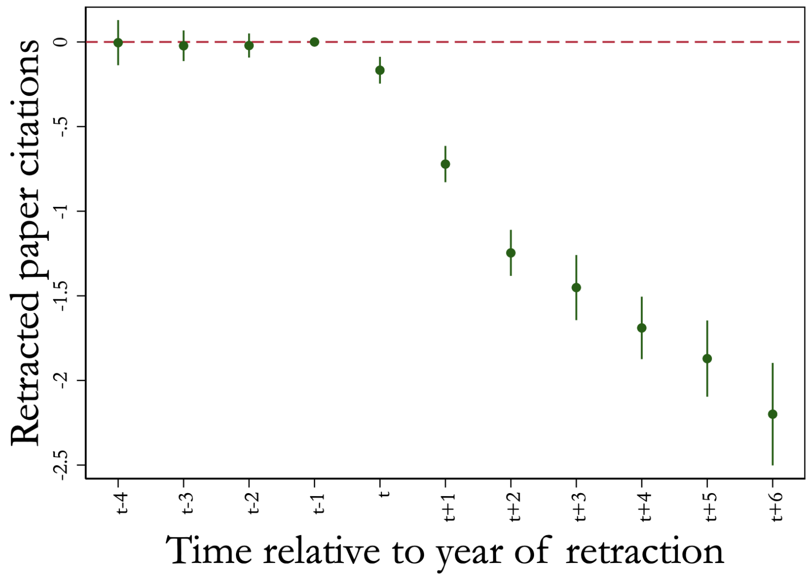
Note: Each line connects the first to the last mention of a single research article on left leaning newspapers (Panel A) or right leaning (Panel B) within the considered time window. Right (left) leaning newspapers have a slam index (GS10) above median. The sample includes retractions in lower ranked journals. Dots represent the number of mentions at a certain point in time. Source of publication date and retraction date: *Altmetric* and *RetractionWatch* respectively.

Figure A.9: US News media coverage of retracted papers by slant (GS10 median and balanced sample).



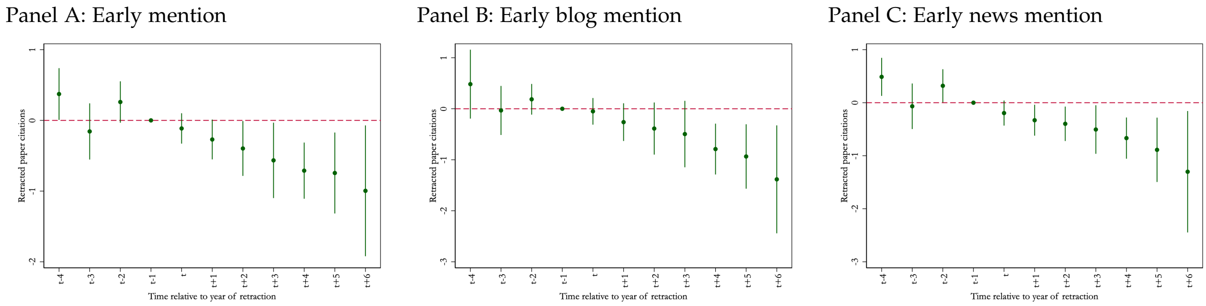
Note: Each line connects the first to the last mention of a single research article on left leaning newspapers (Panel A) or right leaning (Panel B) within the considered time window. Right (left) leaning newspapers have a slam index (GS10) above median. The sample includes retractions in lower ranked journals. Dots represent the number of mentions at a certain point in time. Source of publication date and retraction date: *Altmetric* and *RetractionWatch* respectively.

Figure A.10: Dynamics of retracted papers penalty



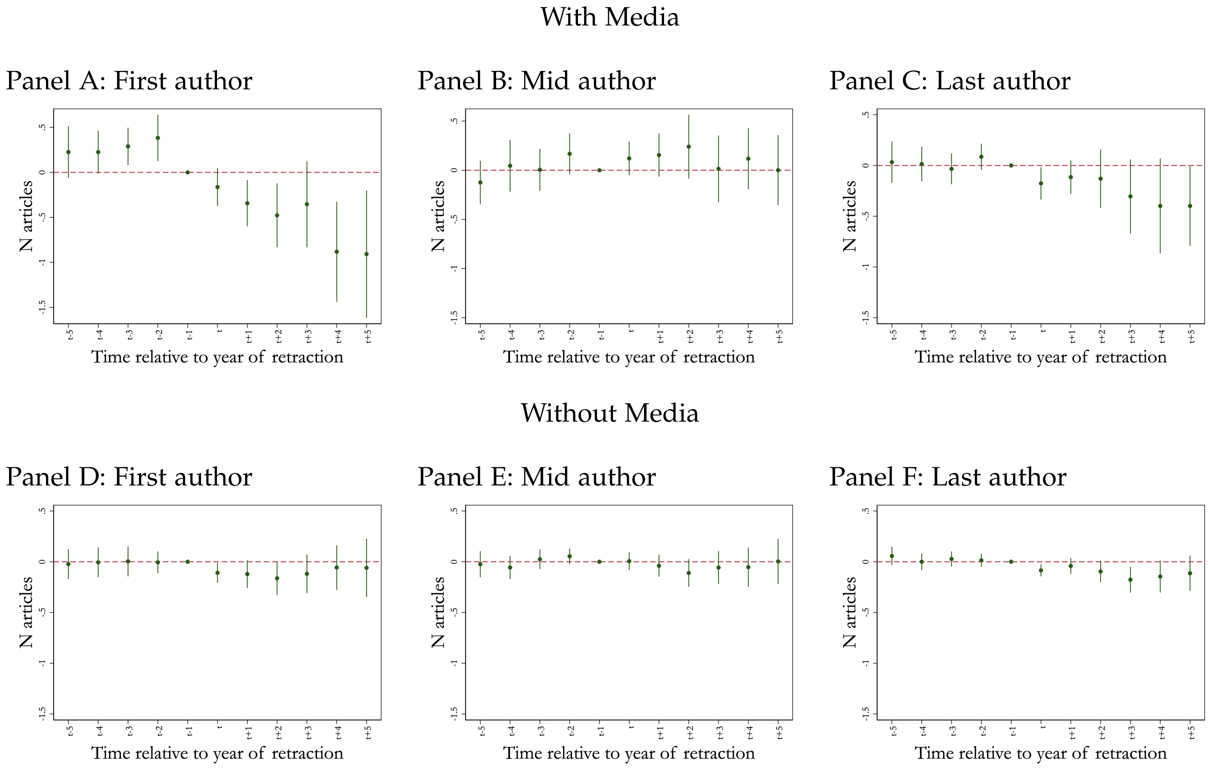
Note: Estimates replicate the model in Table A.3 column (2) but replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t - 1$ excluded). The coefficient displayed are that of the interaction between time dummies and a treatment indicator while the vertical lines represent 95% CI.

Figure A.11: Dynamics of retracted papers penalty with media, using random controls



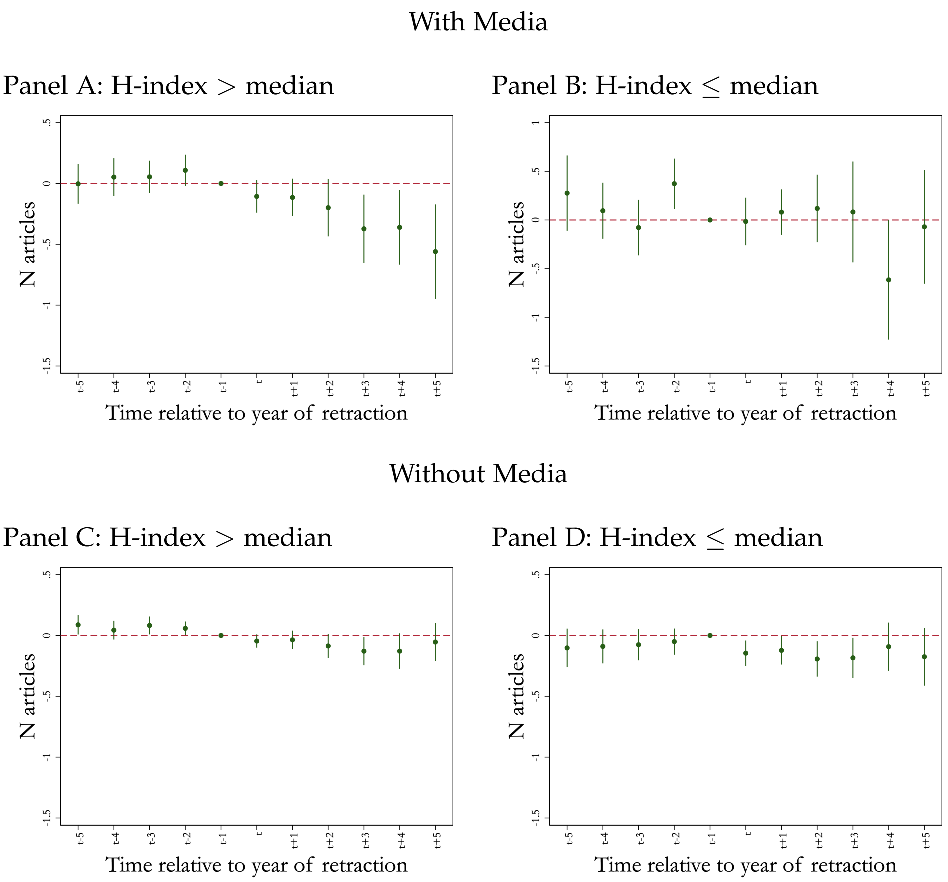
Note: Estimates replicate models in Table A.18 but replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). The coefficient displayed are that of the interaction between time dummies and a treatment indicator while the vertical lines represent 95% CI.

Figure A.12: Author "productivity" by rank (with and without media)



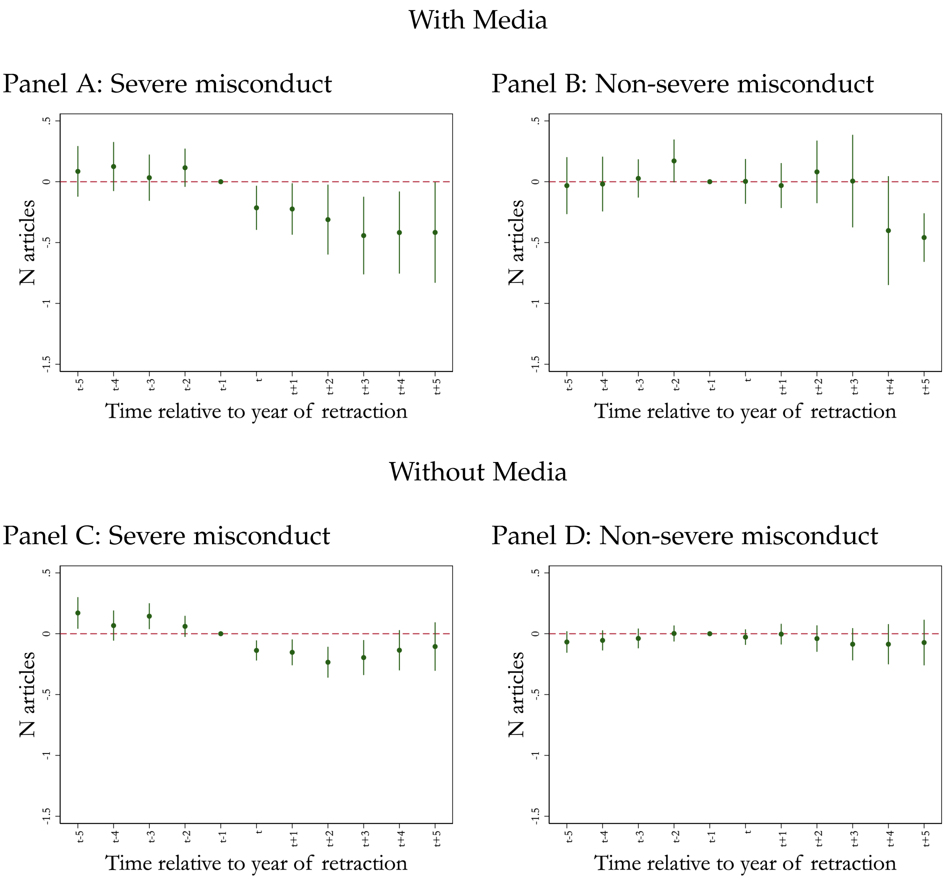
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.13: Author "productivity" by seniority (with and without media)



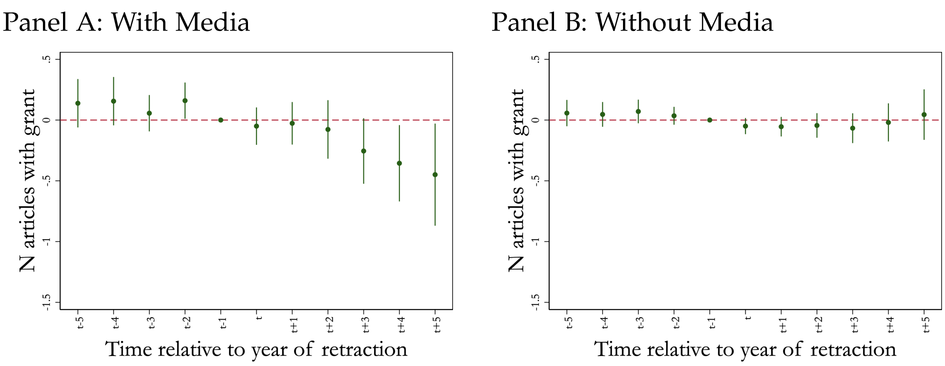
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.14: Author "productivity" by cause of retraction (with and without media)



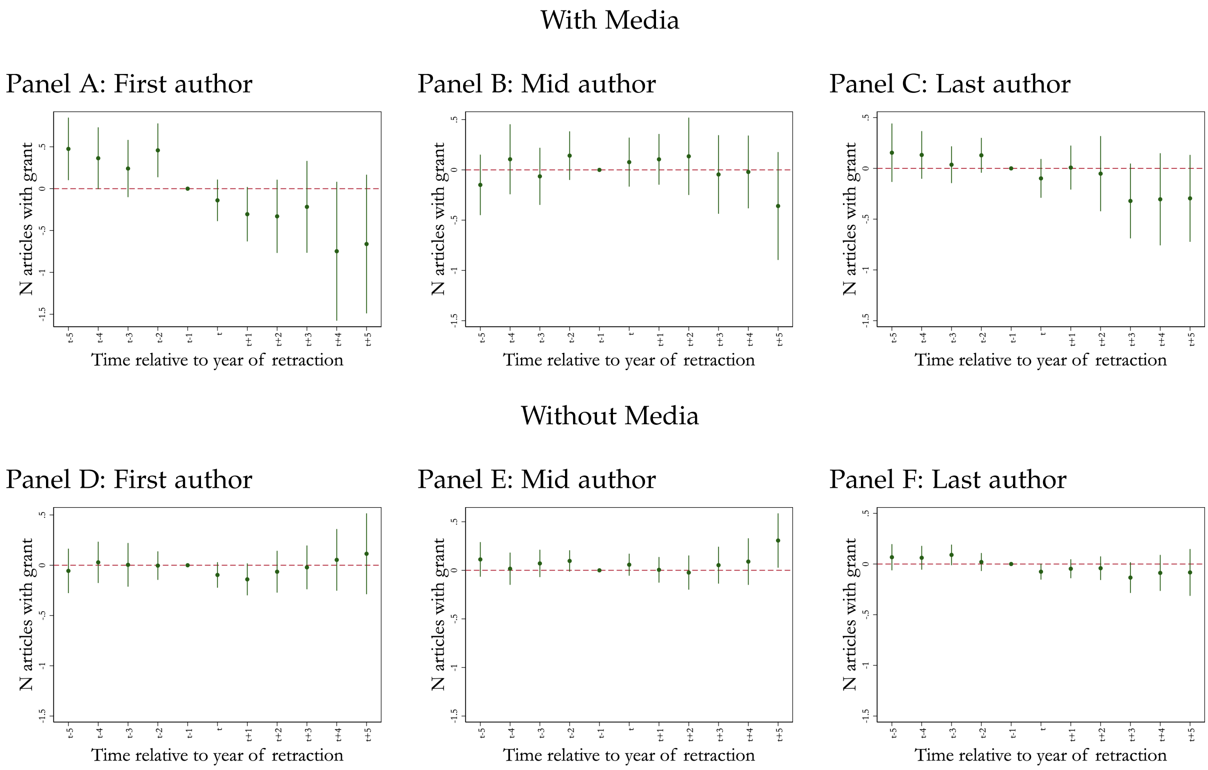
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.15: Author grant supported "productivity" (with and without media)



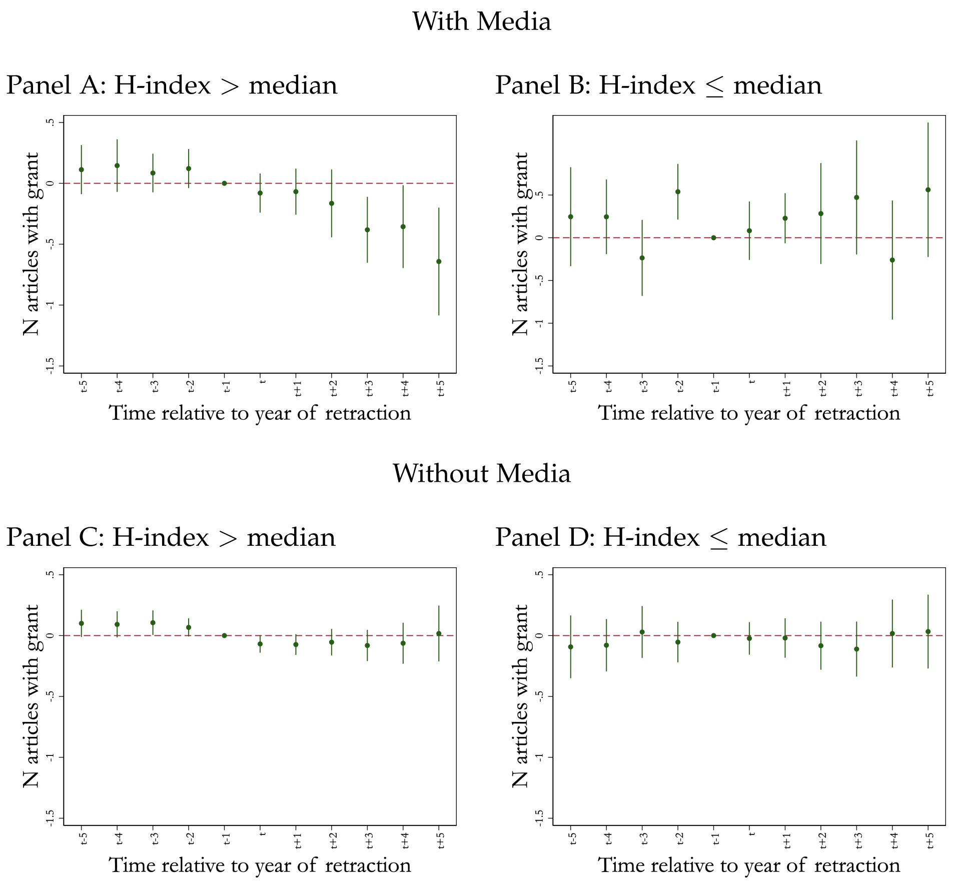
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.16: Author grant supported "productivity" by rank (with and without media)



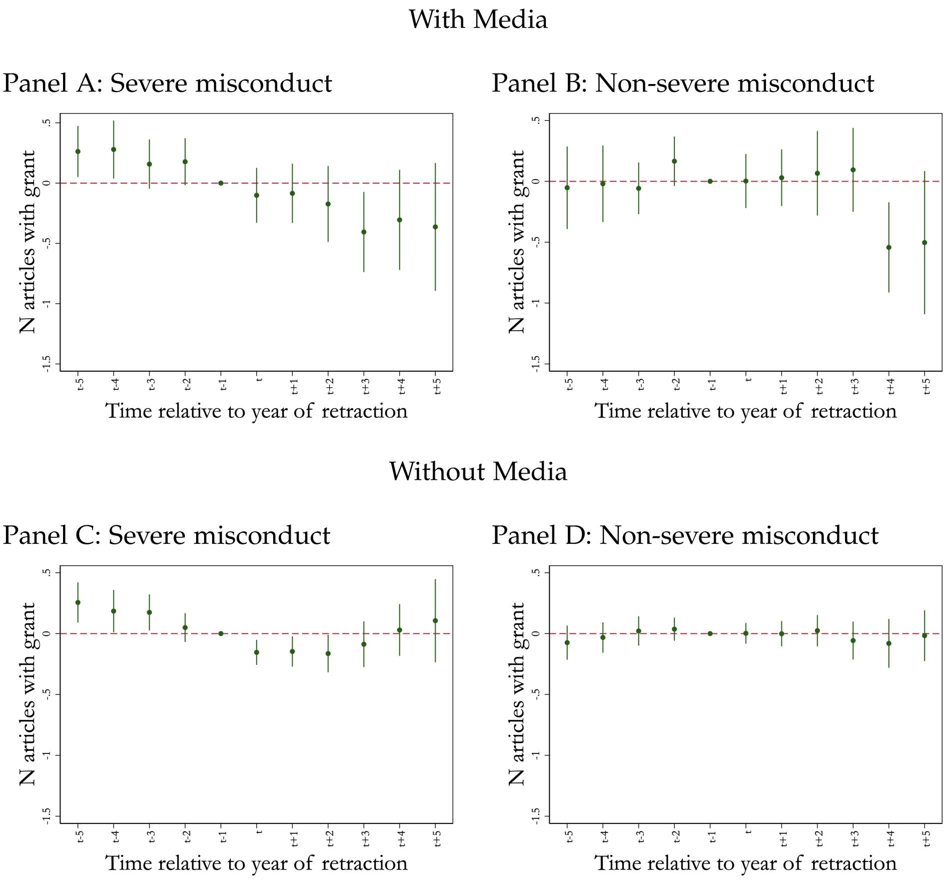
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.17: Author grant supported "productivity" by seniority (with and without media)



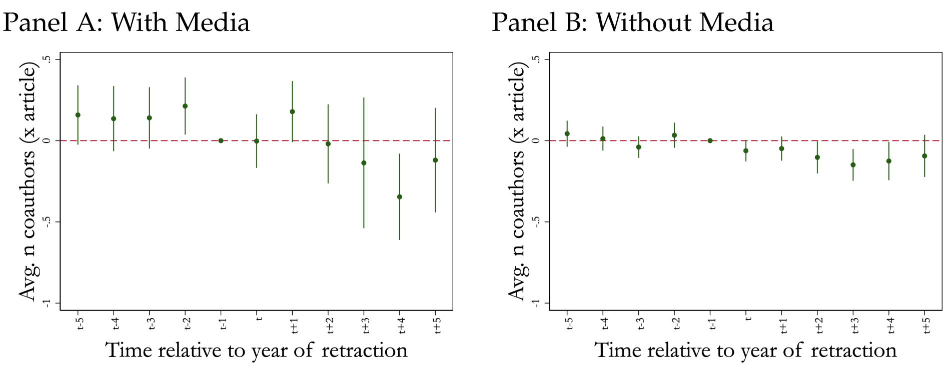
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.18: Author grant supported "productivity" by cause of retraction (with and without media)



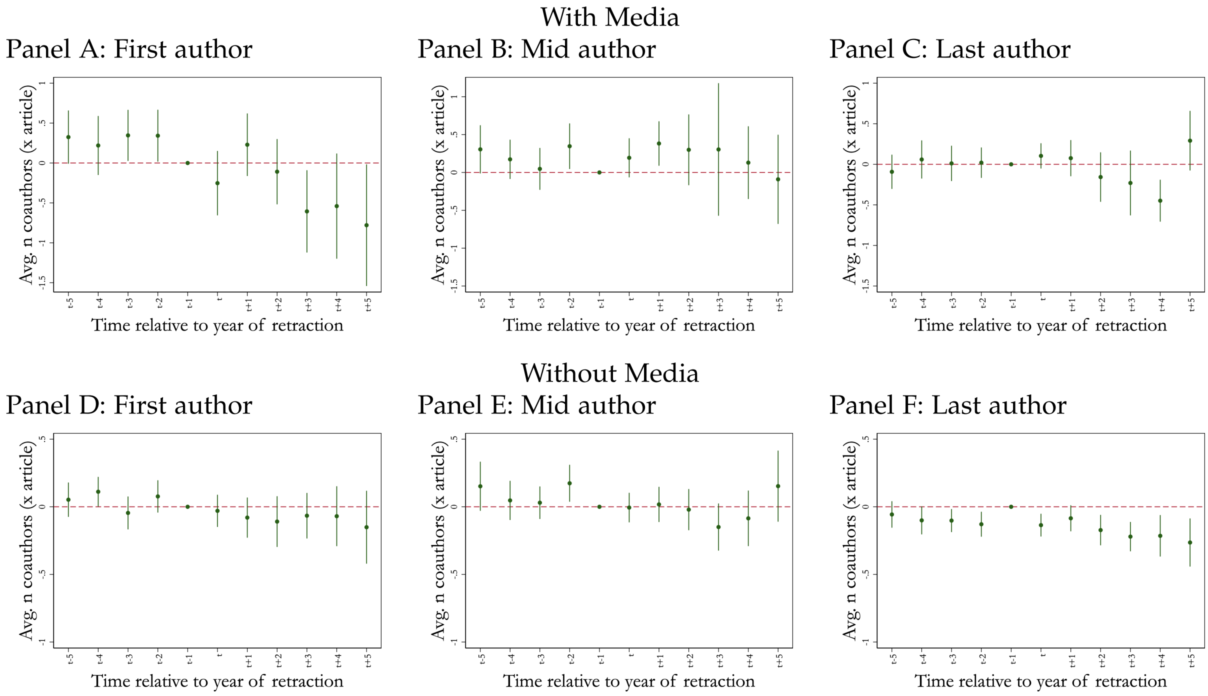
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.19: Author n. of coauthors (with and without media)



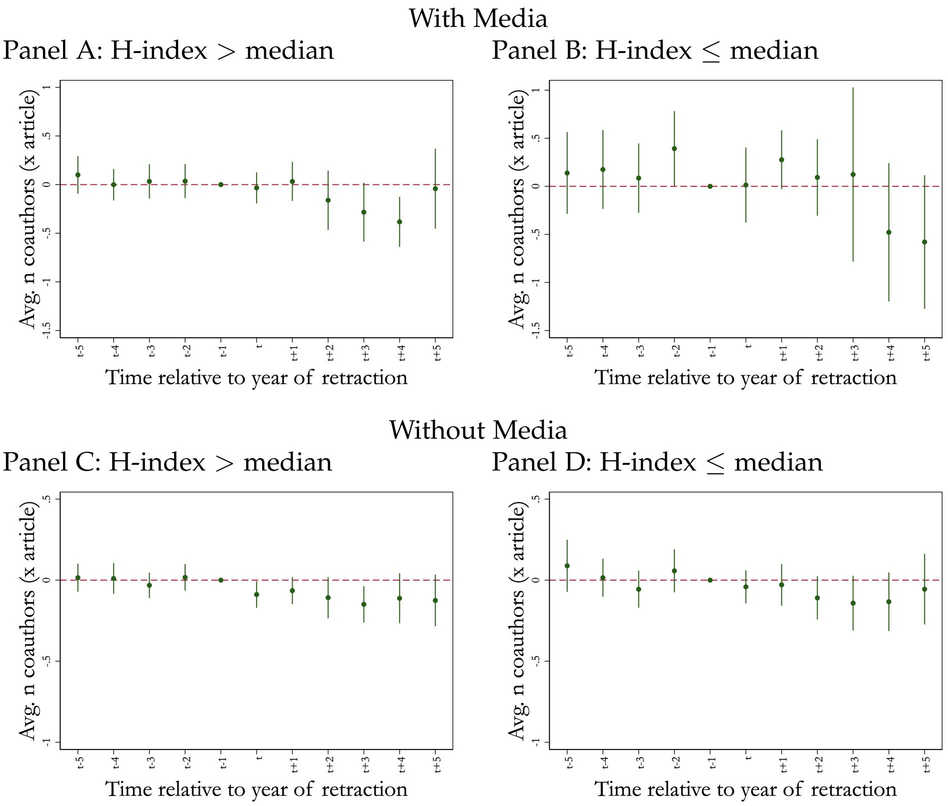
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.20: Author n. of coauthors by rank (with and without media)



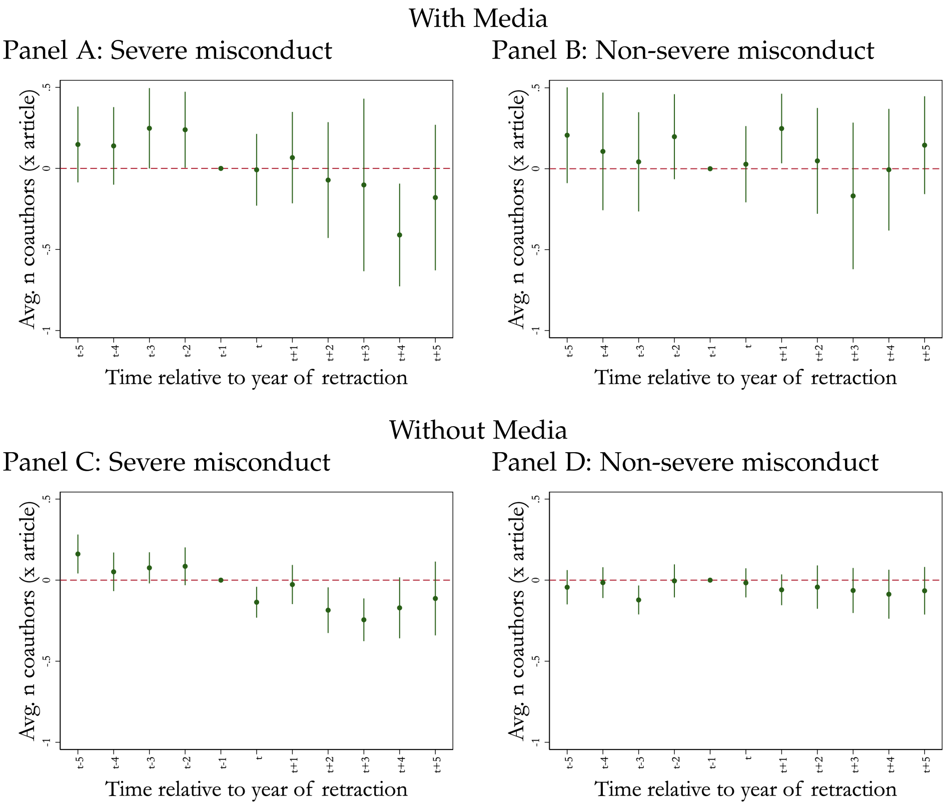
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.21: Author n. of coauthors by seniority (with and without media)



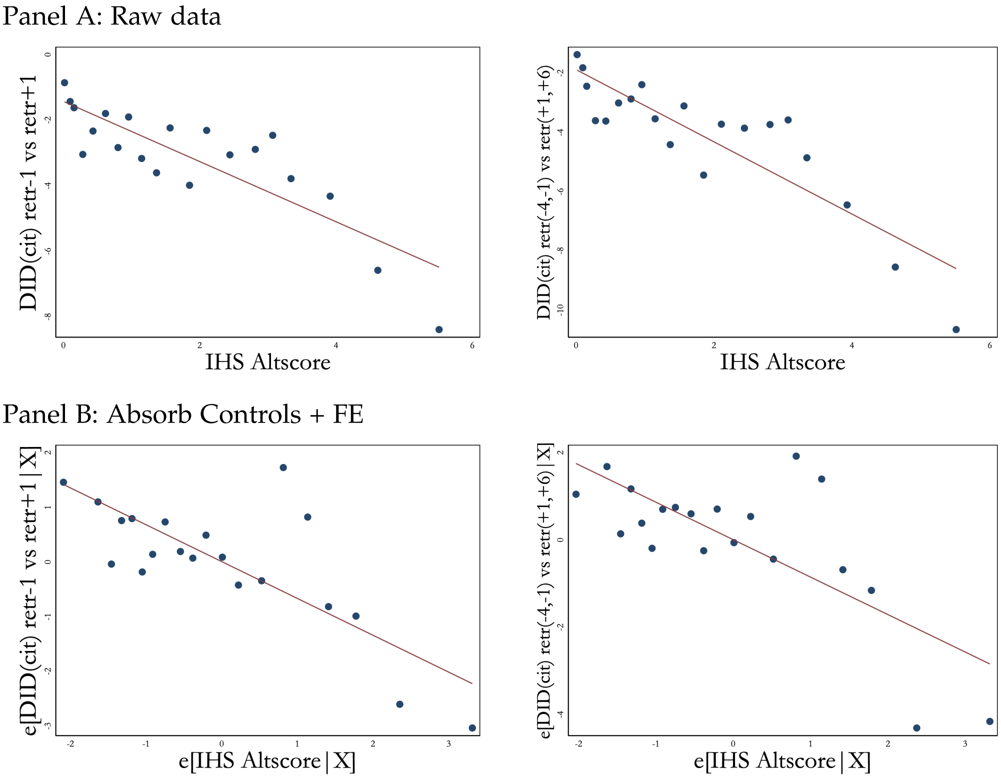
Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.22: Author n. of coauthors by cause of retraction (with and without media)



Note: Estimates compare publication rate for authors of retracted papers vs those of control papers before/after their first (in sample) retraction. Estimates replicate the models from Table A.33. Models are estimated replacing the *Post* indicator with a full set of dummies for each year relative to the retraction ($t-1$ excluded). Vertical lines represent 95% CI.

Figure A.23: Loss in citation and Controls average mentions



Note: The vertical axis represents the individual loss in citations obtained comparing each retracted paper to its selected controls for different pre- and post- time windows. The time window around retraction become larger moving left to right. The orixontal axis represents the inverse hyperbolic sine transformation of journal visibility, measured as the average Altscore of non-retracted papers that appear in the same yournal and year of the retracted one. Controls include the number of non-retracted articles within same journal and year of the treated, the average Euclidean distance of those from the retracted paper, and the level of (non-self) cumulative citations of the retracted paper before retraction. Publication year fixed effects are included.

Tables

Table A.1: Altscore weights

News	8
Blog	5
Policy document (per source)	3
Patent	3
Wikipedia	3
Twitter (tweets and retweets)	1
Peer review (Publons, Pubpeer)	1
Weibo (not trackable since 2015, but historical data kept)	1
Google+ (not trackable since 2019, but historical data kept)	1
F1000	1
Syllabi (Open Syllabus)	1
LinkedIn (not trackable since 2014, but historical data kept)	0.5
Facebook (only a curated list of public Pages)	0.25
Reddit	0.25
Pinterest (not trackable since 2013, but historical data kept)	0.25
Q&A (Stack Overflow)	0.25
Youtube	0.25
Number of Mendeley readers	0
Number of Dimensions and Web of Science citations	0

Table A.2: Author level statistics (within sub-samples)

	Media					No Media				
	(1) N	(2) mean	(3) sd	(4) min	(5) max	(6) N	(7) mean	(8) sd	(9) min	(10) max
					Panel A: All					
N articles	6,100	5.932	7.770	0	69	49,016	5.407	8.488	0	170
N articles with grant	6,100	3.560	5.363	0	61	49,016	2.684	5.640	0	156
N of coauthors	6,100	7.827	9.041	0	100	49,016	5.568	7.200	0	100
					Panel B: First author					
N articles	2,111	3.656	5.075	0	69	16,681	3.330	5.761	0	149
N articles with grant	2,111	2.147	3.537	0	47	16,681	1.588	4.127	0	149
N of coauthors	2,111	7.082	9.513	0	100	16,681	4.923	6.899	0	100
					Panel C: Mid author					
N articles	1,877	4.977	7.378	0	69	14,921	4.442	7.141	0	129
N articles with grant	1,877	3.058	5.200	0	52	14,921	2.212	4.758	0	129
N of coauthors	1,877	7.920	9.353	0	100	14,921	5.696	7.882	0	100
					Panel D: Last author					
N articles	2,112	9.055	9.195	0	68	17,414	8.225	10.70	0	170
N articles with grant	2,112	5.417	6.394	0	61	17,414	4.138	7.097	0	156
N of coauthors	2,112	8.488	8.184	0	82	17,414	6.075	6.814	0	100
					Panel E: H-index above median					
N articles	3,813	8.290	8.822	0	69	25,009	8.320	10.37	0	170
N articles with grant	3,813	5.053	6.189	0	61	25,009	4.285	7.096	0	156
N of coauthors	3,813	9.037	9.485	0	100	25,009	6.902	7.943	0	100
					Panel F: H-index below median					
N articles	2,287	2	2.560	0	27	24,007	2.374	4.134	0	149
N articles with grant	2,287	1.070	1.713	0	20	24,007	1.016	2.654	0	149
N of coauthors	2,287	5.808	7.845	0	100	24,007	4.178	6.028	0	100
					Panel G: Severe cases of misconduct					
N articles	3,496	5.711	7.724	0	69	20,118	5.143	8.091	0	129
N articles with grant	3,496	3.632	5.490	0	61	20,118	2.706	5.823	0	129
N of coauthors	3,496	7.986	9.226	0	100	20,118	5.754	7.324	0	100
					Panel H: Severe cases of misconduct					
N articles	2,604	6.228	7.822	0	56	28,898	5.591	8.748	0	170
N articles with grant	2,604	3.463	5.186	0	44	28,898	2.669	5.509	0	156
N of coauthors	2,604	7.613	8.785	0	100	28,898	5.438	7.109	0	100

Note: All statistics are reported by year. *N of articles* are yearly publications per author. *N of articles with grant* are yearly publications per author that mention any source of funding. *N of coauthors* is the average number of authors across papers published by an author within a year.

Table A.3: Retracted papers penalty

	Exponential	Exponential	OLS
	Citations	Citations	IHS(Citations)
Post	0.122***	0.115***	0.150***
	(0.024)	(0.025)	(0.025)
Post * Treatment	-1.067***	-1.064***	-0.830***
	(0.062)	(0.060)	(0.030)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	N	Y	Y
Pseudo R2 / R2	0.708	0.708	0.772
N	15438	15438	16679
N clusters	966	966	979
N full	16711	16711	16711

Note: First two columns show estimates of pseudo Poisson specifications while third column shows OLS estimation with IHS transformed dependent variable. The dependent variable is the total number of citations (exclusive of self-citations) received by each article in a particular year. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.4: Retracted papers penalty with any media coverage

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post * Treatment	-0.831***	-0.798***	-0.944***	-0.840***
	(0.083)	(0.072)	(0.059)	(0.077)
Post * Treatment * Any social media	-0.325***			
	(0.119)			
Post * Treatment * Any news-blog		-0.434***		
		(0.119)		
Post * Treatment * Any news			-0.377***	
			(0.127)	
Post * Treatment * Any blog				-0.392***
				(0.129)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.709	0.709	0.709	0.709
N	15438	15438	15438	15438
N clusters	966	966	966	966
N full	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Any socialmedia/news/blog is an indicator for papers with at least one overall mention in any of the indicated outlets. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.5: Retracted papers penalty and attention score

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post * Treatment	-0.841***	-0.815***	-0.921***	
	(0.090)	(0.068)	(0.051)	
Post * Treatment * Altscore >p50	-0.283**			
	(0.117)			
Post * Treatment * Altscore >p75		-0.433***		
		(0.118)		
Post * Treatment * Altscore >p90			-0.488***	
			(0.150)	
Post * Treatment * Altscore 3rd quintile				-0.732***
				(0.091)
Post * Treatment * Altscore 4th quintile				-0.918***
				(0.083)
Post * Treatment * Altscore 5th quintile				-1.267***
				(0.088)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.709	0.709	0.709	0.709
N	15438	15438	15438	15438
N clusters	966	966	966	966
N full	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.6: Retracted papers penalty with high cum. citations (pre-retraction)

	(1) Citations	(2) Citations	(3) Citations	(4) Citations	(5) Citations	(6) Citations
Post * Treatment	-1.121***	-1.057***	-0.996***	-1.008***	-0.954***	-0.900***
	(0.119)	(0.074)	(0.058)	(0.107)	(0.071)	(0.058)
Post * Treatment * Pre-retraction citations >p50	0.029			0.026		
	(0.140)			(0.126)		
Post * Treatment * Pre-retraction citations >p75		-0.045			-0.035	
		(0.122)			(0.115)	
Post * Treatment * Pre-retraction citations >p90			-0.226			-0.204
			(0.148)			(0.137)
Post * Treatment * Early mentions				-0.482***	-0.479***	-0.456***
				(0.163)	(0.160)	(0.155)
Article FE	Y	Y	Y	Y	Y	Y
Age FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Pseudo R2	0.709	0.709	0.709	0.709	0.709	0.709
N	15438	15438	15438	15438	15438	15438
N clusters	966	966	966	966	966	966
N full	16711	16711	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. Pre-retraction citations are indicators for papers with relatively higher cumulative citations before the year of retraction. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.7: Retracted papers penalty with high cum. citations (pre-retraction)

	(1) Citations	(2) Citations	(3) Citations	(4) Citations	(5) Citations	(6) Citations
Post * Treatment	-0.921***	-0.838***	-0.752***	-0.896***	-0.842***	-0.767***
	(0.112)	(0.077)	(0.073)	(0.126)	(0.093)	(0.087)
Post * Treatment * Pre-retraction citations >p50	0.117			0.050		
	(0.134)			(0.134)		
Post * Treatment * Pre-retraction citations >p75		0.048			-0.006	
		(0.116)			(0.116)	
Post * Treatment * Pre-retraction citations >p90			-0.178			-0.220
			(0.129)			(0.136)
Post * Treatment * Any news-blog	-0.447***	-0.455***	-0.436***			
	(0.123)	(0.124)	(0.120)			
Post * Treatment * Any social media				-0.335***	-0.328***	-0.319***
				(0.121)	(0.120)	(0.121)
Article FE	Y	Y	Y	Y	Y	Y
Age FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Pseudo R2	0.709	0.709	0.709	0.709	0.709	0.709
N	15438	15438	15438	15438	15438	15438
N clusters	966	966	966	966	966	966
N full	16711	16711	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Any socialmedia/news/blog is an indicator for papers with at least one overall mention in any of the indicated outlets. Pre-retraction citations are indicators for papers with relatively higher cumulative citations before the year of retraction. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.8: Retracted papers penalty with high cum. citations (pre-retraction)

	(1) Citations	(2) Citations	(3) Citations	(4) Citations	(5) Citations	(6) Citations
Post * Treatment	-0.924***	-0.854***	-0.771***	-1.009***	-0.948***	-0.885***
	(0.110)	(0.076)	(0.068)	(0.103)	(0.068)	(0.055)
Post * Treatment * Pre-retraction citations >p50	0.102			0.072		
	(0.133)			(0.128)		
Post * Treatment * Pre-retraction citations >p75		0.052			0.020	
		(0.115)			(0.112)	
Post * Treatment * Pre-retraction citations >p90			-0.173			-0.150
			(0.131)			(0.127)
Post * Treatment * Altscore >p75	-0.448***	-0.453***	-0.429***			
	(0.121)	(0.121)	(0.120)			
Post * Treatment * Altscore >p90				-0.508***	-0.503***	-0.480***
				(0.153)	(0.153)	(0.148)
Article FE	Y	Y	Y	Y	Y	Y
Age FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Pseudo R2	0.709	0.709	0.709	0.709	0.709	0.709
N	15438	15438	15438	15438	15438	15438
N clusters	966	966	966	966	966	966
N full	16711	16711	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. Pre-retraction citations are indicators for papers with relatively higher cumulative citations before the year of retraction. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.9: Retracted papers penalty with early mentions ($Post\ t \geq 0$)

VARIABLES	(1) Citations	(2) Citations	(3) Citations
Post * Treatment	-0.653***	-0.668***	-0.657***
	(0.061)	(0.061)	(0.060)
Post * Treatment * Early mentions	-0.327**		
	(0.138)		
Post* Treatment * Early blog mentions		-0.268	
		(0.170)	
Post * Treatment * Early news mentions			-0.347**
			(0.139)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	Y	Y	Y
Pseudo R2	0.701	0.701	0.701
N	15438	15438	15438
N clusters	966	966	966
N full	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.10: Retracted papers penalty with any media coverage (*Post* $t \geq 0$)

VARIABLES	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post * Treatment	-0.465***	-0.465***	-0.608***	-0.527***
	(0.105)	(0.092)	(0.065)	(0.088)
Post * Treatment * Any social media	-0.346***			
	(0.128)			
Post * Treatment * Any news-blog		-0.414***		
		(0.126)		
Post * Treatment * Any news			-0.355***	
			(0.115)	
Post * Treatment * Any blog				-0.339***
				(0.129)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.701	0.702	0.701	0.701
N	15438	15438	15438	15438
N clusters	966	966	966	966
N full	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Any socialmedia/news/blog is an indicator for papers with at least one overall mention in any of the indicated outlets. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.11: Retracted papers penalty and attention score (*Post* $t \geq 0$)

VARIABLES	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post * Treatment	-0.452***	-0.467***	-0.572***	
	(0.118)	(0.083)	(0.058)	
Post * Treatment * Altscore >p50	-0.339**			
	(0.133)			
Post * Treatment * Altscore >p75		-0.439***		
		(0.118)		
Post * Treatment * Altscore >p90			-0.554***	
			(0.127)	
Post * Treatment * Altscore 3rd quintile				-0.521***
				(0.089)
Post * Treatment * Altscore 4th quintile				-0.607***
				(0.082)
Post * Treatment * Altscore 5th quintile				-0.919***
				(0.072)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.701	0.702	0.702	0.702
N	15438	15438	15438	15438
N clusters	966	966	966	966
N full	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.12: Retracted papers penalty with early mentions (actively cited papers)

	(1) Citations	(2) Citations	(3) Citations
Post * Treatment	-1.036***	-1.078***	-1.064***
	(0.089)	(0.086)	(0.090)
Post * Treatment * Early mentions	-0.428**		
	(0.198)		
Post * Treatment * Early blog mentions		-0.362	
		(0.238)	
Post * Treatment * Early news mentions			-0.399**
			(0.165)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	Y	Y	Y
Pseudo R2	0.733	0.733	0.733
N	7308	7308	7308
N clusters	466	466	466
N full	7662	7662	7662

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.13: Retracted papers penalty with any media coverage (actively cited papers)

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post * Treatment	-0.739***	-0.785***	-1.010***	-0.892***
	(0.124)	(0.113)	(0.093)	(0.113)
Post * Treatment * Any social media	-0.577***			
	(0.162)			
Post * Treatment * Any news-blog		-0.618***		
		(0.162)		
Post * Treatment * Any news			-0.387***	
			(0.139)	
Post * Treatment * Any blog				-0.494***
				(0.171)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.733	0.734	0.733	0.733
N	7308	7308	7308	7308
N clusters	466	466	466	466
N full	7662	7662	7662	7662

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Any socialmedia/news/blog is an indicator for papers with at least one overall mention in any of the indicated outlets. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.14: Retracted papers penalty and attention score (actively cited papers)

	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post * Treatment	-0.704***	-0.805***	-0.967***	
	(0.132)	(0.106)	(0.080)	
Post * Treatment * Altscore >p50	-0.583***			
	(0.163)			
Post * Treatment * Altscore >p75		-0.590***		
		(0.157)		
Post * Treatment * Altscore >p90			-0.544***	
			(0.175)	
Post * Treatment * Altscore 3rd quintile				-0.769***
				(0.149)
Post * Treatment * Altscore 4th quintile				-1.123***
				(0.119)
Post * Treatment * Altscore 5th quintile				-1.317***
				(0.102)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.733	0.733	0.734	0.733
N	7308	7308	7308	7308
N clusters	466	466	466	466
N full	7662	7662	7662	7662

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.15: Retracted papers penalty with early mentions (published in 2011-2017)

VARIABLES	(1) Citations	(2) Citations	(3) Citations
Post*Treatment	-0.989***	-1.015***	-1.008***
	(0.060)	(0.059)	(0.060)
Post*Treatment*Early mentions	-0.438***		
	(0.165)		
Post*Treatment*Early blog mentions		-0.368*	
		(0.194)	
Post*Treatment*Early news mentions			-0.399**
			(0.155)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	Y	Y	Y
Pseudo R2	0.710	0.710	0.710
N	14194	14194	14194
N clusters	776	776	776
N full	15146	15146	15146

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.16: Retracted papers penalty with any media coverage (published in 2011-2017)

VARIABLES	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Post*Treatment	-0.837***	-0.803***	-0.959***	-0.878***
	(0.086)	(0.078)	(0.062)	(0.081)
Post*Treatment*Any social media	-0.359***			
	(0.123)			
Post*Treatment*Any news-blog		-0.459***		
		(0.123)		
Post*Treatment*Any news			-0.415***	
			(0.126)	
Post*Treatment*Any blog				-0.360***
				(0.134)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.710	0.710	0.710	0.710
N	14194	14194	14194	14194
N clusters	776	776	776	776
N full	15146	15146	15146	15146

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Any socialmedia/news/blog is an indicator for papers with at least one overall mention in any of the indicated outlets. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.17: Retracted papers penalty and attention score (published in 2011-2017)

VARIABLES	(1) Citations	(2) Citations	(3) Citations	
Post*Treatment	-0.835***	-0.820***	-0.936***	
	(0.095)	(0.073)	(0.054)	
Post*Treatment*Altscore >p50	-0.325***			
	(0.122)			
Post*Treatment*Altscore >p75		-0.463***		
		(0.121)		
Post*Treatment*Altscore >p90			-0.533***	
			(0.150)	
Post*Treatment*Altscore 3rd quintile				-0.681***
				(0.092)
Post*Treatment*Altscore 4th quintile				-0.940***
				(0.086)
Post*Treatment*Altscore 5th quintile				-1.308***
				(0.087)
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.710	0.710	0.710	0.711
N	14194	14194	14194	14194
N clusters	776	776	776	776
N full	15146	15146	15146	15146

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.18: Retracted papers penalty with early mentions, using random controls

	(1) Citations	(2) Citations	(3) Citations
Post * Treatment	-1.002***	-1.007***	-1.014***
	(0.062)	(0.060)	(0.061)
Post * Early mentions	-0.004		
	(0.079)		
Post * Treatment * Early mentions	-0.361**		
	(0.170)		
Post * Early blog mentions		0.042	
		(0.100)	
Post * Treatment * Early blog mentions		-0.393*	
		(0.205)	
Post * Early news mentions			-0.066
			(0.080)
Post * Treatment * Early news mentions			-0.340**
			(0.156)
Article FE	Y	Y	Y
Age FE	Y	Y	Y
Year FE	Y	Y	Y
Pseudo R2	0.718	0.718	0.718
N (clusters)	15630 (967)	15630 (967)	15630 (967)
N full	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for the year of retraction and all subsequent years. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.19: Retracted papers penalty and attention score by discipline

	(1)	(2)	(3) Hard sciences	(4) Social sciences
	Citations	Citations	Citations	Citations
Post * Treatment * Altscore >p50	1.333**	0.167	-0.321**	0.150
	(0.622)	(0.254)	(0.127)	(0.264)
Post * Treatment * Altscore >p50 * Business/Technology	-1.307*			
	(0.680)			
Post * Treatment * Altscore >p50 * Life sciences	-1.954***			
	(0.650)			
Post * Treatment * Altscore >p50 * Environment	-1.197*			
	(0.717)			
Post * Treatment * Altscore >p50 * Health	-0.948			
	(0.692)			
Post * Treatment * Altscore >p50 * Physics	-1.485**			
	(0.649)			
Post * Treatment * Altscore >p50 * Hard sciences		-0.494*		
		(0.283)		
Article FE	Y	Y	Y	Y
Age FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Pseudo R2	0.710	0.709	0.718	0.595
N	15399	15438	12980	2419
N clusters	964	966	798	166
N full	16672	16711	13837	2835

Note: Estimates derive from pseudo Poisson specifications. Hard sciences: life sciences, environment, health and physical sciences. Social sciences: business and technology, humanities, other social sciences. The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.20: Retracted papers penalty and attention score by severity of misconduct

	All	Minor	Moderate	Severe
	Citations	Citations	Citations	Citations
P * T * Altscore >p50	0.013	-0.023	0.381**	-0.519***
	(0.174)	(0.165)	(0.182)	(0.172)
P * T * Altscore >p50 * Moderate misconduct	0.361			
	(0.249)			
P * T * Altscore >p50 * Severe misconduct	-0.547**			
	(0.245)			
Article, Age & Year FE	Y	Y	Y	Y
Pseudo R2	0.710	0.594	0.694	0.738
N (N clusters)	15438 (966)	4312 (295)	3857 (256)	7269 (415)
N full	16711	4859	4157	7695

Note: Estimates derive from pseudo Poisson specifications. Causes of retractions are classified based on Woo and Walsh (2021). The dependent variable is the total number of citations (exclusive of self-citations) received by each paper in a particular year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citations). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.21: Misconduct classification from Woo and Walsh (2021)

Classification	Coded Retracted Reasons from Retraction Watch database
Minor misconducts	'Withdrawal', 'Author Unresponsive', 'Breach of Policy by Author', 'Breach of Policy by Third Party', 'Complaints about Author', 'Complaints about Company/Institution', 'Complaints about Third Party', 'Civil Proceedings', 'Concerns/Issues about Referencing/Attributions', 'Concerns/Issues about Third Party Involvement', 'Concerns/Issues About Authorship', 'Conflict of Interest', 'Copyright Claims', 'Criminal Proceedings', 'Date of Retraction/Other Unknown', 'Doing the Right Thing', 'Duplication of Article', 'Duplication of Data', 'Duplication of Image', 'Duplication of Text', 'Error by Journal/Publisher', 'Error by Third Party', 'Ethical Violations by Author', 'Ethical Violations by Third Party', 'Euphemisms for Duplication', 'Euphemisms for Misconduct', 'Euphemisms for Plagiarism', 'Informed/Patient Consent – None/Withdrawn', 'Lack of Approval from Author', 'Lack of Approval from Company/Institution', 'Lack of Approval from Third Party', 'Lack of IRB/IACUC Approval', 'Legal Reasons/Legal Threats', 'No Further Action', 'Nonpayment of Fees/Refusal to Pay', 'Notice – Lack of', 'Notice – Limited or No Information', 'Notice – Unable to Access via current resources', 'Objections by Author(s)', 'Objections by Company/Institution', 'Objections by Third Party', 'Plagiarism of Article', 'Plagiarism of Data', 'Plagiarism of Image', 'Plagiarism of Text', 'Publishing Ban', 'Retract and Replace', 'Salami Slicing', 'Temporary Removal', 'Updated to Correction', 'Updated to Retraction', 'Upgrade/Update of Prior Notice', 'Cites Prior Retracted Work'
Moderate misconducts	'Concerns/Issues About Authorship', 'Concerns/Issues About Data', 'Concerns/Issues About Image', 'Concerns/Issues About Results', 'Contamination of Cell Lines/Tissues', 'Contamination of Materials (General)', 'Contamination of Reagents', 'Error in Analyses', 'Error in Cell Lines/Tissues', 'Error in Data', 'Error in Image', 'Error in Materials (General)', 'Error in Methods', 'Error in Results and/or Conclusions', 'Error in Text', 'Investigation by Company/Institution', 'Investigation by Journal/Publisher', 'Investigation by ORI', 'Investigation by Third Party', 'Lack Of Balance/Bias Issues', 'Miscommunication by Author', 'Miscommunication by Company/Institution', 'Miscommunication by Journal/Publisher', 'Miscommunication by Third Party', 'Misconduct by Third Party'
Severe misconducts	'Fake Peer Review', 'Falsification/Fabrication of Data', 'Falsification/Fabrication of Image', 'Falsification/Fabrication of Results', 'Forged Authorship', 'Hoax Paper', 'Manipulation of Images', 'Manipulation of Results', 'Misconduct – Official Investigation/Finding', 'Misconduct by Author', 'Misconduct by Company/Institution', 'Results Not Reproducible', 'Sabotage of Materials', 'Sabotage of Methods', 'Unreliable Data', 'Unreliable Image', 'Unreliable Results'

Table A.22: Selected summary statistics: title ngrams

Most frequent (selected) ngrams					Relevant selected ngrams				
	Mean	Sd	Min	Max		Mean	Sd	Min	Max
Treatment	0.0477	0.213	0	1	# of adult	0.0070	0.084	0	2
Media	0.0945	0.293	0	1	# of algorithm	0.0113	0.107	0	2
Citations $year_p$	0.901	3.217	0	249	# of brain	0.009	0.097	0	1
Total number of words	14.28	5.047	1	54	# of climat	0.005	0.071	0	2
# of base	0.0773	0.272	0	2	# of commun	0.0080	0.091	0	2
# of effect	0.0666	0.252	0	2	# of composit	0.0201	0.142	0	2
# of studi	0.0543	0.228	0	2	# of disord	0.0064	0.083	0	2
# of model	0.0517	0.225	0	2	# of earli	0.0075	0.086	0	1
# of analysi	0.0453	0.209	0	2	# of genom	0.0100	0.101	0	2
# of system	0.0380	0.195	0	2	# of global	0.0065	0.082	0	2
# of induc	0.0314	0.176	0	2	# of graphen	0.0093	0.101	0	3
# of imag	0.0291	0.172	0	3	# of meta_analysi	0.0050	0.070	0	1
# of human	0.0270	0.165	0	2	# of model	0.0517	0.225	0	2
# of perform	0.0256	0.159	0	2	# of network_ETX	0.0082	0.090	0	1
# of mechan	0.0252	0.158	0	2	# of neuron	0.0052	0.075	0	2
# of properti	0.0244	0.155	0	2	# of reveal	0.0092	0.096	0	1
# of oxid	0.0242	0.163	0	3	# of risk	0.0154	0.127	0	3
# of enhanc	0.0238	0.153	0	2	# of stem	0.0095	0.100	0	2
# of regul	0.0238	0.154	0	2	# of STX_structur	0.0057	0.076	0	1
# of associ	0.0236	0.155	0	2	# of trial	0.0108	0.104	0	2
# of respons	0.0233	0.153	0	2	# of vitro	0.0064	0.080	0	1

Note: N-grams represent the number of times the selected espression appears in the title of a research article. All n-grams in the table were selected by one of the lasso procedures. N=20755.

Table A.23: Selected words and media coverage

	Media coverage							
	Linear				Logit			
	EBIC	AICC	CV	Rigourous	EBIC	AICC	CV	Rigourous
# of adult	0.110***	0.093***	0.092***		0.067***	0.044**	0.045**	0.073***
	(0.036)	(0.035)	(0.035)		(0.018)	(0.017)	(0.017)	(0.017)
# of algorithm		-0.047***	-0.048***	-0.049***		-0.186**	-0.185**	
		(0.008)	(0.008)	(0.007)		(0.076)	(0.076)	
# of brain	0.081***	0.068**	0.066**		0.062***	0.041**	0.041**	
	(0.026)	(0.027)	(0.027)		(0.015)	(0.017)	(0.017)	
# of climat	0.228***	0.221***	0.219***		0.134***	0.120***	0.126***	0.132***
	(0.054)	(0.050)	(0.050)		(0.022)	(0.020)	(0.020)	(0.021)
# of commun	0.103***	0.091***	0.089***		0.073***	0.061***	0.061***	0.074***
	(0.034)	(0.034)	(0.035)		(0.019)	(0.020)	(0.020)	(0.019)
# of composit	-0.046***	-0.036***	-0.035***	-0.051***	-0.124***	-0.098***	-0.099***	-0.127***
	(0.006)	(0.007)	(0.007)	(0.007)	(0.034)	(0.033)	(0.033)	(0.034)
# of disord	0.170***	0.162***	0.162***		0.089***	0.080***	0.081***	0.094***
	(0.058)	(0.057)	(0.057)		(0.022)	(0.021)	(0.021)	(0.023)
# of earli	0.090***	0.078**	0.077**		0.066***	0.057***	0.056***	
	(0.034)	(0.033)	(0.033)		(0.019)	(0.019)	(0.019)	
# of genom	0.076**	0.065**	0.065**		0.045**	0.037**	0.036**	0.044**
	(0.033)	(0.033)	(0.033)		(0.017)	(0.017)	(0.017)	(0.017)
# of global		0.070**	0.071**			0.049***	0.049***	
		(0.033)	(0.033)			(0.019)	(0.019)	
# of graphen		0.076**	0.075**			0.087***	0.087***	
		(0.031)	(0.031)			(0.022)	(0.022)	
# of meta_analysi	0.098**	0.120***	0.120***		0.061***	0.080***	0.061***	
	(0.039)	(0.040)	(0.042)		(0.019)	(0.027)	(0.022)	
# of model	-0.043***	-0.036***	-0.037***	-0.044***	-0.074***	-0.065***	-0.064***	-0.075***
	(0.006)	(0.007)	(0.007)	(0.007)	(0.016)	(0.016)	(0.016)	(0.016)
# of network_ETX		-0.063***	-0.063***	-0.061***	-0.183**	-0.181**	-0.181**	
		(0.011)	(0.011)	(0.010)	(0.081)	(0.080)	(0.080)	
# of neuron	0.082**	0.070*	0.069*		0.056***	0.041*	0.039*	
	(0.038)	(0.038)	(0.038)		(0.020)	(0.021)	(0.021)	
# of reveal	0.110***	0.102***	0.102***		0.070***	0.065***	0.059***	0.070***
	(0.029)	(0.029)	(0.029)		(0.016)	(0.015)	(0.015)	(0.016)
# of risk	0.080***	0.068**	0.067**		0.054***	0.034**	0.034**	0.056***
	(0.028)	(0.027)	(0.028)		(0.015)	(0.015)	(0.015)	(0.015)
# of stem	0.092***	0.081***	0.084***		0.060***	0.051***	0.050***	0.059***
	(0.030)	(0.030)	(0.030)		(0.019)	(0.019)	(0.019)	(0.019)
# of STX_structur	0.118***	0.113***	0.112***		0.079***	0.081***	0.079***	
	(0.035)	(0.035)	(0.034)		(0.019)	(0.018)	(0.018)	
# of trial	0.105**	0.074	0.075		0.074***	0.048	0.048	0.078***
	(0.050)	(0.054)	(0.054)		(0.027)	(0.032)	(0.032)	(0.027)
# of vitro		-0.052***	-0.054***	-0.069***		-0.182***	-0.180***	
		(0.009)	(0.009)	(0.008)		(0.058)	(0.058)	
Total # of words	-0.006***	-0.005***	-0.004***	-0.004***	-0.006***	-0.004***	-0.004***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Citations $year_p$	0.018***	0.017***	0.017***	0.018***	0.014***	0.013***	0.013***	0.014***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)
N	20755	20755	20755	20755	20755	20755	20755	20755
Out-of-sample R2	0.095	0.101	0.100	0.077				
R-squared	0.088	0.114	0.115	0.074				
Out-of-sample accuracy	63.47	61.88	60.10	71.08	66.84	63.76	67.37	73.49
Overall accuracy	70.50	63.92	62.25	65.26	76.37	69.36	62.53	71.68
Best cutoff	0.091	0.094	0.091	0.105	0.088	0.080	0.090	0.097
Matthew corr. coeff.	0.373	0.416	0.408	0.378	0.386	0.430	0.449	0.414

Note: Estimates from OLS (columns 1-4) or Logit regression (columns 5-8). The dependent variable *Media coverage* is an indicator for whether a paper attracted online coverage at publication. In column (1)-(4) predictors are selected based on *lasso* while column (5)-(8) predictors are selected based on *lassologit*. Matthews Correlation Coefficient (MCC) is the preferred single metric in the area machine learning with binary classification, especially for imbalanced data. The metric ranges $[-1, 1]$ and takes on the value of zero if the prediction is the same as a random guess. Best cutoff pproximates the optimal positive cutoff using MCC. Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.24: Likelihood of retraction and media coverage (Logit)

	Retraction									
	Logit		EBIC		AICC		CV		Rigourous	
Media coverage	0.009**	0.011***	0.014***	0.016***	0.015***	0.017***	0.015***	0.017***	0.014***	0.016***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Predicted media, Prob.			-0.089***	-0.083**	-0.054***	-0.052***	-0.058***	-0.055***	-0.095**	-0.089**
			(0.034)	(0.034)	(0.017)	(0.017)	(0.018)	(0.019)	(0.038)	(0.039)
Publication year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Subject FE	N	Y	N	Y	N	Y	N	Y	N	Y
N	20755	20393	20755	20393	20755	20393	20755	20393	20755	20393

Note: Estimates from Logit equation. The dependent variable *Retraction* is an indicator for whether a paper was retracted. *Media coverage* is an indicator for whether a paper attracted online coverage at publication. *Predicted media* is media coverage as predicted from the respective *lassologit* procedures. Bootstrap standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.25: Likelihood of retraction and media coverage (Residuals)

	Retraction							
	EBIC		AICC		CV		Rigourous	
Predicted media, Resid.	0.013***	0.022***	0.015***	0.024***	0.015***	0.024***	0.011**	0.021***
	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)
R-squared	0.000	0.002	0.000	0.002	0.000	0.002	0.000	0.001
Publication year FE	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	N	Y	N	Y	N	Y	N	Y
N	20755	20755	20755	20755	20755	20755	20755	20755
N clusters	1008	1008	1008	1008	1008	1008	1008	1008

Note: Estimates from OLS regression. The dependent variable *Retraction* is an indicator for whether a paper was retracted. *Media coverage*, *Resid.* is the residual of the predicted coverage according to different *lassologit* procedures. Bootstrap standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.26: Likelihood of retraction and media coverage (tf-idf)

	Retraction							
	EBIC		AICC		CV		Rigourous	
Media coverage	0.011**	0.020***	0.013***	0.020***	0.013***	0.020***	0.011**	0.020***
	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)
Predicted media	-0.040	-0.032	-0.039**	-0.035*	-0.039**	-0.035*	-0.072	-0.061
	(0.027)	(0.030)	(0.016)	(0.020)	(0.017)	(0.020)	(0.044)	(0.055)
R-squared	0.000	0.001	0.000	0.002	0.000	0.001	0.000	0.001
Publication year FE	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	N	Y	N	Y	N	Y	N	Y
N	20755	20755	20755	20755	20755	20755	20755	20755
N clusters	1008	1008	1008	1008	1008	1008	1008	1008

Note: Estimates from OLS regression. The dependent variable *Retraction* is an indicator for whether a paper was retracted. *Media coverage* is an indicator for whether a paper attracted online coverage at publication. *Predicted media* is media coverage as predicted from the respective *lasso* procedures. Bootstrap standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.27: Likelihood of retraction and media coverage (selection within subjects, publication years and excluding retractions)

	Retraction							
	EBIC		AICC		CV		Rigourous	
Media coverage	0.016***	0.021***	0.018***	0.022***	0.018***	0.022***	0.014***	0.020***
	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)
Predicted media	-0.056***	-0.071**	-0.061***	-0.074***	-0.062***	-0.074***	-0.047***	-0.084**
	(0.018)	(0.032)	(0.016)	(0.025)	(0.016)	(0.025)	(0.015)	(0.041)
R-squared	0.001	0.001	0.001	0.001	0.001	0.002	0.000	0.001
Publication year FE	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	N	Y	N	Y	N	Y	N	Y
N	20393	20393	20393	20393	20393	20393	20393	20393
N clusters	988	988	988	988	988	988	988	988

Note: Estimates from OLS regression. The dependent variable *Retraction* is an indicator for whether a paper was retracted. *Media coverage* is an indicator for whether a paper attracted online coverage at publication. *Predicted media* is media coverage as predicted from the respective *lasso* procedures. Bootstrap standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.28: Months to retraction and Journal-year average visibility

	(1) Time to Retract	(2) Time to Retract	(3) Time to Retract	(4) Time to Retract	(5) Time to Retract	(6) Time to Retract
Altscore	-3.579***					
	(0.565)					
Sh. Blog		-3.376				
		(2.413)				
Blog count		-1.662				
		(1.850)				
Sh. news			-6.287***			
			(1.937)			
News count			2.105			
			(1.384)			
Sh. Tweets				-3.127***		
				(1.117)		
Tweets count				-1.688***		
				(0.645)		
Sh. early blog					-3.331*	
					(1.995)	
Early blog count					-0.423	
					(1.675)	
Sh. early news						-5.453***
						(1.666)
Early news count						1.931
						(1.262)
Observations	967	961	962	968	962	967
R-squared	0.455	0.468	0.468	0.460	0.459	0.455
Publication year FE	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Mean	24.21	24.21	24.21	24.21	24.21	24.21

Note: The dependent variable is the time intercurring between an article publication and its retraction, expressed in months. Covariates represents different measures of journal visibility, measured as the average of non-retracted papers that appear in the same yournal and year of the retracted one. All covariates are standardized and outliers trimmed. Controls include the number of non-retracted articles within same journal and year of the treated, the average Euclidean distance of those from the retracted paper, and the level of (non-self) cumulative citations of the retracted paper before retraction. Publication year fixed effects are included. Journal clustered standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table A.29: Loss in citation and Journal-year average visibility

	(1) DID (-1,1)	(2) DID (-1,1)	(3) DID (-2,2)	(4) DID (-2,2)	(5) DID (-4,4)	(6) DID (-4,4)	(7) DID (-4,6)	(8) DID (-4,6)
Altscore	-1.362***	-1.111***	-1.699***	-1.327***	-1.830***	-1.399***	-1.877***	-1.439***
	(0.187)	(0.221)	(0.201)	(0.202)	(0.224)	(0.221)	(0.236)	(0.232)
Observations	840	840	840	840	840	840	840	840
R-squared	0.120	0.171	0.165	0.230	0.158	0.236	0.152	0.231
Sh. Blog	-1.698**	-1.643**	-1.598**	-1.440**	-1.746**	-1.548**	-1.658**	-1.449**
	(0.738)	(0.686)	(0.740)	(0.677)	(0.767)	(0.687)	(0.764)	(0.686)
Blog count	0.078	0.268	-0.327	-0.088	-0.334	-0.066	-0.467	-0.198
	(0.651)	(0.627)	(0.601)	(0.562)	(0.614)	(0.562)	(0.615)	(0.567)
Observations	831	831	831	831	831	831	831	831
R-squared	0.139	0.188	0.182	0.241	0.175	0.246	0.167	0.239
Sh. News	-1.611***	-1.365***	-1.795***	-1.325**	-1.952***	-1.384**	-1.995***	-1.406**
	(0.519)	(0.501)	(0.582)	(0.559)	(0.633)	(0.599)	(0.650)	(0.611)
News count	0.129	0.107	-0.018	-0.128	-0.051	-0.194	-0.059	-0.214
	(0.446)	(0.421)	(0.480)	(0.461)	(0.515)	(0.489)	(0.525)	(0.495)
Observations	835	835	835	835	835	835	835	835
R-squared	0.125	0.173	0.169	0.228	0.167	0.238	0.161	0.233
Sh. Tweets	-0.650**	-0.540**	-0.764***	-0.558**	-0.856***	-0.601**	-0.894***	-0.627**
	(0.268)	(0.262)	(0.275)	(0.257)	(0.295)	(0.270)	(0.298)	(0.273)
Tweets count	-0.965***	-0.788***	-1.185***	-0.944***	-1.233***	-0.966***	-1.254***	-0.986***
	(0.222)	(0.238)	(0.251)	(0.235)	(0.286)	(0.261)	(0.302)	(0.275)
Observations	842	842	842	842	842	842	842	842
R-squared	0.123	0.173	0.164	0.229	0.156	0.234	0.149	0.228
Pub. year FE	Y	Y	Y	Y	Y	Y	Y	Y
Age at retraction FE	N	Y	N	Y	N	Y	N	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Mean	-3.149	-3.149	-3.707	-3.707	-4.274	-4.274	-4.158	-4.158

Note: The dependent variable is the individual loss in citations obtained comparing each retracted paper to its selected controls for different pre- and post- time windows. Covariates represents different measures of journal visibility, measured as the average of non-retracted papers that appear in the same yournal and year of the retracted one. All covariates are standardized and outliers trimmed. Controls include the number of non-retracted articles within same journal and year of the treated, the average Euclidean distance of those from the retracted paper, and the level of (non-self) cumulative citations of the retracted paper before retraction. Publication year fixed effects are included. Fixed effects for age of the article at retraction are added in even comuns. Journal clustered standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table A.30: Citation statements and early mentions

	(1) Cit. statements	(2) Cit. statements	(3) Mentioning	(4) Mentioning	(5) Contrasting	(6) Contrasting	(7) Supporting	(8) Supporting
Post	0.033	0.038	0.034	0.040	-0.108	-0.132	0.035	0.031
	(0.038)	(0.040)	(0.038)	(0.040)	(0.236)	(0.263)	(0.100)	(0.109)
Post * Treatment	-1.215***	-1.165***	-1.231***	-1.178***	-1.147***	-1.184***	-1.055***	-1.061***
	(0.096)	(0.093)	(0.095)	(0.093)	(0.339)	(0.348)	(0.197)	(0.209)
Post * Early mentions	0.245***	0.171***	0.238***	0.167**	0.738**	0.745**	0.417***	0.306*
	(0.063)	(0.066)	(0.065)	(0.068)	(0.317)	(0.347)	(0.158)	(0.170)
Post * Treatment * Early mentions	-0.411**	-0.486***	-0.396**	-0.478***	-0.595	-0.416	-0.456	-0.455
	(0.164)	(0.170)	(0.167)	(0.173)	(0.684)	(0.730)	(0.414)	(0.440)
Self cit. excluded	N	Y	N	Y	N	Y	N	Y
Article FE	Y	Y	Y	Y	Y	Y	Y	Y
Age FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Pseudo R2	0.713	0.717	0.711	0.715	0.138	0.138	0.267	0.254
N	14594	14158	14536	14089	1701	1421	6370	5761
N clusters	957	948	956	946	201	170	567	521
N full	16711	16711	16711	16711	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations statements received by each paper in a particular year, even columns exclude self citations. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Early mentions is an indicator for papers with at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citation statements). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.31: Citation statements and attention score

	(1) Cit. statements	(2) Cit. statements	(3) Mentioning	(4) Mentioning	(5) Contrasting	(6) Contrasting	(7) Supporting	(8) Supporting
Post	-0.023	0.024	-0.022	0.024	0.050	0.352	0.045	0.115
	(0.052)	(0.055)	(0.053)	(0.056)	(0.344)	(0.360)	(0.131)	(0.142)
Post * Treatment	-1.001***	-0.941***	-0.992***	-0.923***	-1.696***	-1.972***	-1.218***	-1.345***
	(0.121)	(0.119)	(0.121)	(0.119)	(0.634)	(0.681)	(0.351)	(0.373)
Post * Altscore >p50	0.173***	0.090	0.170***	0.091	0.048	-0.367	0.149	0.009
	(0.060)	(0.063)	(0.060)	(0.064)	(0.350)	(0.368)	(0.151)	(0.159)
Post * Treatment * Altscore >p50	-0.415***	-0.441***	-0.441***	-0.477***	0.480	0.808	0.030	0.190
	(0.155)	(0.153)	(0.155)	(0.153)	(0.704)	(0.756)	(0.403)	(0.428)
Self cit. excluded	N	Y	N	Y	N	Y	N	Y
Article FE	Y	Y	Y	Y	Y	Y	Y	Y
Age FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Pseudo R2	0.713	0.717	0.711	0.715	0.136	0.136	0.266	0.253
N	14594	14158	14536	14089	1701	1421	6370	5761
N clusters	957	948	956	946	201	170	567	521
N full	16711	16711	16711	16711	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations statements received by each paper in a particular year, even columns exclude self citations. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citation statements). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.32: Citation statements and attention score extremes

	(1) Cit. statements	(2) Cit. statements	(3) Mentioning	(4) Mentioning	(5) Contrasting	(6) Contrasting	(7) Supporting	(8) Supporting
Post	0.043	0.074	0.046	0.074	0.262	0.567	0.070	0.177
	(0.056)	(0.059)	(0.056)	(0.060)	(0.373)	(0.355)	(0.146)	(0.157)
Post * Treatment	-1.103***	-1.002***	-1.097***	-0.984***	-1.940**	-2.203***	-1.226***	-1.385***
	(0.156)	(0.153)	(0.155)	(0.150)	(0.769)	(0.820)	(0.467)	(0.484)
Post * Treatment * Altscore 3rd qntl	0.428**	0.287	0.438**	0.286	1.369	1.523	0.203	0.315
	(0.211)	(0.217)	(0.212)	(0.218)	(1.254)	(1.342)	(0.654)	(0.684)
Post * Treatment * Altscore 4th qntl	-0.047	-0.105	-0.087	-0.160	0.249	0.468	0.476	0.743
	(0.205)	(0.205)	(0.204)	(0.204)	(1.288)	(1.338)	(0.546)	(0.568)
Post * Treatment * Altscore 5th qntl	-0.456**	-0.510***	-0.472**	-0.538***	0.513	0.899	-0.224	-0.024
	(0.197)	(0.193)	(0.196)	(0.191)	(0.849)	(0.907)	(0.534)	(0.554)
Self cit. excluded	N	Y	N	Y	N	Y	N	Y
Age FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Pseudo R2	0.714	0.717	0.712	0.716	0.142	0.140	0.268	0.255
N	14594	14158	14536	14089	1701	1421	6370	5761
N clusters	957	948	956	946	201	170	567	521
N full	16711	16711	16711	16711	16711	16711	16711	16711

Note: Estimates derive from pseudo Poisson specifications. The dependent variable is the total number of citations statements received by each paper in a particular year, even columns exclude self citations. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Altscore is an aggregate measure of weighted online mentions. All models incorporate article fixed effects, a full suite of calendar-year effects and article age indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (i.e. x% loss in yearly citation statements). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.

Table A.33: Impact on authors' careers (split samples)

	Media								No Media							
	(1) All	(2) First author	(3) Mid author	(4) Last author	(5) H-index >p50	(6) H-index <=p50	(7) Severe misconduct	(8) Non-Severe misconduct	(9) All	(10) First author	(11) Mid author	(12) Last author	(13) H-index >p50	(14) H-index <=p50	(15) Severe misconduct	(16) Non-Severe misconduct
Panel A: N articles																
Post * Treatment	-0.212**	-0.612***	0.102	-0.187*	-0.246***	-0.029	-0.353***	-0.064	-0.092***	-0.090	-0.072	-0.098**	-0.114***	-0.069	-0.222***	-0.017
	(0.085)	(0.159)	(0.116)	(0.105)	(0.091)	(0.133)	(0.113)	(0.106)	(0.035)	(0.059)	(0.055)	(0.046)	(0.041)	(0.054)	(0.051)	(0.046)
Pseudo R2	0.583	0.479	0.617	0.542	0.543	0.301	0.590	0.578	0.597	0.498	0.573	0.597	0.570	0.419	0.594	0.600
N	6055	2082	1867	2102	3788	2266	3459	2595	48672	16549	14776	17344	24968	23699	19931	28740
N clusters	105	105	99	102	100	101	55	50	783	779	761	778	757	766	315	468
N authors	760	266	242	252	451	309	413	347	5905	2061	1823	2021	2874	3031	2382	3523
Panel B: N articles with grant																
Post * Treatment	-0.207**	-0.543***	0.012	-0.165	-0.262***	0.165	-0.332***	-0.019	-0.060	-0.020	-0.022	-0.089	-0.103**	0.008	-0.160**	-0.015
	(0.091)	(0.186)	(0.134)	(0.122)	(0.098)	(0.169)	(0.128)	(0.118)	(0.045)	(0.075)	(0.068)	(0.054)	(0.049)	(0.071)	(0.069)	(0.054)
Pseudo R2	0.552	0.467	0.586	0.518	0.516	0.297	0.558	0.550	0.563	0.481	0.533	0.573	0.546	0.394	0.568	0.562
N	5893	2008	1795	2082	3765	2125	3362	2530	44345	14630	13321	16388	24364	19980	18389	25955
N clusters	104	103	99	101	100	98	55	49	780	766	741	766	754	744	315	465
N authors	735	255	231	249	448	287	399	336	5334	1811	1626	1897	2803	2531	2179	3155
Panel C: Avg. n coauthors (x article)																
Post * Treatment	-0.090	-0.233	0.138	-0.092	-0.119	-0.034	-0.204*	0.046	-0.090***	-0.113**	-0.082*	-0.077**	-0.089***	-0.082*	-0.167***	-0.029
	(0.080)	(0.152)	(0.143)	(0.082)	(0.083)	(0.178)	(0.119)	(0.090)	(0.028)	(0.056)	(0.046)	(0.034)	(0.033)	(0.044)	(0.040)	(0.038)
Pseudo R2	0.384	0.392	0.384	0.395	0.419	0.281	0.392	0.383	0.378	0.365	0.383	0.378	0.390	0.313	0.360	0.391
N	6055	2082	1867	2102	3788	2266	3459	2595	48672	16549	14776	17344	24968	23699	19931	28740
N clusters	105	105	99	102	100	101	55	50	783	779	761	778	757	766	315	468
N authors	760	266	242	252	451	309	413	347	5905	2061	1823	2021	2874	3031	2382	3523
Author FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Career lenght FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Calendar Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N full	6100	2111	1877	2112	3813	2287	3496	2604	49016	16681	14921	17414	25009	24007	20118	28898

Note: Estimates derive from pseudo Poisson specifications. The dependent variables are: N. published articles x author x year; N. published articles with grant support x author x year; or Avg. n collaborators across all author's publications x year. Treatment is an indicator for retracted papers. Post is an indicator for all years strictly after the year of retraction. Media is an indicator for cases where the original publication (either retracted or control papers) had at least one mention (in newspapers and/or blogs) within two weeks from publication. All models incorporate author fixed effects, a full suite of calendar-year effects and carreer lenght indicator variables. Using the following transformation $(1 - exp[\beta]) * 100 = x$ coefficients can be interpreted as elasticities (e.g. x% loss in publication rate). Standard errors in parentheses, clustered around retraction cases. *p < 0.10, **p < 0.05, ***p < 0.01.