Does Registration Reduce Publication Bias? No Evidence from Medical Sciences*

July 14, 2015

Abstract

There is increasing support for the use of research registries in social sciences. One possible advantage of the use of a registry is that it would limit the scope for publication or analysis biases that result from selecting statistically significant results. However, to date, there is surprisingly little evidence for the claim that registration will reduce these biases. We look to historical data from medical publishing for evidence, comparing the distribution of p-values before and after the introduction of registration in prominent journals. We couple this analysis with a pre-analysis survey of medical experts and social scientists to assess their prior expectations of the impact of registration on medical publishing and to assess their perceptions on the specificity and sensitivity of our test of effects. Although there is evidence of publication bias in medical studies, our registered analyses uncovered no evidence that registration affected that bias, leading us to moderately downgrade our confidence in the curative effects of registration.

*__Acknowledgments.__ No funding was used for conducting this research. IRB approval was received for the expert survey data collection (see documentation at [omitted]). Anonymous copies of deidentified data and code used to produce all tables and figures reported in this paper, as well as anonymized copies of the preanalysis plan and the instrument used are archived at `https://osf.io/c6hjv/?view_only=b35e0d852d5c4cd8801cb72a7eac1623`. We registered our pre-analysis plan at [omitted] on December 31st, 2014 with registration number [omitted]. Table 7 in Supplementary Materials documents deviations from the plan as well as a set of clarifications.

1 Introduction

There is now broad recognition that analysis and publication biases permeate academic scholarship. Concerns have been raised in medicine (Ioannidis, 2005), economics (Doucouliagos, 2005), political science (Gerber and Malhotra, 2008a), sociology (Gerber and Malhotra, 2008b), and psychology (Simonsohn et al., 2014). These biases may not only result in a record that is unrepresentative of the full body of evidence available on an issue, they also bring into question the credibility of individual studies. These concerns have ignited a discussion over possible solutions across disciplines.

One approach to reducing analysis and publication bias is to register studies prior to implementation. The core idea is that if scholars articulate a specific study design and analysis plan, there will be less scope to deliberately or inadvertently engage in selective reporting or other questionable practices to generate significant findings (John et al., 2012; Casey et al., 2012). As such, registration may reduce publication bias in individual studies. A second possible advantage is that registration would produce a record of work whether or not it is ultimately published. Thus, registration may alter the sample of reported results even if there these individual results are not affected by the registration procedure. Many groups are now calling for registration in social sciences (see for example Nosek et al. (2015)) for one or both of these reasons and a number of voluntary registration systems have been launched.[1] The appropriateness of registration in political science in particular has been debated in a special forum in *Political Analysis* (Vol 21(1), Winter 2013) and elsewhere.

Although supporters claim the arguments for registration are strong and critics worry about adverse effects, the evidence that registration will have any effect at all is surprisingly absent. This article contributes evidence to the debate over registration by assessing how patterns of published results have or have not changed in response to the institutionalization of a registration system in medicine. Exploiting the fact that patterns of published p-values can be used to assess the presence of publication bias, we analyze how these distributions are different before and after the *International Committee of Medical Journal Editors* (ICMJE) implemented a registration system for randomized control trials in 2005 (De Angelis et al., 2004).[2]

While we bring to bear novel evidence on the impact of registration, we are conscious that analyzing historical observational data on medical registration poses a number of limitations. First, because the data is observational, changes over time could be due to factors other than

[1] See, for example, the American Economic Association Registry, the RIDIE-3ie registry for evaluations in international development, and the Experiments in Governance and Politics Design Registry.

[2] This included the Journal of the American Medical Association, The New England Journal of Medicine, The New Zealand Medical Journal, Norwegian Medical Journal, Canadian Medical Association Journal,The Lancet, MEDLINE, Annals of Internal Medicine, Croatian Medical Journal, Nederlands Tijdschrift voor Geneeskunde (Dutch Journal of Medicine), Journal of the Danish Medical Association, Annals of Internal Medicine, and The Medical Journal of Australia. Note that the British Journal of Medicine, apart of the ICMJE, endorsed a commitment to registration but did not require it for publication in their journal.

registration. Second, it is possible that the change in registration regimes could have positive effects that are not picked up in changes in the distribution of p-values — for example researchers may respond to registration requirements by increasing the power of their studies, which could lead to an increase in the share of significant results (though this kind of change is less plausible for the specific tests employed in Gerber and Malhotra (2008a)). Third, the many ways we could analyze the data mean that as we examine the effects of registration we may fall prey to the kinds of distortions that many hope registration will prevent. Fourth, making inferences from the medical experience to the social science context is made difficult by differences in the type of research that is done as well as the way that registration requirements are operationalized — in particular advocates of registration in social sciences are promoting a more comprehensive registration of analysis plans, including statistical models, than is required for registration of medical studies. Finally, since the data are historical, we could not commit to our analysis strategy before the realization of the data.

To respond to these risks and ambiguities, we registered our own plan for the analysis and conducted an expert survey of medical and social science researchers interested in these issues prior to our analysis. Thirty-six experts, largely researchers in social science and medicine, responded to our survey, providing their predictions for the effects of medical registration and information regarding how they would interpret positive or negative results from our tests.

Following our registered analysis plan, we find no distinguishable effect of registration on measures of bias found in these journals. These results are striking given expert expectations over changes in the distribution of p-values: 72% expected greater shifts than we see around the $p = .05$ threshold and 92% expected larger shifts around the $p = .001$ threshold than we see. Moreover, most experts reported putting at least moderate faith in our test, claiming in advance that a null result would reduce their confidence in the salutary effects of registration.

Ironically, *ex post* analyses can identify patterns in the data that are suggestive of positive effects of registration. In particular, if, rather than restricting analysis to the neighborhood of critical values we examine the *all* p-values we see that the share of these below the 0.05 level drops from 77.1% before 2005 to 70.6% after 2005. This 6.5 point drop is large and a traditional t-test rejects the null of no difference in distributions with a p of one in seven billion. Moreover this *ex post* finding is robust to a large range of strategies for conditioning the data. We suspect that had we not registered our analysis plan we might have concluded that registration had effects, at least for less well-powered studies that employ the 0.05 standard. An approach that forwards all possible models and draws inferences from the collection of results might also favor a positive claim here. However, drawing on our own preregistration, we resist this conclusion, placing more weight on results that could be motivated on *ex ante* principles.

Overall our results cast doubt on the potential of registration to reduce bias. As we document below however, we, and others, saw ours as a specific, but non-sensitive test and in light of that

we only moderately downgrade our confidence in the ability of registration to reduce publication bias.

2 Design and Data

To assess the impact of registration on bias, we analyze the distribution of p-values around two critical thresholds: $p = 0.001$ and $p = 0.05$. We build on the "caliper test" employed by Gerber and Malhotra (2008a) and Gerber and Malhotra (2008b) (henceforth GM) to study bias in political science and sociology, in which one can focus on a region– or "caliper"–around a critical threshold (such as $p = 0.05$) and use a simple binomial test to assess whether more published statistics are above the thresholds than we would expect from a chance process. We refer to a surplus of significant results within a caliper as "critical value bias." Under the assumption that for any given study design, the underlying distribution of test statistics that results from a stochastic data generating process is continuous within a small interval around a given significance level and there will be an equal probability of falling immediately to the left or the right of the cutoff, given that the value falls within the interval. The null of no critical value bias is then assessed by examining the share of outcomes on one side of the threshold and assessing the likelihood of such a share if all units were independently assigned to one side with a 0.5 probability.[3]

Using this test, and under the assumption of a smooth underlying distribution of statistics, the null of no critical value bias in medical studies is strongly rejected for the journals we examine in 2000-2010 (with a p-value well below one in a billion). The assumption of a smooth distribution is however violated by rounding practices in reporting of p-values in medicine which produces multiple spikes. Even still, a jump around the 0.05 level can be seen even in coarsened data. There are for example 372 results with p-values in $(0.03, 0.04]$, 331 in $(0.04, 0.05]$, then there is a drop down to 129 in $(0.05, 0.06]$, 72 in $(0.06, 0.07]$, and 71 in $(0.07, 0.08]$.

Using the logic of the caliper test, we propose two simple tests for the effects of registration. First, we assess whether the share of p-values just above the 0.05 threshold is lower after 2005 than before 2005 in the set of journals that adopted the registration norm in 2005, focusing on results with (implied) z-statistics in the interval $[1.66, 2.26]$. Second and more crudely, we implement a 'global' test that examines the proportion of all p-values that are on either side of 0.001 threshold. Strikingly although results below $p = 0.001$ are very rare in political science they are modal in medical science. However rounding practices in medical sciences result in almost no data on either side of 0.001. Thus rather than rely on the caliper test we employ a global test which has the advantage of engaging more data but has the disadvantage that there is no expectation of an even distribution on each side of the cutoff; thus unevenness does not

[3]See Appendix A for details on the caliper test.

imply bias necessarily and a reduction in unevenness does not imply a reduction in bias. This test nonetheless captures broader shifts in the distribution of reported p-values which we would expect to observe if registration reduced analysis and/or publication bias.

For both tests, we also estimate a regression model to assess whether a given p-value is more likely to be below a critical value after 2005, accounting for a linear time tend and allowing for clustering of standard errors at the article level. We listed this regression analysis as our primary specification in our own analysis plan. Including a time trend has the advantage of accounting for general trends unrelated to the introduction of registration in 2005; it has the disadvantage though of possibly controlling for some of the effects of registration if these entered gradually and increasingly over time.

In light of ambiguities about how the gradual roll-out of registration requirements correspond to article-level registration behavior in mid-2005,[4] we exclude 2005 from the analyses and define the post-registration period 2006-2010 and the pre-registration period as 2000-2004.

To implement this test in medical journals, we draw on data from Jager and Leek (2014), hereafter referred to as JL, which contains p-values scraped from abstracts from 77,430 papers published in *The Lancet*, *The Journal of the American Medical Association*, *The New England Journal of Medicine*, *The British Medical Journal*, and *The American Journal of Epidemiology* between 2000 and 2010.[5] These data are modified in two ways. First we convert p-values to z-scores assuming two sided tests with many degrees of freedom, using the formula $p = 2(1 - F(z))$ where F is the normal distribution function. We use p-values and z-scores interchangeably throughout the article as appropriate. Second, we use filters to focus on values generated from randomized controlled trials (RCTs) only. To identify which studies are randomized control trials, we analyze the abstract text from each article and identify the article as an RCT if the abstract includes any of the following strings: 'Experiment', 'Randomized', 'Randomised', 'Randomly Assigned'.[6] There are several concerns with this data, which were raised in the January 2014 issue of *Biostatistics* containing the Jager and Leek (2014) article on the science-wise false discovery rate.[7] We address the concern of including observational studies by restricting the sample to articles identified as RCTs. While we cannot address the concern of selective reporting of p's in abstracts by using the JL data, the remaining criticisms do not point to reasons to expect differential bias on either side of the 2005 implementation of registration.[8]

[4] As noted by medical journals, "This policy applies to trials that start recruiting on or after July 1, 2005. Because many ongoing trials were not registered at inception, we will consider for publication ongoing trials that are registered before September 13, 2005."

[5] We exclude the *American Journal of Epidemiology*, which did not implement registration requirements.

[6] This coding was determined through iterations until an out-of-sample false negative rate of 0 was generated and an out-of-sample false positive rate of 0.2 was generated. Thus, while predictive, it is not a perfect filter.

[7] For a broad critique see Gelman and O'Rourke (2014); Ioannidis (2014).

[8] See appendix B for a further discussion of the JL data.

3 Expectations

The data we analyze are observational, historical, and noisy. These obvious limitations raise questions about what inferences we can reasonably draw from patterns in this data regarding the effect of registration. One strategy consistent with a Bayesian approach to learning is to assess prior beliefs and prior assessments of the probative value of our evidence from experts working in the field. We implemented this approach and solicited the beliefs and expectations of medical experts and social scientists. We fielded a survey that was publicly accessible and coupled this with encouragements to members of the 2014 editorial board of the 11 ICMJE journals that initially adopted registration requirements and to researchers working on transparency in social sciences. Overall, we received 36 responses, with 44% from experts we contacted directly and the rest from referrals. 44% of respondents were social science researchers, 14% were medical researchers, and the remaining respondents self-classified as graduate students (36%) or professionals (5%). 31% self-described as regular readers of medical journals.[9] We highlight that this is a convenience poll and our sample cannot be considered representative of researchers or experts in this field.

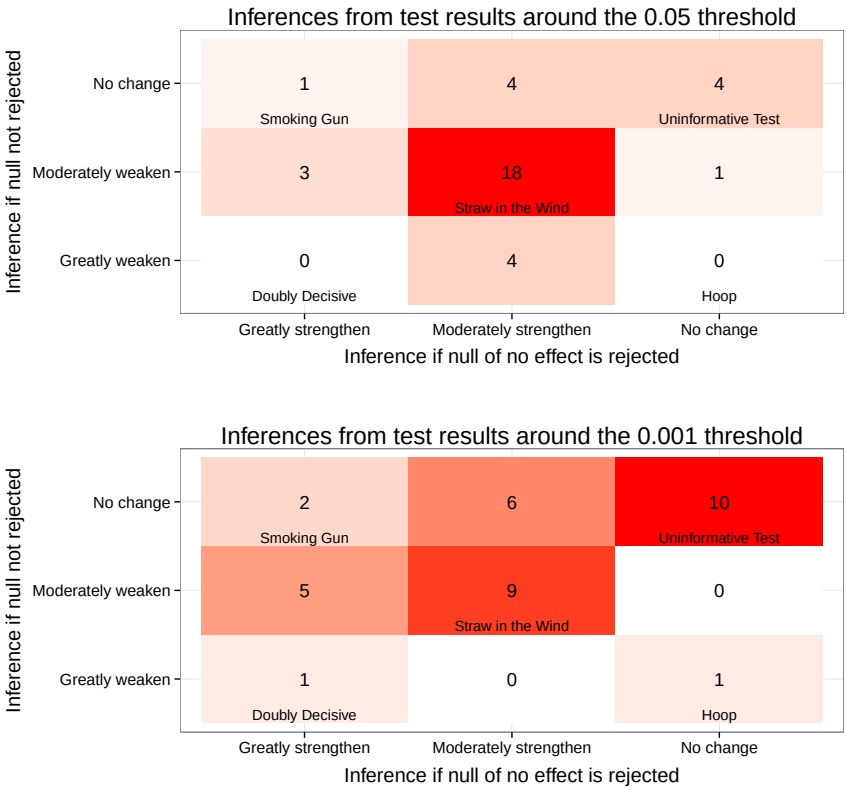
Among our sample, respondents were optimistic about the effects of registration. 9 (25%) judged it likely that registration requirements reduced publication and/or analysis bias, another 18 (50%) judged the impact as somewhat likely, and 8 (22%) unlikely.[10] Confidence in the effects of registration does not however imply confidence that these effects will be visible in changes in the distribution of p-values. To gauge expectations around changes in the distribution we provided experts with information about the distribution of p-values around the critical values aggregated between 2005 and 2010 as well as the bounds on the potential change over time. Overall we see generally optimistic expectations about a shift in shares below critical thresholds, with mean expected shifts of 0.08 around the 0.05 threshold and 0.02 around the 0.001 threshold (see also Table 4 in supplementary materials).

In addition to soliciting beliefs over the effect of registration, we also asked experts how they would update their views on registration in the case of either null results or positive results. Figure 1 reports the responses and relates them to a classification of tests that reflects the sensitivity or specificity of the test for the underlying hypothesis. The figure shows that most experts saw the test around the $p = 0.05$ threshold as a "straw in the wind test" (Van Evera, 1997) suggesting moderate updating following results; more respondents (nearly one third) saw the test around $p = 0.001$ as uninformative, with those that considered it informative classifying it closer to what is sometimes called a "smoking gun" test, that is, specific but not sensitive.

[9] See appendix C for a discussion of the expert survey.

[10] One respondent did not answer this item.

Figure 1: Expert Reported Updating



Note: This figure plots the number of expert respondents who reported they would either not, moderately or greatly update their beliefs about registration in light of either null or significant findings in the analysis.

4 Results

4.1 Non parametric analysis

Table 1 presents the distribution of p-values around each critical threshold before and after registration.[11] For $p = .05$, data is restricted to a narrow caliper and the $p = .001$ analysis includes all data for a global assessment. The table also presents a simple analysis of the impact of registration using a Fisher exact test for equality of proportions.[12] The patterns in the raw data, which do not account for time or journal effects, shows little evidence that registration had a significant impact on the magnitude of publication bias.

The left panel of Figure 2 shows the full distribution of (implied) z-statistics in medical journals before and after registration requirements (medical p-values are converted to z-stats assuming two sided tests and a large number of degrees of freedom). We see broadly that the post-2005 period has relatively more mass below 2 and relatively less mass between 2 and 3.

[11] See Appendix C for further information on the raw data presented to survey respondents.
[12] This assumes that data are independently and identically distributed.

Table 1: Is reporting different before and after 2005?

	I. Pre-2005 (2000-2004)	II. Post-2005 (2006-2010)	III. Total	IV. Test statistic (Difference (II-I))
A Critical Value Bias				
$p \leq .05$	539	487	1026	
$p > .05$	148	160	308	
Total	687	647	1334	
Share ≤ 0.05	0.785	0.753	0.769	d_1 = -0.032 (0.173)
B Global Bias:				
$p \leq .001$	1128	1228	2356	
$p > .001$	2470	2644	5114	
Total	3598	3872	7470	
Share $\leq .001$	0.314	0.317	0.315	d_1 = 0.004 (0.746)

Note: Marginal data taken from the Jager-Leek dataset. p-values for the test statistic are in parentheses. Analysis in upper panel restricted to z-statistics in $[1.66, 2.26]$.

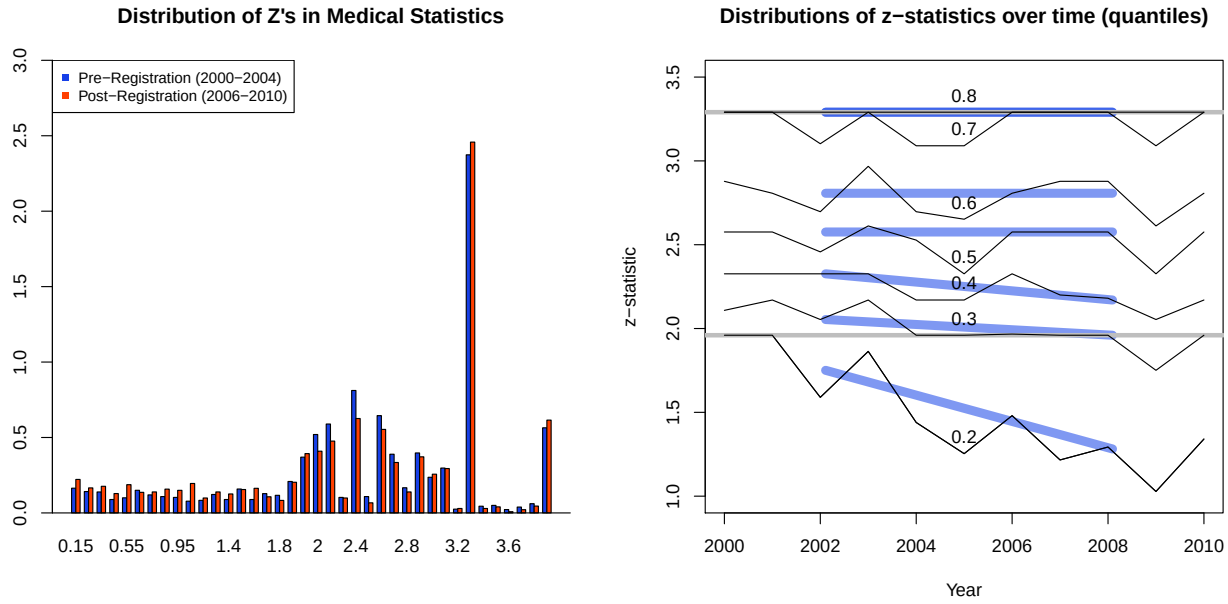
The right figure also shows the changes in the distributions over time; each line here represents a percentile in the distribution of z statistics. We observe a similar pattern: whereas in 2000 the 20th percentile had a z stat of 1.96, in later years this fell to below 1.5. Despite these shifts, there is no visual evidence of a discontinuity around 2005.

We highlight that the lack of evidence for a change associated with 2005 does not mean that there is no bias in either period. Figure 5 in Appendix D reports results from a series of binomial tests using a caliper radius of 0.3 (that is, for each z^*, conditioning analysis on the range $[z^* - 0.3, z^* + 0.3]$). Here, we observe evidence of discontinuities at critical values for which more data exists just right of a critical value than left which is suggestive of publication bias. These discontinuities are consistent with publication bias, particularly for poorly powered studies. Strikingly however, the two figures are almost identical and highlight that overall changes in the distribution are not driven by sharp changes around critical values (see also evidence in Figure 6).

4.2 Primary Analysis

As specified in our pre-analysis plan, our *primary* analysis is conducted using an ordinary least squares model that accounts for time trends and journal fixed effects. Estimating this model we cannot reject the null hypothesis that registration had no distinguishable effect on the prevalence of critical-value bias around either thresholds. For the critical value $z = 1.96$, there is a small

Figure 2: Distribution of Z-Statistics



Note: The left figure shows the distribution of z statistics before and after 2005. The post-2005 period has more mass below 2 and less mass between 2 and 3. The right figure also shows the changes in the distributions over time; each line represents a percentile in the distribution of z statistics. We see that whereas in 2000 the 20th percentile had a z stat of 1.96, in later years this fell to below 1.5. There is no visual evidence of a discontinuity around 2005 however.

and insignificant *increase* of 0.043 ($p = 0.44$) after 2005; for critical value $z = 3.29$, we estimate a similarly small and insignificant increase of 0.039 ($p = 0.24$). Thus the primary analysis produces estimates opposite to the relation we would expect to see if registration reduced publication or analysis bias. For more details as well as some robustness checks, see Table 5 in Appendix E.

4.3 Posterior Inferences

We use three approaches to assess the amount of learning this exercise generates. First we assess posterior distributions for all experts, given their stated priors over the reduction in shares of significant results in given calipers (two thirds of respondents provided detailed information on priors). The results for this exercise are visualized in Figure 8 in Appendix G. We see in most cases a reduction in the expected shifts after 2005, though this is coupled in general (and for the 0.05 analysis in particular) with an increase in the confidence that there was a negative shift. For the priors that we specified for example, our prior expectation for the size of a shift around 0.5 was 0.036 and our posterior was similar at -0.033. However our belief that the shift was negative increased from 61% to 96%.

Second, we calculated a Bayesian analogue of the R^2: using reader priors we estimate an

expected error and compare this to the expected error under our posterior (see Gelman and Pardoe (2006)). Results are provided in Figure 9 in Appendix G, broken down by respondent type.[13] Results suggest a 10-20% reduction in error for both models.

Overall, what conclusion should we draw about the effects of registration? For this we return to our expert assessments of specificity and sensitivity which integrate beliefs both about priors and the imperfections of our test. There we saw 74% of our sample claimed that a null finding for the critical value $z = 1.96$ would either greatly or moderately weaken their confidence that registration can tackle bias; a more skeptical 47% claimed the same would be true for the finding around $z = 3.29$.

While classical inference suggests that no evidence of an effect should not be taken as evidence for no effect, a Bayesian approach suggests the opposite: we do learn from nulls and in this case we learn that our optimism regarding the effects of registration should be dampened.

5 What We Could Have Found

All our analyses of bias around the $z = 1.96$ threshold relied on the choice of a particular caliper radius (of 0.3). The caliper radius was selected to be small enough to exclude multiple critical thresholds so as not to confound our results and wide enough to contain enough data to estimate large effects with reasonable precision. We chose this radius in advance in order to eliminate the possibility of choosing *ex post* in order to produce significant results.

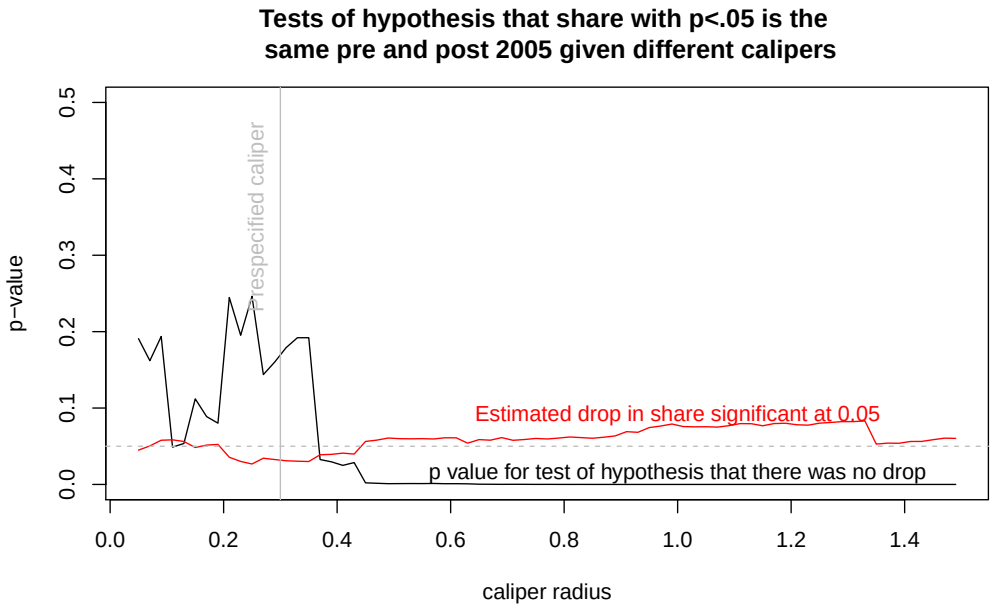
What might have we found had we not pre-committed to the caliper radius of 0.3? Figure 3 illustrates our results across a range of caliper radii.[14] It shows that had we selected a caliper radius of about 0.35 *or larger*, we would have rejected the null of no difference in bias before and after 2005.

Should the non-robustness of our null finding force a reconsideration of our conclusion? One set of arguments contends that all possible models should be analyzed and inferences drawn from all models. However, this approach risks losing sight of the principles upon which test selection occurs in the first place. The caliper we identified in advance for analysis was motivated by the statistical justification for the test: the caliper test is valid in the limit as the caliper radius shrinks to zero. As we extend the caliper we increase our power but threaten the validity of the test. Registering a limited caliper was motivated by a set of principles that ultimately buttress the credibility of the null results.

[13] Note that only one of the five medical researcher respondents gave full information to assess priors and for this respondent, uniquely, our posterior is no more accurate than their prior.

[14] This graph reflects a continuous implementation of the sensitivity tests specified in section 6.4 of the pre-analysis plan.

Figure 3: Sensitivity to alternative caliper radii



Note: This figure plots the percentile of z-statistics within journals by year.

6 Conclusions

Reducing publication and analysis bias in academic scholarship is important for generating credible science. While one can observe evidence of bias in the distribution of p-values in published research in political science and cognate disciplines, there is, to date, little or no evidence that registration will affect this bias. Our analysis contributes to this discussion by examining changes in the distribution of published p-values before and after the introduction of registration requirements for medical journals.

As we noted at the outset, our analysis faces limitations arising from the historical, observational, and noisy nature of our data. Moreover we recognize that various logics could produce risks of false negatives. For example it is possible that registration did reduce analysis bias but that continued publication bias ensured that non-significant findings (that might otherwise have been significant) did not appear in these leading journals.

Nevertheless, despite these shortcomings, the data available from medical sciences is perhaps the best place to look for evidence that registration might make a difference. Our survey of experts suggested that these too placed confidence in the tests we implemented. 63% reported that a null result (for the test around $p = 0.05$) would moderately weaken their confidence in the effectiveness of registration and another 11% that it would greatly weaken it. 47% reported that null results on the $p = 0.001$ test would weaken their confidence in registration.

Implementing these tests, we fail to find evidence for any effects of registration on publication

bias for either cutoff. There are multiple possible reasons for weak effects in medicine. First, medical registries collect sparse data, particularly around analysis procedures; second, by many accounts, registration standards in medicine have not been well enforced Reveiz et al. (2010); Mathieu et al. (2009). While it is possible that these problems will not be as severe for registration in political science and other social sciences, there is no evidence to support this claim. We thus conclude that we should downweight the confidence we place in registration as a solution to bias.

References

Casey, K., R. Glennerster, and E. Miguel (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *The Quarterly Journal of Economics 1755*, 1812.

De Angelis, C., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, J. P. Overbeke, and T. V. Schroeder (2004). Clinical trial registration: A statement from the international committee of medical journal editors. *New England Journal of Medicine 351*(12), 1250–1251.

Doucouliagos, C. (2005). Publication bias in the economic freedom and economic growth literature. *Journal of Economic Surveys 19*, 367–387.

Gelman, A. and K. O'Rourke (2014). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics 15*(1), 18–23.

Gelman, A. and I. Pardoe (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics 48*(2), 241–251.

Gerber, A. and N. Malhotra (2008a). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science 3*, 313–326.

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

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine 2*(8), 696–701.

Ioannidis, J. P. A. (2014). Discussion: Why "an estimate of the science-wide false discovery rate and application to the top medical literature" is false. *Biostatistics 15*(1), 28–36.

Jager, L. R. and J. T. Leek (2014). An estimate of the science-wise false discovery rate and application to top medical literature. *Biostatistics 15*(1), 1–12.

John, L. K., G. Loewenstein, and D. Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 0956797611430953.

Krishnamoorthy, K. and J. Thomson (2004). A more powerful test for comparing two poisson means. *Journal of Statistical Planning and Inference 119*, 23–35.

Mathieu, S., I. Boutron, D. Moher, D. G. Altman, and P. Ravaud (2009, September). Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA : the Journal of the American Medical Association 302*(9), 977–84.

Nosek, B., G. Alter, G. Banks, D. Borsboom, S. Bowman, S. Breckler, S. Buck, C. Chambers, G. Chin, G. Christensen, et al. (2015). Promoting an open research culture. *Science 348*(6242), 1422–1425.

Pryzborowski, J. and H. Wilenski (1940). Homogeneity of results in testing samples from poisson series with an application to testing clover seed for dodder. *Biometrika 31*(3/4), 313–323.

Reveiz, L., A.-W. Chan, K. Krleza-Jerić, C. E. Granados, M. Pinart, I. Etxeandia, D. Rada, M. Martinez, X. Bonfill, and A. F. Cardona (2010, January). Reporting of Methodologic Information on Trial Registries for Quality Assessment: a Study of Trial Records Retrieved From the WHO Search Portal. *PloS One 5*(8), 1–6.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General 143*(2), 534.

Van Evera, S. (1997). *Guide to methods for students of political science*. Ithaca, NY: Cornell University Press.

Does Registration Reduce Publication Bias?

July 14, 2015

Appendices

A Empirical Strategy

The caliper test involves the following steps. First, one specifies a threshold of statistical significance around which the test is conducted. GM specify the critical value $z = 1.96$ for two-tailed tests and $z = 1.645$ for one-tailed tests corresponding to a test size $\alpha = 0.05$. Second, one then specifies calipers, or intervals, of equal size above and below the threshold. In their assessment of publication bias in two top political science journals, GM (2008) specify calipers equal to 10%, 15%, and 20% of the critical value. Third, one compares the number of observed z-scores at the threshold or in the bin above it (n_{over}) to the number of observed z-scores in the bin below the threshold (n_{under}) using the caliper test, by estimating $Pr(X \geq n_{over})$ for $X \sim \text{Binomial}(N = n_{over} + n_{under}, p = 0.5)$. The estimated probability is the probability of obtaining at least as many draws in the "over" bin as actually observed under the assumption that the null hypothesis $H_0 : p = 0.5$ is true.

With respect to both the local and global caliper test employed in this paper, for test j let π_j^0 denote the probability of observing a p-value in a given range below some critical threshold prior to 2005, and π_j^1 the probability afterwards. Let $\delta_j = \pi_j^1 - \pi_j^0$ denote the difference between time periods. We are interested in δ_j, and hypothesize that δ_j is negative.

Caliper Test Assumptions

The validity of the caliper test is threatened by the possible violation of a set of assumptions: that the distribution of test results is continuous; that calipers are sufficiently small so that the distribution is sufficiently flat in bins on either side of a critical threshold; and that test statistics are drawn independently.

We note that neither the assumption of smoothness nor the assumption that the distribution of outcomes across tests are the same on either side of a threshold is innocent. The following simple illustration violates both conditions. Say a researcher assigns one unit to a treatment and nineteen units to a control condition and plans to assess statistical significance using randomization inference (and a one-sided test). Then p-values will come from the set $\{.05, 0.1, 0.15, \ldots, 1\}$, depending on which unit happens to be treated. This illustrates a violation of continuity of test statistics generated by a stochastic data generating process. Here while there is positive mass at 0.05 there is no mass below 0.05. In practice however, we may expect sufficient variation in data structures across studies to justify an assumption of continuity.

Even when continuity holds, it justifies an expectation of equal density only in the immediate neighborhood of a threshold. With wider calipers the shape of the distribution alone may result in bins on either side of a cutoff with unequal density. For example for studies powered at 80%

or 90% we expect the distribution to be centered to the right of 1.96 and thus for the density to be increasing at 1.96, which in principle could give rise to unjustified claims of bias. This concern is alleviated in our study however given our focus on *differences* between periods in the shares below critical thresholds.

Violations of independence of draws may place local concentrations of mass on the empirical distribution of test statistics. This may arise for example if multiple z statistics are generated through small modifications to a common core analysis. In the GM approach this concern is mitigated by admitting only one statistic per study. In our primary analysis we account for this by estimating standard errors clustered at the article level.

Finally we note that the caliper test is based on the idea that the outcome of a given test is stochastic. If, however, researchers are engaged in "data fishing" then they are altering which analyses get reported *conditional* on data — thus any fishing takes place across *analyses* not across studies. It turns out however that this is not a great concern since an alternative conceptualization of publication bias that allows for conditioning on the data yields precisely the same test strategy. Assume that each possible test (from all defensible tests) has a constant, though possibly low, probability of making its way past the authors and publishers and into print. If the distribution of possible results from defensible tests is continuous around the threshold, then under the null of no bias, the expected number of published results should be the same just above and just below thresholds of statistical significance. We can then test this null by comparing counts above and below the threshold. Let X_u and X_o denote the number of observed p-values just under and over a threshold, respectively. Let the mean rate of occurrence of events be λ, then we model the probability distribution of $X_o \sim \text{Poisson}(n_o, \lambda_o)$ and of $X_u \sim \text{Poisson}(n_u, \lambda_u)$ where realizations of X_o and X_u are i.i.d. Let k_u, k_o be observed values of X_u, X_o, respectively, and the null hypothesis of interest is $H_0 : \lambda_o = \lambda_u$ vs. $H_a : \lambda_o \neq \lambda_u$. Under the hypotheses of the "conditional test" (see Pryzborowski and Wilenski (1940)), the conditional distribution of X_o given $X_o + X_u = k$ is binomial with the number of trials k and success probability $n_o c/(n_u + n_o c)$. With c set to unity this is equivalent to the Binomial test.[15]

Caliper Specification

For the GM caliper test, the data are z-statistics and accordingly calipers are equally spaced bins above and below the critical value in "z-space." Because the Jager-Leek data provide p-values, we convert the statistics to z statistics assuming large samples and we implement the test over the resulting distribution of implied z-statistics. We define the z_c-caliper as $[\underline{z}, \overline{z}]$ where $\underline{z} = z_c - w$ and $\overline{z} = z_c + w$. Our caliper radius of 0.3 is chosen to give a range $[1.66 - 2.26]$ that maximizes data use while excluding other critical probability mass points at $p = 0.1$ (with corresponding

[15]For a discussion of this and of more powerful Poisson tests Krishnamoorthy and Thomson (2004).

$z = 1.64$) and at $p = 0.02$ (with corresponding $z = 2.33$). The 0.3 caliper is also approximately the central of three calipers used by GM (.15*1.96).

B The Jager-Leek Data

In an analysis that estimates the science-wise false discovery rate (FDR) in the top medical literature, Jager and Leek (hereafter JL) scraped p-values in abstracts from 77,430 papers published in *The Lancet*, *The Journal of the American Medical Association*, *The New England Journal of Medicine*, *The British Medical Journal*, and *The American Journal of Epidemiology* between 2000 and 2010. The JL dataset only reports p-values that show up numerically in the abstract. For example, while text such as "$p = .02$" would generate an observation, p-values identified as significant or not-significant will not generate an observation. We employ a subset of this data for our analysis with filtering as described in Table 2 below.

Table 2: Sample Filtering

All JL data:	15653
Excluding AJE:	14454
With RCT flag:	8263
Excluding 2005:	7470

Note: Selection of values in JL data used for analysis.

Several concerns have been raised regarding these data in the January 2014 issue of *Biostatistics* containing JL's article on the science-wise FDR. Most of these criticisms about the data are specific to estimating the science-wise FDR. Ioannidis (2014) argues that:

> the data used are the P-values reported in the abstracts of published papers; these P-values are a highly distorted, highly select sample. Besides selective reporting biases, all other biases, in particular confounding in observational studies, are also ignored, while these are often the main drivers for high false-positive rates in the biomedical literature. A reproducibility check of the raw data shows that much of the data Jager and Leek used are either wrong or make no sense: most of the usable data were missed by their script, 94% of the abstracts that reported ≥ 2 P-values had high correlation/overlap between reported outcomes, and only a minority of P-values corresponded to relevant primary outcomes (28).

We address the concern of including observational studies by restricting the sample to articles identified as RCTs (based on the inclusion of key words in abstracts). While we cannot address the concern of selective reporting of p's in abstracts by using the JL data, the remaining criticisms do not point to reasons to expect differential bias on either side of 2005. We believe that the concern of greater selectivity in reporting in abstracts should lead us to overestimates of the

effects of registration though we also note the possibility that registration alters the overall set of results published without altering those that are selected to be highlighted in abstracts.

For further discussion of the data, see *Biostatistics* (2014), volume 15, issue 1.

C Expert Survey

The expert survey was implemented across two samples. First, we encouraged each member of the current (2014) editorial board of the 11 ICMJE journals that initially adopted registration requirements. This includes the Journal of American Medical Association, The New England Journal of Medicine, The New Zealand Medical Journal, the Norwegian Medical Journal, the Canadian Medical Association Journal, the Lancet, the Annals of Internal Medicine, the Croatian Medical Journal, the Dutch Jounal of Medicine, and the Medical Journal of Australia.[16] From the 142 editorial board members, we were able to recover contact information for 128 individuals, or 91% of the universe. We distributed our online survey to these individuals. In addition to this sampling strategy, we also drew on a convenience sample of medical researchers and social scientists that are encouraged through blog and related postings online. The survey instrument may be found at [anonymized version included with submission].

This survey captured beliefs regarding respondent priors on the effects of registration as well as their interpretations of patterns in the data. As specified in the pre-analysis plan, we expected very low response rates for the expert sample and will not try to correct for missing data. The online survey generated 36 responses, of which only 5 (14%) were medical professionals and 31 (86%) were social scientists.

Rather than have experts provide beliefs on the coefficient size of the regression analysis assessing the impact of registration, we provided our sample the marginal distributions of z-statistics (Table 3) and asked them to guess d_1 and d_2.

At the time of registration we had not yet analyzed this data except to produce the marginal distributions shown in Table 1. In particular, we had not produced cross tabulations of the distribution of p-values above and below critical thresholds and over time.

We made an error in our survey for the 0.05 data by providing the margins of a Table that excluded the *British Medical Journal* and that used a narrower caliper radius (of 0.08). Table 1 shows the full sample used in our analysis and Table 3 below shows the table (with completed cells) corresponding to that provided during the survey. The latter is associated with a larger treatment effect again the effct is not significant.

Table 4 reports summary statistics and Figure 4 plots the distribution of guesses for d_1 and d_2.

[16]Note that Medline is not included as it is not a journal and the Danish Medical Association is not included as they do not publicly list their editorial board online.

Table 3: Is reporting different before and after 2005? (Margins used for expert guesses)

A Critical Value Bias	I. Pre-2005 (2000-2004)	II. Post-2005 (2006-2010)	III. Total	IV. Test statistic (Difference (II-I))
$p \leq .005$	116	111	**227**	
$p > .005$	42	62	**104**	
Total	**158**	**173**	**331**	
Share ≤ 0.05	0.734	0.642	0.686	d_1 = -0.093 (0.076)
B Global Bias:				
$p \leq .001$	1128	1228	**2356**	
$p > .001$	2470	2644	**5114**	
Total	**3598**	**3872**	**7470**	
Share $\leq .001$	0.314	0.317	0.315	d_2 = 0.004 (0.746)

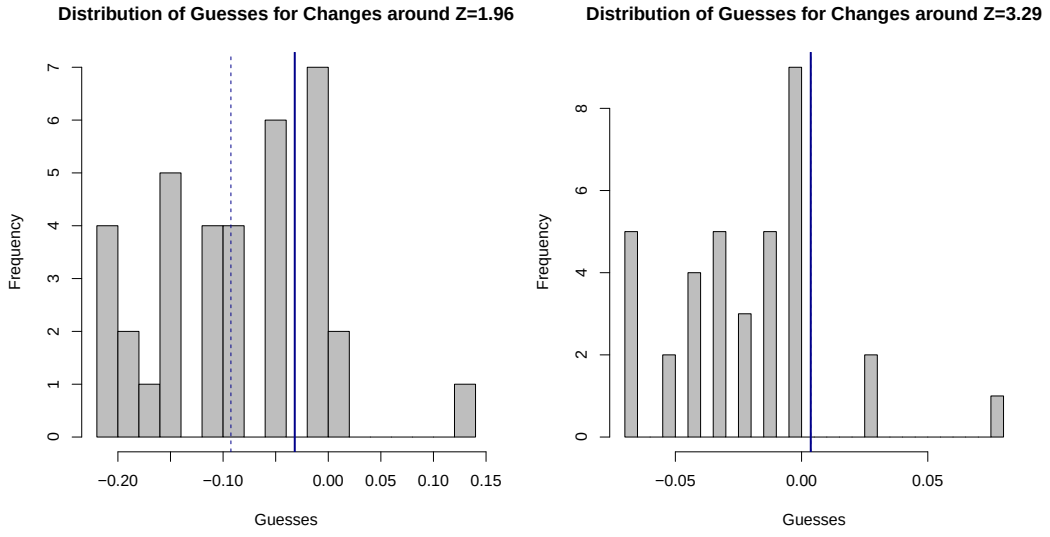
Note: Data taken from the Jager-Leek dataset. Margins (in bold) were those used for eliciting expert guesses. For the 0.05 (upper) panel these differ to what is used in the main analysis in because of an ccidental exlcusion of BMJ datathe use of a narrower caliper of $z \in (1.88, 2.04)$. p values in parentheses.

Table 4: **Expert Beliefs**

	Bounds	**Min**	**Max**	**Mean**	**SD**	*N*
p=.05	(-0.22,0.14)	-0.22	0.14	-0.08	0.09	36
p=.001	(-0.07,0.03)	-0.07	0.08	-0.02	0.03	36

Note: This table reports expert expectations over the effect of registration on publication bias. Experts were provided bounds of the effect size, reported in column 2, to inform their expectations.

Figure 4: Expert Beliefs Over the Impact of Registation

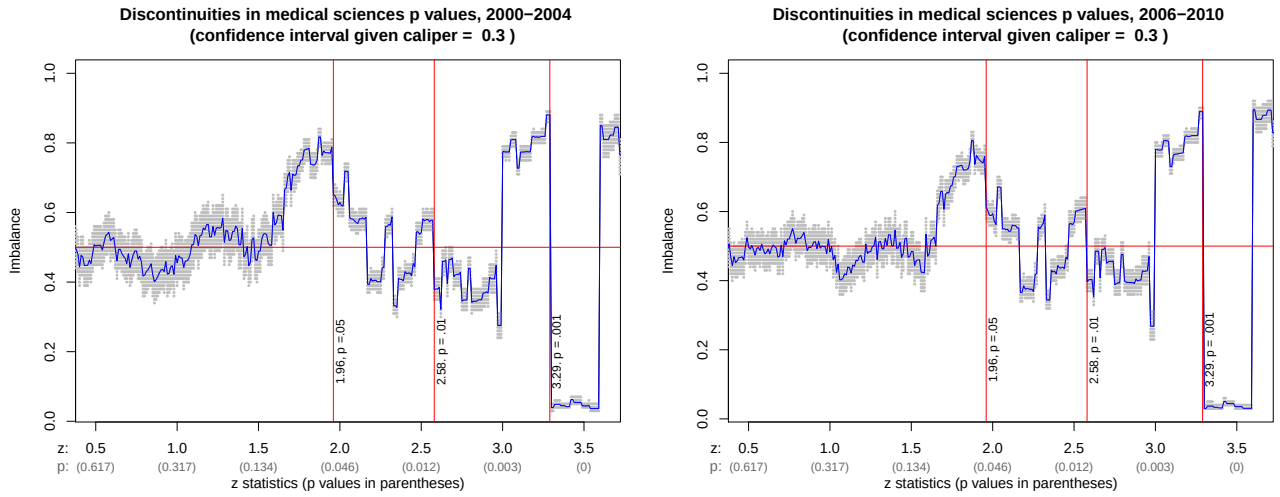


Note: This figure visualizes the distribution of expert estimates of the impact of registration for d_1 and d_2 respectively. The vertical blue lines represent the actual changes: solid for full data (Table 1); dotted for data reported in the pre-analysis plan (Table 3). Responses left of the blue line represent optimistic views of registration while responses right of the blue line represent skeptical beliefs.

D Placebo caliper tests

Figure 5 implements the 0.3 caliper tests for all values of z in $(0.5, 3.5)$.

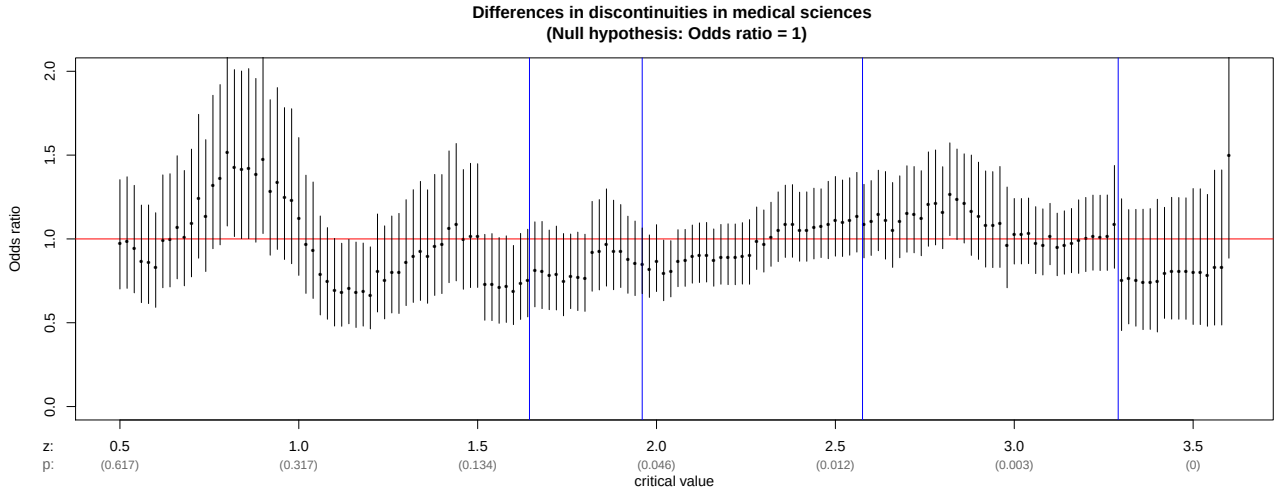
Figure 5: $z-$statistics and discontinuities in the Medical Sciences



Note: Figures show the share of data above z conditional upon being in the ± 0.3 neighborhood of z for each possible critical threshold. Grey dots shows the confidence interval for the underlying probability that a result in this neighborhood is above z. These confidence intervals do not include 0.5 at many critical z values. Data from Jager and Leek (2014).

Figure 6 shows tests of the null of no difference in distributions over time around z for all values of z in $(0.5, 3.5)$, using a caliper of 0.3.

Figure 6: $z-$statistics and differences in discontinuities in the Medical Sciences



Note: Figures show odds ratios for tests constructed for different possible critical values. Confidence intervals constructed from Fisher exact tests. Data from Jager and Leek (2014).

E Regression Analysis: Full Results

Table 5 reports our primary results for the impact of registration on publication bias for critical values of statistical significance, $z = 1.96$ and $z = 3.29$. We present results for the primary analysis as well as robustness checks for other calipers. In all cases we estimate a positive coefficient on the key treatment variable, post-2005.

Table 5: OLS Estimates

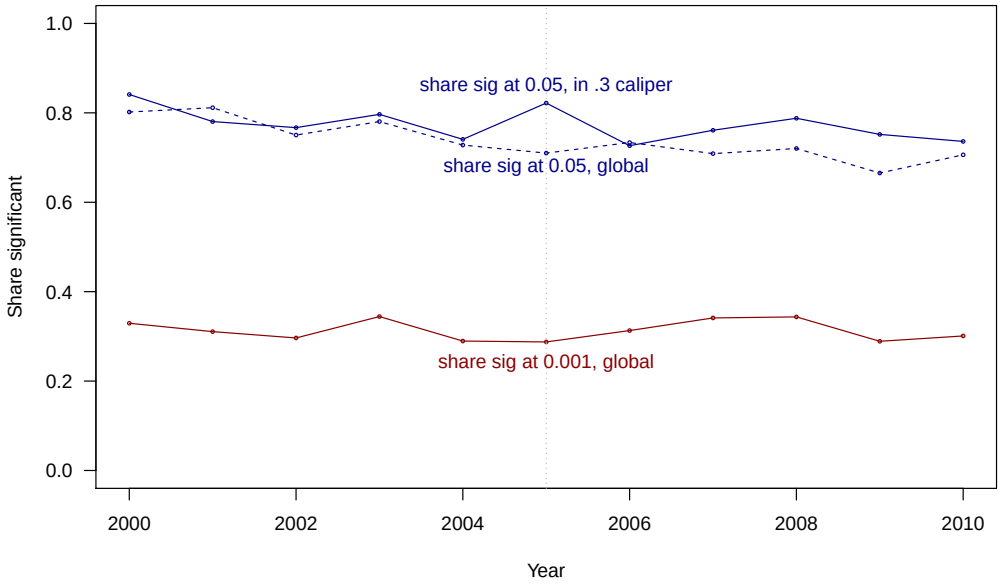
	Dependent Variable: Binary Indicator = 1 if $z \geq z_c$				
	Primary Analysis		**Robustness Check**		
	$z_c = 1.96$	$z_c = 3.29$		$z_c = 1.96$	
	Caliper=.30	No Caliper	Caliper=.15	Caliper=.60	No caliper
	(1)	(2)	(3)	(4)	(5)
Post-2005	0.043	0.039	0.073	0.030	0.012
	(0.056)	(0.033)	(0.071)	(0.046)	(0.032)
Year	−0.012	−0.007	−0.021**	−0.014**	−0.014***
	(0.008)	(0.005)	(0.010)	(0.007)	(0.005)
JAMA	0.022	−0.021	−0.012	−0.015	−0.063***
	(0.036)	(0.026)	(0.044)	(0.032)	(0.024)
Lancet	−0.036	0.012	−0.132***	−0.020	−0.021
	(0.039)	(0.026)	(0.049)	(0.032)	(0.023)
NEJM	−0.019	0.024	−0.026	−0.035	0.001
	(0.037)	(0.025)	(0.044)	(0.030)	(0.022)
Intercept	0.755***	0.292***	0.768***	0.746***	0.758***
	(0.040)	(0.026)	(0.050)	(0.032)	(0.024)
Observations	1,334	7,470	787	2,246	7,470
R^2	0.006	0.002	0.020	0.007	0.011
Adjusted R^2	0.002	0.001	0.014	0.005	0.011

Note: *p<0.1; **p<0.05; ***p<0.01
OLS regression with journal fixed effects and standard errors clustered at the article level. BMJ is the base category. Cols 1 and 2 give primary analyses; 3-5 give robustness checks.

F Temporal Patterns

Figure 7 plots the proportion z-statistics above each of the two thresholds by year.

Figure 7: Bias Over Time



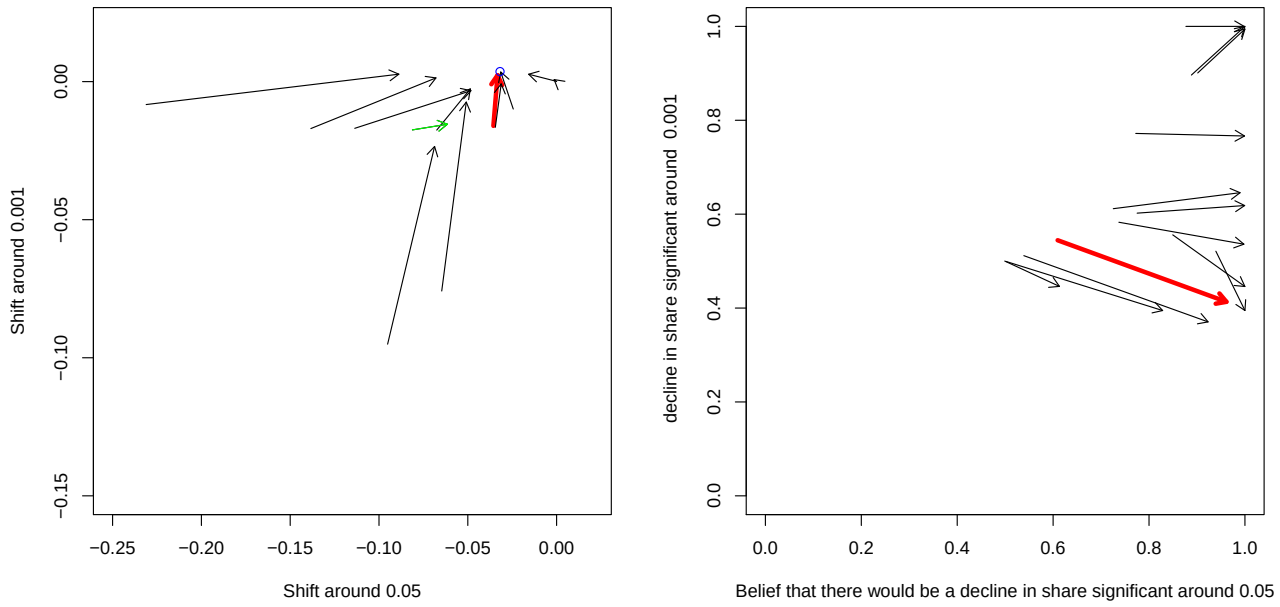
Note: Share above and below critical thresholds at each point in time. Note that there is no change associated with 2005 in any case.

G Learning, given expert priors

Figure 8 shows the changes in expert beliefs implied by the data given reported priors and under the assumption that the experts priors were not already formed by the historical process that gave rise to our data (that is, that the priors are truly priors and not posteriors).

We see two seemingly contradictory trends. On the one hand, respondents generally downgraded the expected change in the share of significant results; on the other, and particularly for the 0.05 threshold, they became *more* confident that there was a decline. These effects arise from a tightening of beliefs around an effect in the expected direction, albeit one that is smaller in magnitude than expected.

Figure 8: Bayesian Updating

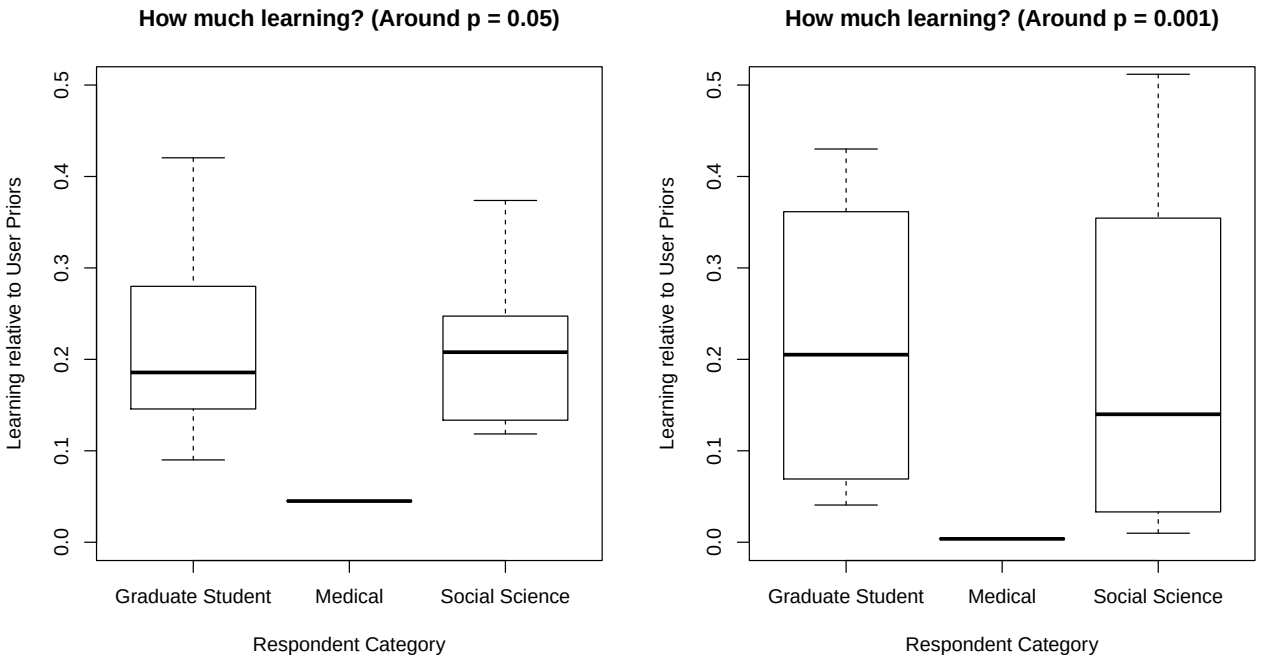


Note: Panel on left shows shift in beliefs regarding likely decline in shares significant for each test. These show general reductions in expectations around the 0.05 threshold and little change for 0.001. One case cannot be seen on this chart because it is an extreme outlier, both unusually accurate and unusually confident. Right panel shows the change the changes in beliefs that there was a decline. Generally these increase for the 0.05 case; the reason being that although for many people the expected decline is lower than they had expected, they are more confident about the decline than they were before. We mark our own learning with the red line in the figure. For two respondents there is also increased confidence in a decline around 0.001; for all of these units there was a drop in the expected size of the decline but the mean shift was small due to high levels of prior confidence.

We assess the amount of learning relative to user priors by calculating a Bayesian analogue of the R^2: using reader priors we estimate an expected error (the mean squared distance from each possible value to the empirical value, weighted by the prior probability) and compare this to the expected error under our posterior (Gelman and Pardoe, 2006). Results are given in Figure 9.

Note that only one medical researcher completed this survey item. With the exception of a sole medical researcher respondent, all groups we see an approximate 10% - 20% reduction in error for both models.

Figure 9: Learning



Note: Reduction in error between expectations using user priors and expectations from our model as measured using a Bayesian analogue of the R^2 statistic. Note that only one medical researcher responded to this item.

H Expert Beliefs on Probative Value

Figure 1 reports responses to our survey question regarding likely updating in response to positive or negative test results and relates them to a classification of tests that reflects the sensitivity or specificity of the test for the underlying hypothesis. In doing so we highlight the distinction between inferences around a particular test and inferences regarding the underlying hypothesis that the test seeks to inform, which is the quantity that is ultimately of interest. To illustrate, one might test positive for a given disease but posterior beliefs about whether one actually has the disease depends on the specificity and sensitivity of the test as well as prior beliefs.

In our case we do not assess a repeatable test and so the sensitivity and specificity of the tests are not known. For such problems qualitative researchers have emphasized inferences based on subjective beliefs about event probabilities. They provide a useful labeling of the types of test that are implied by the search for different types of evidence (Van Evera, 1997). For example a "smoking gun test" seeks information that may be unlikely to be observed if the proposition is true but is *very* unlikely to be observed if the proposition if false; for such a test a positive result provides strong evidence in favor of the proposition but a negative result provides only weak evidence against. In contrast, a "hoop" test seeks information that is likely to be observed even if the proposition is false but is *very* likely if the proposition is true: a positive result gives only modest support for a proposition, a negative result speaks strongly against. A "doubly decisive" test is both hoop and smoking gun and can shift beliefs strongly no matter what is found; a "straw in the wind" test does not discriminate as strongly between propositions and results in moderate updating in one or other direction. Most experts saw the test around $p = 0.05$ as a straw in the wind test, suggesting moderate updating following results; more respondents (nearly two thirds) saw the test around $p = 0.001$ as uninformative, with those that considered it informative classifying it closer to a smoking gun test, that is, specific but not sensitive.

Note that we can use our data to assess whether the experts responding are thinking like Bayesians. If the expert respondents behave like Bayesians, then we should observe the following two correlations. Those with a stronger prior belief that registration reduces bias would have a smaller reduction in confidence in registration effects in the case of a null result, relative to those with a weaker prior belief in registration effects. Similarly those with a weaker prior belief that registration reduces bias should have a larger increase in confidence in registration effects, relative to those with a stronger prior belief in registration effects in case of a positive result. In other words, in both scenarios there should be a negative correlation between the strength of prior beliefs in registration effects and the expected magnitude of change in confidence in registration effects. Table 6 presents the frequencies of respondents from the expert survey by prior and their self-reported posteriors given potential findings for each test.

Somewhat surprisingly, we find that respondents from the expert survey do not form expectations that are consistent with Bayesian updating. For the test at $p = 0.05$ (see panel A) there

Table 6: Are our experts' expected interpretations consistent with Bayesian updating?

	Self-reported interpretation if we:					
	Fail to reject null			Find a positive result		
Expert's prior that registration reduced publication and/or analysis bias	*Weakened* confidence that registration can tackle bias			*Strengthened* confidence that registration can tackle bias		
	Greatly	Moderately	None	Greatly	Moderately	None
A. For a test at the p=0.05 cutoff						
Likely that it has	0	7	2	3	5	1
Somewhat likely that it has	2	14	2	1	17	0
Very unlikely that it has	2	1	5	0	4	4
	correlation: 0.08			*correlation:* 0.49		
B. For a test at the p=0.001 cutoff						
Likely that it has	1	2	6	3	2	4
Somewhat likely that it has	0	10	7	6	9	3
Very unlikely that it has	1	2	5	0	4	4
	correlation: -0.04			*correlation:* 0.17		

Note: Data are from the expert survey. Cells contain the number of respondents by stated prior and stated interpretation given potential findings. Below each quadrant we show paired c orrelations between expert priors and self-reported interpretations. To calculate these correlations, expert priors are coded 1=Likely, 0.5=Somewhat likely, 0=Very unlikely, and interpretations are coded 1=Greatly, 0.5=Moderately, 0=None. We expect those with strong priors in favor of registration to be most likely to change beliefs in light of negative evidence, and least likely to change in light of positive evidence. Thus we expect a positive correlation in the first column and a negative correlation in the second.

is a positive correlation between the direction of expert priors and the direction of updating regardless of whether we find a null result or a positive result; this correlation seems to be stronger under the condition if there is a positive result than if there is a null result. For the test at $p = 0.001$ (see panel B) we find a positive correlation between the direction of expert priors and the direction of updating if we find a positive result and a weakly negative correlation if we find a null result.[17] These findings suggest the possibility of confirmation bias among the sample in the expert survey.

[17]If we recode these factor variables as numeric variables such that the values for expert priors are coded 1=Likely, 0.5=Somewhat likely, 0=Very unlikely, and the values for the interpretations are coded 1=Greatly, 0.5=Moderately, 0=None, then we observe correlations between expert priors and the expected direction of updating to be 0.49 for the test at $p = 0.05$ given a positive result and 0.08 for the test at $p = 0.05$ given a null result. We obtain a correlation of 0.17 given a positive result and a correlation of -0.04 given a null result for the test at $p = 0.001$.

I Description of deviations from the Pre-Analysis Plan

We registered our pre-analysis plan at Experiments in Governance and Politics on December 31st, 2014 (Link to *Non-Anonymous* Pre-Analysis Plan: [LINK] Table 7 documents deviations from the plan as well as a set of clarifications.

Table 7: Pre-Analysis Plan

	Analysis Plan	**Inconsistency / Clarification**
Observational analysis	Caliper specification	1. Clarification: There was an inconsistency in the caliper specification in the analysis plan documents. In most instances a caliper radius of 0.3 was specified. In one prominent place we erroneously wrote 0.03. The 0.3 caliper radius was used for analysis (though results are provided with multiple calipers).
Expert survey analysis	Primary analysis	1. Inconsistency: We specified that we would stratify results by sampling strategy. We do not do this due to the small sample of experts.
	Bayesian analysis	2. Inconsistency: We specified that we would use a grid approximation procedure. Instead, because of underflow problems, we implement Bayesian analysis using stan. 3. Clarification: The expression used for the mapping from expert inputs (m_0, m_1, σ, ρ) to priors should have been written: $\begin{pmatrix} \pi_1^0 \\ \pi_1^1 \end{pmatrix} \sim \text{logit}^{-1}\left(\text{N}\left(\begin{bmatrix} 4(m_0-.5) \\ 4(m_1-.5) \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)\right)$ and not: $\begin{pmatrix} \pi_1^0 \\ \pi_1^1 \end{pmatrix} \sim \text{logit}^{-1}\left(\text{N}\left(\begin{bmatrix} m_0 \\ m_1 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)\right)$ The transformation of the mean is used to keep the mean of the priors close to m_0 and m_1. This transformation was provided in advance to survey respondents via an online app. With these transformation the values we provided in advance of $m_0 = 0.8$ and $m_1 = 0.75$ which yields an expected difference of -0.036 for the 0.05, $m_0 = 0.33$ and $m_1 = 0.30$ which yields an expected difference of -0.0158 for the 0.001 test.

Note: This table notes any deviation from the planned analyses.