Does Registration Reduce Publication Bias? Evidence from Medical Sciences

Registration Document

Albert Fang, Grant Gordon, and Macartan Humphreys
Department of Political Science, Columbia University
December 31, 2014

1 Motivation

There is broad recognition that reporting and publication biases operate at scale in academic scholarship.\textsuperscript{1} For example, published results in our field, political science, are much more likely to report \textit{p} values just below the 0.05 critical value than just above it (Gerber and Malhotra \textsuperscript{2008a}). This could be due to publication practices or to researcher degrees of freedom that allow them to select models that produce significant results. There are discussions inside the discipline, and more broadly in social sciences, about whether the introduction of a registration system like that prevailing in medicine would mitigate these effects, especially effects associated with researcher degrees of freedom. The basic idea is that if researchers precommit to a particular specification then they have less scope to (deliberately or inadvertently) report disproportionately results from analyses that yield significant findings (see for example Casey, Glennerster, and Miguel \textsuperscript{2011}; Humphreys, de la Sierra, and van der Windt \textsuperscript{2013}).

Though the arguments for registration are simple and strong, there is surprisingly little evidence that it makes any difference. Do registration requirements really change reporting norms and statistical practices? To contribute evidence to the debate, we seek to assess levels of bias before and after critical registration dates for journals that required and did not require registration. As such, this study aims to provide a systematic assessment of the impact of registration requirements on ‘critical value’ publication bias.

The study faces two challenges however. We seek to understand the effects of registration on prospective experimental research yet our own analysis uses observational historical data. The fact that the data is observational makes it difficult to establish causal effects. We focus here on assessing whether patterns of published statistics are less indicative of bias after those journals adopted norms than before. Since the data is historical it is hard to pre-register our analysis in a credible way and to ensure that results we report are truly tests or are themselves fished results. These features generate interpretational difficulties for any results we find. We hope to gain clarity regarding what is learned from our analysis by surveying prior beliefs from readers about publication practices and what we should expect to see if registration did or did not have a causal effect.

2 Registration

In 2005, the International Committee of Medical Journal Editors (ICMJE) started requiring that all clinical trials be registered in order for results to be considered for publication.\textsuperscript{2} Researchers would be obliged to register their trials and publicly catalog information related to the design, treatment, and empirical strategy of a given project. After two years, the ICMJE conducted a review of registration and noted that while registration “precipitated much angst” in the field, it had ultimately been quickly and widely adopted.\textsuperscript{3} Of the four journals for which we have \textit{p} value data, three adopted the norm together while the fourth, the British Medical Journal, adopted the requirement separately (due to a disagreement over the criteria for a registry to be recognized \textsuperscript{Abbasi 2004}).

Registration may have many benefits. It might counter publication bias by ensuring that there is a record of studies that are undertaken, even if they are not ultimately published. In addition by introducing analytic transparency it may also affect what gets reported in any given study. For example it becomes difficult to report significant effects that are found in some subpopulation if \textit{ex ante} the study was not focused on that subpopulation in particular.

3 Hypothesis

This project seeks to assess whether medical registration reduced ‘critical value bias’ in ICMJE journals that adopted registration requirements. We define critical value bias as the extent to which the mass of \(p\)-values around critical thresholds \(\alpha = 0.001\) and 0.05 is non-uniform; this is discussed further in section 4. To examine the effect of this change, we forward the following hypothesis:

\[ H_1 \text{ There is less critical value bias in medical publishing in journals after medical registration requirements were implemented than before they were implemented.} \]

4 Strategy to Assess Bias and Changes in Bias

We propose two simple tests that build on the same strategy to contribute evidence to this debate. 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 around a critical threshold (such as \(z = 1.96\)) and use a simple binomial test to assess whether more published statistics are above the thresholds than we would expect from a chance process, which we refer to as “critical value bias.” The key idea behind the GM caliper test is the following. 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, 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. This can be implemented exactly using a binomial test.

We leverage this test and assess whether the share just above the 0.05 threshold is lower post 2005 than pre 2005 in the set of journals that adopted the registration norm in 2005. Thus rather than implementing a simple binomial test we implement a Fisher test of differences in proportions. Moreover as we describe below we implement a regression-based analogue of this that allows us to take account of journal fixed effects, time trends, and article-level clustering of errors.

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 and assess whether this proportion is different before and after 2005. This second test has the advantage of engaging more data — and much more of the relevant data for medical studies — but has the disadvantage that there is no expectation of an even distribution around any cutoff; thus unevenness does not imply bias necessarily and a reduction in unevenness does not imply a reduction in bias. Again we employ both a regression-based approach and a proportions test to assess our hypothesis.

With respect to both strategies, for test \(j\) let \(\pi_0^j\) denote the probability of observing a \(p\) value in a given range below some critical threshold prior to 2005, and \(\pi_1^j\) the probability afterwards. Let \(\delta_j = \pi_1^j - \pi_0^j\) denote the difference between time periods. We are interested in \(\delta_j\), and hypothesize that \(\delta_j\) is negative.

4.1 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 \{0.05, 0.1, 0.15, ..., 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 across studies we may expect sufficient variation in data structures 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 fishing. 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 level of the article.

Finally we note that the GM 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 as 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.

5.1 P-Values Data


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 cumulative normal. Second, we use filters to focus on values generated from 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: ‘Randomized Controlled Trial’, ‘Randomised Controlled Trial’, ‘Experiment’, ‘Randomized’, ‘Randomised’, ‘Randomly Assigned’. 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 .2 was generated. Thus, while predictive, it is not a perfect filter.

4For a discussion of this and of more powerful poisson tests, see Krishnamoorthy and Thomson (2004).

5Note that we exclude the American Journal of Epidemiology from this dataset, which never implemented registration requirements.
5.1.1 Treatment Period Definition

In light of ambiguities about how the gradual roll-out of registration requirements correspond to article-level registration behavior in mid-2005, we exclude 2005 from the analyses and define the post-registration period 2006-2010 and the pre-registration period as 2000-2004. Doing so minimizes the chance of bias resulting from two-way treatment crossover.

5.1.2 Known Issues with 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 nor will general relationships. There are several known issues with these data which were raised in the January 2014 issue of Biostatistics containing JL’s article on the science-wise FDR was published. 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. 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. For further discussion of the data, see Biostatistics (2014), volume 15, issue 1.

5.2 Expert Survey

We also will gather data from a convenience sample of experts regarding their priors on the effects of registration as well as their interpretations of patterns in the data. The survey instrument may be found at http://tinyurl.com/p-priors. Responses from this survey will be analyzed to assess expert’s priors over the effect of registration as well as the sensitivity of our test.

This survey is implemented across two samples. First, we encourage 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. From the 142 editorial board members, we were able to recover contact information for 128 individuals, or 91% of the universe. We will distribute our online survey to these individuals. In addition to this sampling strategy, we also draw on a convenience sample of medical researchers and social scientists that are encouraged through blog and related postings online. We will stratify analysis by sampling strategy.

Note that we expect very low response rates for the expert sample and will not try to correct for missing data. Indeed, we view this analysis speculative. We do specify our expectations in advance however: We believe that medical experts expect that registration generates little or no reduction in publication bias and that social scientists expect a small effect.

---

6 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.”

7 For a broad critique see (Gelman and O’Rourke, 2014)

8 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.
5.3 Researcher exposure to data

At the time of registration we have not yet analyzed this data except to produce the marginal distributions shown in Table 2. In particular, we have not produced cross tabulations of the distribution of \( p \) values above and below critical thresholds and over time.

6 Analysis

This section articulates the empirical strategy using simulated data for all specific tests and visualizations. The Jager and Leek (2014) data is only analyzed and presented in table 2.

6.1 Visualization

We first visually assess the rate of critical value bias over time in figure 1, which plots the proportion of bias over time for \( Z = 3.29, 1.96 \), using a set of narrow and wide calipers.

**Figure 1: Proportion of Bias Over Time**

Note: This figure visualizes the proportion of bias over time for \( \alpha = .001 \) and \( \alpha = .05 \). The grey line at 2005 indicates registration. This figure uses simulated data.

6.2 Analyses of Historical Data

Our primary analysis uses 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 trend and allowing for clustering at the article level, using the following equation:

\[
Y_{ij} = \beta_0 + \beta_1 T_i + \lambda z_i + \alpha_j + \epsilon_i
\]  

where \( Y \) is a binary variable that takes a value of 1 if study \( i \) in journal \( j \) reports a \( p \) value below the critical value and 0 otherwise. \( T \) is a dummy variable indicating whether the article was published after 2005, \( z_i \) is the year of article publication, and \( \alpha_j \) are a set of journal fixed effects. \( \beta_1 \) is the coefficient of interest that captures the effect of registration regime on critical value bias. Standard errors are clustered by article for the primary regression analysis. Table 1 presents results that illustrates this analysis using simulated data.

A second, more simple analysis uses a Fisher exact test for equality of proportions, assuming independence of \( p \) values. To assess the effect of the change in registration regime on the distribution of reported \( p \) values we employ two simple test statistics, shown as \( d_1 \) and \( d_2 \) in table 2. Here \( d_j \) is the difference in the share of \( p \) values below the critical value in question before and after 2005. Statistic \( d_1 \) is calculated using data for which \( z \in [1.66, 2.26] \) with the critical value given by \( z = 1.96 \) (corresponding to \( p = 0.05 \)). This range corresponds to a local caliper; the test is repeated globally as well. Statistic \( d_2 \) is calculated using all data, with the critical value given by \( p = 0.001 \).

The Fisher test is then equivalent to testing the hypothesis that \( \delta_j = 0 \). Although our hypothesis is one-sided we will report results from a two-sided test using \( \alpha = 0.05 \) as is standard in the literature. Graphically, we will show the probability that a statistic published is in the range \( [z, z + c] \) given that the statistic is in the range \( [z - c, z + c] \).
Table 1: Registration Effects (Simulated Data)

<table>
<thead>
<tr>
<th></th>
<th>(3.29)</th>
<th>(1.96)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>.001</td>
<td>.05</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.874***</td>
<td>0.458***</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>Treatment</td>
<td>−0.123**</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.057)</td>
</tr>
<tr>
<td>Year</td>
<td>0.015*</td>
<td>−0.005</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Journal FE</td>
<td>✓ ✓</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1,001</td>
<td>1,632</td>
</tr>
<tr>
<td>R²</td>
<td>0.007</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Note: * significant at p < .10; **p < .05; ***p < .01. Standard errors clustered at the article level.

where \( z \) is any given critical value and \( c \) is a caliper. To form a confidence interval around the probability at each \( z \) given \( c \), we conduct a binomial test estimating \( Pr(n_{\text{over}} \leq X) \) for \( X \sim \text{Binomial}(N = n_{\text{over}} + n_{\text{under}}, p') \), where \( n_{\text{over}} \) is the number of observed \( z \)-statistics in \([z, z + c]\) and \( n_{\text{under}} \) is the number of observed \( z \)-statistics in \([z - c, z]\), for all values of \( p' \) ranging from 0 to 1 in 0.005 intervals. We then identify the \( p' \)'s for which the \( p \)-value from the binomial test is greater than 0.05.

Table 2: Is reporting different before and after 2005?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A Critical Value Bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p ≤ .05 (within caliper)</td>
<td>?</td>
<td>?</td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>p &gt; .05 (within caliper)</td>
<td>?</td>
<td>?</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>158</td>
<td>173</td>
<td>331</td>
<td></td>
</tr>
<tr>
<td>Share ≤ 0.05</td>
<td>?</td>
<td>?</td>
<td>0.69</td>
<td>( d_1 = ? )</td>
</tr>
<tr>
<td>B Global bias:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p ≤ .001</td>
<td>?</td>
<td>?</td>
<td>2356</td>
<td></td>
</tr>
<tr>
<td>p &gt; .001</td>
<td>?</td>
<td>?</td>
<td>5114</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3598</td>
<td>3872</td>
<td>7470</td>
<td></td>
</tr>
<tr>
<td>Share ≤ 0.001</td>
<td>?</td>
<td>?</td>
<td>0.32</td>
<td>( d_2 = ? )</td>
</tr>
</tbody>
</table>

Note: What are \( d_1 \) and \( d_2 \)? Marginal data taken from the Jager-Leek dataset. Only \( z \)-statistics corresponding to articles identified as RCTs are included. Values for \( d_1, d_2 \) in the ranges \([-0.66, 0.60]\) and \([-0.61, 0.65]\) are consistent with the marginals respectively.

6.3 Analysis of Expert Survey Data

In addition to these core analysis of the historical data we undertake two analyses of data derived from an expert survey. First we gather user priors on the impact of registration from an online survey and assess how much is learned from our analysis relative to reader priors. We then form a posterior \( \pi_0 \) and \( \pi_1 \) employing the following as a prior for \( \alpha = .05 \) (test 1) and \( \alpha = .001 \) (test 2 respectively):

\[
\left( \begin{array}{c} \pi_0^0 \\ \pi_1^0 \end{array} \right) \sim \text{logit}^{-1} \left( \text{N} \left( \left[ \begin{array}{c} 0.80 \\ 0.75 \end{array} \right], 2 \left[ \begin{array}{c} 1 \\ 0.6 \\ 1 \end{array} \right] \right) \right)
\]

\[
\left( \begin{array}{c} \pi_0^1 \\ \pi_1^1 \end{array} \right) \sim \text{logit}^{-1} \left( \text{N} \left( \left[ \begin{array}{c} 0.33 \\ 0.31 \end{array} \right], 2 \left[ \begin{array}{c} 1 \\ 0.6 \\ 1 \end{array} \right] \right) \right)
\]

Under these priors, the mean of the distribution of \( \delta_1 \) is approximately −.035 for \( \delta_1 \) and −0.015 for \( \delta_2 \). We then estimate a posterior on \((\pi_0, \pi_1)\) using Bayes rule, implemented over a fine grid.

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 and compare this to the expected error under our posterior (see
We report results for three categories of respondents, data permitting: sample medical researchers, non-medical researchers, and a convenience sample. Figure 2 illustrates. Second, for respondents that are not regular readers of medical journals, we estimate our results using their priors rather than ours.

6.4 Sensitivity Tests: Simulated Data

We conduct a series of sensitivity tests to assess the robustness of our results. To examine whether our results are driven by the size of the calipers selected, we increase their size and re-analyze the results. Results that assess the effect of registration using these larger calipers on simulated data are reported in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>(3.29)</th>
<th>(1.96)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(3.95, 3.29)</td>
<td>(2.12, 1.96)</td>
</tr>
<tr>
<td></td>
<td>[3.29, 2.63]</td>
<td>[1.96, 1.81]</td>
</tr>
<tr>
<td></td>
<td>(P = .001)</td>
<td>(P = .05)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.766***</th>
<th>0.432***</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.128***</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Year</td>
<td>0.015**</td>
<td>-0.005</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>Journal FE</td>
<td>(\checkmark)</td>
<td>(\checkmark)</td>
</tr>
<tr>
<td>(N)</td>
<td>1,683</td>
<td>3,017</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.007</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Note:** * significant at \(p < .10\); ** \(p < .05\); *** \(p < .01\).

Standard errors clustered at the article level.

6.5 Illustration of results from simple tests

Implementing the Fisher test and the Bayesian model on all possible realizations of the data for \(d_1\) and \(d_2\) provides results as shown in Figure 3.
Figure 3: Test results from all possible realizations of $d_1$. Black circles show estimated difference in means, these are shaded when the null of $\delta = 0$ is rejected. Red circles show posterior estimates of $\delta^1$, shaded when the credibility interval excludes zero. (Cell(1,1) refers to the number of units with low $p$ pre-2005; all other elements of the distribution can be calculated if this cell’s contents are known)
References


