Mandating Access: Assessing the NIH's Public Access Policy

Staudt, Joseph

Ohio State University

November 2017

Mandating Access: Assessing the NIH's Public Access Policy*

Joseph Staudt†

November 2017

Abstract

In 2008, the National Institutes of Health (NIH) mandated that the full text of NIH-supported articles be made freely available on PubMed Central (PMC) – the largest and most commonly used repository of biomedical literature. This paper examines how this "PMC mandate" impacted publishing patterns in biomedicine and researcher access to the biomedical literature. Using ~ 1 million NIH articles and several matched comparison samples, I find that NIH articles are more likely to be published in traditional subscription-based journals (as opposed to "open access" journals) after the mandate. This indicates that the mandate did not induce widespread discrimination, by subscription-based journals, against NIH articles. I also find that the mandate did not increase the number of forward citations to NIH articles published in subscription-based journals. This is consistent with researchers having widespread access to the biomedical literature prior to the mandate, leaving little room for the mandate to increase access.

Keywords: economics of science, open access, nih, nih public access policy, policy evaluation.
JEL Classification Numbers: 031, 034, 038

*I gratefully acknowledge financial support from NIH Grant P01 AG039347. I would also like to thank Bruce Weinberg, David Blau, Daeho Kim, Richard Steckel, Robert Munk, Garrett Senney, Wei Yang Tham, Wei Cheng, Elisabeth Perlman, Martha Stinson, and workshop participants at Ohio State University and the U.S. Census Bureau for many helpful comments and discussions. All remaining mistakes are my own.

†Ohio State University, Department of Economics, Arps Hall 410, 1945 N. High St., Columbus, Ohio 43210.

1 Introduction

Economists have long argued that scientific advancement is crucial for economic growth. Since science is a cumulative process (Aghion et al., 2008; Aghion and Howitt, 1992; Mokyr, 2002; Murray et al., 2009; Romer, 1990; Scotchmer, 1991), its advancement depends on researchers having broad access to the scientific literature. This insight has facilitated the increasingly common practice, by departments, universities, and funding agencies, of mandating that affiliated scientific articles be made open access – "digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber, 2012, p. 4).[1] In keeping with this trend, the National Institutes of Health (NIH) implemented an open access mandate in 2008, which stipulates that the full text of any NIH-funded article must be made freely available, within 12 months of publication, on PubMed Central (PMC) – the largest and most commonly used full-text repository of biomedical research. Since medical research is estimated to have large economic benefits (Murphy and Topel, 2003) and the NIH is the largest funder of biomedicine in the world (Chakma et al., 2014), the PubMed Central (PMC) mandate is arguably the most consequential open access mandate ever enacted. This paper examines how the NIH's PMC mandate impacted publishing patterns in biomedicine as well as how it impacted researcher access to the biomedical literature.[2]

I first examine how the PMC mandate impacts the likelihood that NIH articles are published in open access journals relative to traditional subscription-based (toll access) journals. To the extent that authors value "open science", the PMC mandate reduces one of the main costs of publishing an NIH article in a toll access journal – restricted access for potential readers and citers – by ensuring that these articles are freely available on PubMed Central. Thus, the mandate may induce authors of NIH articles to submit to toll access journals at higher rates than authors of non-NIH articles. From a journal's point-of-view, the PMC mandate reduces proprietary control over NIH articles – again, by making them freely available on PubMed Central. Since proprietary control is necessary to earn subscription revenue, the PMC mandate may have made toll access journals less inclined to publish NIH articles.[3]

[1] See the Registry of Open Access Repositories Mandatory Archiving Policies (ROARMAP) for a list of open access mandates.

[2] Contemporaneous work by Bryan and Ozcan (2016) is the only other article (to my knoweldge) that examines how the PMC mandate impacts access to the biomedical literature. They are primarily interested in how the mandate impacted citations from patents to the biomedical literature. They find that patent-to-article citations increase by between 25 to 51 percent. They find that article-to-article citations do not significantly increase, which is consistent with my findings. More broadly, the current paper fits into an "open science" literature that examines the extent to which restrictions on access to scientific literature and materials can impede scientific progress (Furman and Stern, 2011; Murray et al., 2009; Sampat and Williams, 2015; Williams, 2013)

[3] It is notable that the Association of American Publishers, which represents all major publishers in biomedicine, strongly opposed the PMC mandate, specifically warning that it would undermine its members'

However, if journal editors (rather than owners) make publication decisions, this disinclination may be attenuated. Moreover, the NIH's lack of enforcement until 2013 (Van Noorden, 2014) may have allowed journals to publish NIH articles without fear of losing control over journal content. Thus, the net effects of the PMC mandate will depend on the relative strengths of these "author-side" and "journal-side" effects.

To assess these effects, I first identify all ($\sim$ 1 million) NIH-funded articles published between 2003 and 2013 in the MEDLINE database – the largest repository of biomedical literature in the world. Second, in order to pin down counterfactual outcomes, I construct several sets of comparison articles that are not NIH-funded. Third, I use the Directory of Open Access Journals (DOAJ) to determine whether an article is published in an open access or toll access journal. Finally, I use difference-in-differences (DID) and propensity score methods to examine the impact of the PMC mandate on the probability that an article is published in a traditional toll access journal.

Most results indicate that the PMC mandate increased the probability that an NIH article is published in a toll access journal by 0-3 percentage points, with the more credible estimates at the upper end of this range. This is consistent with author-side effects dominating journal-side effects. At the very least, there is no evidence that the PMC mandate induced widespread discrimination against NIH articles by toll access journals, which may be due to editors, not owners, making publication decisions.[4] To further check for evidence of discrimination, I examine whether the mandate changed the probability that an NIH article is published in the journals *Science* or *Nature*. Given their policies against publishing articles available elsewhere (such as PubMed Central), evidence of discrimination, if it exists, should be most pronounced with these journals. The evidence is ambiguous, but weakly suggests that journal-side effects may dominate for particular subsets of journals.

Next, I examine whether the PMC mandate increased researcher access to the biomedical literature. If, prior to the mandate, some researchers had difficulty accessing NIH articles, then, after the mandate, we should observe an increase in the rate at which NIH articles are

economic incentives by making their content available online (www.publishers.org/issues/5/9). At least one member of the publishing industry explicitly suggested that "Another possible implication (of the PMC mandate) is that journals may no longer be willing to review and accept articles with unsustainable terms attached" (McMullan, 2008). Though prior to the PMC mandate, Seamans (2001) found that, in a sample of mostly non-profit journals, 17.64 percent expressed reservations about accepting submissions of theses and dissertations available on the web. Howard (2011) documents several university press editors' reluctance to publish theses and dissertations that can be found "immediately on Google or by going to the university page and just clicking and downloading it...". Finally, as noted by (Suber, 2012, p. 173), medicine is a field particularly likely to follow the "Ingelfinger Rule" and refuse to accept articles that have circulated online.

[4]It is worth noting that the "Cost of Knowledge Protest" by academics against Elsevier's journals began in 2012. Since the protest began near the end of my sample period (2003-2013), it is unlikely that it substantially impacts my main results. However, to the extent that the protest does impact my results, it likely makes my estimates smaller than they would otherwise be.

cited in follow-on research. In particular, there should be an increase in citations to NIH articles published in toll access journals, which prior to the mandate (and unlike NIH articles published in open access journals), would have been unavailable to researchers without a subscription to the journal.

Using triple differences (DDD) and propensity score methods, I find that the PMC mandate did not increase overall access to NIH articles. This is consistent with recent work showing that making articles open access does not increase citations to those articles.[5] I also examine how citations by researchers in poor/developing countries and researchers affiliated with commercial enterprises are impacted by the PMC mandate. While researchers affiliated with universities or research hospitals in rich countries may have access to the entire biomedical literature, researchers in poor countries or at commercial enterprises (especially start-ups on a shoestring budget) may have more limited access in the absence of the mandate (Ware and Monkman, 2009; Houghton et al., 2011). However, even for these subgroups, I find little evidence that the mandate substantially increased access. It is important to note, however, that these results do not indicate whether the mandate increased access for other consumers of the biomedical literature such as doctors or inventors.

Given the non-rivalrous nature of ideas, ideas' importance for economic growth, and the cumulative nature of science, it is crucial to understand the impacts of restrictions (and the lifting of restrictions) on access to scientific research. This is especially true for biomedical research, which is estimated to yield large returns. This paper analyzes how the abolition of a particular set of restrictions (the PMC mandate) impacts publishing patterns in biomedicine and access to the biomedical literature.

Moving forward, the paper is organized as follows. Section 2 describes the data. Section 3 discusses the details of the PMC mandate, my econometric strategy, and the plausibility of the identifying assumptions underlying this strategy. Section 4 presents the results, and Section 5 concludes.

[5]Though the best evidence does not suggest that open access increases citations, early studies find a large impact. Lawrence (2001) paved the way using a sample of articles from computer science conferences. This finding is replicated in Antelman (2004) and Davis and Fromerth (2007) for mathematics, by Antelman (2004) for philosophy, political science, and engineering, by Schwarz and Kennicutt Jr (2004) and Metcalfe (2005, 2006) for astrophysics, by Harnad and Brody (2004) for physics, by Walker (2004) for oceanography, and by Eysenbach (2006) for multidisciplinary science. Craig et al. (2007) provide a useful review of this early literature. Since these early studies all use observational cross-sections, it is difficult to draw reliable causal conclusions from their results. A more recent literature, that explicitly attempts to account for the endogeneity of open access, has found much more modest effects. Using panel data, Evans and Reimer (2009) and McCabe and Snyder (2014) find that open access increases citations by approximately 8 percent. Using an instrumental variables strategy, Gaule and Maystre (2011) do not find a statistically significant impact of open access on citations. Finally, using evidence from randomized controlled trials, Davis et al. (2008) and Davis (2011) fail to find a statistically significant increase in the number of citations to open access articles.

2 Data

The data used in this paper are obtained from three main sources: MEDLINE, Web of Science (WOS), and the Directory of Open Access Journals (DOAJ). MEDLINE is a bibliographic database maintained by the National Library of Medicine and is the most comprehensive index of the biomedical literature.[6] It includes a large amount of information about each indexed article, including the year and journal in which it is published and any grants that support it (in particular, NIH grants). WOS is maintained by Clarivate Analytics' and indexes the references of MEDLINE articles. It enables the tracking of citation relationships between MEDLINE articles. DOAJ is the most comprehensive index of open access journals across all fields, and it enables the labeling of journals in MEDLINE as open or toll access. For more details on all data, see Appendix A.

2.1 Outcome Variables

The first set of outcomes are designed to measure how the PMC mandate impacted publishing patterns in biomedicine. I first examine whether an article is published in an open access or toll access journal. A journal is identified as as open access if it is indexed by the Directory of Open Access Journals (DOAJ). Otherwise, it is classified as toll access.[7] Next, I examine whether an article is published in the journals *Science* or *Nature*, which are identified using the unique journal identifier in MEDLINE (see Appendix A). The time frame of analysis for these two variables is 2003-2013 – 5 years before and after the 2008 PMC mandate.

The second set of outcomes are designed to measure whether the PMC mandate increased access to the biomedical literature. I first examine the total number of 2-year forward citations that an article receives. Next, I examine the number of 2-year forward citations that an article receives from particular subsets of authors. First, I restrict the count of 2-year forward citations to those received by articles with authors at commercial enterprises. Second, I restrict the count to those received by authors at institutions located in poor or

[6] Technically, MEDLINE is a subset of a larger database called PubMed (distinct from PubMed Central). However, the data in MEDLINE have undergone rigorous quality control and are readily available for use by researchers.

[7] For several reasons, this classification scheme is imperfect. First, some journals in MEDLINE may be open access, but are not indexed in DOAJ. Such journals will be falsely classified as toll access. Second, a journal may be open access (or toll access) at one point in time, and then change status. Such journals will be correctly classified in some years and falsely classified in others. Third, some journals are neither fully open nor fully toll access. For instance, some journals allow authors to pay a fee to make their article open access in an otherwise toll access journal. Other journals allow some forms of "green" open access, which allow researchers to put non-final versions of their articles on personal webpages or repositories like PubMed Central. Still other journals embargo articles for a time and then open them to all researchers. Unfortunately, the DOAJ data do not allow me to address these nuances.

developing countries.[8] Since the WOS citation data end in mid-2014, I end the analysis for 2-year forward citation measures in 2011, which ensures that all articles have a full 2 years to accrue citations.

2.2 NIH Articles and Comparison Articles

There are 2,050,044 articles in MEDLINE tagged as being funded by the National Institutes of Health (NIH).[9] 956,801 of these are published between 2003 and 2013 (the time frame for the analysis of publishing patterns) and 745,076 between 2003 and 2011 (the time frame for the analysis of researcher access).

To pin down counterfactual outcomes, I construct several sets of comparison articles using non-NIH articles. The first comparison sample is the set of all non-NIH articles in MEDLINE published between 2003 and 2013 (2011). There are 7,482,563 (5,809,078) such articles. I refer to this set of comparison articles as the "MEDLINE" comparison sample. The second set of comparison articles is the set of all non-NIH articles published in the same journal and year as at least one NIH article.[10] There are 5,792,555 (4,455,039) such articles. I refer to this set of comparison articles as the "Journal" sample.

The third set of comparison articles is constructed using the PubMed Related Citations Algorithm (PRCA) which identifies, for any given article, a set of "similar" articles.[11] First, I use the PRCA to harvest similar articles for each NIH article. After restricting the set of harvested articles to those that are published in the 2003-2013 (2011) period and are not themselves NIH articles, there are a total of 3,171,838 (2,542,714) unique comparison articles. I refer to this set of comparison articles as the "Full PRCA" sample.

The final set of comparison articles is a subset of the Full PRCA sample. Taking advantage of the fact that the PRCA delivers a similarity score for each harvested article, I am able to identify the particular comparison article that most closely matches each NIH article and implement a 1-to-1 matching (without replacement) algorithm. I refer to this set

[8] I identify articles affiliated with a commercial enterprise using MapAffil and articles from poor/developing countries using a combination of MapAffil and the United Nations National Accounts. See Appendix A for details.

[9] MEDLINE contains two pieces of information that are used to identify NIH articles. The first is the list of grants contained in the MEDLINE element GrantList and the second is information contained in the element PublicationTypeListist. The National Library of Medicine (NLM) generates the GrantList using the grants listed by authors on an article. Thus, it is possible that some articles are NIH-supported, but are not identified as such because authors do not list the NIH grant. The PublicationTypeList is curated by librarians at NLM in the same way that MeSH terms are curated.

[10] Ideally, I would use journals published in the same journal issue. However, the journal issue element in the MEDLINE data is often missing, making this strategy infeasible.

[11] The harvested articles are obtained from PubMed, a superset of the MEDLINE database. The algorithm defines similarity using overlapping MeSH terms and text in titles and abstracts.

of comparison articles as the "1-to-1 PRCA" sample.

In sum, I construct four comparison samples in order to estimate counterfactual outcomes: The "MEDLINE", "Journal", "Full PRCA", and "1-to-1 PRCA" samples.

2.3 Covariates

The data allow me to construct a rich set of article-level covariates. These include the number of backward citations, the number of backward citations published in open access journals, the number of unique n-grams (1-, 2-, and 3-) that an article uses in either the title or abstract, the number of top n-grams an article "originates", the number top n-grams for which an article is an "early adopter",[12] the number of "Descriptor" Medical Subject Heading (MeSH) terms[13] that tag an article, the number of "Qualifier" MeSH terms that tag an article, the number of authors, an indicator for whether the author is a corporate entity, a set of indicator variables for the type of institution to which the first author belongs (e.g. university, hospital, etc.), a set of fixed effects for the country in which the first author works, a set of indicator variables characterizing an article's "Publication Type" (e.g. whether the article is a review article, a clinical trial, etc.), a set of indicator variables characterizing the languages in which the article is published, and the number of non-NIH grants that support the article. I also use the MeSH terms that tag each article to construct a variable characterizing the field of each article, which allows the inclusion of field-fixed effects in regression models (see Appendix C for details). When the outcome variable is 2-year forward citations, I also estimate specifications that include journal fixed effects. Journal fixed effects cannot be included when the outcome variable is the toll access indicator because there is no within-journal variation in this outcome. See Appendix A for details on all covariates.

2.4 Summary Statistics

Appendix Tables A.1.a through A.1.d present summary statistics for each of the four comparison samples. In the MEDLINE sample, the covariates are fairly imbalanced across the NIH and comparison groups. In both the pre- and post-mandate periods, NIH articles tend to cite more articles, use more top n-grams, are tagged with more MeSH terms, have more authors, are less likely to have a corporate authors, are more likely to be a "Journal Article",

[12] An article "originates" an n-gram if the article uses the n-gram in the n-gram's vintage year (first year the n-gram appears in the MEDLINE corpus). An article is an "early adopter" of an n-gram if the article uses the n-gram within 5 years of the n-gram's vintage. A "top" n-gram is one that, compared to all other n-grams in its vintage, is in the top 0.01% in terms of total mentions in the MEDLINE corpus.

[13] Medical Subject Heading (MeSH) terms are used to classify the content of each record indexed in MEDLINE. Librarians at the National Library of Medicine (NLM) read each article and determine which MeSH terms apply to that article.

are less likely to be an "Irregular Article", and are more likely to be published in English. Though these differences persist in the other three comparison samples, the magnitudes tend to decrease, and the covariates in the journal, full PRCA, and 1-to-1 PRCA samples become increasingly more balanced. Indeed, in the 1-to-1 PRCA sample, many of the covariates are quite similar across the NIH and comparison articles, suggesting that this is the most appropriate comparison group.

3 Research Design

3.1 Details of the NIH Public Access Policy ("PMC Mandate")

On February 3, 2005 the National Institutes of Health (NIH) issued a policy statement that requested all NIH-supported articles to be submitted to PubMed Central. This request became effective on May 2, 2005.[14] Despite the request, a 2006 NIH report to Congress revealed that voluntary compliance with this request was below 4 percent.[15] Thus, Congress instructed the NIH to change the request to a mandate – the "PMC mandate". On January 11, 2008 the NIH announced that the full text of all NIH-supported articles accepted for publication on or after April 7, 2008 were to be submitted, in final peer-reviewed form, to PubMed Central immediately upon acceptance for publication.[16] However, authors have the option to embargo their work for up to one year. By 2012 compliance stood at 75 percent.[17]

3.2 Econometric Strategy

The treatment of interest is the requirement to submit an article to PubMed Central immediately upon acceptance for publication. Depending on the type of outcome variable, I

[14]Specifically, the policy statement read: "beginning May 2, 2005, NIH-funded investigators are requested to submit to the NIH National Library of Medicine's (NLM) PubMed Central (PMC) an electronic version of the author's final manuscript upon acceptance for publication, resulting from research supported, in whole or in part, with direct costs from NIH." http://grants.nih.gov/grants/guide/notice-files/NOT-OD-05-022.html.

[15]http://legacy.earlham.edu/ peters/fos/nihfaq.htm

[16]The Public Access Policy was the NIH's response to Division G, Title II, Section 218 of PL 110-161 (Consolidated Appropriations Act, 2008), which states: "The Director of the National Institutes of Health shall require that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: Provided, That the NIH shall implement the public access policy in a manner consistent with copyright law." Note that, though the article must be submitted to PubMed Central immediately upon acceptance for publication, the author retains the option of embargoing the article for up to 12 months after publication.

[17]http://www.whitehouse.gov/sites/default/files/microsites/ostp/public_access-final.pdf

use difference-in-differences (DID), triple differences (DDD), and propensity score methods to identify the effect of this treatment.

3.2.1 Difference-in-Differences

Each article in the sample is either an NIH article ($N_i = 1$) or a comparison article ($N_i = 0$) and is either published before ($P_i = 0$) or after ($P_i = 1$) the PMC mandate. In the difference-in-differences (DID) framework, the impact of the mandate can be identified by estimating the following regression equation:

$$Y_i^{obs} = \beta_t + \gamma N_i + \delta(P_i \times N_i) + \rho X_i + \epsilon_i. \quad (1)$$

Y_i^{obs} is an observed outcome variable (for instance, a toll access indicator or an indicator for publication in *Science* or *Nature*), β_t is a full set of publication year fixed effects, P_i is an indicator that turns on in 2009, N_i indicates whether an article is NIH-supported, and X_i is a vector of article-level covariates. δ is the parameter of interest, measuring the impact of the mandate under this model.

3.2.2 Triple Differences

When the outcome variable is the count of 2-year forward citations, it can be analyzed in the DID framework above. However, articles (NIH or comparison) published in open access journals are unlikely to be impacted by the PMC mandate because any researcher can access and cite these articles, by definition, both before and after the mandate. Thus, if we observe an increase in citations to NIH articles, we should observe them accruing to articles published in toll access journals.

This observation suggests that a triple differences (DDD) framework is a more appropriate for the citation outcomes. In this setting, an article is either published in a toll access ("subscription") ($S_i = 1$) or open access ($S_i = 0$) journal, and the impact of the mandate can be identified by estimating the following regression equation:

$$Y_i^{obs} = \beta_t + \gamma_1 N_i + \gamma_2 S_i + \gamma_3 (S_i \times N_i) + \gamma_4 (P_i \times N_i) + \gamma_5 (P_i \times S_i) + \delta(P_i \times S_i \times N_i) + \rho X_i + \epsilon_i. \quad (2)$$

S_i indicates whether article i is published in a toll access journal, and the other variables are defined as in equation (1). Again, δ is the parameter of interest, measuring the impact of the PMC mandate under the DDD model.

Figure 1 assesses the common trends assumption that underlies the DID and DDD models by plotting pre-mandate trends for the two main outcome variables of interest: an indicator

for whether an article is published in a toll access journal and the count of 2-year forward citations. Since the count of 2-year forward citations is analyzed in a DDD framework, the trends of this variable are also broken down by whether an article is published in a toll or open access journal. The graphs suggest that, for both outcomes and across all four comparison samples, the common trends assumption is reasonably satisfied during the pre-mandate period.

3.2.3 Stratified Difference-in-Differences

An additional assumption underlying the DID/DDD models is that the composition of groups is constant over time (Blundell and Dias, 2009). Concern over whether this assumption holds is especially salient when the data are repeated cross-sections and different units continually enter and leave the sample (Stuart et al., 2014). To deal with the possibility of changing group composition, I use a method that combines propensity score stratification and DID/DDD regression.[18]

As a first step, I estimate the propensity scores for the DID and DDD settings, $P(N_i = 1, P_i = 1|X_i)$ and $P(N_i = 1, P_i = 1, S_i = 1|X_i)$, as function of article-level covariates. Second, I algorithmically partition the articles into strata that have similar propensity scores. Third, I estimate equations (1) and (2) *within each stratum*. Finally, I combine the estimates within the strata into composite estimates of the the impact of the mandate.

Since the propensity scores within a stratum are relatively similar, the within-stratum covariate distributions of the four DID and eight DDD groups are also relatively similar. This helps alleviate concerns about composition changes within groups over time. Any additional differences in the covariate distributions across groups are at least partially accounted for by continuing to include covariates X_i when equations (1) and (2) are estimated within each stratum. I also present estimates obtained from trimmed samples in which articles with estimated propensity scores above 0.9 and below 0.1 are dropped. These articles have covariate values at which there is very limited overlap between treated and control articles (Crump et al., 2009). See Appendix D for details on the propensity score estimation and stratification.

[18]For alternative procedures that use the propensity score in DID framework to make the constant composition assumption more plausible, see Stuart et al. (2014) or Blundell et al. (2004).

4 Results

4.1 Publishing Patterns in Biomedicine

Figure 2 displays the point estimates, along with 95% confidence intervals,[19] of the PMC mandate's impact on the probability that an NIH article is published in a toll access journal (δ in equation 1). For each of the four comparison samples, estimates are obtained with and without covariates, with and without stratification, and on the trimmed and untrimmed samples.

All estimates, across all four comparison samples, are positive. When the entire (untrimmed) MEDLINE sample is used without stratification or covariates, the estimate is quite imprecise. Adding covariates slightly increases precision, while trimming and stratifying greatly increase precision. Overall the estimates from the MEDLINE sample range from 0.004 to 0.016. The estimates obtained using the journal sample follow a very similar pattern, ranging from 0.006 to 0.018. In addition to being positive, all estimates from the two PRCA samples are statistically significant. Adding covariates and trimming does little to improve precision, but stratifying significantly increases precision. Overall, the estimates from the two PRCA samples range from 0.012 to 0.030. Thus, these estimates suggest that, after the PMC mandate, the probability that an NIH article is published in a toll access journal increases by between 0.4 to 3 percentage points.

These estimates do not support the hypothesis that the PMC mandate decreased the probability that an NIH article is published in a toll access journal. That is, there is no evidence that journal-side effects dominate. There is stronger evidence that the impact is positive – that is, author-side effects dominate. Since the 1-to-1 PRCA sample is generated in a way that increases the similarity between the NIH and comparison group, I view these estimates as the most credible. Thus, the most credible evidence suggests that, after the PMC mandate, the probability that an NIH article is published in a toll access journal increased by 2-3 percentage points.

Table 1 assesses the magnitudes of these point estimates. In 2009, the year after the PMC mandate, there were a total of 92,646 NIH-funded articles published – 5,697 (6.1%) in open access journals and 86,949 (93.9%) in toll access journals. If the mandate increased the probability that an NIH article is published in a toll access journal by 1 percentage point (effect size: 0.01), then, in the absence of the mandate, only 86,023 (92.9%) NIH articles would have been published in toll access journals. Thus, under this scenario, the mandate shifted $86,949 - 86,023 = 926$ NIH articles from open access to toll access journals in

[19] The confidence intervals are computed using standard errors clustered at the aggregated field level. See Appendix C for details on field construction.

2009. If we instead assume that the mandate increased the probability that an NIH article is published in a toll access journal by 2 or 3 percentage points, then the mandate shifted 1,853 or 2,779 articles from open access to toll acccess journals. This computation can be carried out for all years between 2009 and 2013. Overall, these estimates suggest that between 5,076 and 15,253 articles were shifted from open to toll access journals over the 2009-2013 period, which is between 10 ($= 5,076/50,981 * 100$) and 30 ($= 15,253/50,981 * 100$) percent of the NIH articles actually published in open access journals.

Figure 3 displays how these effects evolve dynamically over time. All estimates show persistent increases in the probability that an NIH article is published in a toll access journal after the mandate. Indeed, these increases are permanent over the period we analyze (which helps to justify the assumption of a permanent level shift when assessing magnitudes in Table 1). However, the increase estimated from the MEDLINE sample is very slight. It is clear that these trends are much more pronounced and precisely estimated for the two PRCA samples.

The evidence above suggests that there is no evidence of widespread discrimination, by toll access journals, against publishing NIH articles after the PMC mandate. However, it is possible that certain subsets of journals discriminated against NIH articles. To probe this possibility, I change the outcome variable in equation (1) from an indicator for publication in a toll access journal to an indicator for publication in the journals *Science* or *Nature*. These journals have strict policies against publishing an article that has already been circulated in any form. Thus, if discrimination exists, it should be observable with these journals.

As will become clear, it is useful to first examine the dynamics of this outcome variable, which are displayed in Figure 4. Across all four comparison samples, the probability of NIH articles (relative to comparison articles) being published in *Science* or *Nature* is trending upward in the pre-mandate period. Then, in the post-mandate period, this trend suddenly stops. This suggests that there are group-specific trends that are not accounted for by the standard DID model. Indeed, since the relative probability of an NIH article being published in *Science* or *Nature* is higher in the post-mandate period, a standard DID model will yield misleading positive estimates of the effect of the mandate.

To deal with this issue, I estimate both the standard DID model and a model that adds a linear time trend interacted with the NIH indicator ($N_i \times t$) to equation 1. Figure 5 displays these estimates. As expected, the two models generate conflicting results. The positive estimates from the standard DID model suggest that author-side effects dominate and the negative estimates from the linear trend DID model suggest that journal-side effects dominate. Since there are clearly group-specific trends in the outcome, and the linear trend DID model allows for such trends, I view these estimates as more credible. If correct, this

suggests that journal-effects may dominate for journals, such as *Science* or *Nature*, that have very strict policies against publishing articles that appear elsewhere. However, given the sensitivity of these results to model selection, they should be interpreted cautiously.

In sum, the evidence suggests fairly strongly that the PMC mandate did not induce widespread discrimination against NIH articles by toll access journals. Across a wide variety of samples and models, the results suggest that the mandate actually *increased* the probability that an NIH article is published in a toll access journal, most likely by 0-3 percentage points. This result is consistent with the PMC mandate increasing the accessibility of NIH articles published in toll access journals and thus inducing researchers who value "open science" to submit their work to such journals at higher rates. However, there is some less robust evidence that the mandate decreased the probability of NIH articles being published in toll access journals that have particularly strict policies against publishing articles that appear elsewhere.

4.2 Access to the Biomedical Literature

Figure 6 displays the triple differences (DDD) point estimates, along with 95% confidence intervals,[20] of the PMC mandate's impact on the count of 2-year forward citations received by an NIH article published in a toll access journal (δ from equation 2). As in the previous section, for each of the four comparison samples, estimates are obtained with and without covariates, with and without stratification, and on the trimmed and untrimmed samples.

With a few exceptions, the point estimates are positive. For instance, when the entire (untrimmed) MEDLINE sample is used without stratification or covariates, the estimate suggests that the PMC mandate increased the number of 2-year forward citations by a statistically significant 1.5. When article-level covariates are included, the estimate becomes smaller (about 0.5), more precise, and statistically indistinguishable from 0.

This pattern is common across the comparison samples and estimators – adding article-level covariates tends to increase precision and attenuate the magnitude of point estimates. When covariates are not included, the point estimates range from 0.06 to 1.93, and these are often statistically significant. However, when covariates are included, the point estimates range from -0.16 to 0.73, and are typically clustered around 0. Indeed, when covariates are included for the two PRCA samples, the estimates range from -0.16 to 0.19. Given that the mean NIH article published in a toll access journal prior to the mandate received 10.42 2-year forward citations, these effects are quite small.

[20] The confidence intervals are computed using standard errors clustered at the journal level. The results are very similar when they are clustered at the aggregated field level.

Figure 7 displays the dynamics for 2-year forward citations. Across all four comparison samples, the count of forward citations received by NIH articles published in toll access journals is slightly higher in the post mandate period. However, especially for the 1-to-1 PRCA sample, this count may have been trending upward even prior to the mandate.

Overall, these estimates suggest that the PMC mandate did not substantially increase the total number of 2-year forward citations to NIH articles published in toll access journals. This is consistent with most researchers having broad access to the biomedical literature prior to the mandate, leaving little room for the mandate to increase overall access.

Though the previous results suggest that the PMC mandate did not increase *overall* access to the biomedical literature, it is possible that particular sub-groups of researchers had limited access to NIH articles prior to the mandate, gaining access only after the mandate. If so, we should be able to detect increases in 2-year forward citations from these sub-groups to NIH articles published in toll access journals after the mandate. I examine citations from authors affiliated with a commercial enterprise and authors affiliated with an institution located in a poor/developing country. Commercial enterprises may be more sensitive than research universities or hospitals to high journal subscription costs, limiting the access of their researchers to articles published in toll access journals (Ware and Monkman, 2009; Houghton et al., 2011). This is especially plausible for start-ups, which may be on a shoestring budget. Similarly, institutions located in poor/developing countries may lack the resources to purchase broad access to the literature for their researchers.

Figures 8 and 9 display the DDD estimates when the outcome variable is 2-year forward citations from researchers affiliated with a commercial enterprise and an institution located in a poor/developing country. The patterns of the results are strikingly similar to the results for total 2-year forward citations. In particular, in specifications without article-level covariates the estimates are positive and often statistically significant. Adding covariates tends to attenuate the magnitude of the estimates (but does not greatly increase precision). The estimates for commercial enterprises range from -0.008 to 0.049 overall and from -0.008 to 0.018 when covariates are included. The estimates for poor/developing countries range from -0.021 to 0.056 overall and -0.021 to 0.022 when covariates are included. Given means of 0.204 and 0.205, the upper end estimates are quite large, but more plausible estimates are quite modest. Overall, the evidence that the PMC mandate increased access for these sub-groups of researchers to NIH articles published in toll access journals is weak.

5 Conclusion

This paper examined the impacts of the National Institutes of Health's 2008 PubMed Central (PMC) Mandate on publishing patterns in biomedicine and researcher access to the biomedical literature. Three main findings emerge from the analysis.

First, I find no evidence that the PMC mandate induced widespread discrimination, by toll access journals, against NIH articles. In contrast, the best evidence suggests that the probability of an NIH article being published in a toll access journal *increases* after the mandate. If researchers value "open science", then, all else equal, they will prefer to publish in journals that provide widespread access. Prior to the mandate, this preference may have induced researchers to publish some articles in open access journals that would otherwise have been published in toll access journals. Since the PMC mandate provided universal access to NIH articles *regardless of where they are published*, researchers no longer had to take the openness of the journal into account when deciding where to publish their work. This, along with journal editors (not owners) making publication decisions, could account for the increase the the relative proportion of NIH articles that are published in toll access journals after the mandate.

Second, there is weak evidence that journals with particularly strict policies against publishing material that has appeared elsewhere may have discriminated against NIH articles. For such journals, the incentive to maintain propriety control over journal content may be strong enough to reject NIH articles that would have been published in the absence of the mandate. This may be especially true for prestigious journals such as *Science* or *Nature* that receive many high quality non-NIH submissions.

Third, I find little evidence that the PMC mandate increased researcher access to the biomedical literature. Indeed, this holds even for subgroups of researchers that, a priori, may have been suspected to have limited access to the literature. This finding is consistent with most researchers having broad access to the biomedical literature prior to the mandate, providing little scope for the mandate to increase access. This finding is consistent with findings in Bryan and Ozcan (2016). However, it must be stressed that this result does not provide evidence on whether the mandate impacted access for doctors, inventors, or other consumers of the biomedical literature. Indeed Bryan and Ozcan (2016) find that the mandate did increase patent citations to the NIH articles after the mandate. In follow-up work, I will further explore the relationship between the PMC mandate and patent-to-article citations.

References

Aghion, P., M. Dewatripont, and J. C. Stein (2008). Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics 39*(3), 617–635.

Aghion, P. and P. Howitt (1992). A model of growth through creative destruction. *Econometrica 60*(2), 323–351.

Antelman, K. (2004). Do open-access articles have a greater research impact? *College & research libraries 65*(5), 372–382.

Blundell, R. and M. C. Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources 44*(3), 565–640.

Blundell, R., M. C. Dias, C. Meghir, and J. Reenen (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association 2*(4), 569–606.

Bryan, K. A. and Y. Ozcan (2016). The impact of open access mandates on invention.

Chakma, J., G. H. Sun, J. D. Steinberg, S. M. Sammut, and R. Jagsi (2014). Asia's ascent: global trends in biomedical R&D expenditures. *New England Journal of Medicine 370*(1), 3–6.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295–313.

Craig, I. D., A. M. Plume, M. E. McVeigh, J. Pringle, and M. Amin (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics 1*(3), 239–248.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

Davis, P. M. (2011). Open access, readership, citations: A randomized controlled trial of scientific journal publishing. *The FASEB Journal 25*(7), 2129–2134.

Davis, P. M. and M. J. Fromerth (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics 71*(2), 203–215.

Davis, P. M., B. V. Lewenstein, D. H. Simon, J. G. Booth, M. J. Connolly, et al. (2008). Open access publishing, article downloads, and citations: Randomised controlled trial. *BMj 337*, a568.

Evans, J. A. and J. Reimer (2009). Open access and global participation in science. *Science 323*(5917), 1025–1025.

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS biology 4*(5), e157.

Furman, J. L. and S. Stern (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review 101*(5), 1933–1963.

Gaule, P. and N. Maystre (2011). Getting cited: Does open access help? *Research Policy 40*(10), 1332–1338.

Harnad, S. and T. Brody (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-lib Magazine 10*(6).

Houghton, J., A. Swan, and S. Brown (2011). Access to research and technical information in denmark.

Howard, J. (2011). The road from dissertation to book has a new pothole: The internet. *The Chronicle of Higher Education*.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature 411*(6837), 521–521.

McCabe, M. and C. M. Snyder (2014). Identifying the effect of open access on citations using a panel of science journals. *Economic Inquiry 52*(4), 1284–1300.

McMullan, E. (2008). Open access mandate threatens dissemination of scientific information. *Journal of Neuro-Ophthalmology 28*(1), 72–74.

Metcalfe, T. S. (2005). The rise and citation impact of astro-ph in major journals. *Bulletin of the American Astronomical Society 37*, 555–557.

Metcalfe, T. S. (2006). The citation impact of digital preprint archives for solar physics papers. *Solar Physics 239*(1-2), 549–553.

Mokyr, J. (2002). *The gifts of Athena: Historical origins of the knowledge economy*. Princeton University Press.

Murphy, K. M. and R. H. Topel (2003). The economic value of medical research. *Measuring the gains from medical research: An economic approach 15*(30), 125–146.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern (2009). Of mice and academics: Examining the effect of openness on innovation. Technical report, National Bureau of Economic Research.

Packalen, M. and J. Bhattacharya (2015, January). Age and the trying out of new ideas. Working Paper 20920, National Bureau of Economic Research.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy 98*(5 pt 2).

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association 79*(387), 516–524.

Sampat, B. and H. L. Williams (2015). How do patents affect follow-on innovation? evidence from the human genome. Technical report, National Bureau of Economic Research.

Schwarz, G. J. and R. C. Kennicutt Jr (2004). Demographic and citation trends in astrophysical journal papers and preprints. *Bulletin of the American Astronomical Society 36*, 1654–1663.

Scotchmer, S. (1991). Standing on the shoulders of giants: Cumulative research and the patent law. *The Journal of Economic Perspectives*, 29–41.

Seamans, N. (2001). Electronic theses dissertations: 2001 survey of editors and publishers.

Staudt, J., H. Yu, R. P. Light, G. Marschke, K. Borner, and B. A. Weinberg (2017). High-impact and transformative science (HITS) metrics: Definition, exemplification, and comparison. Working paper.

Stuart, E. A., H. A. Huskamp, K. Duckworth, J. Simmons, Z. Song, M. E. Chernew, and C. L. Barry (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology 14*(4), 166–182.

Suber, P. (2012). *Open access*. MIT Press.

Torvik, V. I. (2015). Mapaffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. In *D-Lib magazine: the magazine of the Digital Library Forum*, Volume 21. NIH Public Access.

Van Noorden, R. (2014). Funders punish open-access dodgers. *Nature*.

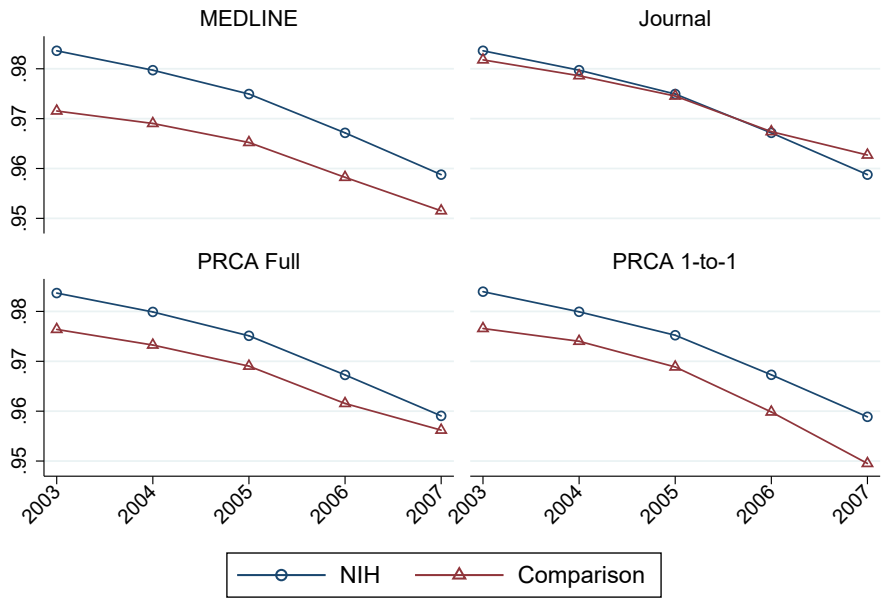
Walker, T. (2004). Open access by the articles: An idea whose time has come? *Nature Web Focus*.

Ware, M. and M. Monkman (2009). *Publishers Research Consortium Report*.

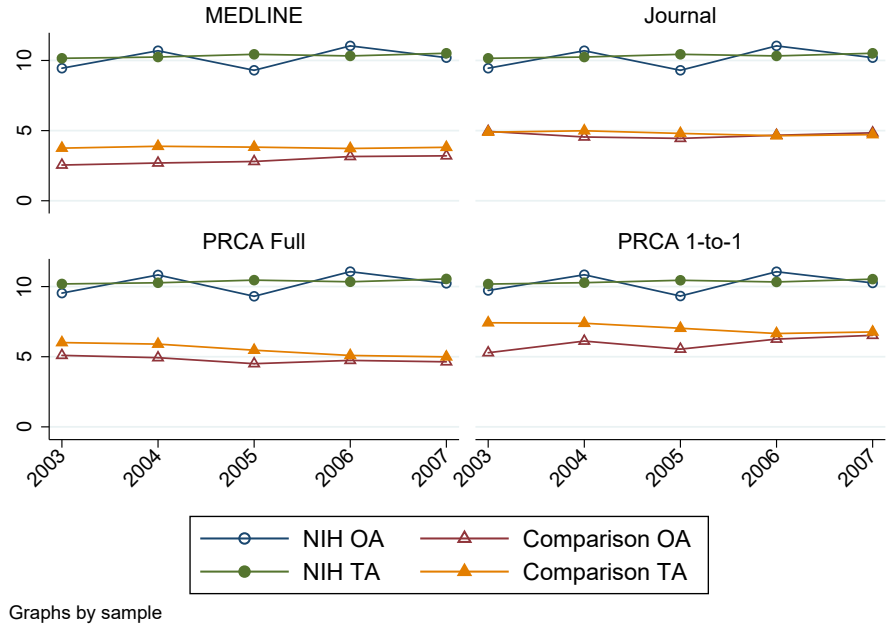
Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy 121*(1), 1–27.

Figure 1: Mean Trends of Outcome Variables Prior to the PMC Mandate.

Panel A: Toll Access Indicator

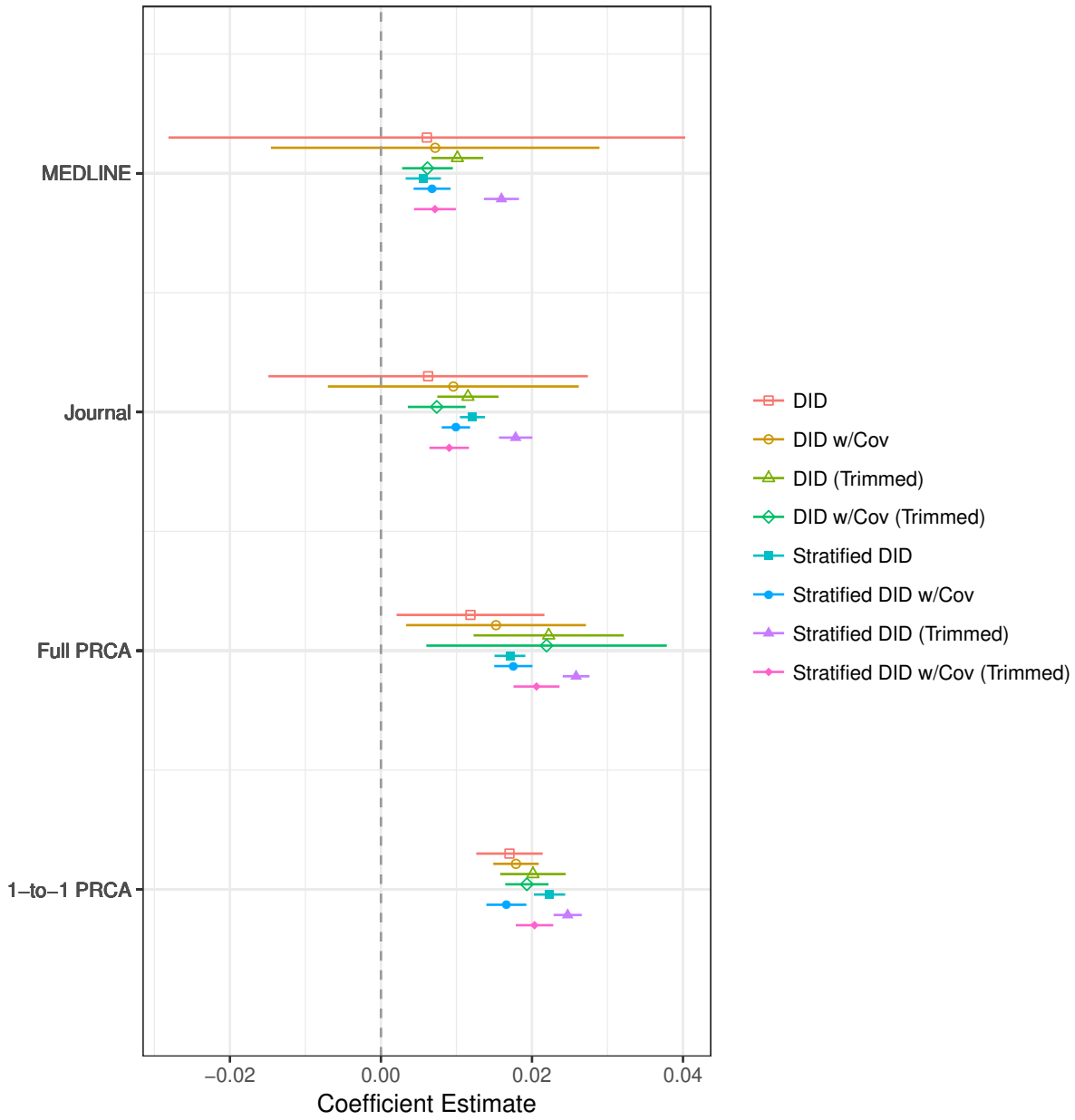


Panel B: 2-Year Forward Citation Count



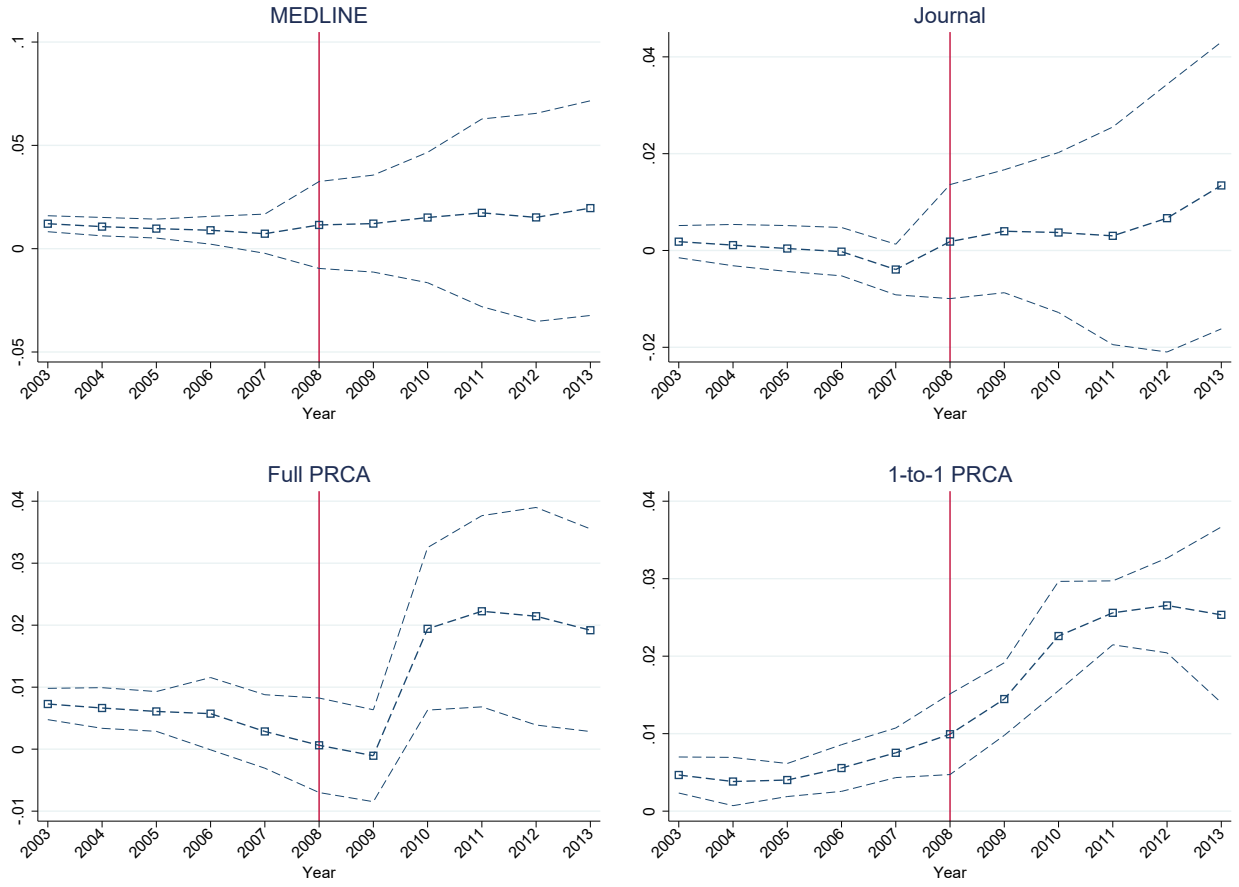
Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score.

Figure 2: Estimated Impacts of the PMC Mandate on an NIH Article's Probability of Being Published in a Toll Access Journal.



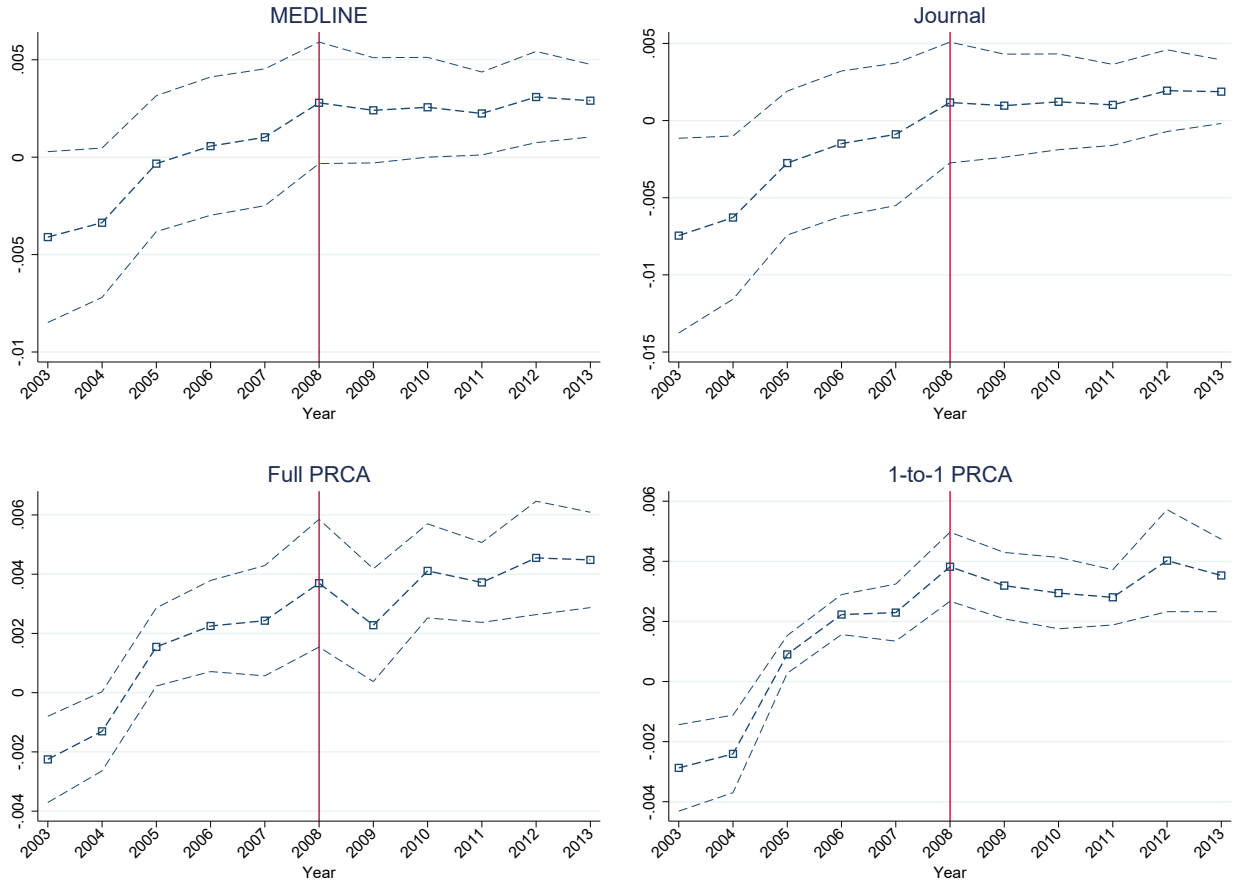
Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. The 47 article-level covariates include backward citations, text-based metrics, MeSH term counts, author counts, and sets of indicators for whether the author is a corporate entity, institution type, publication type, language, and grant support. Also included are country and field fixed effects. Propensity scores are estimated, separately for each comparison sample, using logistic regression on the 47 covariates. The trimmed samples eliminate articles with propensity scores outside (0.1, 0.9). The stratified DID estimates are obtained by stratifying the samples on the propensity scores (Appendix D.2), estimating DID models within each stratum, and then combining the within-stratum estimates into a final composite estimate. The 95% confidence intervals are computed using standard errors clustered at the aggregated field level (Appendix C).

Figure 3: Dynamic Impacts of the PMC Mandate on an NIH Article's Probability of Being Published in a Toll Access Journal.



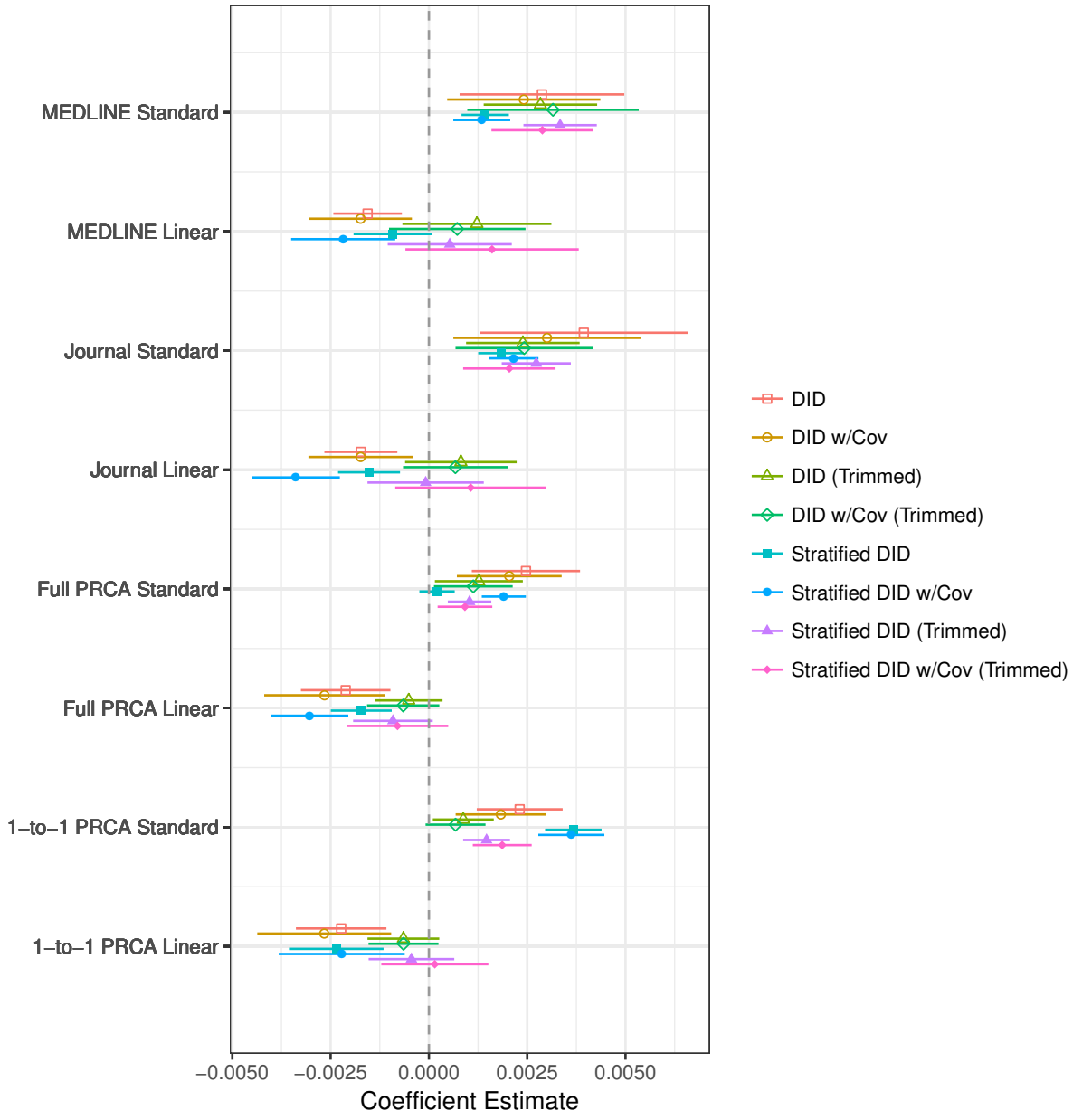
Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. Hollow circles denote point estimates, for a particular year, of the impact of NIH funding on the probability that an article is published in a toll access journal. Light blue bands indicate 95% confidence intervals, which are constructed using standard errors clustered at the aggregated field level (Appendix C). The red vertical line indicates 2008 – the year in which the PMC mandate went into effect.

Figure 4: Dynamic Impacts of the PMC Mandate on an NIH Article's Probability of Being Published in *Science* or *Nature*.



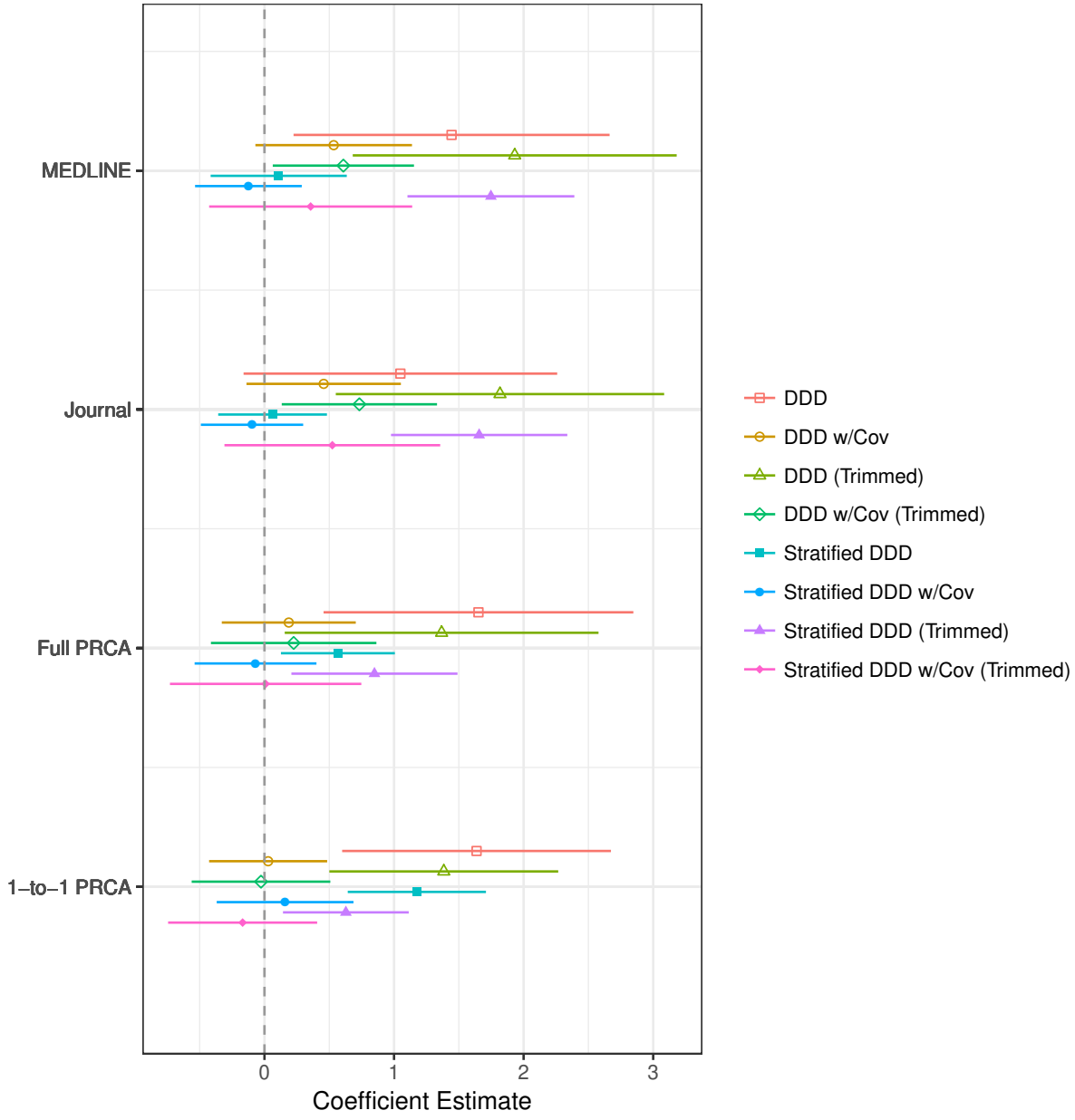
Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. Hollow circles denote point estimates, for a particular year, of the impact of NIH funding on the probability that an article is published in a toll access journal. Light blue bands indicate 95% confidence intervals, which are constructed using standard errors clustered at the aggregated field level (Appendix C). The red vertical line indicates 2008 – the year in which the PMC mandate went into effect.

Figure 5: Estimated Impacts of the PMC Mandate on an NIH Article's Probability of Being Published in *Science* or *Nature*.



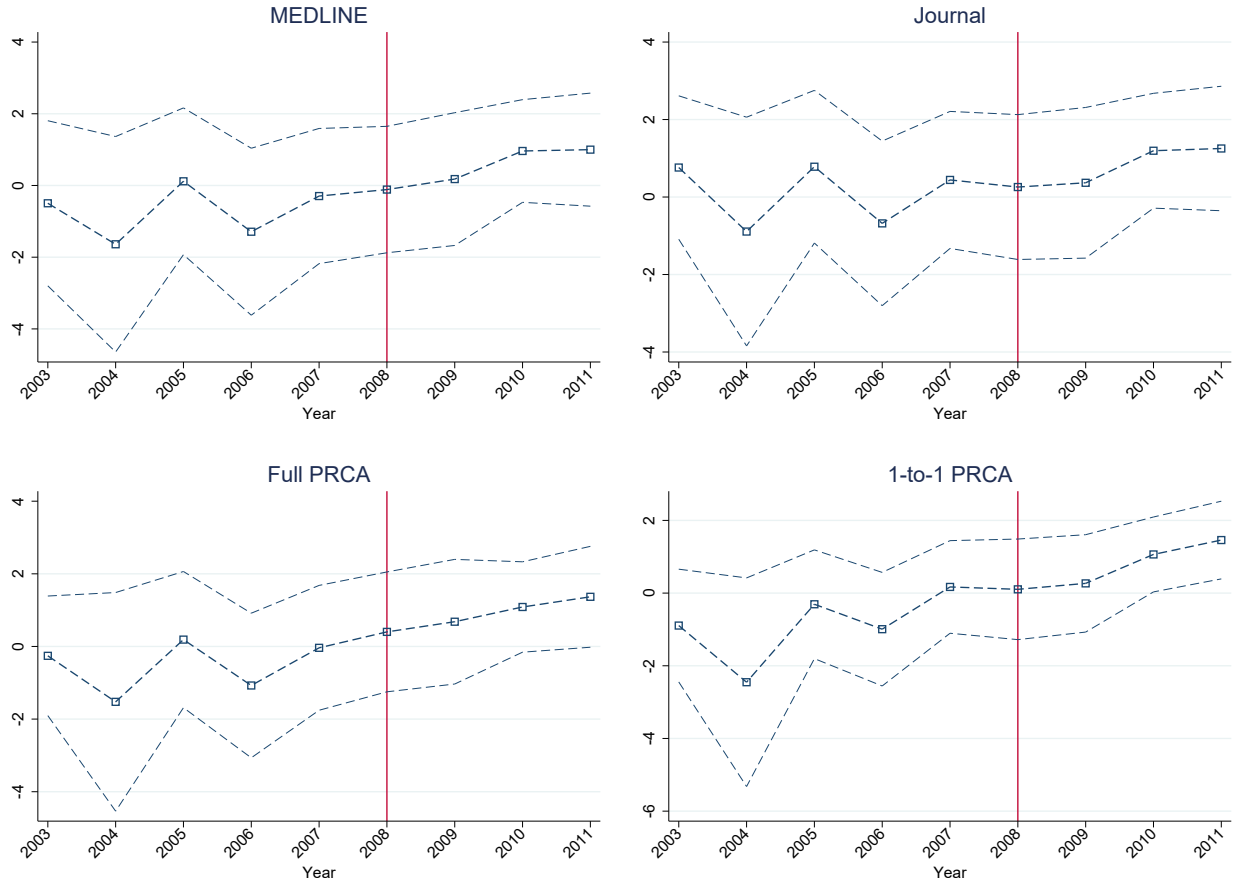
Notes–The estimates labeled "Standard" are obtained by estimating equation (1) and the estimates labeled "Linear" are obtained by adding a linear time trend interacted with the NIH indicator ($N_i \times t$) to equation (1). The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. The 47 article-level covariates include backward citations, text-based metrics, MeSH term counts, author counts, and sets of indicators for whether the author is a corporate entity, institution type, publication type, language, and grant support. Also included are country and field fixed effects. Propensity scores are estimated, separately for each comparison sample, using logistic regression on the 47 covariates. The trimmed samples eliminate articles with propensity scores outside (0.1, 0.9). The stratified DID estimates are obtained by stratifying the samples on the propensity scores (Appendix D.2), estimating DID models within each stratum, and then combining the within-stratum estimates into a final composite estimate. The 95% confidence intervals are computed using standard errors clustered at the aggregated field level (Appendix C).

Figure 6: Estimated Impacts of the PMC Mandate on the Count of 2-Year Forward Citations.



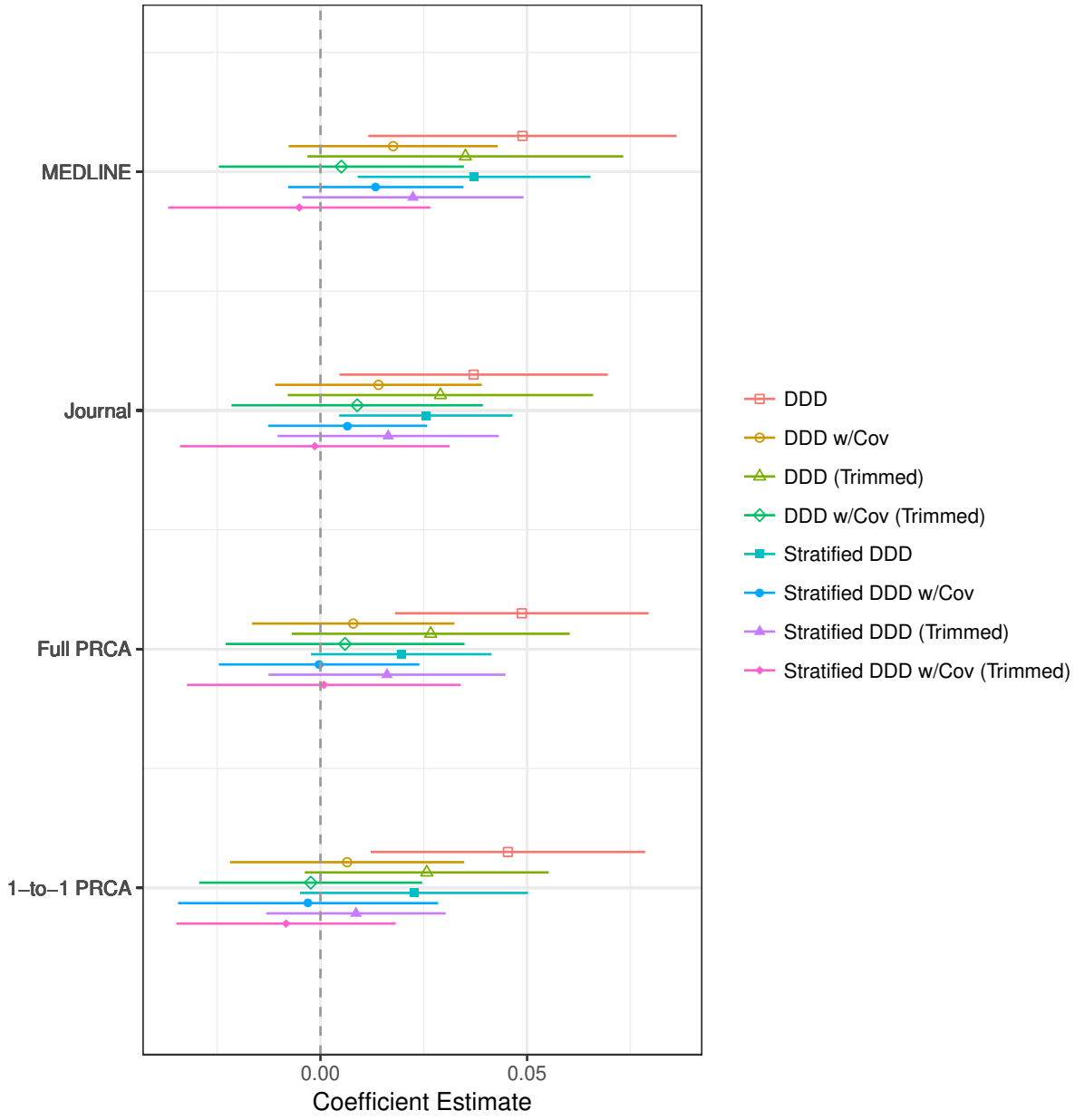
Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. The 47 article-level covariates include backward citations, text-based metrics, MeSH term counts, author counts, and sets of indicators for whether the author is a corporate entity, institution type, publication type, language, and grant support. Also included are country, field, and journal fixed effects. Propensity scores are estimated, separately for each comparison sample, using logistic regression on the 47 covariates. The trimmed samples eliminate articles with propensity scores outside (0.1, 0.9). The stratified DID estimates are obtained by stratifying the samples on the propensity scores (Appendix D.2), estimating DID models within each stratum, and then combining the within-stratum estimates into a final composite estimate. The 95% confidence intervals are computed using standard errors clustered at the journal level.

Figure 7: Dynamic Impacts of the PMC Mandate on the Number of 2-Year Forward Citations an Article Receives.



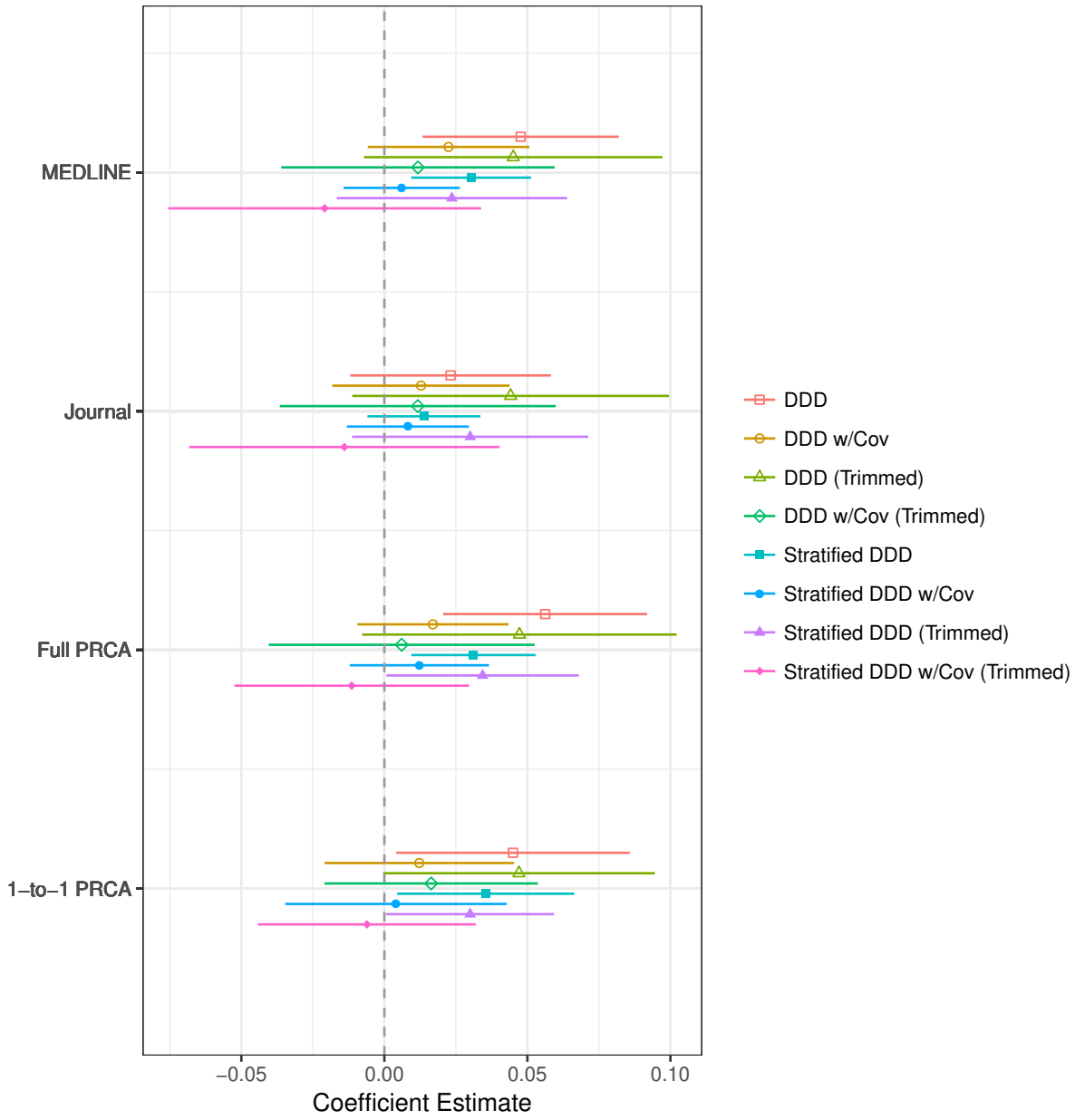
Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. Hollow circles denote point estimates, for a particular year, of the impact of NIH funding on the probability that an article is published in a toll access journal. Light blue bands indicate 95% confidence intervals, which are constructed using standard errors clustered at the journal level. The red vertical line indicates 2008 – the year in which the PMC mandate went into effect.

Figure 8: Estimated Impacts of the PMC Mandate on the Count of 2-Year Forward Citations from Authors Affiliated with Commercial Enterprises.



Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. The 47 article-level covariates include backward citations, text-based metrics, MeSH term counts, author counts, and sets of indicators for whether the author is a corporate entity, institution type, publication type, language, and grant support. Also included are country, field, and journal fixed effects. Propensity scores are estimated, separately for each comparison sample, using logistic regression on the 47 covariates. The trimmed samples eliminate articles with propensity scores outside (0.1, 0.9). The stratified DID estimates are obtained by stratifying the samples on the propensity scores (Appendix D.2), estimating DID models within each stratum, and then combining the within-stratum estimates into a final composite estimate. The 95% confidence intervals are computed using standard errors clustered at the journal level.

Figure 9: Estimated Impacts of the PMC Mandate on the Count of 2-Year Forward Citations from Authors Located in Poor/Developing Countries.



Notes–The MEDLINE sample contains all NIH and non-NIH articles published between 2003 and 2013 that are indexed in the 2016 MEDLINE baseline files. The journal sample contains non-NIH articles published in the same journal-year as at least one NIH article, the full PRCA sample contains non-NIH articles harvested using the PubMed Related Citations Algorithm (PRCA), and the 1-to-1 PRCA sample contains non-NIH articles most similar to each NIH article on the basis of the PRCA similarity score. The 47 article-level covariates include backward citations, text-based metrics, MeSH term counts, author counts, and sets of indicators for whether the author is a corporate entity, institution type, publication type, language, and grant support. Also included are country, field, and journal fixed effects. Propensity scores are estimated, separately for each comparison sample, using logistic regression on the 47 covariates. The trimmed samples eliminate articles with propensity scores outside (0.1, 0.9). The stratified DID estimates are obtained by stratifying the samples on the propensity scores (Appendix D.2), estimating DID models within each stratum, and then combining the within-stratum estimates into a final composite estimate. The 95% confidence intervals are computed using standard errors clustered at the journal level.

Table 2: Magnitudes of Estimated Impact of the PMC Mandate on Publishing Patterns in Biomedicine.

	2009	2010	2011	2012	2013	All
Total NIH Articles	92,646	99,968	104,136	104,176	107,549	508,475
NIH Articles in OA Journals	5,697	7,514	9,968	12,774	15,028	50,981
NIH Articles in TA Journals	86,949	92,454,	94,168	91,402	92,521	457,494
Proportion of NIH Articles in TA Journals	0.939	0.925	0.904	0.877	0.860	0.900
Effect Size: 0.01						
Predicted Proportion	0.929	0.915	0.894	0.867	0.850	
Predicted Count	86,023	91,454	93,127	90,360	91,446	452,410
Predicted Difference	926	1,000	1,041	1,042	1,076	5,076
Effect Size: 0.02						
Predicted Proportion	0.919	0.905	0.884	0.857	0.840	
Predicted Count	85,096	90,455	92,085	89,318	90,370	447,324
Predicted Difference	1,853	1,999	2,083	2,084	2,151	10,170
Effect Size: 0.03						
Predicted Proportion	0.909	0.895	0.874	0.847	0.830	
Predicted Count	84,170	89,455	91,044	88,277	89,295	442,241
Predicted Difference	2,779	2,999	3,124	3,125	3,226	15,253

A Data Sources

I begin with six sources of raw data: 1) MEDLINE, 2) Web of Science, 3) the Directory of Open Access Journals (DOAJ), 4) the MeSH vocabulary, 5) MapAffil, and 6) United Nations National Accounts. From MEDLINE, I obtain a list of unique article IDs and information about each article's journal, grant support, publication date, publication type, author count, MeSH terms, title, and abstract. The unique article IDs are called PubMed identifiers (PMIDs), which are assigned to articles by the National Library of Medicine. From Web of Science, I obtain a list of *citing* PMIDs. I also obtain, for each *citing* PMID, a list of *cited* PMIDs (the citing PMID's references). From the Directory of Open Access Journals (DOAJ), I obtain a list of journals identified as being "open access". From the MeSH vocabulary data set, I obtain the tree structure of MeSH terms that the NLM uses to classify articles in MEDLINE. From MapAffil, I obtain, for each PMID, information on the affiliation of the first author, including country and type of affiliation (e.g., university, hospital, etc.). From the United Nations National Accounts, I obtain data on per capita GDP for a panel of countries. The following subsections will explain each data set in more detail.

A.1 MEDLINE

MEDLINE is a bibliographic database created and maintained by the U.S. National Library of Medicine (NLM). The database can be downloaded by anyone, free of charge.[21] This paper uses the 2016 baseline files.[22] These are distributed by the NLM as 812 compressed Extensible Markup Language (XML) files.[23]

I wrote a series of Perl scripts to extract data from the XML files and place them into tab-delimited text files.[24] The elements that I extract are:[25]

1. "Status" attribute
2. PMID (and the "Version" attribute)
3. NlmUniqueID
4. MeshHeadingList
5. GrantList
6. PublicationTypeList

[21] http://www.nlm.nih.gov/bsd/licensee/medpmmenu.html

[22] https://www.nlm.nih.gov/bsd/licensee/2016_stats/baseline_med_filecount.html

[23] XML is a markup language that organizes data into a format that is both human-readable and machine-readable.

[24] These scripts (and the rest of the code used to produce the results in this paper) are freely-available in the following GitHub respository: https://github.com/EconJoe/NIHMandate. The parsers rely heavily on the XML::Simple module from the Comprehensive Perl Archive Network (CPAN).

[25] See https://www.nlm.nih.gov/bsd/mms/medlineelements.html for a description of all elements in MEDLINE.

7. PubDate
8. MedlineDate
9. ArticleDate
10. ArticleTitle
11. Abstract and AbstractText
12. Language

The top-level element for each record (article) in the MEDLINE XML files is MedlineCitation. This element has four attributes, but I am only interested in the "Status" attribute. This attribute indicates how thoroughly the record's information has been vetted. I only use records with the status "MEDLINE" as these have undergone the most rigorous quality review and are the only true MEDLINE records.

The PMID, or PubMed ID, is a unique identifier for every record in MEDLINE. The PMID element also contains an attribute called "Version". This attribute is included to deal with the "versioning" publishing model, in which multiple versions of the same article are published.[26] The PMID element is crucial for linking the MEDLINE and Web of Science data sets. There are 24,358,442 records in the 2016 baseline files. Because 317 PMIDs have several "versions", there are only 24,358,073 unique PMIDs.

The NLMUniqueID element is a seven, eight, or nine character identifier that uniquely identifies the journal in which a record is published. It is crucial for linking jounal-level information within MEDLINE and other NLM sources. There are 23,395 unique NLMUniqueID in the 2016 baseline files. The mean NLMUniqueID contains 1,041 articles and the median contains 89. 2,579 NLMUniqueID contain only a single article and *The Journal of Biological Chemistry* contains 170,684 articles. In addition to using the NLMUniqueID as a linking variable, I also use it to estimate journal fixed effects and to cluster the standard errors at the journal level in some of the models in the paper.

Unfortunately, other sources of journal-level data, such as DOAJ, do not use the NLMUniqueID. Instead, they use the International Standard Serial Number (ISSN) to identify journals. Thus, to link journal-level information in MEDLINE to these other data sources, I need to use the ISSN. The ISSN is an eight-character value that uniquely identifies periodical publications, including journals. It is assigned by ISSN National Centers, not the NLM. Thus, it is more universal and more useful than the NLMUniqueID for linking to non-NLM sources. If a journal has both a print and electronic format, then each format will receive a separate ISSN. Fortunately, MEDLINE typically include all formats, which allows me to link data at the journal-level regardless of which ISSN format is used in non-NLM sources. The ISSNLinking element is an ISSN that links all formats of the same journal. This element also helps to uniquely identify journals with multiple ISSNs.

The MeshHeadingList element contains a list of all MeSH (Medical Subject Heading) terms assigned to the record. MeSH terms are used to classify the content of each record indexed in MEDLINE. NLM librarians read each article and determine which MeSH terms

[26] PLoS Contents is the only journal indexed in MEDLINE that uses the versioning model of publishing.

apply to that article. Thus, they are librarian-supplied, not author-supplied. This eliminates concerns about authors strategically choosing MeSH terms. The MeshHeadingList contains the following elements: DescriptorName and QualifierName, each of which have the attribute "MajorTopicYN". As suggested by the names, DescriptorName describes the record content, QualifierName qualifies the description, and "MajorTopicYN" indicates whether the MeSH term is a major or minor topic of the article. For instance, "Fetal Growth Retardation" might be a descriptor and "complications" might qualify the descriptor. The MeSH terms are crucial for linking MEDLINE and the MeSH vocabulary. The article-level covariates computed using the MeSH terms are: the 1) total number of descriptor terms and 2) total number of qualifying terms that tag each article. I also use the MeSH terms to group articles into fields and estimate field fixed effects – see Appendix C for the details.

The GrantList element contains a list of all grants that are acknowledged by a record. It includes the grant number as well as the funding agency. I use the funding agency to identify which are records are NIH-funded. The article-level covariates computed using the grant list are: 1) an indicator for whether an article is NIH funded and 2) the count of non-NIH grants that support an article.

The PublicationTypeList element contains a list of all publication types that characterize an article. Like MeSH terms, these publication types are librarian-supplied. Examples of publication types include "Journal Article", "Review", and "Retracted Publication". There are XX publication types[27], and I combine them into 21 groups to include as article-level covariates in models that I estimate. Two of the publication types are "Research Support, N.I.H., Extramural" and "Research Support, N.I.H., Intramural". I use this as an additional source of information about which records are NIH funded.

MEDLINE has three date elements that I use to determine the publication date of each record: PubDate, MedlineDate, and ArticleDate. PubDate follows a standard dating format, making it very easy to identify the Year element. When dates do not follow this standard format, they are found in the element MedlineDate. For these non-standard dates, I manually code the year. In some cases, there is a year range instead of a single year. For these cases, I take the first year in the range as the publication year. The element ArticleDate contains the date that a publisher first publishes an electronic version of an article. ArticleDate always follows a standard dating format, making it easy to identify the Year element. Often, the date information in the PubDate or MedlineDate elements differs from the date information in the ArticleDate element. This is because the electronic and print versions of articles are often published on different dates. I take the minimum year as the relevant year of publication. Typically, the PubDate and MedlineDate Year elements do not differ by more than a year from ArticleDate Year element. I use the publication year to estimate a set of year fixed effects in all models and also to define the pre and post PMC mandate periods (before and after 2008).

The element ArticleTitle contains the complete English title for each record. If the article is originally published in a different language, it is translated to English. The elements Abstract and AbstractText contain the abstract for each record published in an English language journal. Unlike titles, abstracts are not translated if they are originally published in another language. The titles and the abstracts for each record are used to construct text

[27]See here for the full list: https://www.nlm.nih.gov/mesh/pubtypes.html

metrics that are included as article-level covariates in estimated models. See Appendix B for additional information on processing title and abstract text.

The element Language contains information on the language in which an article is published. I create 10 indicator variables for 10 languages, which serve as article-level covariates in models that I estimate. These languages are: English, German, French, Russian, Japanese, Spanish, Italian, Chinese, and Other. I also include an additional indicator for articles whose language is undetermined.

A.2 Web of Science

Clarivate Anlytics Web of Science (WOS) is a citation indexing database. Indeed, it is the most widely used source of citation data.[28] Unlike the rest of the data used in this paper, the WOS data is not freely available. Instead, access to the data was negotiated in 20XX and the data were delivered in December 2014. The data were delivered as 32 XML files and include all articles published between 1950 and 2014 that are indexed in both WOS and MEDLINE. There are 13,878,957 citing articles. The mean number of references is 22.76, the median is 17, maximum of 6,310, and the standard deviation is 25.27. There are a total of 14,328,197 cited articles. These receive an average of 22.04 citations, a median of 8, a maximum of 251,686, and a standard deviation of 114.98.

WOS provides a wide variety of information about each article. However, for each article, I only extract the PMID along with all of the PMIDs cited by the article (i.e, each PMID and its references). The PMID allows me to link WOS records to MEDLINE records. The references for each PMID allow me to construct various citation measures for each article and author. Specifically, it allows me to construct the following outcome variables: total 2-year forward citations, 2-year forward citations received by articles associated with a commercial firm (see MapAffil data below), and 2-year forward citations received by articles associated with poor/developing countries (see UN National Accounts data below). These data also allow me to construct the following article-level covariates: count of backward citations and count of backward citations to articles published in open access journals.

A.3 Directory of Open Access Journals (DOAJ)

The Directory of Open Access Journals (DOAJ) is an online directory that indexes peer-reviewed open access journals. It began as a project at Lund University in 2002, but is now an independent organization. The database can be downloaded by anyone, free of charge.[29] The database is updated daily, and past versions are not readily available. I downloaded the file on November 11, 2016, and will make it available upon request. The database is distributed as a CSV (comma-separated) file. I use journals' International Standard Serial Number (ISSN) to match DOAJ data to the MEDLINE data.[30]

[28] Another common citation indexing database is Elsevier's Scopus.

[29] Go to http://doaj.org/faqmetadata, and click "Download the file to your computer".

[30] See http://doaj.org/faqsearchresults for the fields contained in the DOAJ data file. This data allows me to construct one of the main outcome variables of interest: an indicator variable for whether an article is published in a toll access journal.

A.4 MeSH Vocabulary

The MeSH vocabulary is a small set of XML files that contains all MeSH terms and information about each term (e.g., the date it was introduced). These files are freely available from the National Library of Medicine (NLM).[31] I extract the following information for each MeSH term: 1) the term itself, 2) a unique ID assigned to each MeSH term, and 3) the branches of the MeSH tree on which the term is located. The MeSH terms can map to multiple branches on the MeSH tree. I use MeSH branches to characterize the field of articles, which is described in Appendix C.

A.5 MapAffil

MapAffil (Torvik, 2015) is a data set containing information on the affiliation of MEDLINE articles' authors. The 2016 tranche of data consist of 37,412,190 PMID-authors. I extract information on the country and type of institution that characterize each affiliation and use the PMID to link this information to MEDLINE.

There are 929 countries in the MapAffil data. Country information is used to compute author country fixed effects. Each affiliation is categorized into eight institution types: commercial, educational, hospital, educational/hospital, government, military, other organization, or unknown. Institution type information is used to construct a set of indicator variables characterizing the type of author affiliation for each article.

I also use the country and institution type information to construct citation measures that only include citations from authors affiliated with particular countries or institution types. In particular, I am able to identify citations that come from authors in poor/developing countries and who are affiliated with commercial enterprises.

A.6 United Nations National Accounts

The UN National Accounts main aggregates are updated yearly by Economic Statistics Branch of the UN Statistics Division. I use country aggregates on per capita GDP at current prices (U.S. dollars).[32] The data contain yearly GDP information on 220 countries between 1970 and 2015, though I only use data between 2003 and 2013. When data is missing for a particular country-year, I linearly interpolate the value. For 2003-2013, the mean per capita GDP is $14,681 (SD=$23,057) and the median is $4,809.

I use this data to classify each country into per capita GDP quintiles by year. I then link this country-year level data to MapAfill, which enables me to link GDP quintile information to each MEDLINE article. I use this information to identify citations that come from authors in poor/developing countries. In this case, I define a country as poor/developing for a particular year if it is in one of the bottom two quintiles of the per capita GDP distribution in that year.

[31] http://www.nlm.nih.gov/mesh/filelist.html

[32] See: http://data.un.org/Data.aspx?q=GDP+per+capita&d=SNAAMA&f=grID%3a101%3bcurrID%3aUSD%3%3a1

B Processing Title and Abstract Text

This section draws heavily on Staudt et al. (2017), which itself draws heavily on Packalen and Bhattacharya (2015). As noted in Appendix A, I use a Perl script to extract the ArticleTitle, Abstract, and AbstractText elements from each record (article) indexed in the 812 MEDLINE 2016 Baseline Files. After extraction, the script indexes all words, word pairs and word triplets (1-, 2-, and 3-grams). It then processes each n-gram by performing the following operations:

1. Convert all text to lower-case.
2. Eliminate 2- and 3-grams with words that cross the following characters: ,.?!;:)(}{][–.
3. Eliminate all remaining characters that are not alphanumeric.
4. Eliminate all n-grams that contain words appearing in the stopword list provided by the NLM at this address: http://mbr.nlm.nih.gov/Download/2009/WordCounts/wrd_stop
5. Eliminate all n-grams that contain the following character sequences: web, www, http, pubmed, medline, clinicaltrialsgov.
6. Eliminate all n-grams that contain more than two adjacent numbers.
7. Eliminate all n-grams that have a length of less than three characters.
8. Keep all 1-grams with character length 3-29, 2-grams with character length 7-59, and 3-grams with character length 11-89.
9. Stem each word from each n-gram using the module Lingua::Stem from the Comprehensive Perl Archive Network (CPAN).
10. Index all the processed n-grams from each title and abstract into 812 tab-delimited text files corresponding to the 812 MEDLINE XML files.

Once they are processed, I identify each n-gram's "vintage" ("birth") year – that is, the year the n-gram first appears in the MEDLINE corpus. After an n-gram appears in the MEDLINE corpus, I am able to identify all articles that use the n-gram in a title or abstract. I use this information to identify, for every vintage, a set of "top" n-grams. An n-gram is a top n-gram if it is in the top 0.01 percent of all n-grams in its vintage, in terms of the total number articles that mention it after birth. Top n-grams are identified *within vintage* because n-grams from earlier vintages will have more time to accumulate article mentions than n-grams from later vintages. Thus, it does not make sense to compare n-grams that have different vintages.

I use this information to construct three article-level covariates. First, I compute the count of top n-grams that an article originates. An article originates a top n-gram if it uses the top concept in its vintage year. If multiple articles use a top n-gram in its vintage year, then that particular n-gram has multiple originators. Second, I compute the count of top n-grams that an article adopts early – i.e. within 5 years of the n-gram's vintage. Finally, I compute the total number of n-grams, regardless of vintage or "top" status, that an article uses in its title or abstract.

C Aggregating MeSH Terms to Construct Fields

This section draws heavily on the Appendix of Staudt et al. (2017). They devise an algorithm which uses the Medical Subject Headings (MeSH) that tag most articles in MEDLINE to characterize the fields to which each article belongs. Note that Staudt et al. (2017) use the 2014 MEDLINE baseline files, but the current paper uses the 2016 MEDLINE baseline files.

There are 27,883 raw terms in the 2016 MeSH vocabulary and they vary widely in their descriptive detail. For instance, some articles are tagged with general terms such as *Body Regions* and some are tagged with more detailed terms such as *Peritoneal Stomata*. Thus, in order to construct comparable fields, I aggregate all MeSH terms to a similar level of descriptive detail.

To understand the aggregation method, it is important to first understand how MeSH terms are organized. MeSH terms have a hierarchical structure. At the top of the hierarchy (first-level terms) are 16 very general terms such as *Anatomy*, *Organisms*, and *Diseases*. Each of these 16 first-level terms are identified by a unique capital letter. For instance, *Anatomy* is identified by the letter A, *Organisms* is identified by B, and so on. Beneath each of these first-level MeSH terms is a group of second-level MeSH terms. For instance, *Body Regions* is a second-level MeSH term beneath the top-level term *Anatomy*. Each second-level MeSH term is identified by the capital letter of the first-level MeSH term it is beneath and by two numbers. For instance, *Body Regions* is identified by A01. Beneath each second-level MeSH term is a group of third-level MeSH terms identified by the capital letter of the first-level term it is beneath, the two numbers of the second-level term it is beneath, and three subsequent numbers. For instance, *Anatomic Landmarks* is a third-level MeSH term under *Body Regions* and is identified as A01.111. This structure continues to depths of up to 12 levels.

Aggregating MeSH terms (that is, classifying lower level MeSH terms as a part of higher level MeSH terms) is complicated by the fact that most MeSH terms fall beneath multiple higher level MeSH terms. Consider the MeSH term *Asthma*. This term has four separate identifiers: C08.127.108, C08.381.495.108, C08.674.095, and C20.543.480.680.095. Thus, *Asthma* falls under the first level MeSH term *Diseases* (identified by C). It also falls under the second-level terms *Respiratory Tract Diseases* (C08) and *Immune System Diseases* (C20). The problem arises because MEDLINE records only contain the MeSH terms themselves, not their identifiers. For instance, if a MEDLINE record is tagged with the MeSH term *Asthma*, it is not clear whether this is the *Asthma* that is beneath *Respiratory Tract Diseases* (C08) or *Immune System Diseases* (C20).

Consider aggregating the raw MeSH term *Asthma* to the second-level – i.e., splitting it between the second-level terms *Respiratory Tract Diseases* and *Immune System Diseases*. I opt to simply assign half to each higher level term. Thus, an article originally tagged with the raw term *Asthma* is now tagged with two second-level terms, each weighted by 1/2.

Now consider aggregating the raw MeSH term *Asthma* to the fourth-level. In this case, *Asthma* must be split between the following fourth-level terms:

- *Lung Diseases, Obstructive* [C08.381.495] from C08.381.495.108
- *Hypersensitivity, Immediate* [C20.543.480] from C20.543.480.680.095

- *Asthma* [C08.127.108] from C08.127.108
- *Asthma* [C08.674.095] from C08.381.495.108

In this case, a quarter of the raw term *Asthma* is assigned to each of these four fourth-level terms. Thus, overall, 1/4 will be assigned to *Lung Diseases, Obstructive*, 1/4 to *Hypersensitivity, Immediate*, and 1/4+1/4=1/2 to *Asthma* itself. Thus, an article originally tagged with the raw term *Asthma* is now tagged with three fourth-level terms, two weighted by 1/4 and one weighted by 1/2.

A last complication is that most article are tagged by multiple raw MeSH terms. As an example, suppose that, in addition to being tagged with *Asthma*, an article is also tagged with the raw terms *Neck* (identified by A01.598) and *Health Information Exchange* (identified by L01.700.253, L01.399.500.500, L01.313.500.500, and E05.318.308.940.968.625.500.500). By the process discussed above, 1/4 of *Health Information Exchange* will be assigned to each of the four fourth-level MeSH terms: *Health Information Exchange* itself (L01.700.253), *Health Information Management* (L01.399.500), *Medical Informatics* (L01.313.500), and *Data Collection* (E05.318.308). Since the lowest level of aggregation for *Neck* is the third-level, it cannot be assigned to a fourth-level term. In this *Neck* is simply eliminated – it is too highly aggregated.

Each of the original remaining MeSH terms, *Asthma* and *Health Information Exchange*, are assumed to receive equal weight in characterizing the article. Under this assumption, the article will be apportioned to each fourth level MeSH term as follows:

- 1/2*1/4=1/8 to *Lung Diseases, Obstructive*
- 1/2*1/4=1/8 to *Hypersensitivity, Immediate*
- 1/2*1/4=1/8 to *Asthma*
- 1/2*1/4=1/8 to *Asthma*
- 1/2*1/4=1/8 to *Health Information Exchange*
- 1/2*1/4=1/8 to *Health Information Management*
- 1/2*1/4=1/8 to *Medical Informatics*
- 1/2*1/4=1/8 to *Data Collection*

Obviously 1/8+1/8+1/8+1/8+1/8+1/8+1/8+1/8=1. Thus, an article that was originally tagged by the three raw MeSH terms *Asthma*, *Neck* and *Health Information Exchange* is now apportioned between seven different fourth-level MeSH terms – *Asthma* receiving a weight of 1/8+1/8 = 1/4 and the other six receiving a weight of 1/8 each.

In general, each MEDLINE article is apportioned across aggregated MeSH terms in two stages. First, the original MeSH terms are equally apportioned across the higher-level MeSH terms of which they are a part (e.g. apportion *Asthma* equally across *Lung Diseases, Obstructive*, *Hypersensitivity, Immediate*, *Asthma*, and *Asthma*. Second, the higher-level MeSH terms are weighted by the inverse of the number of original MeSH terms of the proper

level that tag the article (e.g. the hypothetical article was tagged by three original MeSH terms, but only two at the proper level of aggregation, and so each is weighted by 1/2).

Each article is assigned to the most highly weighted fourth-level MeSH term. In the example above, the article would be assigned to *Asthma*, which received a weight of 1/4. Ties are broken randomly. Thus, each article is assigned to a single aggregated "field". These fields are used to compute field fixed effects and cluster standard errors.

I also use raw MeSH terms to develop and alternative characterization of an article's field. In particular, I first identify the major Descriptor MeSH terms for each article. If there are multiple major MeSH terms, I choose the first listed as the raw term to characterize the field. This provides a less aggregated alternative way to compute field fixed effects.

D Propensity Score Estimation and Stratification

D.1 Estimating the Propensity Score

As is usual in the literature, I estimate propensity scores using logistic regression. This is done separately for each of the four comparison samples – MEDLINE, Journal, Full PRCA, and 1-to-1 PRCA. In the base specification, I estimate the models allowing all 46 coviariates to enter linearly. As a robustness check, I also estimate a model that includes, in addition to the 46 linear terms, 982 second order terms (i.e. all squared terms and pair-wise interactions).[33]

D.1.1 DID Setting

In the difference-in-differences (DID) setting, the treated group is the set of NIH-funded articles that are published after the PMC mandate. Thus, I estimate $e^{DID}(x) = P(W_i = 1|X_i = x) = P(N_i = 1, P_i = 1|X_i = x)$. The DID estimates can be obtained for all outcome variables. Recall that the toll access and *Science/Nature* indicators are analyzed for the period 2003-2013, and the 2-year forward citation outcomes are analyzed for the period 2003-2011. Thus, for each of the four comparison samples, I separately estimate the propensity score for these two time periods. This results in a total of eight sets of estimated propensity scores – two for each of the four comparison samples.

D.1.2 DDD Setting

In the triple differences (DDD) setting, the treated group is the set of NIH-funded articles, published in toll access journals, after the PMC mandate. In this case, I estimate $e^{DDD}(x) = P(W_i = 1|X_i = x) = P(N_i = 1, P_i = 1, S_i = 1|X_i = x)$. Since there is no within-journal variation in the toll access and *Science/Nature* indicators, the DDD estimates can only be obtained for the 2-year forward citation variables. Thus, there is only a need to obtain estimated propensity scores for the 2003-2011 period, which results in a total of four sets of estimates – one for each of the four comparison samples.

[33] Since there are 46 covariates, there are $46 * (46 + 1)/2 = 1,081$ second order terms. However, squared indicator variables are redundant, and so are not included in the model, leaving 982 second order terms.

D.2 Stratification on the Propensity Score

The idea of stratifying units by covariates goes back to Cochran (1968) and the idea of stratifying on the propensity score dates to Rosenbaum and Rubin (1983, 1984). The basic idea is that, by grouping observations with similar propensity scores into strata, a stratified randomized experiment can be mimicked. That is, within strata, analyses can be performed *as if* the treatment is randomly assigned. By stratifying on the propensity score instead of the covariates, the dimensionality of the problem is reduced to manageable levels.[34] As noted by Imbens and Rubin (2015), stratification is less dependent on the correct specification of the propensity score than is propensity score weighting. This is important because in the present application, there is virtually no a priori information about the true propensity score.

Stratifying the sample on the propensity score requires two main decisions. First, the number of strata must be chosen. Second, the cut points for each stratum must be chosen. I automate these decisions by using an algorithm adapted from Imbens and Rubin (2015) to the difference-in-differences (DID) and triple differences (DDD) settings. Specifically, let B_{ij} be a variable indicating whether article i is in stratum $j = 1, ..., J$. That is,

$$B_{ij} = \begin{cases} 1 & \text{if } b_{j-1} \leq \hat{e}(X_i) < b_j \\ 0 & otherwise \end{cases}$$

$$B_{iJ} = \begin{cases} 1 & \text{if } b_{J-1} \leq \hat{e}(X_i) \leq b_J \\ 0 & otherwise \end{cases}$$

where b_{j-1} and b_j are the upper and lower cut points for each stratum. The algorithm chooses the number of strata J and the ranges for each strata $(b_{j-1}, b_j]$. It starts with a single stratum $J = 1$. Within this stratum, the algorithm conducts a one-way ANOVA test for the equality of the means of the log odds ratio for the four DID groups or the eight DDD groups. This test yield an F-value. If this F-value is large, the propensity scores are quite different, and the algorithm attempts to split this stratum into two strata that are more similar. If the stratum is split, it will be split at the median value of the propensity score. However, before splitting, the algorithm must ensure that the two new candidate strata each have a sufficient number of observations to conduct analyses within these strata. That is, before splitting, the algorithm checks whether the potential new strata have $N_g > N_{min,g}$ for each DID group ($g = 1, 2, 3, 4$) or DDD group ($g = 1, 2, 3, 4, 5, 6, 7, 8$). It also checks whether the total number of observations is sufficient $N > N_{min}$. If the F-statistic is sufficiently small *and* there are a sufficient number of observations within each of the new candidate strata, then the single stratum is split in two. Otherwise, the stratum is not split. The algorithm then repeats these steps within every new stratum created. The algorithm continues to split strata until all F-statistics are sufficiently small or further splitting would result in strata with too few observations to conduct the analysis.

The inputs for this algorithm are the threshold F-statistc for splitting a stratum, F_{max}, the minimum number of observations in each of the DID or DDD groups, $N_{min,g}$, and the

[34] Given K dichotomous covariates, the number of strata is 2^K, so the number of strata grows exponentially. This means that some strata may have no observations, and others may have only treated or control observations. Analysis cannot be carried out for such strata.

minimum number of total observations within each stratum, N_{min}. For the baseline analysis, I choose $F_{max} = 2$, $N_{min,g} = 100$ for all g, and $N_{min} = 1000$.

D.3 Obtaining Stratified DID/DDD Estimates

Once the articles in each sample are stratified, it is straight forward to estimate equations (1) and (2) within each strata, and then combine these estimates into a single composite estimate. Let $\hat{\delta}_j$ be the estimate, within stratum j, of δ from equation (1) or (2). Let N_j be the number of observations within stratum j. Then the ATT is $\hat{\delta} = \sum_1^J (N_j/N)\hat{\delta}_j$, where J is the total number of strata and N is the total number of observations. Let $\hat{V}_j$ be the estimated variance of the ATT estimator within stratum j. Then the variance of the ATT estimator is given by $\hat{V} = \sum_1^J (N_j/N)^2 \hat{V}_j$.

Table A1.a: Summary Statistics for the MEDLINE Sample.

	NIH Pre		Comp. Pre		NIH Post		Comp. Post		All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Outcome Variables										
TA Journal	0.97	0.18	0.96	0.2	0.9	0.3	0.88	0.32	0.92	0.27
Science/Nature	0.01	0.08	0.01	0.08	0.01	0.09	0	0.07	0.01	0.08
2-Yr For. Cites (Total)	10.41	17.97	3.78	10.55	7.13	18.85	2.6	8.20	3.79	10.98
2-Yr Forward Cites (Com. Enterprise)	0.2	0.8	0.1	0.6	0.1	0.57	0.05	0.39	0.08	0.53
2-Yr Forward Cites (Dev. Country)	0.21	0.81	0.15	1.48	0.07	0.66	0.06	0.69	0.11	1.11
Covariates										
Backward Cites	36.15	30.92	17.11	24.3	37.62	35.18	19.39	26.24	20.41	27.02
OA Backward Cites	0.52	1.41	0.25	0.98	1.34	2.7	0.67	1.83	0.52	1.6
Age 0 Top Concepts	0.02	0.2	0.01	0.16	0.03	0.22	0.02	0.19	0.02	0.18
Age ≤ 5 Top Concepts	0.33	0.94	0.19	0.72	0.27	0.84	0.17	0.67	0.2	0.72
Total Concepts	138.42	47.04	101.7	60.68	120.35	67.83	91.69	68.29	100.17	65.16
Total MeSH Descriptors	13.86	5.24	10.75	5.62	13.22	5.37	10.27	6.07	10.84	5.89
Total MeSH Qualifiers	9.01	5.8	6.44	5.06	8.68	5.93	6.26	5.35	6.63	5.35
Author Count	5.34	3.89	4.32	6.71	6.17	6.12	4.89	13.18	4.75	10.11
Corporate Author	0	0.01	0.01	0.11	0	0.01	0.01	0.09	0.01	0.1
Journal Article	0.99	0.12	0.91	0.29	0.98	0.15	0.91	0.28	0.92	0.27
Research Support, U.S. Gov't, Non-P.H.S.	0.14	0.35	0.03	0.16	0.13	0.34	0.03	0.16	0.04	0.19
Research Support, U.S. Gov't, P.H.S.	0.39	0.49	0.01	0.09	0.02	0.14	0	0.05	0.03	0.16
Research Support, ARRA	0	0	0	0	0	0.05	0	0.01	0	0.01
Research Support, Non-U.S. Gov't	0.48	0.5	0.34	0.47	0.48	0.5	0.39	0.49	0.38	0.49
Review Article	0.11	0.32	0.12	0.33	0.12	0.32	0.1	0.3	0.11	0.31
English Abstract	0	0.02	0.07	0.26	0	0.02	0.06	0.23	0.06	0.23
Case Report	0.01	0.09	0.08	0.28	0.01	0.1	0.07	0.26	0.07	0.25
Comparative Study	0.12	0.32	0.09	0.29	0.06	0.24	0.06	0.23	0.08	0.26
Meta-Analysis	0	0.05	0	0.06	0.01	0.07	0.01	0.08	0	0.07
Evaluation Studies	0.02	0.14	0.02	0.15	0.01	0.11	0.02	0.13	0.02	0.14
Guideline	0	0.02	0	0.05	0	0.03	0	0.04	0	0.04
Multicenter Study	0.02	0.13	0.01	0.12	0.02	0.14	0.01	0.12	0.01	0.12
Observational Study	0	0	0	0	0	0.03	0	0.04	0	0.03
Randomized Controlled Trial	0.03	0.16	0.02	0.15	0.03	0.17	0.02	0.15	0.02	0.15
Technical Report	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01
Twin Study	0	0.05	0	0.02	0	0.04	0	0.02	0	0.02
Validation Studies	0.01	0.1	0.01	0.09	0.01	0.08	0.01	0.08	0.01	0.09
Clinical Trial	0.03	0.16	0.03	0.16	0.02	0.13	0.01	0.12	0.02	0.14
Irregular Article	0.02	0.14	0.12	0.32	0.04	0.19	0.11	0.31	0.1	0.31
Other Language	0	0.01	0.02	0.14	0	0.01	0.01	0.12	0.02	0.12
English	1	0.02	0.89	0.31	1	0.02	0.92	0.26	0.92	0.27
German	0	0	0.01	0.11	0	0	0.01	0.1	0.01	0.1
French	0	0.01	0.02	0.12	0	0.01	0.01	0.1	0.01	0.11
Russian	0	0.01	0.01	0.1	0	0.01	0.01	0.09	0.01	0.09
Japanese	0	0	0.01	0.11	0	0	0.01	0.09	0.01	0.09
Spanish	0	0.01	0.01	0.11	0	0.01	0.01	0.1	0.01	0.1
Italian	0	0	0	0.06	0	0	0	0.04	0	0.05
Chinese	0	0.01	0.02	0.15	0	0.01	0.02	0.14	0.02	0.13
Other Grant Count	0.07	0.37	0.02	0.21	0.17	0.88	0.05	0.39	0.05	0.38
Commercial Affiliation	0.01	0.1	0.02	0.15	0.01	0.09	0.02	0.14	0.02	0.14
Educational Affiliation	0.65	0.48	0.45	0.5	0.65	0.48	0.49	0.5	0.49	0.5
Eductional/Hospital Affiliation	0.16	0.37	0.16	0.37	0.15	0.36	0.17	0.37	0.16	0.37
Government Affiliation	0	0.05	0.01	0.07	0	0.05	0.01	0.07	0	0.07
Hospital Affiliation	0.07	0.25	0.11	0.32	0.07	0.26	0.1	0.3	0.1	0.3
Military Affiliation	0	0.03	0	0.04	0	0.03	0	0.04	0	0.04
Organization Affiliation	0.09	0.29	0.1	0.3	0.1	0.29	0.09	0.29	0.1	0.29
Unkown Affiliation	0.02	0.13	0.15	0.35	0.02	0.15	0.12	0.33	0.12	0.33
Observations	448,326		3,605,437		508,475		3,877,126		8,439,364	

Table A1.b: Summary Statistics for the Journal Sample.

	NIH Pre		Comp. Pre		NIH Post		Comp. Post		All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Outcome Variables										
TA Journal	0.97	0.18	0.97	0.18	0.9	0.3	0.89	0.31	0.93	0.26
Science/Nature	0.01	0.08	0.01	0.1	0.01	0.09	0.01	0.08	0.01	0.09
2-Yr For. Cites (Total)	10.41	17.97	4.77	11.89	7.13	18.85	3.13	9.08	4.58	12.1
2-Yr Forward Cites (Com. Enterprise)	0.2	0.8	0.12	0.68	0.1	0.57	0.06	0.43	0.1	0.58
2-Yr Forward Cites (Dev. Country)	0.21	0.81	0.18	1.7	0.07	0.66	0.07	0.77	0.13	1.23
Covariates										
Backward Cites	36.15	30.92	20.83	25.66	37.62	35.18	22.66	27.47	23.95	28.18
OA Backward Cites	0.52	1.41	0.31	1.09	1.34	2.7	0.78	1.97	0.62	1.74
Age 0 Top Concepts	0.02	0.2	0.02	0.17	0.03	0.22	0.02	0.2	0.02	0.19
Age ≤ 5 Top Concepts	0.33	0.94	0.23	0.77	0.27	0.84	0.19	0.71	0.22	0.76
Total Concepts	138.42	47.04	108.7	58.84	120.35	67.83	95.93	68.55	105.73	64.53
Total MeSH Descriptors	13.86	5.24	11.22	5.69	13.22	5.37	10.77	6.02	11.34	5.86
Total MeSH Qualifiers	9.01	5.8	6.84	5.21	8.68	5.93	6.64	5.42	7.03	5.45
Author Count	5.34	3.89	4.63	7.55	6.17	6.12	5.15	14.71	5.03	11.21
Corporate Author	0	0.01	0.01	0.08	0	0.01	0.01	0.07	0	0.07
Journal Article	0.99	0.12	0.91	0.29	0.98	0.15	0.91	0.28	0.92	0.27
Research Support, U.S. Gov't, Non-P.H.S.	0.14	0.35	0.03	0.18	0.13	0.34	0.03	0.17	0.05	0.21
Research Support, U.S. Gov't, P.H.S.	0.39	0.49	0.01	0.1	0.02	0.14	0	0.06	0.03	0.18
Research Support, ARRA	0	0	0	0	0	0.05	0	0.01	0	0.01
Research Support, Non-U.S. Gov't	0.48	0.5	0.4	0.49	0.48	0.5	0.45	0.5	0.44	0.5
Review Article	0.11	0.32	0.11	0.32	0.12	0.32	0.1	0.3	0.11	0.31
English Abstract	0	0.02	0.01	0.11	0	0.02	0.01	0.09	0.01	0.09
Case Report	0.01	0.09	0.07	0.26	0.01	0.1	0.06	0.24	0.06	0.23
Comparative Study	0.12	0.32	0.1	0.3	0.06	0.24	0.06	0.23	0.08	0.27
Meta-Analysis	0	0.05	0	0.06	0.01	0.07	0.01	0.09	0.01	0.07
Evaluation Studies	0.02	0.14	0.03	0.16	0.01	0.11	0.02	0.14	0.02	0.15
Guideline	0	0.02	0	0.04	0	0.03	0	0.04	0	0.04
Multicenter Study	0.02	0.13	0.02	0.12	0.02	0.14	0.02	0.13	0.02	0.13
Observational Study	0	0	0	0	0	0.03	0	0.04	0	0.03
Randomized Controlled Trial	0.03	0.16	0.03	0.16	0.03	0.17	0.03	0.16	0.03	0.16
Technical Report	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01
Twin Study	0	0.05	0	0.02	0	0.04	0	0.02	0	0.03
Validation Studies	0.01	0.1	0.01	0.09	0.01	0.08	0.01	0.09	0.01	0.09
Clinical Trial	0.03	0.16	0.03	0.17	0.02	0.13	0.01	0.12	0.02	0.15
Irregular Article	0.02	0.14	0.11	0.32	0.04	0.19	0.11	0.31	0.1	0.3
Other Language	0	0.01	0	0.06	0	0.01	0	0.05	0	0.05
English	1	0.02	0.98	0.12	1	0.02	0.99	0.11	0.99	0.11
German	0	0	0	0.03	0	0	0	0.03	0	0.03
French	0	0.01	0	0.05	0	0.01	0	0.05	0	0.05
Russian	0	0.01	0	0.04	0	0.01	0	0.03	0	0.03
Japanese	0	0	0	0.02	0	0	0	0.03	0	0.02
Spanish	0	0.01	0	0.05	0	0.01	0	0.06	0	0.05
Italian	0	0	0	0.02	0	0	0	0.01	0	0.02
Chinese	0	0.01	0	0.07	0	0.01	0	0.05	0	0.05
Other Grant Count	0.07	0.37	0.03	0.23	0.17	0.88	0.07	0.43	0.06	0.42
Commercial Affiliation	0.01	0.1	0.03	0.16	0.01	0.09	0.02	0.14	0.02	0.14
Educational Affiliation	0.65	0.48	0.5	0.5	0.65	0.48	0.52	0.5	0.53	0.5
Eductional/Hospital Affiliation	0.16	0.37	0.16	0.36	0.15	0.36	0.16	0.37	0.16	0.36
Government Affiliation	0	0.05	0.01	0.07	0	0.05	0.01	0.07	0.01	0.07
Hospital Affiliation	0.07	0.25	0.1	0.3	0.07	0.26	0.09	0.28	0.09	0.28
Military Affiliation	0	0.03	0	0.05	0	0.03	0	0.04	0	0.04
Organization Affiliation	0.09	0.29	0.11	0.31	0.1	0.29	0.1	0.3	0.1	0.3
Unkown Affiliation	0.02	0.13	0.11	0.31	0.02	0.15	0.1	0.3	0.09	0.29
Observations	448,326		2,714,446		508,475		3,078,109		6,749,356	

Table A1.c: Summary Statistics for the Full PRCA Sample.

	NIH Pre		Comp. Pre		NIH Post		Comp. Post		All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Outcome Variables										
TA Journal	0.97	0.18	0.96	0.19	0.9	0.3	0.88	0.32	0.93	0.26
Science/Nature	0.01	0.08	0.01	0.07	0.01	0.09	0	0.06	0.01	0.07
2-Yr For. Cites (Total)	10.45	18.05	5.33	12.06	7.17	19.03	3.89	11.17	5.59	13.69
2-Yr Forward Cites (Com. Enterprise)	0.21	0.8	0.14	0.73	0.11	0.58	0.07	0.5	0.12	0.65
2-Yr Forward Cites (Dev. Country)	0.21	0.81	0.18	0.76	0.07	0.67	0.08	1.03	0.14	0.87
Covariates										
Backward Cites	36.09	30.95	23.64	27.76	37.54	35.17	26.83	30.64	27.78	30.54
OA Backward Cites	0.52	1.41	0.36	1.16	1.35	2.71	1.01	2.32	0.73	1.93
Age 0 Top Concepts	0.02	0.2	0.02	0.17	0.03	0.22	0.02	0.21	0.02	0.19
Age ≤ 5 Top Concepts	0.33	0.94	0.25	0.81	0.27	0.85	0.24	0.8	0.26	0.82
Total Concepts	138.39	47.1	117.79	56.95	120.12	68.02	105.82	70.59	116	63.37
Total MeSH Descriptors	13.87	5.26	12.32	5.54	13.25	5.38	12.46	5.8	12.65	5.61
Total MeSH Qualifiers	9.02	5.82	7.65	5.41	8.68	5.94	7.87	5.76	8	5.66
Author Count	5.34	3.9	4.72	3.25	6.19	6.16	5.4	3.83	5.21	4.01
Corporate Author	0	0.01	0.01	0.08	0	0.01	0	0.06	0	0.06
Journal Article	0.99	0.12	0.94	0.24	0.98	0.15	0.95	0.22	0.95	0.21
Research Support, U.S. Gov't, Non-P.H.S.	0.14	0.35	0.03	0.18	0.13	0.34	0.03	0.18	0.06	0.23
Research Support, U.S. Gov't, P.H.S.	0.39	0.49	0.01	0.11	0.02	0.14	0.01	0.07	0.05	0.22
Research Support, ARRA	0	0	0	0	0	0.05	0	0.01	0	0.02
Research Support, Non-U.S. Gov't	0.48	0.5	0.46	0.5	0.49	0.5	0.53	0.5	0.49	0.5
Review Article	0.11	0.32	0.13	0.34	0.12	0.33	0.12	0.33	0.12	0.33
English Abstract	0	0.02	0.05	0.22	0	0.02	0.04	0.19	0.04	0.18
Case Report	0.01	0.1	0.04	0.19	0.01	0.1	0.03	0.17	0.03	0.17
Comparative Study	0.12	0.32	0.11	0.31	0.06	0.24	0.07	0.25	0.09	0.28
Meta-Analysis	0	0.05	0	0.06	0.01	0.07	0.01	0.09	0	0.07
Evaluation Studies	0.02	0.14	0.03	0.16	0.01	0.11	0.02	0.14	0.02	0.15
Guideline	0	0.02	0	0.04	0	0.03	0	0.04	0	0.03
Multicenter Study	0.02	0.13	0.02	0.12	0.02	0.14	0.02	0.14	0.02	0.13
Observational Study	0	0	0	0	0	0.03	0	0.04	0	0.03
Randomized Controlled Trial	0.03	0.16	0.02	0.15	0.03	0.17	0.03	0.17	0.03	0.16
Technical Report	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01
Twin Study	0	0.05	0	0.02	0	0.04	0	0.03	0	0.03
Validation Studies	0.01	0.1	0.01	0.1	0.01	0.08	0.01	0.09	0.01	0.09
Clinical Trial	0.03	0.16	0.03	0.17	0.02	0.13	0.02	0.13	0.02	0.15
Irregular Article	0.02	0.14	0.07	0.26	0.04	0.19	0.07	0.25	0.06	0.24
Other Language	0	0.01	0.01	0.11	0	0.01	0.01	0.09	0.01	0.09
English	1	0.02	0.93	0.25	1	0.02	0.95	0.21	0.96	0.21
German	0	0.01	0.01	0.09	0	0	0.01	0.07	0.01	0.07
French	0	0.01	0.01	0.09	0	0.01	0.01	0.08	0.01	0.08
Russian	0	0.01	0.01	0.08	0	0.01	0	0.07	0	0.07
Japanese	0	0	0.01	0.08	0	0	0	0.07	0	0.07
Spanish	0	0.01	0.01	0.08	0	0.01	0.01	0.08	0	0.07
Italian	0	0	0	0.04	0	0	0	0.03	0	0.03
Chinese	0	0.01	0.02	0.12	0	0.01	0.01	0.11	0.01	0.1
Other Grant Count	0.07	0.37	0.03	0.26	0.17	0.89	0.11	0.55	0.08	0.5
Commercial Affiliation	0.01	0.1	0.03	0.16	0.01	0.09	0.02	0.14	0.02	0.14
Educational Affiliation	0.65	0.48	0.51	0.5	0.65	0.48	0.53	0.5	0.55	0.5
Eductional/Hospital Affiliation	0.16	0.37	0.16	0.37	0.15	0.36	0.17	0.38	0.16	0.37
Government Affiliation	0	0.05	0.01	0.07	0	0.05	0.01	0.07	0	0.07
Hospital Affiliation	0.07	0.25	0.09	0.29	0.07	0.26	0.08	0.28	0.08	0.28
Military Affiliation	0	0.03	0	0.04	0	0.03	0	0.04	0	0.04
Organization Affiliation	0.09	0.29	0.11	0.31	0.1	0.29	0.11	0.31	0.1	0.31
Unkown Affiliation	0.02	0.13	0.09	0.29	0.02	0.15	0.08	0.27	0.07	0.26
Observations	437,941		1,707,488		494,907		146,4350		4,104,686	

Table A1.d: Summary Statistics for the 1-to-1 PRCA Sample.

	NIH Pre		Comp. Pre		NIH Post		Comp. Post		All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Outcome Variables										
TA Journal	0.97	0.18	0.96	0.19	0.9	0.3	0.88	0.33	0.92	0.27
Science/Nature	0.01	0.08	0.01	0.08	0.01	0.09	0	0.06	0.01	0.08
2-Yr For. Cites (Total)	10.43	18.04	7.77	15.31	7.12	19.12	5.13	15.75	7.53	17.25
2-Yr Forward Cites (Com. Enterprise)	0.21	0.8	0.19	0.85	0.11	0.58	0.09	0.57	0.15	0.71
2-Yr Forward Cites (Dev. Country)	0.21	0.82	0.23	0.81	0.08	0.68	0.09	1.69	0.15	1.09
Covariates										
Backward Cites	36.02	31	31.41	29.07	37.38	35.16	33.23	32.41	34.55	32.17
OA Backward Cites	0.52	1.41	0.48	1.36	1.35	2.73	1.27	2.53	0.93	2.17
Age 0 Top Concepts	0.02	0.2	0.02	0.19	0.03	0.22	0.03	0.23	0.03	0.21
Age ≤ 5 Top Concepts	0.34	0.94	0.35	0.97	0.27	0.85	0.3	0.9	0.31	0.92
Total Concepts	138.34	47.2	140.09	51.48	119.7	68.3	120.89	71.56	129.24	61.82
Total MeSH Descriptors	13.87	5.26	14.65	5.48	13.23	5.39	14.19	5.57	13.97	5.45
Total MeSH Qualifiers	9.01	5.82	9.78	6.07	8.65	5.94	9.58	6.26	9.25	6.05
Author Count	5.34	3.9	5.25	3.33	6.19	6.19	5.87	4	5.68	4.55
Corporate Author	0	0.01	0	0.04	0	0.01	0	0.04	0	0.03
Journal Article	0.99	0.12	0.98	0.14	0.98	0.15	0.97	0.16	0.98	0.14
Research Support, U.S. Gov't, Non-P.H.S.	0.14	0.35	0.04	0.2	0.13	0.34	0.03	0.18	0.09	0.28
Research Support, U.S. Gov't, P.H.S.	0.39	0.49	0.02	0.14	0.02	0.14	0.01	0.08	0.1	0.31
Research Support, ARRA	0	0	0	0	0	0.05	0	0.01	0	0.03
Research Support, Non-U.S. Gov't	0.48	0.5	0.6	0.49	0.49	0.5	0.64	0.48	0.55	0.5
Review Article	0.11	0.32	0.12	0.32	0.12	0.33	0.12	0.32	0.12	0.32
English Abstract	0	0.02	0.03	0.18	0	0.02	0.02	0.16	0.01	0.12
Case Report	0.01	0.1	0.01	0.11	0.01	0.1	0.01	0.12	0.01	0.11
Comparative Study	0.12	0.32	0.13	0.33	0.06	0.24	0.07	0.25	0.09	0.29
Meta-Analysis	0	0.05	0	0.06	0.01	0.07	0.01	0.09	0	0.07
Evaluation Studies	0.02	0.14	0.03	0.16	0.01	0.12	0.02	0.13	0.02	0.14
Guideline	0	0.02	0	0.03	0	0.03	0	0.03	0	0.02
Multicenter Study	0.02	0.13	0.02	0.13	0.02	0.15	0.02	0.14	0.02	0.14
Observational Study	0	0	0	0	0	0.03	0	0.04	0	0.02
Randomized Controlled Trial	0.03	0.16	0.02	0.15	0.03	0.17	0.03	0.17	0.03	0.16
Technical Report	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01
Twin Study	0	0.04	0	0.03	0	0.04	0	0.03	0	0.04
Validation Studies	0.01	0.1	0.01	0.1	0.01	0.09	0.01	0.08	0.01	0.09
Clinical Trial	0.03	0.16	0.03	0.17	0.02	0.13	0.02	0.13	0.02	0.15
Irregular Article	0.02	0.14	0.03	0.17	0.04	0.19	0.04	0.2	0.03	0.18
Other Language	0	0.01	0.01	0.08	0	0.01	0	0.07	0	0.05
English	1	0.02	0.97	0.18	1	0.02	0.97	0.16	0.98	0.12
German	0	0.01	0	0.06	0	0	0	0.05	0	0.04
French	0	0.01	0	0.06	0	0.01	0	0.06	0	0.04
Russian	0	0.01	0	0.05	0	0.01	0	0.05	0	0.04
Japanese	0	0	0	0.06	0	0	0	0.05	0	0.04
Spanish	0	0.01	0	0.05	0	0.01	0	0.05	0	0.04
Italian	0	0	0	0.03	0	0	0	0.02	0	0.02
Chinese	0	0.01	0.01	0.11	0	0.01	0.01	0.1	0.01	0.07
Other Grant Count	0.07	0.36	0.05	0.32	0.17	0.87	0.15	0.66	0.11	0.61
Commercial Affiliation	0.01	0.1	0.03	0.16	0.01	0.09	0.02	0.15	0.02	0.13
Educational Affiliation	0.65	0.48	0.57	0.49	0.64	0.48	0.57	0.5	0.61	0.49
Eductional/Hospital Affiliation	0.16	0.37	0.15	0.36	0.15	0.36	0.17	0.37	0.16	0.36
Government Affiliation	0	0.05	0	0.07	0	0.05	0.01	0.07	0	0.06
Hospital Affiliation	0.07	0.25	0.07	0.25	0.07	0.26	0.07	0.26	0.07	0.26
Military Affiliation	0	0.03	0	0.04	0	0.03	0	0.04	0	0.03
Organization Affiliation	0.09	0.29	0.13	0.34	0.09	0.29	0.12	0.32	0.11	0.31
Unkown Affiliation	0.02	0.13	0.04	0.2	0.02	0.15	0.05	0.22	0.03	0.18
Observations	431,647		431,647		481,002		481,002		1,825,298	