The Unfolding SPIRIT of Science: Evidence from P-Hacking in Academic Clinical Trials

Jorge Guzman
Columbia University
NBER

Zubin Jelveh
University of Maryland

Bruce Kogut
Columbia University

Abstract

The norms of transparency and rigor are central to the institution of open science, incentivizing scientists to compete to produce knowledge and yet adhere to Mertonian norms (Merton, 1973). Nowhere are these norms more critical than in clinical trials, where the development of new medical interventions directly impacts human lives. In this paper, we first present evidence that professional norms in science have not prevented questionable research practices in academic clinical trials. Analyzing 10,072 primary p-values, we find little evidence of p-hacking among industry trials, a domain where private gains serve as the overarching incentive. We cannot, however, reject the hypothesis that p-hacking is prevalent in academic trials, which comprise 22% of the sample. We exploit the impact of an institutional intervention aimed at promoting Mertonian norms labeled SPIRIT 2013. Spirit is a set of guidelines endorsed by leading journals aimed at improving the creation of clinical trial study protocols. We find trials initiated after SPIRIT 2013 exhibited lower levels of p-hacking. Linguistic analysis shows that trial registrations more closely adhering to SPIRIT guidelines had a lower incidence of p-hacking. Additionally, in a dataset of over 300 thousand medical publications, only those relied upon clinical trials demonstrated a reduction in p-hacking following the introduction of SPIRIT 2013. Finally, while SPIRIT 2013 appears to reduce the incidence of p-hacking, it does not eliminate it entirely. Our findings underscore the relevance of institutional interventions to improving the quality of scientific findings, particularly in the critical domain of medical research.

keywords: p-hacking, clinical trials, career concerns, universities, institutions, innovation.

1 Introduction

Clinical trials are federally regulated processes to assess causally the safety and efficacy of new health interventions involving human subjects. Their results are the primary

evidence used by the U.S. Food and Drug Administration (FDA) to approve new treatments and by physicians to prescribe these treatments to patients. The integrity of statistical evidence from clinical trials depends on the quality of subject recruitment, sample size, randomization of treatment assignment, and data analysis. For decades, some scientists have expressed their ongoing concerns over the potential for the selective reporting of evidence in clinical trials (e.g., Rotonda, 1990; Dickersin et al., 1987; Dickersin, 1997).

In the broader natural and social sciences, forensic studies on statistical procedures have revealed that research results reported as valid are often false positives. One major factor contributing to false positives stem from the cases when researchers explore numerous analytical approaches or subsets of data until they find a statistically significant result without properly accounting for multiple comparisons (Simmons et al., 2011; Simonsohn et al., 2014). Another issue arises when researchers consciously or unconsciously select the most favorable results among many possible analyses (Chan et al., 2004). Concerns over these practices, commonly referred to as p-hacking, have existed for nearly 200 years (Babbage, 1830).

In the past decade, studies have demonstrated that P-hacking is prevalent in many scientific fields (Simonsohn et al., 2014; Head et al., 2015; Brodeur et al., 2016; Elliott et al., 2022b). Evidence on p-hacking in clinical trials is less abundant, though prior analysis of industry-sponsored clinical trials found evidence of p-hacking limited to small company sponsors, implying contextual variables influence the willingness to p-hack (Adda et al., 2020).

In this paper, we focus on clinical trials run by academics whose incentive structures and institutional rivalry differ from those in industry settings. Academic researchers often face strong pressures to publish in high-impact journals, secure funding, and achieve tenure that together leads to perverse incentives to engage in research practices like p-hacking.

We suggest three different channels through which the introduction of SPIRIT can

reduce the incidence of p-hacking by academics. These channels align the behavior of scientists with Mertonian norms for the operation of open science (Merton, 1973). First, the introduction of a standard for reporting results compliant with protocols increases public scrutiny and the verifiability of research, enabling a more effective organized skepticism and hence more clarity on the evaluation of truth claims made by researchers. With respect to reduced p-hacking, verifiability also increases the risk of retraction for bad or careless science, which can have significant career cost for individual scientists (Azoulay et al., 2015, 2017). Second, at a more mechanical level, study protocols reduce the degrees of freedom for researchers to change the analysis or select the results that best fit a more favorable hypothesis (Gelman and Loken, 2014; Simmons et al., 2011). Third, to the extent that following the SPIRIT guidelines is required by institutions that exert influence over a researcher's career path (e.g. journals, funders, etc.), SPIRIT may lead to the exit of researchers who are more prone to p-hacking.

Our first result confirms the results of previous studies that p-hacking in clinical trials is evident for those run by academics and not by industry researcher using econometric specifications designed to test irregularities in the non-monotonicity and discontinuities found in the statistical distributions, especially in the range around the $p = 0.05$ threshold. The statistical findings indicate that p-hacking is found for primary outcomes and not for the secondary ones, which is consistent with the pressures of publishing leading to statistical manipulation of those estimates that are scientifically important. Consistent with previous work, we do not observe evidence of p-hacking in clinical trials sponsored by industry. These results are also consistent with Balasubramanian et al. (2022), who use a specific sub-sample of clinical trials (Phase 2 oncology) to document similar differences between non-profit and for-profit organizational incentives.

Given the evidence of p-hacking in academiC-run clinical trials, we next assess the impact of an institutional intervention aimed at promoting Mertonian norms proposed in a set of guidelines SPIRIT first proposed in 2013, an acronym for "Standard Protocol Items:

Recommendations for Interventional Trials". SPIRIT 2013, as we will call the guidelines, provides a 33-item checklist for standardizing the elements of a study protocol consisting of a detailed plan outlining the design, objectives, methodology, statistical plan, and organization of a clinical trial. The purpose of a study protocol is to ensure that the trial is conducted in a consistent, rigorous, and transparent manner and that the results are scientifically valid and reliable. Crucially for our purposes, a study protocol should be completed prior to the start of a clinical trial.

Endorsed by leading journals, the SPIRIT guidelines promote the standardization of statistical analysis plans (SAPs), emphasizing the pre-registering analysis plans, defining clear outcome measures and their usage, and clarifying the sample used in the study. By doing so, SPIRIT aims to reduce the possibility of altering the study during the analysis phase. Institutional responses similar to SPIRIT have become increasingly common across social and medical sciences (Olken, 2015; Ofosu and Posner, 2021), and professional guidelines have been shown to be effective in improving professional practice (Grimshaw and Russell, 1993; Abaluck et al., 2020; Zhang, 2023). Although evidence on the effectiveness of pre-analysis plans is less abundant, recent unpublished work by Brodeur et al. (2022a) finds that pre-registration of analysis plans reduces p-hacking in a sample of randomized control trials published in economics journals.

We find that clinical trials that began SPIRIT 2013 exhibited lower levels of p-hacking. Finally, we conduct an innovative exploration of the impact of SPIRIT by drawing the toolkit of natural language processing (NLP) to measure the similarity in the language used in describing the design of clinical trials relative and the language of the SPIRIT 2013 proposal. We find that linguistic similarity between the clinical trial and SPIRIT protocol is associated with lower p-hacking. To increase the validity of our measure, we replicate this analysis on a large sample of published academic papers in medical journals that include both clinical and non-clinical trials. Similarity to SPIRIT for this sample also predicts lower p-hacking only for studies using clinical trials published after the release of SPIRIT.

Additional analyses show the robustness of our results to differences in specification.

2 Norms, Institutions, and P-Hacking

The development of formal and normative institutions on setting research standards, rewards, and penalties grew out of the historical recognition of the effects of bad research practices on the quality of scientific outcomes. This argument lies at the core of our study, that is, improvement in the doing of science occurs often as a response to discovery of poor research practices. A few historical examples will help to reinforce the evidence to this perspective.

The history of modern science begins with the emergence of scientists and their public discoveries, and the development of secular institutions and norms prescribing the conduct of scientific research. As a scholar of this emergence, the sociologist Robert K. Merton described modern science as "legitimized in terms of institutional values" of universalism, open science, disinterestedness, and organized skepticism that functionally provided the possibility for the inspection and replication of scientific results (Merton, 1973). By the 17th century, secular institutions such as royal societies, academies of science, and universities came to set norms, and resolve quarrels over the assignment of credit for original discoveries by some of the greatest scientists of that time. As suggested by these quarrels, the new norms focused on recognizing, and thereby incentivizing, scientists for their contributions to the public good.

Yet, modern science has another side concerning the deviations from these norms, often motivated by competitive incentives. In 1830, Charles Babbage despaired of the "impositions that have been practiced in science", listing four types of fiddling in science: hoaxing, or making up the results; forging, which is copying another's manuscript; trimming, or lopping off "little bits here and there which differ most in excess from the mean"; and

cooking, or "to make multitudes of observations, and out of these to select those only which agree" (Babbage, 1830; Merton, 1957). That is, with the growing importance of science came the threat of a degradation of the doing of science because of the ever-growing competition among scientists and increasingly unregulated commercial interests.

In response to this danger, regulatory authorities and institutions developed. For example, in the aftermath of the regulatory and institutional responses that arose from a history of drug scandals and shoddy science in the U.S. case,[1] the U.S. government created, in 1887, what became the predecessor to the National Institutes of Health (NIH). Following the death of 13 children in Saint Louis, the U.S. Congress later passed the 1902 Biologics Control Act that empowered the federal government to oversee the production processes of biological products and also their inspection. In 1906, the Food and Drug Administration (FDA), was founded with the responsibility to permit only new drugs to go to the market through its regulatory powers by requiring the use of clinical trials.

The interwar years in the United States saw the rise of the American Medical Association's (AMA) efforts to self-regulate the industry by refusing to advertise drugs that failed their standards in the medical association's journal. However, self-regulation was complicated by the increasing cooperation between large pharmaceutical companies and universities (such as Harvard and the University of Pennsylvania) to commercialize academic drug discoveries, raising concerns of corrupting science (Urquiola, 2020). The resulting outcry contributed to the passage of the 1938 Food, Drug, and Cosmetic Act that required the FDA to regulate and to review both pre-clinical and clinical test results for new drugs (Rasmussen, 2005). It was only in 1962 that the FDA assumed the responsibility to regulate mandatory testing by "adequate and well-controlled" scientific trials (White Junod, 2008).

[1] A listing of regulatory responses to specific illnesses, such as pertussis, polio, measles, AIDs, is provided through the website of the FDA: https://www.fda.gov/about-fda/histories-product-regulation/science-and-regulation-biological-products.

Meanwhile, the statistical foundations to clinical trials made great progress due to the research by British statistician Ronald Fischer and to the formalizing of clinical trial protocols by A.B. Hill in the 1950s (Hill, 1951). Starting already in the 1920s and 1930s, academics and universities were engaged by pharmaceutical companies to develop new drugs and measure their morbidity risk. Industry and universities took note of the commercial success of better science and sought to create relationships with academic faculty. By the 1920s, Princeton took the lead in offering important elements of tenure to attract and retain faculty. Other private universities followed, as did many state universities and colleges, such that "by 1950, the full package is often commonplace" (Urquiola, 2020, p. 142). Metrics for the success of these investments followed. By the 1960s, the Institute of Scientific Ideas (ISI) created the citation index to trace publications, and their citation counts, as new measures for scientific output and the quality of science (Cole and Cole, 1974).

In addition to tenure, well-funded laboratories at top institutions were also an important carrot for attracting top scientists. American universities turned to federal funding that increased rapidly through the post-World War II era. In 2022 constant dollars, the National Science Foundation and NIH grew, between 1976 and 2022, from $2.662 billion to $7.59 billion and $9.00 billion to $43.84 billion, respectively (AAAS, 2023). Moreover, funding from, and collaboration with, industry became an even more important resource for academic and industry scientists to pay for their expensive research. By 2015, private R&D spending on medical research was five times greater than federal spending (drugcostsfacts.org, 2016).

Academic scientific research evolved into an intensely competitive business of securing funding for labs and personnel, incentives for publication and citations, earning faculty tenure with the prospect of attracting subsequent funding for their research. At the heart of the system are the journals who play institutional roles in the decisions to accept and reject research that are central to the success and failure of academic research careers. Concerned over the quality of scientific research, journals in some fields, such as pharmaceuticals,

also often seek to improve the production of science by publicly supporting standards and protocols of scientific research practices, such as reflected in the public support by medical journals for the 2013 SPIRIT codification for the design of random controlled trials and the public analysis for data analysis.[2]

Thanks to innovative improvements in the forensic study of reported research results, a robust tool kit has recently emerged for the statistical analysis of errant scientific research, of which the detection of p-hacking of statistical significance results is highly prevalent. To provide a calibration of whether p-hacking is common in academic medical trials research, we pose the forensic question of whether academic incentives to publish or industry incentives to pass the test of the market leads to more or less p-hacking. Having established that p-hacking is more prevalent among academics, we examine if new research guidelines deterred the use of p-hacking in clinical trial reporting. For this purpose, we analyze whether the introduction of SPIRIT 2013 guidelines decreased the incidence p-hacking in clinical trials.

3 Data

We obtained data on all clinical trials available on ClinicalTrials.gov as of November 2021. ClinicalTrials.gov is a platform launched by the National Library of Medicine, an institute of N.I.H., in 2000 to keep a centralized repository of clinical trials. Since 2007, an FDA amendment has required that all clinical trials of drugs, biologics, and devices regulated by the FDA be registered on this website. Various journals and funders also require trials to be registered.

We limit our sample of trials to the 15,181 interventional clinical trials on ClinicalTrials.gov that have reported results with p-values, comprising 178,207 results. These include

[2]Listing of journal signatories can be found at https://spirit-statement.org/about-spirit/spirit-endorsement/.

trials of all phases and types of interventions. Interventions that advance through the phases of the therapy pipeline are each reported as independent clinical trials. We include all trials, regardless of FDA approval or other market outcomes.

Next, we only retain p-values that are reported with three or more digits of significance in order to focus directly on variation across the 0.05 significance threshold.[3] We drop clinical trials with values that are not numbers or erroneously outside the range [0,1] and keep only trials that started in 2020 or earlier. Finally, we keep only trials sponsored by academics or by industry. Medical schools and university hospitals are included in academic clinical trials while big pharmaceutical firms and small biotechnology startups are some of the types of firms in industry.

Of the 8,059 trials that fit these criteria, 2,339 are academic clinical trials.[4] We choose to focus on primary outcomes as the main piece of our analysis since they are the outcomes most relevantly evaluated to assess the effectiveness of new medical therapies. As the p-hacking tests we employ focus on the 0.05 significance boundary, we further limit our sample of p-values to those within the 0.001 and 0.15 interval. The final sample comprises 10,072 primary p-values. Among these, 22% are p-values from academic clinical trials and the remainder from industry. Summary statistics for these data are presented in Table A1 in the Supplement. For comparison, we also extract 40,097 secondary p-values from the original data.

The included academic clinical trials span the gamut of types of trials that occur in the medical field. 55% of clinical trials have a 'Not Applicable' phase, which is used to describe behavioral and device interventions. 12% are phase 2, and 25% are phase 3 or 4. The average sample size for the trials is 107 subjects, but this number is significantly

[3]We prefer to retain these digits approach over using de-rounding methods since recent studies show that de-rounding itself can create spurious bunching of p-values at the threshold (Kranz and Putz, 2022) See further discussion section S2 in the Supplement including evidence that observable trial characteristics are not correlated with significant digit reporting.

[4]See Supplement for statistics on the top 20 most prominent academic organizations.

right-skewed. 68% of the trials are in the United States. In terms of purpose, 63% of academic clinical trials are a treatment intervention for a condition, and an additional 23% prevention or supportive care. Trials also vary on the type of statistical test reported, including 2-sided tests (13%), t-tests (17%), and regression analysis (23%). Finally, trials represent both superiority tests (i.e., that the trials need to be significantly better than current practice), and non-superiority tests (that they only need to be as good as current practice). Table S3 in the Supplement reports a list of the 20 organizations with the most trials.

4 Methods: Estimating P-Hacking

Statistical analyses testing the validity of empirical results have taken several forms over time. For example, early precocious work by Rosenthal (1979) used simulations to explain how low power can lead to false conclusions. Dickersin et al. (1987) focused on selective reporting by comparing the distribution of outcomes in reported versus unreported clinical trials. Like Rosenthal,Ioannidis (2005) emphasized the lack of power, but this time, in the context of randomized controlled trials.

More recently, the main area of concern in evaluating the fidelity of clinical trials is the 'file-drawer problem'. This problem exists when "undisclosed flexibility in data collection and analysis allows presenting anything as significant" (Simonsohn et al., 2014). For example, researchers may vary their sample size, choose among multiple potential dependent variables, add or remove covariates, or report only a subset of the conditions analyzed. Together, this tendency to report only analyses that 'work' has come to be called *p-hacking* (Simonsohn et al., 2014).

In Figure 1, we report a simulation in which we estimate the p-value of a regression, and compare it to a p-hacked regression that drops the 10% of all data points with the highest

negative residual (i.e., the worst fit the model) and then chooses this p-hacked p-value only in the case where it is significant and the true model is not. Two differences are apparent: there is a discontinuity at $p = 0.05$, indicating that a series of p-values have moved from non-significant to significant, and there is additional mass across most p-values lower than 0.05. Because many of the p-hacked regressions do not have p-values close to 0.05, they create mass at lower values, shifting the whole distribution.

Building on the principles illuminated through this simulation, the literature has developed a menu of different ways to estimate the existence of this p-hacking. The majority of these tests rely on the intuition in Simmons et al. (2011) that, for true effects, the distribution of significant p-values should have right skewness—that is, there should be more p-values at $p < .01$ than for $0.01 < p < 0.02$, and so on. A valid test of p-hacking is to assess a monotonic decline in the distribution of p-values through the sample. The simplest test of this nature is a binomial test, typically done comparing whether the share of p-values between 0.04 and 0.045 is larger or smaller than those between 0.045 and 0.05. If it is larger, it is considered evidence for p-hacking.

Since this work, two additional approaches have been developed. Elliott et al. (2022b) show that distribution of p-values, when constructed from commonly used tests, should be continuous. Thus, a second approach relies on methods centered on the discontinuity at critical p-values. Tests developed by Cattaneo et al. (2020) (which build on McCrary (2008)) can be employed to examine the discontinuity at a threshold to test for manipulation, comparing the density to the left and right of it. For example, the discontinuity approach compares differences in p-values right above and below $p = 0.05$ and evaluates the difference in density between the two and the statistical significance of this difference.

The third method is a distribution-based approach that evaluates manipulation of p-values by the steepness of their distribution. In this approach, as shown in Elliott et al. (2022a), when researchers focus specifically on picking the most significant results among

many potential specifications (and not simply those above a threshold), p-value distributions that are too steep can be indicative of p-hacking. Cox and Shi (2023) subsequently introduced distribution tests evaluating both this assumption and general monotonicity in the distribution of p-values.

We utilize the first and second method, focusing on p-hacking that is detectable near the threshold of $p = .05$ rather than across the distribution of all p-values. For discontinuity tests, we apply the density discontinuity tests in Cattaneo et al. (2020), and to investigate whether the p-curve is non-increasing we employ binomial tests. We focus on evidence of p-hacking local to the $p = 0.05$ significance threshold for three reasons. First, we hypothesize that the scope for p-hacking is limited in randomized trials, which comprise close to 90% of the academic trials in our data, since, if randomization did not fail, then treatment assignment should be uncorrelated with other predictors, and techniques to lower p-values that rely on changing specifications should be less effective. This "robustness" to p-hacking of randomized trials predicts that p-hacked p-values are more likely to bunch close to significance thresholds.[5] Second, since most academic trials involve small sample sizes (the median sample size in our data is 68), we hypothesize that it is relatively more challenging to obtain p-values close to zero.[6] Third, the distribution-based tests outlined by Elliott et al. (2022b) require that p-values be derived from t-tests; however, t-tests are not abundant in our data. Of the p-values from academia-led trials, 20% of p-values are derived from t-tests. The figure is 4% for industry-led trials.

[5] See Elliott et al. (2022a) for a demonstration.

[6] In the clinical trials data, each outcome is associated with a unique sample size value that may vary across different outcomes in the same trial. Our definition of sample size at the trial level is an average of the sample sizes across the outcomes in the trial. For example, the median sample size of industry trials is 248 and the share of p-values from these trials that is less than 0.001 is 15.7%, while the share is 4.1% for academia-led trials.

5 Results

5.1 Estimates of P-Hacking in Academia and Industry

Panels A and B of Figure 3 provide our first result by comparing p-hacking in academia and industry. Panel A shows the plot of the distribution of primary outcome p-values for clinical trials conducted by academics. There is a clear discontinuity in the density of p-values at $p = 0.05$, rising before the threshold on the left side but not rising on the right side. The mean difference in the density is -9.96 and significant below the 0.01 level. Panel B is industry. As in prior work (Adda et al., 2020), we observe a mostly continuous distribution on both sides of the threshold for industry-sponsored trials. The difference in density values is much smaller and with the opposite sign, at 2.1, and not statistically significant. While there is no significant level of p-hacking in industry, the incidence of p-hacking in academia seems substantial.

Panel C in Figure **??** replicates the results using binomial tests. The average share of academia p-values in the $(0.4, 0.5)$ interval above 0.045 is 62%. This distribution is significantly different from the null hypothesis of 50% as well as from the distribution of industry values. The average share of industry values above 0.045 is 51%, lower than academia and not significantly different from the null. These estimates are consistent with the notion that academia p-hacks more than industry.

Panel D plots the results of our discontinuity tests for primary and secondary outcomes for academia- and industry-led trials. (See Table S4 in the Supplement for analogous binomial tests.) We posit that the differences in p-values observed for primary outcomes are driven by academics choosing analyses that are most likely to lead to a publication, whereas we are less likely to observe these differences for isecondary outcomes that are less pivotal in publication decisions. This is indeed what we find. We cannot reject the null of no p-hacking for secondary outcomes in academia-sponsored trials. We see that for

industry trials, the density difference is similar for primary and secondary outcomes.

5.2 The Impact of SPIRIT

We next assess, in Figure 3, the impact of the introduction of the *SPIRIT 2013* guidelines on the distribution of p-values.

Panel A uses data outside of our sample to ask whether protocols are mentioned in academic publications more broadly after the introduction of SPIRIT.[7] To do so, we count the number of publications per year reported in Google Scholar that contain the strings "study protocol" and "clinical trial" in the same article. Each annual count is then divided by the corresponding number of Google Scholar results that contain the string "clinical trial" in that year. This ratio is plotted in red in Panel A. We perform a complementary analysis using PubMed search results and plot this series in black in Panel A.[8] An increase in these ratios can be observed for clinical trial protocols after 2013. For the Google Scholar series, the trend before 2013 is mostly flat but rises sharply following 2013. For the PubMed series, the sharp rise begins slightly prior to 2013, suggesting the growing emphasis on protocols may have pre-dated the publication of the SPIRIT guidelines, potentially 'priming the pump.'

The centerpiece of our analysis provides an assessment of the differences in evidence of p-hacking before and after the introduction of SPIRIT. Panel B reports the density difference in the p-values of academic primary outcomes across three sub-samples. The first sub-sample consists of all the trials started before 2013, and hence before SPIRIT was introduced. We observe a significant discontinuity in the density with an average difference of -11.65 and significantly different from zero, indicating evidence for p-hacking. The second is all

[7] As we describe in the next section, trials were only required to submit protocol documents on clinitrials.gov after an FDA rule change which went into effect in 2017, precluding us from directly estimating how SPIRIT impacted the text of protocols on clinicaltrials.gov around the time of its introduction in 2013.

[8] The PubMed search is limited to the title and abstract fields.

trials started after 2013, and therefore after SPIRIT. The size of the discontinuity is smaller, with an average value of -4.68 and a standard deviation of 2.43. A z-test of equality of means rejects these two density distributions are the same at $p = .08$.

The third row of Panel B investigates the discontinuity in p-values only for those trials registered before SPIRIT but whose results are reported afterward. If we were simply observing secular changes in the practice of science, then trials reported after the introduction of SPIRIT should show a reduction in the discontinuity in p-values. However, there is no perceptible reduction in this discontinuity for trials reported after 2013 but started beforehand. Panel C replicates these results with a binomial test. The results of panels B and C indicate that the introduction of SPIRIT plausibly increased compliance to new protocol standards in clinical trial research post 2013 and subsequently reduced the incidence of p-hacking.

Lastly, Panel D shows the plot of the over-time evidence of p-hacking and is created by running the binomial test on samples over years. Specifically, we consider all p-values for academic primary outcomes occurring from year t to year $t + 3$. We use this longer window for generating a larger sample size. Using these data, we plot the distribution of the p-value *of this test* by year for different types of subsamples. These subsamples include our main sample, a sample that also includes two-digit p-values with a de-rounding procedure, a sample of only superiority tests, only randomized allocations, and a fourth sample excluding one-sided tests.

The results show a consistent pattern. Before 2010, when all trials considered in the sample were *before* SPIRIT, all tests reject the null of no p-hacking. Then, starting in 2010, there is a gradual increase in the plotted p-values of these tests in all samples up to 2013 when they then stabilize. After 2013, the binomial test does not show evidence of p-hacking. While existing work has focused mostly on differences in regulation across contexts (Balasubramanian et al., 2022; Adda et al., 2020), or the disclosure of primary

outcomes, our paper is the first to show evidentiary results of the impact of norms on changing academic incentives and institutions.

5.3 Differences Across Trials with Submitted Study Protocols

We next consider the study protocols more directly and how variation within them predicts p-hacking. To do so, we designed an analysis that relies upon creating a natural language processing (NLP) measure by obtaining the text of protocols submitted in each clinical trial for a subsample of trials. This sample of protocols was made possible by a 2017 FDA rule change required that trials with primary completion dates after January 18, 2017 upload study protocols when they submit results.[9] Compliance with this rule is high: 98% of trials with results and with primary completion dates after 2017 submitted the text of protocols to clinicaltrials.gov. We observe, however, that 65% of academia-run trials (and 69% of industry-run trials) submit protocols that are dated after the start date of the trial. On the other hand, the vast majority of submitted protocols are dated prior to the primary completion date (89% for academia and 84% for industry).[10] These statistics suggest room for improvement in ensuring that protocols are verifiably written prior to the start of a clinical trial. However, in light that most protocols (and associated statistical analysis plans) are dated prior to the primary completion date,there is still potential for protocols to dissuade p-hacking in the statistical analysis phase of trials.

Using this corpus of text, we measure the similarity between the language in the submitted protocol and the language in the published SPIRIT guidelines. We estimate the similarity between documents by calculating a term-frequency inverse-document frequency (TF-IDF) vector for each document and then calculating the similarity between the vector of the protocol and vector of the SPIRIT guidelines. Shorter distances indicate

[9] See: https://clinicaltrials.gov/ct2/manage-recs/fdaaa

[10] We are not able to verify if the submitted version of a protocol is in fact the first version of the protocol.

greater similarity. Our conjecture is that trials whose protocol text is more similar to the text in the SPIRIT guidelines exhibit greater compliance and are less likely to be p-hacked.

Figure 4 plots a binned scatterplot for all observations with a p-value between 0.04 and 0.05 for trials started after 2013. For each bin, the y-axis reports the share of these observations between 0.045 and 0.05, making this, in essence, a version of a binomial test. We observe a negative and significant relationship. Trials whose protocol text is more similar to the SPIRIT 2013 guidelines are significantly more likely to have p-values further away from 0.05, suggesting they p-hack less.

Table S5 in our supplement reports the OLS regression of this relationship along with additional robustness specifications. The coefficient of our measure of similarity between a clinical trial and SPIRIT is negative and statistically significant. Changing a protocol from the 10th percentile of similarity to the 90th percentile of similarity in our measure implies a decrease of 13 percentage points in the share of p-values in the critical range.

6 Evidence from Publications

We bring our analysis full circle by considering the effect of SPIRIT on medical publications. There are a number of advantages to examining publications. First, as our publication data spans the introduction of SPIRIT, analyzing the relationship between adopting SPIRIT-like language and p-hacking before and after the introduction of SPIRIT can shed light on the potential impact of SPIRIT. If we find a negative relationship between sounding like SPIRIT and p-hacking in both the pre- and post-SPIRIT periods for clinical trials, it would suggest that adopting SPIRIT-like language serves as a proxy for good research practice. Importantly, this would suggest SPIRIT had little impact on the propensity to p-hack. Alternatively, if we observe a positive or no relationship before SPIRIT and a negative relationship after SPIRIT, it would provide more compelling evidence for the impact of the

guidelines. This pattern would suggest that the formal adoption of SPIRIT led to a change in research practices, with researchers who adhere to the guidelines becoming less likely to engage in p-hacking after their introduction.

Second, our publication data includes clinical trial papers along with papers on other health-related topics. By comparing across these two groups of papers, we can differentiate between changes that impact clinical trials specifically and those that affect medical research more broadly. If SPIRIT guidelines directly influenced clinical trial research practices, we should observe a differential effect for publications reporting clinical trials compared to those reporting other types of studies. Finally, publication data serves as an additional validation sample for our analysis, mitigating the concern that institutional elements of the clinical trial registration process (such as the selective reporting of outcomes to ClinicalTrials.gov) could be driving our results.

To develop a dataset of p-values in medical publications, we download all papers available through the NIH PubMed Central Author Manuscript Dataset: Open Access Subset (of Medicine, 2022). Using regular expressions, we obtain all p-values reported in each paper as long as they are reported in a standard format. We measure the importance of p-values based on where they appear in a paper, rather than labeling them as primary or secondary outcomes. We assume that p-values reported in the abstract are likely to refer to the primary outcome. In contrast, p-values in the results section likely reflect a mix of main and secondary analyses. This differs from our clinical trials data, where primary and secondary outcomes are defined as part of the study protocol. Our main analysis here focuses on p-values from abstract, representing the primary analysis. However, our results also hold when including the broader set of p-values from the results section (see Supplement). As before, we only retain p-values reported to at least three decimal places.

The database of published papers does not indicate which papers present results of clinical trials, thus we employ a combination of natural language processing (NLP) and

supervised machine learning techniques to classify papers. We outline our procedure here and provide a full accounting in the Supplement. First, we generate vector representations of the methods section text for each paper using SPECTER, a pre-trained transformer-based model specifically designed for generating document-level embeddings of scientific papers (Cohan et al., 2020). We focus on the methods section as we hypothesize that it contains the most relevant information for identifying clinical trial papers and their adherence to SPIRIT guidelines.

The first measure defines a paper as a clinical trial if it contains a reference to an NCT number, the unique identifier for a trial registered with clinicaltrials.gov. To further identify potential clinical trials, we train gradient boosting models using the SPECTER embeddings as features and the presence of an NCT number as the binary outcome. The second measure employs an unsupervised approach also using the SPECTER embeddings. We construct a "clinical trial" vector by generating embeddings of a document containing the words 'clinical', 'trial', 'random', and 'prospective', and then subtracting the embeddings of a document containing the words 'metaanalysis', 'systematic', 'review', and 'retrospective'. This process aims to capture the semantic meaning of a clinical trial while avoiding papers that perform meta-analyses or retrospective analyses on clinical trial data. We then compute the cosine similarity between each paper's embedding and this constructed "clinical trial" vector, with higher similarity scores indicating a greater likelihood of the paper reporting clinical trial results.

We assess the performance of our measures in various ways and report the results of these analyses in the Supplement. To test the robustness of our results, we also present results in the Supplement for embeddings generated using doc2vec (Le and Mikolov, 2014). Our findings demonstrate a strong positive relationship between predictions and the likelihood that a paper presents results from a clinical trial. We categorize as a clinical trial all papers scoring at the top 20-percent of both clinical trial measures, and as not clinical trials those scoring at the bottom 20-percent of both measures and focus our

analysis only on comparing these two groups.[11]

Our main sample of study focuses only on p-values between 0.04 and 0.05. The data contains 4,589 papers with 5,911 primary p-values of which 3,662 p-values are categorized as coming from papers describing the results of clinical trials and 2,249 as not clinical trials. Section S8 in the Supplementary Materials reports summary statistics of our data, and additional details on its construction such as the preprocessing and parameters of the SPECTER, doc2vec and gradient boosting models and specific Python libraries used.

We report the results in Figure 6.

Panel A. plots a histogram of the primary p-values of clinical trials in our data. Consistent with our analysis of clinical trial data, we observe significant differences in the p-values reported right below and right above p=0.05. The total density of p-values drops to almost a quarter after 0.05. This could be driven by many mechanisms, including p-hacking and publication bias—where editors prefer to publish papers with statistically significant results. We also note that there is no decreasing trend in p-values between 0.03 and 0.05, suggesting a bunching closer to 0.05. On the other hand, there is a decreasing trend after 0.05. There is no reason publication bias would cause such bunching. Differences in trends are likely driven by p-hacking.

Panel B. studies only the p-values between 0.04 and 0.05. It plots the slope of a regression between similarity to SPIRIT2013 and primary p-values. We split the analysis before and after the release of SPIRIT and between clinical trials and non-clinical trials. If SPIRIT helped decrease p-hacking, then we would expect that papers with text more similar to SPIRIT, which are presumably following it more closely, to have a lower concentration of p-values close to 0.05—i.e., less p-hacking.

Plot i. is the relationship of similarity to p-values before the release of SPIRIT. It is positive. Papers that are written with text more similar to SPIRIT are, if anything, *more*

[11] In the Supplement, we show that our results are robust to different choices of these thresholds.

likely to p-hack. This is the case for both clinical and non-clinical trial. Potentially, in the absence of a protocol, the use of text related to better practices is used more intensively to justify marginally significant p-values.

Plot ii. shows the relationship of similarity to p-values after the release of SPIRIT. For clinical trials, the slope is now negative and significant. Consistent with the benefits of better institutions, clinical trial papers with text more similar to SPIRIT are *less* likely to p-hack. The relationship for non-clinical trials, on the other hand, remains positive.

Table S6 in the Supplement reports regression estimates of these correlations. The results remain similar even after including journal and year fixed effects, and when we focus on secondary, rather than primary, outcomes. This evidence suggests a consistent conclusion: SPIRIT reduced p-hacking, only for clinical trial research.

Panel C explores whether our results are robust given the many choices we incorporated in defining our variables through a specification test. We repeat each regression in Panel B with different variations of our parameters, including all possibilities of: allowing the Doc2Vec vector to be 50, 100, or 200 length; allowing the SPIRIT similarity vector to be 50, 100, or 200 length; including and excluding the use of the variable *Doc2Vec Similarity* in defining clinical trials; and allowing the share used to define papers that are clinical trial (or not) to be the top 10%, 20% or, 25%. We then report the estimates of all these regressions for each time period (before and after SPIRIT) sorted by magnitude.

The results show our effects are robust. Before SPIRIT, we do not find any specification that predicts that similarity to SPIRIT reduces p-hacking. There is also no appreciable difference between the estimates for clinical trials and non-clinical trials. The majority of specifications predict that similarity to SPIRIT reduces p-hacking in clinical trials. This is not the case for non-clinical trials, there is not estimate that is negative and significant, and the distribution of these estimates is shifted to the right of those for clinical trials. Similarity to SPIRIT only predicts lower p-hacking for clinical trials.

To further assess whether our estimate is robust, we report in Table S9 specification curve robustness estimates following the method of Simonsohn et al. (2020). In essence, we create a series of potential null dependent variable values as the residual of predictions using our specifications, and then bootstrap sampling to ask how unique is the relationship we document between similarity to SPIRIT and decreasing p-values, compared to the estimate performed using the null dependent variable. Our estimate is more negative than over 99 percent of the null estimates for clinical trials after SPIRIT. This is not the case with the correlations we document before SPIRIT or for non-clinical trials. That is, consistent with our emphasis on the importance of SPIRIT in shaping clinical trials, we find its relationship robust for post-SPIRIT clinical trial papers while any other correlation appears to be simply driven by random chance.

7 Conclusion

P-hacking is a recent manifestation of the false scientific claims that Babbage criticized almost 200 years ago. In turn, SPIRIT 2013 stands in a long history of institutional responses to the many horrific scandals caused by the production and sale of elixirs, drugs, and cures. Yet, bad research practices continue to persist and evolve, as do the institutional innovations made in response.

Red Queen competition is a particularly apt expression for the co-evolution between normative institutions and scientists who, for reasons of ambition or cognitive bias, manipulate their findings (van Valen, 1973, 1977; Smith, 1976; Solé, 2022; Barnett and Hansen, 1996). Much like the Red Queen in Lewis Carroll's *Through the Looking Glass*, science appears to move two steps forward, one step back (Carroll, 2012).[12]. The introduction of computing capable of accelerating the ability to run many specifications overwhelmed

[12]"A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"

the regulatory capabilities to detect and deter negligent scientific research. To close this gap, institutional innovations, e.g., SPIRIT 2013, embraced a low-tech supervision: ex-ante posting of research plans and statistical models, the ex-post sharing of data and code, as well as other information pertinent for diligent reviewers and readers to validate the reported results.

Many journals in the medical sciences as well as other fields have adopted some of these compliance policies on their own initiative. One recent study in economics Brodeur et al. (2022b) found suggestive evidence indicating that requiring pre-analysis plans led to a reduction in p-hacking.[13] In their finding on the positive correlation of publications and citations with replication in psychology, Youyou et al. (2023) implies that institutional insistence to disclose may improve science incorporate better scientific reputations into career concerns over. Our study contributes to the evidence-base on the effectiveness of these interventions for the medical sciences, which have heretofore been scant.

There is cause for hope that the advance of data sciences has the potential to automate the assessment of research practices. Using machine learning applied to over 14,000 papers in psychology, Youyou et al. (2023) reports that replication rates for experimental papers are significantly inferior to non-experimental studies and that the variation in replication rates are highly correlated with sub-fields within psychology. Jelveh et al. (2023) link the text of academic papers written by economists with data on political behavior to show a robust correlation between research finding and predicted political ideology. In this paper, we show that, using natural language processing, similarity between statistical analysis protocols and the institutional guidelines such as SPIRIT are predictive of differences in the likelihood that studies are p-hacked.

In this study, we did not empirically consider *why* academics should be more subject to

[13]In economics, funding foundations played critical roles in highlighting the problem and solutions through financing research on the efficacy of the policies, personal communication, Danny Goroff, National Science Foundation and Alfred P. Sloan Foundation.

the temptation to p-hack relative to industry. The simplest conjecture is that industry has other mechanisms by which to discipline abuses in clinical trial research, particularly in the form of economic consequences should flawed drugs be released to the market, as well as regulatory oversight by agencies such as the FDA. Without directly facing a market that values the application of research results, academics may respond more to professional incentives divorced from application, e.g., attaining tenure or competing over funding. The competition for these rewards can inspire scholars to work hard and to create a sorting based on the ability to achieve these rewards (Urquiola, 2020, p. 141-144).

Yet, at what cost comes highly incentivized research? Hill and Stein (2021) find that research scientists in the field of structural genomics trade-off research quality against speed to publish, the more competitive the environment. In an early study of whether cited research papers lead to more cited patents, Gittelman and Kogut (2003), for example, find that the impact (measured by citations) of a scientific paper is negatively associated with the impact of innovations in patent space. Industry scientists create patents that are more market-oriented and pay higher acknowledgement, and royalties, to successful industry patents. These dynamics indicate qualitative differences in the incentive structures between academics and industry.

The adoption of protocols after SPIRIT has been far from perfect, with recent estimates placing it at about only 50 percent of clinical trials (Spence et al., 2020; Schönenberger et al., 2022). Among those trials adopting a protocol, a meaningful proportion of the studies evidence discrepancies between the original analysis plan and the eventual publication (Cro et al., 2020; Kahan et al., 2020). The historical record of institutional responses to improve standards in medical research has not come to its end of history.[14] Altogether, the history of science suggests that the Red Queen cycle persists, from fudging statistics that provoke institutional responses, then followed by new ways to ”trim” and ”cook”. An

[14]For example, in 2017, the FDA announced its The Final Rule to improve compliance of reporting for clinical trials,(Zarin et al., 2016).

alternative is to shift attention from the regulatory compliance rules that relies heavily on improving the forensic identification of wrongdoing, towards systemically changing the incentives and norms for researchers to succeed by other means than inducing socially-undesirable research practices.

References

AAAS. Historical trends in federal r&d, 2023. URL https://www.aaas.org/programs/r-d-budget-and-policy/historical-trends-federal-rd.

Jason Abaluck, Leila Agha, David C Chan Jr, Daniel Singer, and Diana Zhu. Fixing misallocation with guidelines: Awareness vs. adherence. Technical report, National Bureau of Economic Research, 2020.

Jérôme Adda, Christian Decker, and Marco Ottaviani. P-hacking in clinical trials and how incentives shape the distribution of results across phases. *Proceedings of the National Academy of Sciences*, 117(24):13386–13392, 2020.

Pierre Azoulay, Jeffrey L Furman, and Fiona Murray. Retractions. *Review of Economics and Statistics*, 97(5):1118–1136, 2015.

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

Charles Babbage. *Reflections on the Decline of Science in England: And on Some of Its Causes, by Charles Babbage (1830). To which is Added On the Alleged Decline of Science in England, by a Foreigner (Gerard Moll) with a Foreword by Michael Faraday (1831).*, volume 1. B. Fellowes, 1830.

Parasuram Balasubramanian, Lamar Pierce, and Trey Cummings. Research validity across organizational forms: Evidence from phase 2 oncology clinical trials. 2022.

William P Barnett and Morten T Hansen. The red queen in organizational evolution. *Strategic management journal*, 17(S1):139–157, 1996.

Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32, 2016.

Abel Brodeur, Nikolai Cook, Jonathan Hartley, and Anthony Heyes. Do pre-registration and pre-analysis plans reduce p-hacking and publication bias? *Available at SSRN 4180594*, 2022a.

Abel Brodeur, Nikolai Cook, and Anthony Heyes. Methods matter: p-hacking and publication bias in causal analysis in economics: Reply. *American Economic Review*, 112(9): 3137–39, 2022b.

Lewis Carroll. *Alice's Adventures in Wonderland and through the Looking Glass*. The Penguin English Library, London, England, 2012.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455, 2020.

An-Wen Chan, Asbjørn Hróbjartsson, Mette T Haahr, Peter C Gøtzsche, and Douglas G Altman. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama*, 291(20):2457–2465, 2004.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*, 2020.

Jonathan R Cole and Stephen Cole. *Social stratification in science*, volume 42. American Association of Physics Teachers, 1974.

Gregory Cox and Xiaoxia Shi. Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *The Review of Economic Studies*, 90(1): 201–228, 2023.

Suzie Cro, Gordon Forbes, Nicholas A Johnson, and Brennan C Kahan. Evidence of unexplained discrepancies between planned and conducted statistical analyses: a review of randomised trials. *BMC medicine*, 18(1):1–8, 2020.

Kay Dickersin. How important is publication bias? a synthesis of available data. *AIDS education and prevention*, 9:15–21, 1997.

Kay Dickersin, SS Chan, TC Chalmers, HS Sacks, and H Smith Jr. Publication bias and clinical trials. *Controlled clinical trials*, 8(4):343–353, 1987.

drugcostsfacts.org. What role does private sector r&d play compared to the national institutes of health?, 2016. URL `https://www.drugcostfacts.org/public-vs-private-drug-funding`.

Graham Elliott, Nikolay Kudrin, and Kaspar Wüthrich. (when) can we detect *p*-hacking? *arXiv preprint arXiv:2205.07950*, 2022a.

Graham Elliott, Nikolay Kudrin, and Kaspar Wüthrich. Detecting p-hacking. *Econometrica*, 90(2):887–906, 2022b.

Andrew Gelman and Eric Loken. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist*, 102(6):460, 2014.

Michelle Gittelman and Bruce Kogut. Does good science lead to valuable knowledge? biotechnology firms and the evolutionary logic of citation patterns. *Management Science*, 49(4):366–382, 2003.

Jeremy M Grimshaw and Ian T Russell. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *The Lancet*, 342(8883):1317–1322, 1993.

Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106, 2015.

A. Bradford Hill. The clinical trial. *British Medical Bulletin*, 7(4):278–282, 1951.

Ryan Hill and Carolyn Stein. Race to the bottom: Competition and quality in science. Technical report, Working Paper, 2021.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8): e124, 2005.

Zubin Jelveh, Bruce Kogut, and Suresh Naidu. Political language in economics. *Economic Journal*, 2023.

Brennan C Kahan, Tahania Ahmad, Gordon Forbes, and Suzie Cro. Public availability and adherence to prespecified statistical analysis approaches was low in published randomized trials. *Journal of Clinical Epidemiology*, 128:29–34, 2020.

Sebastian Kranz and Peter Putz. Methods matter: P-hacking and publication bias in causal analysis in economics: Comment. *American Economic Review*, 112(9):3124–36, 2022.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714, 2008.

Robert K Merton. Priorities in scientific discovery: a chapter in the sociology of science. *American sociological review*, 22(6):635–659, 1957.

Robert K Merton. *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press, 1973.

National Library of Medicine. Pmc open access subset. 2022.

George K Ofosu and Daniel N Posner. Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, pages 1–17, 2021.

Benjamin A Olken. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80, 2015.

Nicholas Rasmussen. The drug industry and clinical research in interwar america: Three types of physician collaborator. *Bulletin of the History of Medicine*, 79(1):50–80, 2005. ISSN 00075140, 10863176. URL `http://www.jstor.org/stable/44448153`.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

Tavola Rotonda. Underreporting research is scientific misconduct. *Jama*, 263:1405–8, 1990.

Christof Manuel Schönenberger, Alexandra Griessbach, Ala Taji Heravi, Dmitry Gryaznov,

Viktoria L Gloy, Szimonetta Lohner, Katharina Klatte, Nilabh Ghosh, Hopin Lee, Anita Mansouri, et al. A meta-research study of randomized controlled trials found infrequent and delayed availability of protocols. *Journal of Clinical Epidemiology*, 149:45–52, 2022.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, 143(2):534, 2014.

Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020.

J Maynard Smith. A comment on the red queen. *The American Naturalist*, 110(973):325–330, 1976.

Ricard Solé. Revisiting leigh van valen's "a new evolutionary law"(1973). *Biological Theory*, pages 1–6, 2022.

O'Mareen Spence, Kyungwan Hong, Richie Onwuchekwa Uba, and Peter Doshi. Availability of study protocols for randomized trials published in high-impact medical journals: a cross-sectional analysis. *Clinical Trials*, 17(1):99–105, 2020.

Miguel Urquiola. Markets, minds, and money. In *Markets, Minds, and Money*. Harvard University Press, 2020.

Leigh van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30, 1973.

Leigh van Valen. The red queen. *The American Naturalist*, 111(980):809–810, 1977.

Suzanne White Junod. Fda and clinical drug trials: A short history. 2008.

Wu Youyou, Yang Yang, and Brian Uzzi. A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proceedings of the National Academy of Sciences*, 120(6):e2208863120, 2023.

Deborah A Zarin, Tony Tse, Rebecca J Williams, and Sarah Carr. Trial reporting in clinicaltrials. gov—the final rule. *New England Journal of Medicine*, 375(20):1998–2004, 2016.

Jonathan Zhang. Can educational outreach improve experts' decision making? evidence from a national opioid academic detailing program. *Evidence from a National Opioid Academic Detailing Program (March 6, 2023)*, 2023.

Acknowledgements

We thank for useful feedback to participants at the Conference of Academic Lobbying at Columbia University, and members of the Macro Research Lab at Berkeley University. We also thank Sebastian Calonico and Suresh Naidu. All errors and omissions are our own.

Supplementary Materials

- S1. Clinical Trial Characteristics
- S2. Significant Digit Reporting
- S3. Summary Statistics
- S4. Top 20 Academic Organizations
- S5. Subgroup Analysis
- S6. Bootstrapping Discontinuity Results
- S7. Text Similarity Regressions

Figure 1: Simulation of P-Value Distribution with and Without P-Hacking

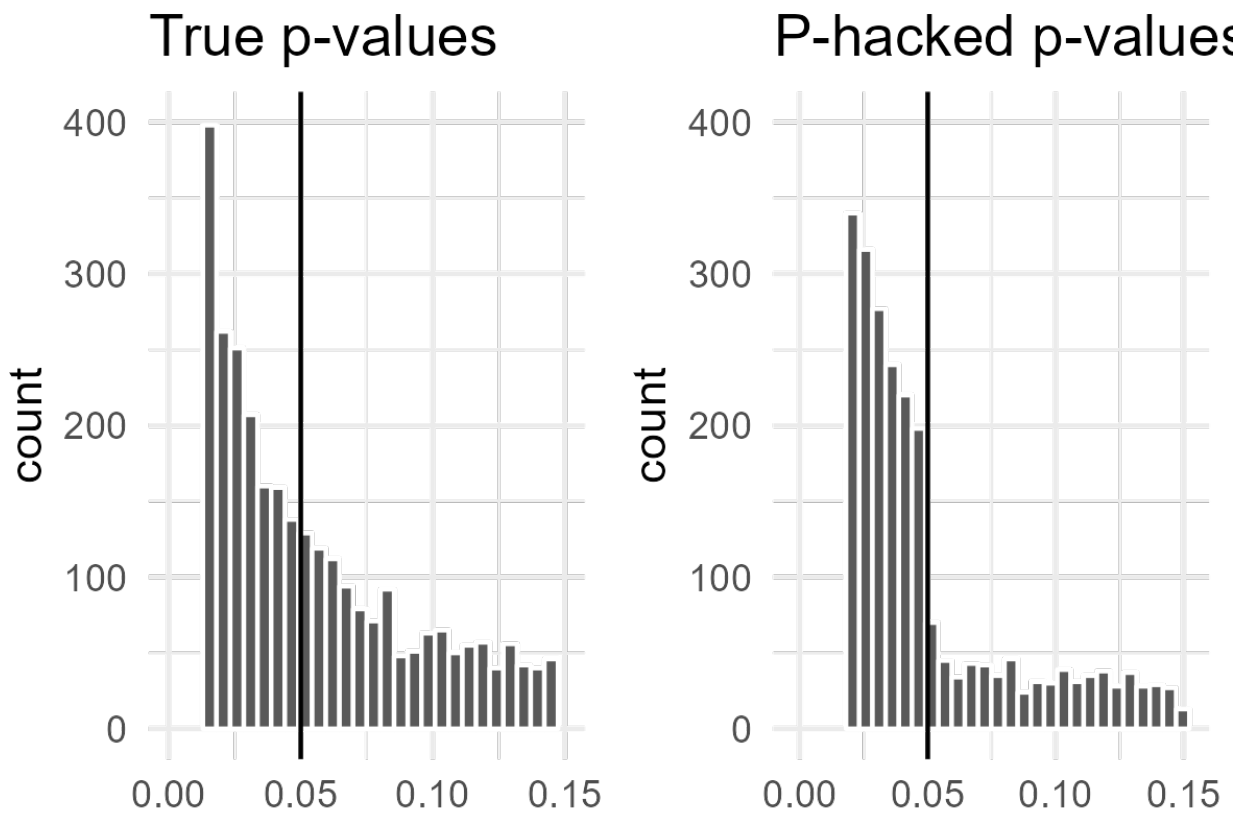


Figure 2: Description of Data Construction and SPIRIT Protocol

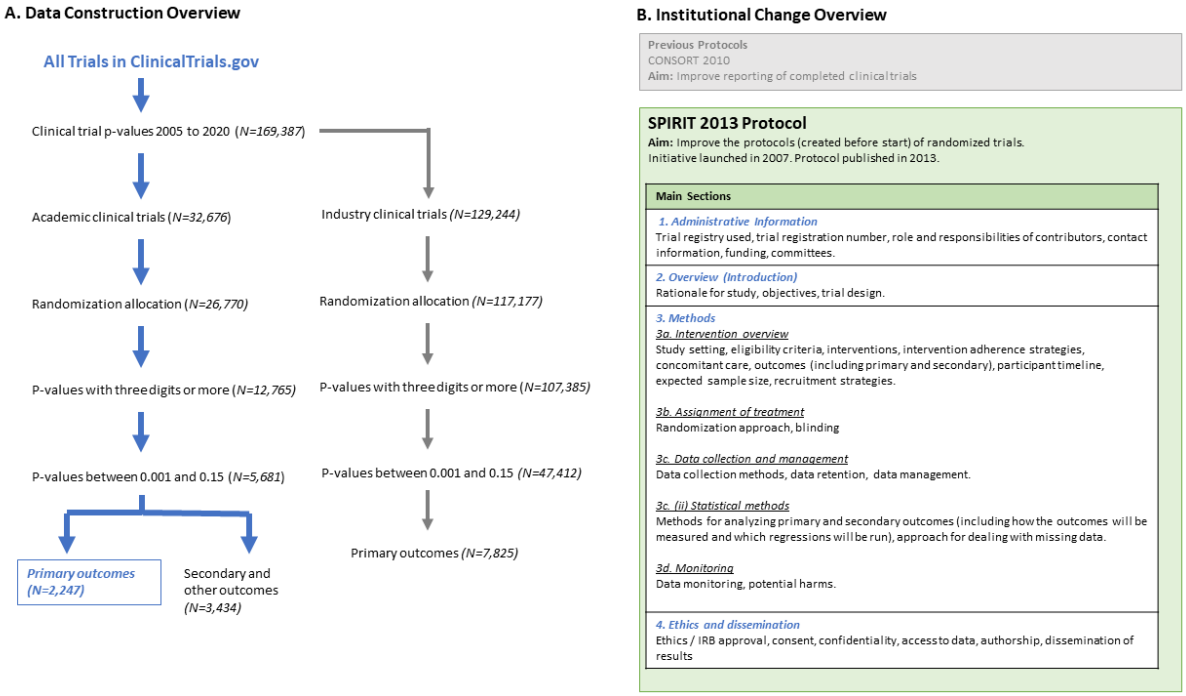
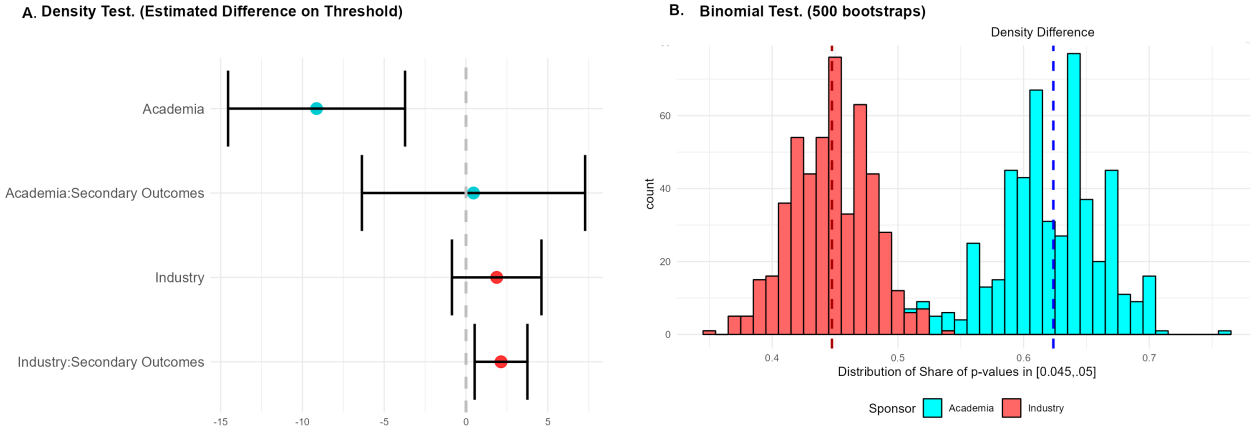
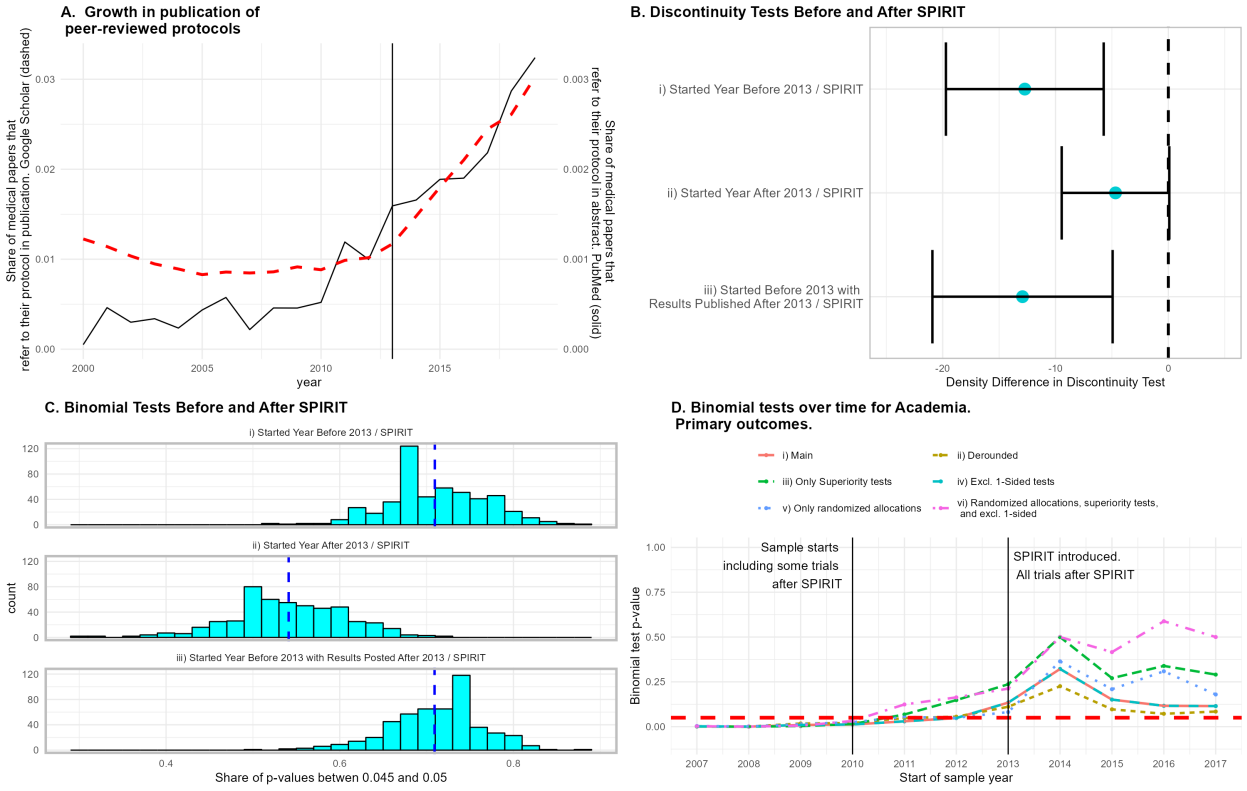


Figure 3: Distribution of P-Values by Clinical Trial Sponsor



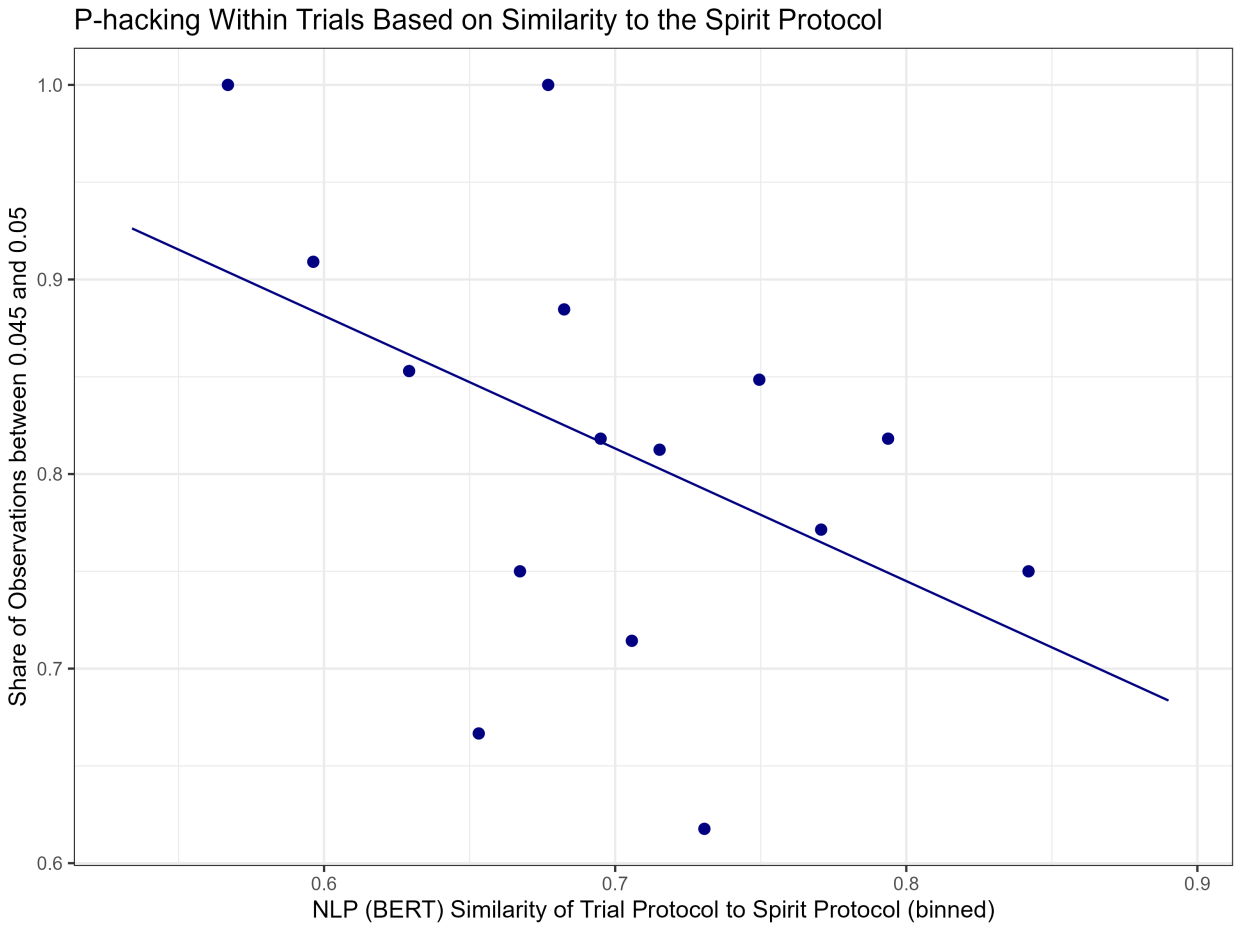
Notes: Primary outcomes only unless otherwise stated in (D). Density plots use the (Cattaneo et al., 2020) methodology estimated for p-values in the range of 0.001 to 0.15, and plotted from 0.02 to 0.11. Histogram width for (A) and (B) pre-set to 0.005, all other values are chosen by default through the *rddensity* command. Figure (C) plots the histogram of the mean of p-values in the range of 0.045 to 0.05, within those between 0.04 to 0.05, for 500 bootstrapped samples. Figure (D) plots the density difference between the density from the left, and the density from the right, in the discontinuity test of four sub-samples.

Figure 4: The Adoption of the SPIRIT Protocol



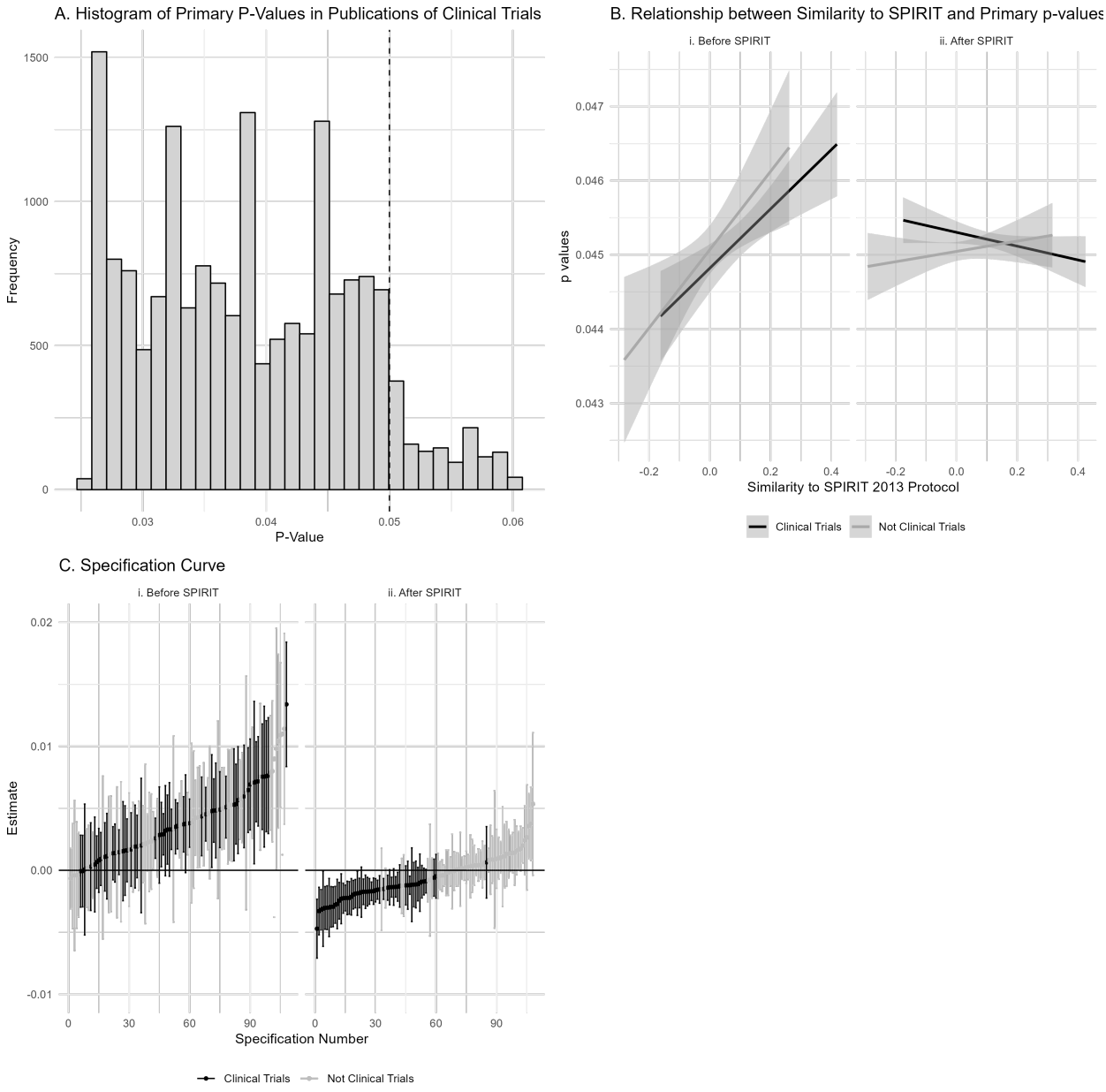
Notes: Panel (A) plots the distribution of the share of medical papers referring to their study protocol by year, for both Google Scholar and PubMed. Panels (B) and (C) plot both the density difference at $p = 0.05$, and the share of p-values in the p-hacking bin in a binomial test. We observe significant differences for trials started before the introduction of SPIRIT 2013 to those started after. The results hold, no matter whether the reporting of p-values occurs before or after the introduction of SPIRIT. Panel (D) plots the incidence of p-hacking over time. For each year, we pick a window from t to $t + 3$ and report the p-value of a binomial test of p-hacking. The samples we use are (i) all p-values with three digits of significance, (ii) including de-rounded two-digit p-values, (iii) only superiority tests (i.e., that the test must prove this treatment is superior to standard of care), (iv) excluding one-sided p-values, (v) only randomized allocations, and (vi) satisfying the criteria in (iii), (iv), and (v).

Figure 5: Similarity of Protocols to SPIRIT and incidence of p-hacking in binomial test



Notes: We plot the binned scatterplot of the relationship of text-based similarity between a trial's protocol and the SPIRIT guidelines to the share of p-values that are close to $p = 0.05$. The sample is all primary outcome three-digit p-values between 0.04 and 0.05. The dependent variable is a binary indicator equal to 1 if the p-value is in the range of 0.045 to 0.05 and 0 otherwise. Table S2 reports the OLS version of this scatterplot, the coefficient for our measure of similarity is -1.24 and significant at the 5 percent level.

Figure 6: The Relationship of SPIRIT to Publucation Outcomes



Notes: Data is *publications* disclosed by NIH. Panel i. plots the distribution of p-values in the data between 0.02 ad 0.06. Panel ii. is the relationship of the similarity to the SPIRIT protocol measure (developed through NLP) and the p-value, for p-values within 0.04 and 0.05. Panel iii. reports many specifications replicating these values. Bands represents 95% confidence intervals, robust standard errors reported.