Mandating Access: An Empirical Assessment of the NIH's Public Access Policy

Joseph Staudt*

Department of Economics, Ohio State University (OSU)

February 2014

Abstract

In 2008, the National Institutes of Health (NIH) mandated that articles supported by a NIH grant be made freely available on PubMed Central. We hypothesize that this reduced the incentives of toll access journals to publish NIH-supported articles but did not impact the incentives of open access journals. We use difference-in-differences estimators and a large sample of articles in the life sciences to test this hypothesis. Our estimates strongly suggest that the NIH mandate shifted the publication of NIH-supported articles from toll access to open access journals. Since toll access journals tend to be higher quality than open access journals, we also estimate that the mandate decreased the quality of journal in which NIH-supported articles are published. These results suggest that the NIH should consider policies to reduce the disincentive effects of the mandate and that other funding agencies should proceed with caution as they design and implement their own open access mandates.

Keywords: economics of science, open access, nih, nih public access policy, policy evaluation.

JEL Classification Numbers: 031, 034, 038

*Address: 1945 N. High St., 410 Arps Hall, Columbus OH, 43210, USA, e-mail: *staudt.8@osu.edu*. I am grateful to Bruce Weinberg for excellent guidance throughout the life of this project. I would also like to thank David Blau, Daeho Kim, Richard Steckel, Robert Munk, and Garrett Senney for many helpful comments and discussions. Finally, thanks to Jaroslav Horvath, Heejong Kim, Neslihan Sakarya, Tatsuro Senga, and Jing Zhang for translation assistance. All remaining mistakes are my own.

1 Introduction

The production of new ideas is essential for economic growth (Lucas, 1988; Romer, 1986). One important source of new ideas is science. Because science is cumulative[1], scientists must have broad access to journal articles that are relevant to their research. Sharply increasing journal prices in recent decades[2], along with stagnant library acquisition budgets, have caused concern about a narrowing of scientists' access to journal articles (Peek, 2008; RIN, 2011). A common response to these concerns has been louder calls for all scientific articles to become open access–"digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber, 2012, p. 4).[3] One important tool used to increase the number of open access articles has been open access mandates by departments, universities, and funding agencies, which typically stipulate that authors affiliated with or receiving funding from the mandating organization must take certain steps to ensure that their research is available to anyone, free of charge.[4] Figure 1 shows that, as late as 2003, there was only a single open access mandate. By 2013, there were nearly 350 such mandates. This paper seeks to better understand how open access mandates impact scientific publishing by analyzing the impacts, on publishing patterns in the biomedical sciences, of arguably the most important of these mandates: the 2008 National Institutes of Health's Public Access Policy.

The NIH Public Access Policy stipulates that any article reporting research funded by the NIH must be made freely available on PubMed Central–a digital full-text repository hosted by the NIH.[5] We hypothesize that the mandate reduced the incentive of toll access journals to publish NIH-supported articles but did not impact the incentive of open access journals. This is due to the different ways toll and open access journals earn revenue. Toll access journals earn revenue by selling subscriptions, which partially relies on the exclusivity of journal content. Open access journals earn revenue by charging submission or publication fees to authors, which does not rely on content exclusivity. The NIH mandate limits the exclusivity of NIH-supported articles by making them freely available on PubMed Central.

[1] Aghion et al. (2008); Aghion and Howitt (1992); Mokyr (2002); Murray et al. (2009); Romer (1990); Scotchmer (1991).

[2] Albee and Dingley (2000, 2001, 2002); ARL (2011); Bergstrom (2001); Bosch et al. (2011); Bosch and Henderson (2012, 2013, 2014); Henderson and Bosch (2010); Dingley (2003, 2004, 2005).

[3] The first formal public statement calling for and defining open access was the Budapest Open Access Initiative of 2002. This was quickly followed by the Bethesda Statement on Open Access and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, both in 2003.

[4] Consult the Registry of Open Access Repositories Mandatory Archiving Policies (ROARMAP) for details on these mandates.

[5] PubMed Central was developed by the National Center for Biotechnology Information (NCBI), which is part of the National Library of Medicine (NLM), which is a branch of the NIH. NIH-supported articles must be submitted, in final peer-reviewed form, to PubMed Central immediately upon acceptance for publication, but can be embargoed for up to 12 months after publication.

Thus, we predict that toll access journals are less likely to publish NIH-supported articles after the mandate.[6]

Using difference-in-differences estimators, along with data on 7 million articles and 5,600 journals from MEDLINE and the Directory of Open Access Journals (DOAJ), this paper presents strong evidence in favor of this hypothesis. Our main set of results show that the mandate increased the probability that a NIH-supported article is published in an open access journal by approximately 0.13 to 0.14 percentage points per year. This implies that, absent the mandate, approximately 41 percent of NIH-supported articles published in open access journals between 2009 and 2012 would have been published in toll access journals. We also show that the mandate did not impact the probability that an article supported by other grant types is published in an open access journal. The fact that the mandate impacted NIH-supported articles but not articles supported by other grants bolsters the causal interpretation of our results. We additionally show that, while the impact of NIH-support on the probability of an article being published in an open access journal was trending downward prior to the mandate, this trend suddenly stopped in 2008–the year of the NIH mandate. Since the trend break occurs precisely when we would expect, we interpret it as further validation that our main results are causal.

If the mandate caused toll access journals to be more reluctant to publish NIH-supported articles, we would expect that it also caused them to require the fewer NIH-supported articles that they do publish to clear a higher quality hurdle. Indeed, our second set of results show that the mandate increased the average number of citations per year to NIH-supported articles published in toll access journals by approximately 0.40. Since the median number of citations per year for NIH-supported articles in our sample is about 3.33, this impact is quite large. In contrast, there is no evidence that citations to NIH-supported articles published in open access journals were impacted by the mandate.

The foregoing results are important because, as we will show, toll access journals tend to be higher quality than open access journals. Thus, by shifting some NIH-supported articles

[6]There is some anecdotal evidence that toll access journals are less likely to publish NIH-supported articles after the mandate. The Association of American Publishers, which represents all major publishers in biomedicine, strongly opposes the NIH mandate. The association specifically warns that the mandate undermines its members' economic incentives by making their content available online (www.publishers.org/issues/5/9). Moreover, at least one member of the publishing industry explicitly suggests that "Another possible implication (of the NIH mandate) is that journals may no longer be willing to review and accept articles with unsustainable terms attached" (McMullan, 2008). Though prior to the NIH mandate, Seamans (2001) found that, in a sample of mostly non-profit journals, 17.64 percent expressed serious 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.

from toll access to open access journals, the NIH mandate actually caused NIH-supported articles to be published in lower quality journals. Indeed, our third set of results show that the mandate caused NIH-supported articles to be published in journals approximately 1.2 to 2.4 percentiles lower in the journal quality distribution (as measured by average citations per document).[7] This implies that the mandate caused the typical NIH-supported article to be published in journals receiving about 8 percent fewer citations per document. Insofar as articles published in higher impact journals receive more attention, the mandate may have, ironically, decreased the exposure of researchers to NIH-supported articles by shifting a greater proportion of those articles into open access journals. Moreover, since NIH-supported articles tend to be higher quality than non NIH-supported articles, this shift may have actually lowered the average quality of article that researchers read.

The results outlined above may be of little concern if the NIH mandate achieved its primary goal: increasing access to biomedical literature. However, a set of studies that econometrically analyze the impacts, on citations, of making an article open access casts doubt on whether this goal was achieved. Early studies indicate that open access has a very large impact on citations.[8] A large impact indicates that there is a large pool of researchers that do not have access to articles unless they are made open access. Unfortunately, it is difficult to draw reliable causal conclusions from these early studies because they all use observational cross-sectional data. Indeed, a more recent wave of studies that explicitly attempt to deal with the endogeneity of open access have all found a much smaller estimate of the impact of open access on citations.[9] Thus, overall, there is very little evidence that

[7] In fact, this decrease in the quality of journal in which NIH-supported articles are published could have happened through at least two different channels. First, as emphasized here, by causing toll access journals to discriminate against NIH-supported articles, the mandate shifted some NIH-supported articles from toll access (higher quality) to open access (lower quality) journals. However, this discrimination could also cause NIH-supported articles to be published in lower quality toll access journals.

[8] Lawrence (2001) paved the way using a sample of articles from computer science conferences. He finds that that open access increased citations by an astounding 336 percent. This large impact was quickly validated in many other disciplines. Antelman (2004) and Davis and Fromerth (2007) demonstrate the large impact on citations of open access for mathematics. Antelman also demonstrates the impact for philosophy, political science, and engineering. Schwarz and Kennicutt Jr (2004) and Metcalfe (2005, 2006) do so for astrophysics, Harnad and Brody (2004) do so for physics, and Walker (2004) does so for oceanography. Finally, Eysenbach (2006) demonstrates the large impact on citations of open access in multidisciplinary science. Craig et al. (2007) provide a useful review of this early literature.

[9] Evans and Reimer (2009) and McCabe and Snyder (2014) use panel data to estimate the impact of open access on citations. Both studies find that open access increases citations by approximately 8 percent. Gaule and Maystre (2011) attempt to deal with the endogeneity of open access using an instrumental variables approach. Specifically, they use a sample of biology papers published in the Proceedings of the National Academy of Sciences (PNAS), which allows authors to pay an optional $1,000 fee in order to make their article available, at no charge to the reader, immediately upon publication. They instrument for the endogenous decision of the authors to pay the fee using a variable indicating whether the article was published in the last quarter of the fiscal year. They reason that departments are more willing to pay for low priority items like publication fees when they are about to lose unused budgets at the end of the fiscal year. They do not find

open access increases citations, which is consistent with researchers having extensive access to toll access content through institutional subscriptions. In this light, it is unlikely that the NIH mandate substantially increased researcher access to the biomedical literature.[10]

Uncovering how the NIH Public Access Policy impacts publishing patterns in biomedicine is important for several reasons. First, the NIH is the largest funder of medical research in the world–by a wide margin. Chakma et al. (2014) document that in the year 2012, the NIH's biomedical R&D expenditures were $30.9 billion. This accounted for about 63 percent of publicly funded biomedical R&D expenditures and 26 percent of total biomedical expenditures in the United States. In contrast, the 2012 total of publicly funded biomedical R&D across all European countries was $28.1 billion.[11] In Japan, Australia, Canada, and China, it was $9.5 billion, $4.7 billion, $3.3 billion, and $2.0 billion. Thus, if any open access mandate is able to impact publishing patterns, it is likely to be the NIH's.

Second, while some open access mandates require articles to be posted on a researcher's personal homepage or an institutional repository, the NIH mandate requires articles to be submitted to PubMed Central.[12] This is significant because, as Guedon (2004) points out, open access does not necessarily imply accessibility. Articles dispersed across author homepages and difficult-to-search institutional repositories are often not very accessible, even if they are technically open access. In contrast, PubMed Central is an exceptionally easy to use repository, with many useful search features–including the use of medical subject headings (MeSH), which classify articles by topic, and are specifically designed to expedite literature searches.

Finally, the NIH mandate is currently serving as a model for upcoming open access mandates from other federal funding agencies. Indeed, in February 2013 the U.S. Office of Science and Technology Policy issued a memorandum instructing all federal agencies that have over $100 million in annual R&D expenditures to develop an open access policy similar to the NIH's Public Access policy.[13] Thus, understanding how the NIH mandate impacted

a statistically significant impact of open access on citations. Finally, Davis et al. (2008) and Davis (2011) conduct randomized controlled trials. The first study randomly assign articles published in 11 journals owned by the American Physiological Society to be open access. The second study does the same for articles published in 36 journals owned by a variety of publishers. Neither study found a statistically significant increase in the number of citations to open access articles.

[10]Note, however, that these citation studies say nothing about increased access for physicians, nurses, or lay people. Though researchers are the primary consumers of scientific research, these other groups may derive value from greater access to the biomedical literature.

[11]Europe is defined as the EU countries plus Switzerland, Norway, and Iceland.

[12]The European Research Council (ERC) is the most prominent example of a funding organization whose mandate allows the author to deposit the article into a university repository. Other examples include the Canadian Institutes of Health Research (CIHR), the Australian Research Council (ARC), and the Austrian Science Research Fund (FWF). See the Registry of Open Access Repositories Mandatory Archiving Polices (ROARMAP) for details.

[13]Specifically, the memo states that, "The Office of Science and Technology Policy (OSTP) hereby directs

publishing patterns in biomedicine will shed light on how we can expect new mandates to impact the publishing patterns in the fields that they fund.

The rest of this paper is organized as follows. Section 2 discusses the fundamental differences between open access and toll access journals. It also discusses the different types of open access. Section 3 discusses, in detail, the NIH Public Access Policy. It also discusses the econometric strategy and estimation procedure we use to estimate the impact of the NIH mandate on publishing patterns in the biomedical sciences. Section 4 discusses our data, section 5 presents our results, and section 6 concludes.

2 Open Access Versus Toll Access

Before proceeding, it is useful to briefly describe the essential differences between open access journals and toll access (subscription-based) journals. Traditionally, journals' primary source of revenue has been subscriptions. Under this arrangement, a journal charges libraries (or individuals) a fee to access the journal's content. In contrast, open access journals charge no fee to readers who wish to access content. Rather, these journals' revenue is primarily earned by charging the article authors to either submit or publish their work.[14] Thus, toll access journals earn revenue from reader-side fees and open access journals earn revenue from author-side fees.

This distinction is important for this paper because revenue from reader-side fees partially relies on the exclusivity of journal content while revenue from author-side fees does not. If journal content is available at a lower price elsewhere (such as a price of zero at PubMed Central), then readers (or institutions) will be less inclined to purchase the same content at a higher price from the journal itself, causing the journal to lose revenue. Thus, journals that use reader-side fees as their main revenue source (toll access journals) are less likely to publish content that is non-exclusive (such as NIH-supported articles). In contrast, journals that use author-side fees (open access journals) already charge a subscription price of zero, and so do not lose revenue if journal content is also available elsewhere at a price of zero.

In reality, there is not always a neat distinction between open and toll access journals. Some journals are full open access, which means they offer immediate universal access to all content at no charge. Other journals offer the option of making an article open access if the authors are willing to pay a sometimes substantial fee. In this case, some articles within a journal are open access and other articles are toll access. Still other journals embargo content

each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government." See: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

[14] In practice, these fees are often paid by author institutions or funding organizations.

for months or years, and then make it freely available after the embargo expires. There are also different "shades" of open access. The "gold route" to open access is achieved when authors publish their work in open access journals. An alternative route is the "green route", which means the author self-archives on a personal webpage, an institutional repository, or a field repository.[15] The NIH mandate forces NIH-supported articles to take the green route to open access by forcing authors to submit these articles to PubMed Central.

This paper can be thought of as exploring interaction between the green and gold routes. Specifically, we examine how the NIH's mandate that NIH-supported articles become open access via the green route (articles must be submitted to PubMed Central) impacts articles' propensity to become open access via the gold route (be published in open access journals). Due to data limitations, we will restrict our attention to full open access journals.

3 Research Design

3.1 Details of NIH Public Access Policy

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.[16] This request became effective on May 2, 2005.[17] Despite the request, a 2006 NIH report to Congress stated that voluntary compliance with this request was below 4 percent.[18] Thus, Congress instructed the NIH to change the request to a mandate. On January 11, 2008 the NIH announced that 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.[19] Though the article must be submitted to PubMed Central

[15]The terms "gold route" and "green route" were coined by Stevan Harnad (Suber, 2012, p. 53). For an extensive list of publisher policies on self-archiving, see SHERPA/RoMEO, based at the University of Nottingham.

[16]PubMed Central is the digital archive of the National Library of Medicine. It contains approximately 3 million full-text articles. See: http://www.ncbi.nlm.nih.gov/books/NBK153388/.

[17]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.

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

[19]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."

immediately upon acceptance for publication, the author retained the option of embargoing the article for up to 12 months after publication. By 2012 compliance stood at 75 percent.[20] This increase is suggested in figure 2 by the jump in the number of manuscripts submitted to PubMed Central after the announcement that the request had become a mandate. It is crucial to recall that the NIH does not mandate that NIH-supported articles be published in an open access journal. Rather, it mandates that, regardless of whether an article is published in an open or toll access journal, it must be submitted to PubMed Central where it will be made freely available. Thus, the mandate essentially limits journals' proprietary control over NIH-supported articles.

3.2 Econometric Strategy

Ideally, to test the impact on publishing patterns of an open access mandate, we would randomly mandate that some pre-publication articles be submitted to PubMed Central upon acceptance for publication. Moreover, to mimic the impact of the NIH mandate, we would inform the potential publisher of this condition. This would identify the causal effects of an open access mandate on publication patterns.

Obviously we cannot carry out this experiment. However, the size and discreteness of the NIH mandate, coupled with the fact that it applied to NIH-supported articles but not to other articles, gives us an opportunity to reduce omitted variable bias. Indeed, this strategy enables us to credibly identify the causal effects of the NIH mandate on publishing patterns in the biomedical sciences.

As a first pass at identifying the effect of the mandate, we estimate the following regression equation:

$$oa_{ajt} = \beta_t nihgrant_{ajt} + \delta_t controls_{ajt} + \nu_j + \alpha_t + \varepsilon_{ajt}. \quad (1)$$

The variable oa_{ajt} is an indicator variable for whether article a was published in an open access journal j during year t. The variable $nihgrant_{ajt}$ is an indicator for whether the article was NIH-supported. $controls_{ajt}$ is a vector of article-level and journal-level control variables, such as article topic, article citations, and journal quality, that will be explained further in sections 4 and 5. α_t are year fixed effects and ν_j is a journal-specific error component.

The coefficients of interest are the β_t, which measure, for each year, the impact of NIH-support on the probability of an article being published in in an open access journal. If we see a break in the trend of the β_t around 2008–the year in which the NIH mandate was implemented–we will have evidence that the mandate impacted publishing patterns in the biomedical sciences. In particular, we will have evidence that the mandate impacted the

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

probability that a NIH-supported article is published in an open access journal.

Estimation of equation (1) produces some very illustrative graphs that show how the mandate caused changes in publishing patterns in the biomedical sciences. However, our main results are obtained by estimating the following regression equation:

$$\begin{aligned} oa_{ajt} &= \beta nihgrant_{ajt} + \gamma Post08_t + \\ &\quad \delta(nihgrant_{ajt} * Post08_t) + \\ &\quad \eta(nihgrant_{ajt} * t) + \theta(Post08_t * t) + \\ &\quad \lambda(nihgrant_{ajt} * Post08_t * t) + \\ &\quad controls_{ajt} + \nu_j + \alpha_t + \varepsilon_{ajt}. \end{aligned} \tag{2}$$

In this equation, $Post08_t$ indicates years after 2008. The rest of the variables are defined as in equation (1). This is simply a difference-in-differences specification with a linear time trend. Our coefficient of interest is λ, which measures the causal effect of the mandate on the variable oa_{ajt}. Note that the time trend allows this impact to change linearly over time. Since the mandate only applies to NIH-supported articles published after 2008, this set of articles serves as our treatment group. NIH-supported articles published before the mandate, and non-NIH-supported articles published both before and after the mandate, serve as our control group.

3.3 Estimation Procedure

Given the structure of our data, the ordinary least squares (OLS) estimator is extremely inefficient. The primary reason for this is the very large variation in the number of articles published in different journals. Indeed, the standard deviation of the number of articles published across journals is approximately 2,312, which is nearly twice the the mean of 1,295. For instance, our estimation sample contains 70,788 articles published in The Journal of Biological Chemistry while 157 journals published only a single article.

With this extreme variation in journal size (combined with a within-journal residual correlation of approximately 0.95), Scott and Holt (1982) show that the proportionate efficiency loss of the OLS estimator relative to a feasible generalized least squares (FGLS) estimator that assumes within-journal equicorrelated errors can be up to 99 percent. Thus, we opt to estimate equations (1) and (2) using a FGLS estimator that assumes within-journal equicorrelated errors[21] rather than the OLS estimator.

[21] This formulation is equivalent to assuming a random effects model, which is easily implemented in Stata and other statistical packages.

Of course, the trouble with using a FGLS estimator is that we must specify the structure of the error covariance matrix–in our case, within-journal equicorrelated errors. If this structure is misspecified then the standard errors of the FGLS estimator will be incorrect. However, as suggested in Cameron and Miller (2013), we will use standard errors that are clustered at the journal-level to guard against arbitrary misspecification of the error covariance matrix.[22]

It it is worth noting that the direction of the estimated impacts are the same for both the OLS and FGLS estimators. Indeed, the estimated impacts using the OLS estimator tend to be a bit larger, but are much less precise for reasons discussed above. The OLS estimates will be made available upon request.

3.4 Discussion of Mechanisms

A priori, the impact of the mandate on publishing patterns is ambiguous. As we have emphasized, the publisher of a toll access (non open access) journal would, ceteris paribus, be more reluctant to publish a NIH-supported article after the mandate. However, it is the editor of the journal, not the publisher, who makes the decision of whether or not to publish a particular article. If the editor has a different objective than the publisher, then the mandate's impact on the publisher's actions may be blunted.

Of course, even if the editor has a different objective function than the publisher, the publisher can decrease the impact of the editor's decisions on profits in at least four ways. First, and most extreme, the publisher could terminate the editor and replace her with an editor more amenable to pursuing the objective of the publisher.[23] Second, the publisher could pressure the editor to favor some articles over others.[24] Third, the publisher could choose to switch a journal's business model from toll access to open access. Finally, the publisher could start new open access journals in areas of biomedicine that are particularly

[22](Wooldridge, 2009, p. 287) suggests a similar procedure in the case of heteroskedastic errors. In particular, he suggests using the weighted least squares estimator, which is more precise than the OLS estimator, and then using heteroskedasticity-robust standard errors to guard against arbitrary misspecification of the variance function.

[23]Though extreme, this has actually happened in a number of high profile cases where the opinions of the editor and publisher diverged. In 1999, the American Medical Association fired George Lundberg, the editor of its flagship journal JAMA. This was the result of Lundberg's decision, in light of Bill Clinton's impeachment trial, to publish a study of college students' attitudes about whether oral sex is "sex" (see Smith, 1999b). Also in 1999, the Massachusetts Medical Society (MMS) fired Jerome P. Kassirer, editor of the New England Journal of Medicine. The MMS wanted to use the brand name of its flagship journal to promote its other journals. Kassirer resisted, arguing that such a move would possibly tarnish the name of the journal he edited. He was subsequently terminated (see Smith, 1999a).

[24]In a survey of 33 editors of medical journals owned by not for-profit publishers, Davis and Müllner (2002) found that, while most editors report having a high level of editorial freedom, a substantial minority (42 percent) have received at least some pressure from their publisher over editorial content.

likely to attract NIH-supported research.

On the other side of the submission relationship is the author. Ceteris paribus, an article will have a larger impact if it is freely available for anyone to read and cite (although as discussed in the introduction, this impact may be quite small). Also, an article will have a larger impact if it is published in a more prestigious journal. Given the fact that high-prestige journals tend to be toll access, the author faces a tradeoff between openness and prestige prior to the mandate. However, after the mandate, the author will be more likely to submit her article to a high-prestige toll access journal because the article will also be freely available on PubMed Central. In effect, after the mandate, the author no longer faces a tradeoff between prestige and openness.

In the end, we are agnostic with respect to the mechanism in operation. We would like to identify the net effects of the mandate on publication patterns in the biomedical sciences.

4 Data

Our first step is to identify every article and journal indexed in the MEDLINE database from 1999 through 2012.[25] MEDLINE is a bibliographic database for the National Library of Medicine (NLM), and includes nearly every meaningful article published in the biomedical sciences. It is the most comprehensive index of the biomedical literature that exists, and so is the best source of data for the present analysis. Our next step is to use a variety of sources to collect information on these articles and journals. This is discussed over the next several sub-sections.[26]

4.1 Open Access Journals

We use the Directory of Open Access Journals (DOAJ) to determine whether a journal is open access or toll access.[27] DOAJ is by far the most comprehensive index of open access journals, listing 9,709 journals as of April, 2014. Crucially, DOAJ documents when journals that were not born open access became open access. It also documents the more rare case of a journal switching from open access to toll access.

Using a journal's International Standard Serial Number (ISSN)–which uniquely identifies the journal–we are able to determine whether a journal indexed in MEDLINE is also indexed in DOAJ. If a journal is indexed in both MEDLINE and DOAJ, we know that the journal is currently an open access journal or was an open access journal at some point. We are able

[25] This is the time period for which we have data on all variables of interest.

[26] The Data Appendix to this paper discusses our data in much greater detail.

[27] DOAJ began as a project at Lund University in 2002, and is now an independent organization.

to use the dating information in DOAJ to identify when a journal switched from toll access to open access and vice versa. If a journal is indexed in MEDLINE, but is not indexed in DOAJ then the journal is classified as toll access.[28]

4.2 Grant Receipts

We use MEDLINE to extract data on whether an article was supported by an NIH grant. Every entry for a MEDLINE article includes a list of grants that supported the research discussed in the article. This enables us to identify all NIH-supported articles.

Unfortunately, there is the possibility of non-acknowledgment of NIH support. That is, some articles–whether by mistake or for strategic reasons–may fail to acknowledge that a NIH grant supported the article. In this case, we will mistakenly identify an article as not having received NIH support, when it did in fact receive such support. Since acknowledgment of taxpayer funding is required by federal law, we hope that non-acknowledgment is a minor problem.[29] However, we have not been able to find evidence suggesting what proportion of NIH-supported research fails to acknowledge this support.

An additional concern is that some grant lists in the MEDLINE database are incomplete. That is, some article entries only contain a partial list of grants. Fortunately, MEDLINE indicates which grant lists are complete and incomplete, allowing us to compute that only 0.4 percent of grant lists are incomplete. Thus, as long as the grant was acknowledged by the author and the National Library of Medicine did not make a mistake, we are able to identify the vast majority of NIH grants.

4.3 Controls

We use a variety of journal-level and article-level control variables to account for time-varying heterogeneity that may cause our estimators to be inconsistent. First, we use a journal's average citations per document in a two-year period[30] as a measure of journal quality. This measure was obtained from Scimago, which is compiled using Elsevier's Scopus database.[31]

[28] Of course, another possibility is that the journal is open access, but it has not yet been indexed by DOAJ. We attempt to account for this possibility by using UlrichsWeb data on open access journals to cross validate the DOAJ data. Overall, the two sources are consistent. This is discussed in the data appendix.

[29] See http://grants.nih.gov/grants/guide/notice-files/not98-013.html.

[30] This is computed, "...considering the number of citations received by a journal in the current year to the documents published in the two previous years, –i.e. citations received in year X to documents published in years X-1 and X-2" (http://www.scimagojr.com/help.php#understand_journals).

[31] In results not shown, we experiment with two alternative measures of journal quality: the Scimago Journal Rank (SJR) index and the Hirsch index. The results were very similar to those presented in section 5, and will be made available upon request.

Second we use the NLM's medical subject headings (MeSH) to control for the article topic. The MeSH terms are extracted from the MEDLINE database. MeSH terms have a hierarchical structure and can be very detailed.[32] However, we confine ourselves to top-level and second-level MeSH terms to control for article topic. The structure of MeSH terms is discussed in further detail in the appendix.

Third, we control for article quality using citations from Thomson Reuters Web of Science (WOS). Specifically, we use the total number of citations that each article has ever received from all other articles indexed in the Web of Science. Unfortunately, we do not have citations separated by year for each article, which prevents us from calculating measures such as 2-year forward citation counts. Since older articles will obviously have more citations than younger articles, we standardize the number of citations within each cohort of articles. Using standardized citations, we are able to control for each article's relative quality within its cohort.

Finally, we control for whether a journal is owned by a commercial or non-profit publisher. We use three sources of data to obtain information about whether a journal is owned by a commercial or non-profit publisher: the NLM's List of Serials Indexed for Online Users (LSIOU), the NLM Catalog, and UlrichsWeb Global Serials Directory. All three sources contain information on the publishers of journals indexed in MEDLINE.

Using a journal's NLMID–an identification number that uniquely identifies each journal in MEDLINE–we are able to match publisher information from LSIOU and the NLM Catalog to each journal indexed in MEDLINE. Using the ISSN, we are able to match publisher information from UlrichsWeb to each journal indexed in MEDLINE.

Once publisher information is assigned to each journal, we manually identified whether each publisher was a commercial or non-profit organization. This was a very tedious process that utilized Google searches, e-mail correspondence, and translation work from several colleagues. This process is explained in more detail in the appendix.

4.4 Sample Restrictions and Estimation Sample

MEDLINE classifies all articles into different publication types such as "Journal Article", "Clinical Trial", "News Article", etc.[33] We only analyze articles that are classified as a "Journal Article". We then exclude all articles with data missing for MeSH category, article quality, journal quality, and publisher type. The losses of articles and journals from each restriction are listed in table 1. Overall, we begin with 7,952,412 articles published in 8,258

[32]Perhaps the best way to understand the structure of MeSH terms is to browse them at: http://www.nlm.nih.gov/cgi/mesh/2014/MB_cgi.

[33]For a full list of publication types in MEDLINE, see www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T42/

journals. Sample size reductions due to missing MeSH term or article quality data are negligible. Concern arises from the loss in the number of articles, and especially the number of journals, due to missing data for journal quality and publisher type. However, as we will demonstrate in section 5.5, very little changes when we drop these journal-level variables as controls and obtain our results using the entire sample. This suggests that sample selection, due to missing data, is not a significant problem in our sample. In the end, our estimation sample is a repeated cross-section of 7,271,545 articles published in 5,613 journals between the years 1999 and 2012.

5 Results

5.1 Summary Statistics

Before presenting the main results, we discuss some general characteristics of our data. Figure 3 shows, for each year between 1999 and 2012, the total number of articles and the total number of journals used in our estimation sample. It is clear that there was a precipitous increase in the number of both articles and journals over this time period. Indeed, in 1999 our sample contains about 345 thousand articles and by 2012 it contains about 645 thousand articles–a nearly 87 percent increase. The number of journals rose from 3,337 in 1999 to 4,877 in 2012–a more modest increase of approximately 46 percent.

Figure 4 shows, for each year, the increase in the proportion of articles and journals in our estimation sample that are open access. The number of articles published in open access journals increased 11-fold–from 6,931 to 79,533–between 1999 and 2012. This rapid growth caused the number of articles published in open access journals to increase from about 2.00 percent to 12.32 percent of total articles over this time period. The number of open access journals increased 5-fold–from 88 to 444. This caused the number of open access journals to increase from about 2.63 percent to 9.10 percent of total journals between 1999 and 2012. Thus, in our estimation sample, there has been a dramatic increase in open access literature, whether measured by articles or journals. Moreover, open access literature makes up a much larger share of all literature than it used to.

Figure 5 shows, for each year, the proportion of articles in our estimation sample that received NIH support. Overall, this proportion has stayed relatively constant, ranging from just under 0.12 to just over 0.13.

5.2 Graphical Analysis

Table 2 and figure 6, present the main results from the FGLS estimation of the β_t's from equation (1). First, consider column (1) of table 2. In 1999 we estimate that, relative to articles not supported by a NIH grant, a NIH-supported article is about 0.86 percentage points more likely to be published in an open access journal. In 2000 this drops to about 0.73 percentage points and continues to steeply drop–becoming negative in 2006–until 2008 at which point the decline begins to slow. After 2010, the downward trend reverses and by 2012, NIH-supported articles are about 0.35 percentage points *less* likely to be published in an open access journal.

This trend in the estimates of the β_t's is more easily observed by examining figure 6. Each dot marker in this graph corresponds to the estimates in column (1) from table 2. This graph makes very clear the downward trend in the estimated impacts of NIH support prior to 2008 and its slowdown and eventual reversal after 2008. Fitting OLS lines through the dot markers in figure 6, and allowing the slopes and intercepts of these lines to change after 2008, reveals that the impact of NIH support was trending downward at a rate of 0.13 percentage points per year until 2008. After 2008, the trend has an estimated slope of approximately zero (slightly positive). The estimated difference of 0.0013 between the pre-post slopes indicates that the NIH mandate increased the probability that NIH-supported articles are published in open access journals by 0.13 percentage points per year.

This finding is remarkably robust. It holds up when we control for top-level MeSH terms, second-level MeSH terms, article quality, journal quality, and publisher type. Indeed, graphs analogous to figure 6 for each of these specifications are virtually indistinguishable from figure 6 itself. This can be seen by examining each column of table 2. Thus, we have uncovered strong evidence that the NIH mandate increased the probability of NIH-supported articles being published in open access journals.

It is important to note that when we fit OLS lines through the dot markers in figure 6, we impose that the intercepts and slopes be allowed to change after 2008–the year of the NIH mandate. However, allowing the slopes and intercepts to change in a different time period may improve the fit of these lines. To examine this possibility, we allow the data to determine which two line segements best characterize the relationship between calendar time and the dot markers by estimating, for each year $\bar{t}$, the following regression equation:

$$\hat{\beta}_t = \theta_1 + \theta_2 I(t \leq \bar{t}) + \pi_1 t + \pi_2 t * I(t \leq \bar{t}) + \varepsilon_t. \tag{3}$$

In this equation, the $\hat{\beta}_t$'s are the estimated β_t's from equation (1) that are found in each column of table 2. $I(t \leq \bar{t})$ equals 1 if t is less than or equal to $\bar{t}$ and equals 0 otherwise.

We define the $\bar{t}$ that produces the smallest residual sum of squares as the "break year". For every specification in table 2, the estimated break year is 2008–the year of the NIH mandate. Thus, the break year that we *impose* in figure 6 turns out to be the break year that we *estimate* from equation (3). These results are shown in the row labeled "Break Year" in table 3. We use a bootstrap method to compute the precision of the break year estimator. We see from table 3 that the break years are extremely precisely estimated. Indeed, the row labeled "Bootstrap Proportion" shows that only a trivial number of bootstrap samples yield estimated break years that are not 2008.[34] If the estimated break year had occurred prior to the mandate or was estimated imprecisely, doubt would have been cast on a causal interpretation of our estimates. However, since the break year occurred exactly when we would expect–2008, the year of the NIH mandate–we have much greater confidence that the mandate did, indeed, have a causal impact on publishing patterns in the biomedical sciences.

5.3 Difference-in-Differences Estimates

We now turn to the main results of this paper–the FGLS estimates of λ from equation (2) in section 3.2. These results are displayed in table 4. First consider column (1). We see that, prior to the mandate, the impact of NIH support on the probability of being published in an open access journal (relative to being published in a toll access journal) was trending down at a rate of about 0.128 percentage points per year. This can be seen from the estimate of θ, the coefficient on the interaction of the NIH support indicator and calendar time. After the mandate, this impact essentially became flat, increasing at a rate of 0.0011 percentage points per year. This can be seen from the sum of the estimates of θ and λ. Thus, the difference (λ) is estimated to be about 0.138 percentage points per year. This is quite consistent with our preliminary analysis in the previous sub-section. Overall, our estimates of λ are remarkably robust across all specifications, ranging from 0.00137 to 0.00142, and are all precisely estimated. Thus, we estimate that the NIH mandate increased the probability of an NIH-supported article being published in an open access journal by approximately 0.13 to 0.14 percentage points per year.

How meaningful is this effect? In 2008, approximately 4.69 percent of NIH-supported articles were published in open access journals. If pre-mandate trends had continued past 2008, we predict that in 2009 only 4.69-0.14=4.55 percent of NIH-supported articles would have been published in open access journals. Since, in 2009, there were 79,551 NIH-supported articles published across all journals, we would predict that 0.0455*79,551=3,620 NIH-supported articles would have been published in open access journals. In fact, 4,478 NIH-supported

[34] We discuss the bootstrap method in more detail in the appendix.

articles were published in open access journals in 2009. Thus, we estimate that the mandate shifted about 4,478-3,620=858 NIH-supported articles into open access journals in 2009. This suggests that, absent the mandate, approximately 100*(858/4,478)=19 percent of articles published in open access journals in 2009 would have been published in toll access journals. As seen in table 5, these estimates can be computed for all subsequent years. Overall, between 2009 and 2012, we estimate that the mandate shifted about 12,475 NIH-supported articles into open access journals, which is approximately 41 percent of all NIH-supported articles published in open access journals.

5.4 Estimates Using Placebos

To further assess the whether the results presented in sections 5.2 and 5.3 are causal, we produce a set of "placebo" estimates in which we replace the indicator for NIH-support in equation (2) with a variable indicating whether an article is supported by other grant types but is not supported by a NIH grant. The first placebo is a variable that indicates whether an article is supported by a grant from any U.S. funding agency but is not supported by a NIH grant. The second placebo is a variable that indicates whether an article is supported by a grant from from any non-NIH organization under the auspices of the Department of Health and Human Services (HHS) but is not supported by a NIH grant.

The results from estimating λ in equation (2) using these placebos are presented in table 6. The placebo estimates are nearly an order of magnitude smaller than the main estimates presented in table 4. Moreover, none of the placebo estimates are statistically significant. Thus, there is no evidence that the NIH mandate impacted the probability that articles supported by non-NIH grants are published in an open access journal. Since the mandate only applied to articles supported by NIH grants, this is precisely what we would expect. Thus, the placebo results are further validation that the results in table 4 and figure 6 are, indeed, causal estimates.

5.5 Sample Restrictions and Sample Selection

As noted in section 4.4, and displayed in table 1, we lose a concerning number of observations when we drop observations that do not have journal quality or publisher type data. As mentioned, our measure of journal quality–citations per document from the previous two years–is obtained from Scimago, which is compiled using Elsevier's Scopus database. Thus, when we drop observations without journal quality data we lose all journals (and articles published in those journals) that are indexed in MEDLINE but are not indexed in Scopus. In contrast, publisher type data was coded manually using publisher data from several sources.

Thus, when we drop observations without publisher type data, we lose all journals (and articles published in those journals) that are indexed in MEDLINE but were unable to be identified as being owned by a commercial or non-profit publisher. More specifically, we lose approximately 30 percent of the journals in our original sample when we drop observations without journal quality data and about 11 percent when we drop observations without publisher type data. This data loss motivates us to examine whether our results are robust to sample selection.

To assess whether sample selection is, in fact, a problem, we re-estimate equation (2) using all of the observations that each specification will allow. For instance, when we include no controls, we are able to use all 7,952,412 articles and 8,258 journals in our starting sample from table 1. These results are displayed in table 7. Fortunately, none of the results substantively change. Specifications that include citations per document as a control yield somewhat larger estimates–ranging from about 0.00152 to 0.00153. All other specifications are very similar to their counterparts in table 4 (note that the specifications in columns (11) and (12) are identical, by construction, to their counterparts in table 4). Overall, we conclude that the sample restrictions used to obtain our main estimation sample do not drive our results.

5.6 Additional Evidence

The previous several sub-sections have presented evidence that the NIH mandate decreased toll access journals' desire to publish NIH-supported articles. If this is the case, we would also expect the quality of NIH-supported articles published in toll access journals to increase after the mandate. Essentially, after the mandate, a NIH-supported article must clear a higher quality hurdle in order for a toll access journal to justify its publication. To examine this implication, we estimate the following equation:

$$\begin{aligned} avgcites_{ajtg} &= \alpha_g + \beta_g nihgrant_{ajtg} + \gamma_g Post08_t + \delta_g (nihgrant_{ajtg} * Post08_t) + \\ &\quad \eta_g controls_{ajtg} + \nu_j + \alpha_t + \varepsilon_{at}. \end{aligned} \tag{4}$$

In this equation, $avgcites_{ajtg}$ is the average number of citations that article a, which was published in journal j during year t, received over the years since its publication. All remaining variables are defined as in equation (2). Note that the subscript g indexes whether the journal is open access or toll access. This equation is simply a difference-in-differences specification without a linear time trend as in equation (2). The parameters of interest in equation (4) are the δ_g, which measure the impact of the mandate on the average yearly citations for journal type g.

Panel A of table 8 displays the OLS estimates of equation (4). We see that the results are perfectly consistent with our main results from table 4 and figure 6. In particular, we see that the mandate had no statistically significant impact on the quality of NIH-supported articles published in open access journals (except a marginally significant negative effect when we do not include any control variables). In contrast, we estimate that the mandate increased the quality of NIH-supported articles in toll access journals. Specifically, we estimate that the mandate increased the average number of yearly citations that NIH-supported articles published in toll access journals receive by between 0.55 and 0.65.

Panel B of table 8 displays the estimates when we include journal fixed effects in the regression. The magnitude of the point estimates tend to decrease, but the overall pattern remains unchanged. In particular, we find that the mandate had no quality impact for NIH-supported articles published in open access journals, but increased the average number of citations to NIH-supported articles by 0.37 to 0.40 in toll access journals.

Since approximately 15 percent of the articles in our estimation sample have zero average yearly citations, it is worth re-estimating equation (4) under the assumption that it is a Tobit model. In this case, we assume that the error terms are normally distributed conditional on the regressors, and use maximum likelihood to estimate the parameters of equation (4).

Panel C of table 8 displays the Tobit estimates. Specifically, it displays the average partial effect of the mandate on the *unconditional* expected number of average yearly citations. It is reassuring that nothing of substance changes when we explicitly account for the fact that a large proportion of articles in our estimation sample have zero citations. Indeed, we still find that the mandate had no quality impact for NIH-supported articles published in open access journals, but did have a positive and statistically significant impact on NIH-supported articles published in toll access journals. Moreover, the magnitudes of the estimates are similar to the OLS estimates.

Overall, the results of this section suggest that the NIH mandate increased the average number of citations to NIH-supported articles by between 0.367 and 0.648. The median NIH-supported article in our sample receives approximately 3.33 citations per year. Thus, we estimate that the mandate boosted citations for the typical NIH-supported article by between 11 and 19 percent.

5.7 Journal Quality in Open and Toll Access Journals

In the last several sub-sections, we presented strong evidence that the NIH's limitation of journals' proprietary control over NIH-supported articles caused toll access journals to be less inclined to publish such articles after the mandate. Moreover, we presented evidence that the

fewer articles that toll access journals do publish must clear a higher quality threshold. These results are important because open access journals tend to be lower quality than toll access journals. Indeed, consider figure 7 which displays the empirical cumulative distribution function (CDF), separately for open access and toll access journals, of the natural log of citations per document.

This CDF clearly shows that toll access journals tend to be higher quality than open access journals. Consider journals that have one citation per document, which corresponds to having a natural log of zero. We see that, relative to toll access journals, a greater proportion of open access journals have less than or equal to one citation per document. Indeed, this pattern is true for all but very small values of citations per document.

The fact that toll access journals tend to be higher quality than open access journals raises an intriguing question: Did the NIH mandate, by shifting NIH-supported articles from toll access to open access journals, actually decrease the quality of journals in which NIH-supported research is published? If every NIH-supported article that, because of the NIH mandate, is rejected by a toll access journal is able to get published in an open access journal of equal quality, then the mandate should have no impact on the quality of journals in which NIH-supported articles are published. On the other hand, if such articles are unable to get published in an open access journal of equal quality, and therefore must settle for a lower quality open access journal, then the mandate will decrease the quality of journals in which NIH-supported articles are published.

To examine this possibility, we estimate the following regression equation:

$$\begin{aligned} citesperdoc_{ajtg} &= \alpha_g + \beta_g nihgrant_{ajtg} + \gamma_g Post08_t + \delta_g (nihgrant_{ajtg} * Post08_t) + \\ &\quad \eta_g controls_{ajtg} + \nu_j + \alpha_t + \varepsilon_{at}. \end{aligned} \tag{5}$$

In this equation $citesperdoc_{ajtg}$ is the percentile rank of the journal j, in which article a was published at time t. We opt to use the percentile rank because citations per document are extremely right skewed and we are unable to use a log transformation because of the large number of journals that received zero citations over the previous two years.

Panel A of table 9 displays the OLS estimates of equation (5). We see that the mandate did decrease the quality of journals in which NIH-supported articles are published. Indeed, the OLS estimates suggest that, after the mandate, NIH-supported articles were published in journals 1.2-2.4 percentiles lower in the journal quality distribution. Thus, there is strong evidence that the NIH mandate, by shifting NIH-supported articles from toll access to open access journals, caused these articles to be published in lower quality journals.

The median NIH-supported article in our sample is published in a journal with 3.97

citations per document over the previous two years, which is at the 89.1th percentile of all journals in our sample. Moving down 1.2 percentiles puts an article in a journal that received 3.76 citations per document over the previous two years. Moving down 2.4 percentiles puts an article in a journal that received 3.58 citations per document over the previous two years. Thus, we estimate that the mandate caused the typical NIH-supported article to be published in a journal that receives between 5 and 10 percent fewer citations per document over the previous two years.

6 Conclusion

In this paper, we have presented evidence that the NIH's 2008 Public Access Policy altered publication patterns in the biomedical sciences. Specifically, by limiting journals' proprietary control over the content of NIH-supported articles, the NIH mandate made toll access journals more reluctant to publish these articles, causing more NIH-supported articles to be published in open access journals. Moreover, the fewer NIH-supported articles that toll access journals do publish must clear a higher quality hurdle to justify publication.

This is problematic if the NIH wants to ensure that researchers have broad exposure to the research it supports. We establish that toll access journals tend to be higher quality than open access journals and that, by shifting NIH-supported articles from toll access to open access journals, the NIH mandate decreased the quality of journals in which these articles are published. If we make the assumption that researchers focus most of their limited attention on higher quality journals, then the NIH mandate may have decreased the exposure of researchers to NIH-supported articles. Moreover, since NIH-supported articles tend to be higher quality than non NIH-supported articles, this may have actually caused a decline in the average quality of article that researchers read.

Such shifts in publication patterns in biomedicine may be justified if the NIH mandate did, in fact, increase access to NIH-supported articles. However, the lack of a citation boost from open access suggests that most researchers had access to NIH-supported articles published in toll access journals prior to the mandate. Thus, it is quite likely that the mandate had little, if any, impact on researchers' access to the biomedical literature. However, experiments by Davis et al. (2008) and Davis (2011) suggest that, while open access has no impact on citations, it does have a positive impact on the number of article downloads, which is a proxy for the number of individuals that read a given article. This suggests that the mandate may have increased access to NIH-supported research for patients or physicians, which may have conferred substantial benefits to these individuals.

Given the results of this paper, it would make sense for the NIH to examine ways of

reducing toll access journals' reluctance to publish NIH-supported research. One possibility is to increase the allowed embargo time past 12 months. This would extend the period over which publishers have exclusive proprietary control over the content published in their journals, and thus decrease the incentive to reject NIH-supported articles.

It would also be wise for the funding agencies, recently instructed by the OSTP to develop their own open access mandates, to carefully consider how their specific mandate will impact the particular fields that they support. In fields such as physics and computer science, which have long traditions of extensive self-archiving in field repositories like arXiv and CiteSeerX[35], the mandate may have a very small impact. However, in fields such as the biological sciences, which have a less open tradition, the impacts may be similar to the NIH's Public Access Policy.

References

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein (2008), "Academic freedom, private-sector focus, and the process of innovation." *The RAND Journal of Economics*, 39, 617–635.

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

Albee, Barbara and Brenda Dingley (2000), "U.s. periodical prices–2000." *Association for Library Collections Technical Services*.

Albee, Barbara and Brenda Dingley (2001), "U.s. periodical prices–2001." *Library resources & technical services*.

Albee, Barbara and Brenda Dingley (2002), "U.s. periodical prices–2002." *Library resources & technical services*.

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

ARL (2011), "Monograph & serial costs in arl libraries." URL http://www.arl.org/storage/documents/monograph-serial-costs.pdf.

Bergstrom, Theodore C (2001), "Free labor for costly journals?" *Journal of Economic Perspectives*, 183–198.

[35]Technically, authors do not self-archive on CiteSeer. They self-archive on their personal webpages, and CiteSeer harvests the articles on these webpages and archives them in CiteSeer.

Bosch, Stephen and Kittie Henderson (2012), "Coping with the terrible twins: Periodicals price survey 2012." *Library Journal*, 137, 31.

Bosch, Stephen and Kittie Henderson (2013), "The winds of change: periodicals price survey 2013." *Library Journal*, 25.

Bosch, Stephen and Kittie Henderson (2014), "Steps down the evolutionary road: periodicals price survey 2013." *Library Journal*, 25.

Bosch, Stephen, Kittie Henderson, and Heather Klusendorf (2011), "Periodicals price survey 2011: under pressure, times are changing." *Library Journal*.

Cameron, A Colin and Douglas L Miller (2013), "A practitioner's guide to cluster-robust inference." *Forthcoming in Journal of Human Resources*, 221–236.

Chakma, Justin, Gordon H Sun, Jeffrey D Steinberg, Stephen M Sammut, and Reshma Jagsi (2014), "Asia's ascent: global trends in biomedical r&d expenditures." *New England Journal of Medicine*, 370, 3–6.

Craig, Iain D, Andrew M Plume, Marie E McVeigh, James Pringle, and Mayur Amin (2007), "Do open access articles have greater citation impact?: a critical review of the literature." *Journal of Informetrics*, 1, 239–248.

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

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

Davis, Philip M, Bruce V Lewenstein, Daniel H Simon, James G Booth, Mathew JL Connolly, et al. (2008), "Open access publishing, article downloads, and citations: randomised controlled trial." *BMj*, 337, a568.

Davis, Ronald M and Marcus Müllner (2002), "Editorial independence at medical journals owned by professional associations: a survey of editors." *Science and engineering ethics*, 8, 513–528.

Dingley, Brenda (2003), "Us periodical prices–2003." *Library resources & technical services*, 47, 192–207.

Dingley, Brenda (2004), "Us periodical prices–2004." *Library resources & technical services*.

Dingley, Brenda (2005), "Us periodical prices–2005." *Library resources & technical services*.

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

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

Gaule, Patrick and Nicolas Maystre (2011), "Getting cited: does open access help?" *Research Policy*, 40, 1332–1338.

Guedon, Jean-Claude (2004), "The green and gold roads to open access: The case for mixing and matching." *Serials review*, 30, 315–328.

Harnad, Stevan and Tim Brody (2004), "Comparing the impact of open access (oa) vs. non-oa articles in the same journals." *D-lib Magazine*, 10.

Henderson, Kittie S and Stephen Bosch (2010), "Seeking the new normal: Periodicals price survey 2010." *Library Journal*, 135, 36–40.

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

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

Lucas, Robert E (1988), "On the mechanics of economic development." *Journal of Monetary Economics*, 22, 3–42.

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

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

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

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

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

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern (2009), "Of mice and academics: Examining the effect of openness on innovation." Technical report, National Bureau of Economic Research.

Peek, Robin (2008), "Harvard faculty mandates oa." *Information Today*, 25, 15–17.

RIN (2011), "Overcoming barriers: access to research information." URL http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/overcoming-barriers-access-research-information.

Romer, Paul M (1986), "Increasing returns and long-run growth." *The journal of political economy*, 94, 1002–1037.

Romer, Paul M (1990), "Endogenous technological change." *Journal of Political Economy*, 98.

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

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

Scott, Andrew J and D Holt (1982), "The effect of two-stage sampling on ordinary least squares methods." *Journal of the American Statistical Association*, 77, 848–854.

Seamans, N (2001), "Electronic theses dissertations: 2001 survey of editors and publishers." URL http://lumiere.lib.vt.edu/surveys/results/.

Smith, Richard (1999a), "Another editor bites the dust: trust is needed to balance editorial independence and accountability." *BMJ: British Medical Journal*, 319, 272.

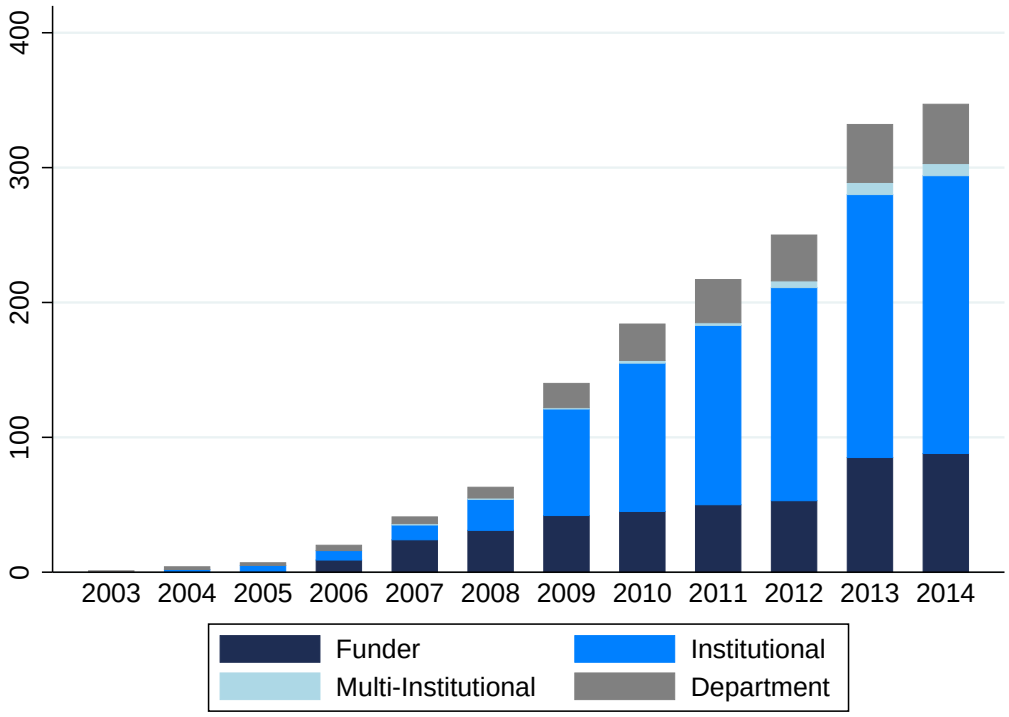
Smith, Richard (1999b), "The firing of brother george: The ama has damaged itself by sacking jamas editor." *BMJ: British Medical Journal*, 318, 210.

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

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

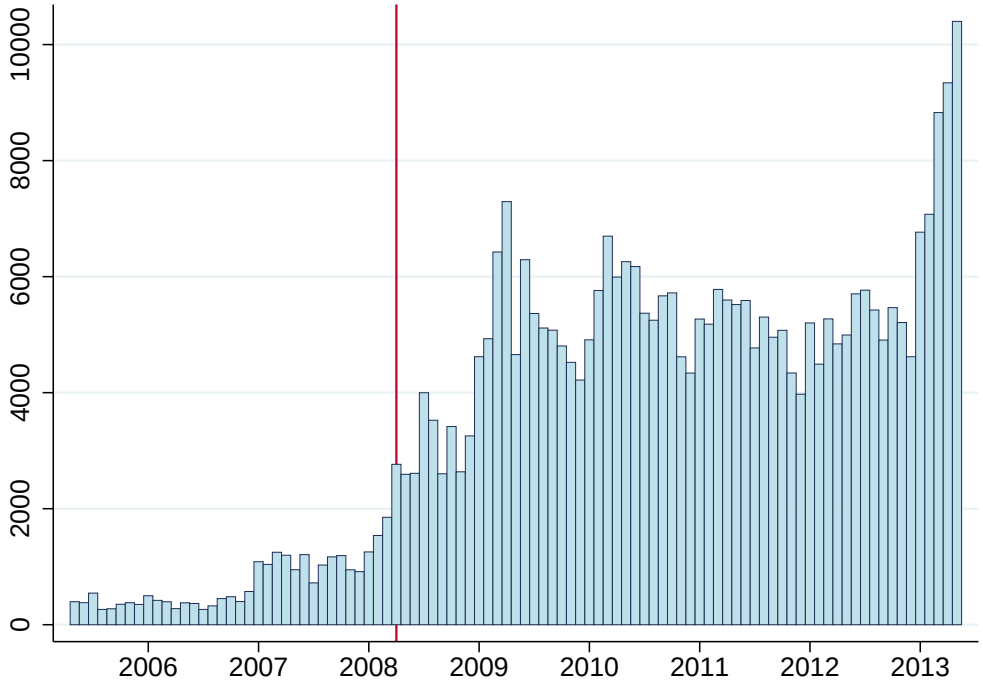
Wooldridge, Jeffrey (2009), *Introductory econometrics: A modern approach*. Cengage Learning.

Figure 1: Cumulative number of open access mandates.



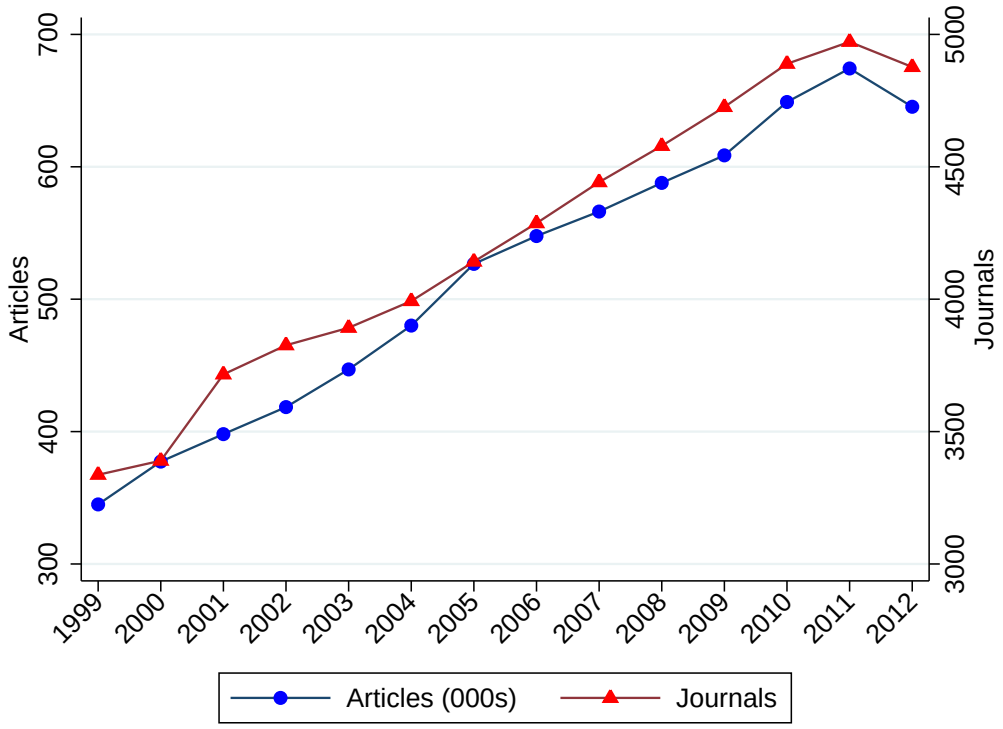
Notes–The data were obtained from the Registry of Open Access Repositories Mandatory Archiving Policies (ROARMAP).

Figure 2: Number of manuscript submissions to PubMed Central.



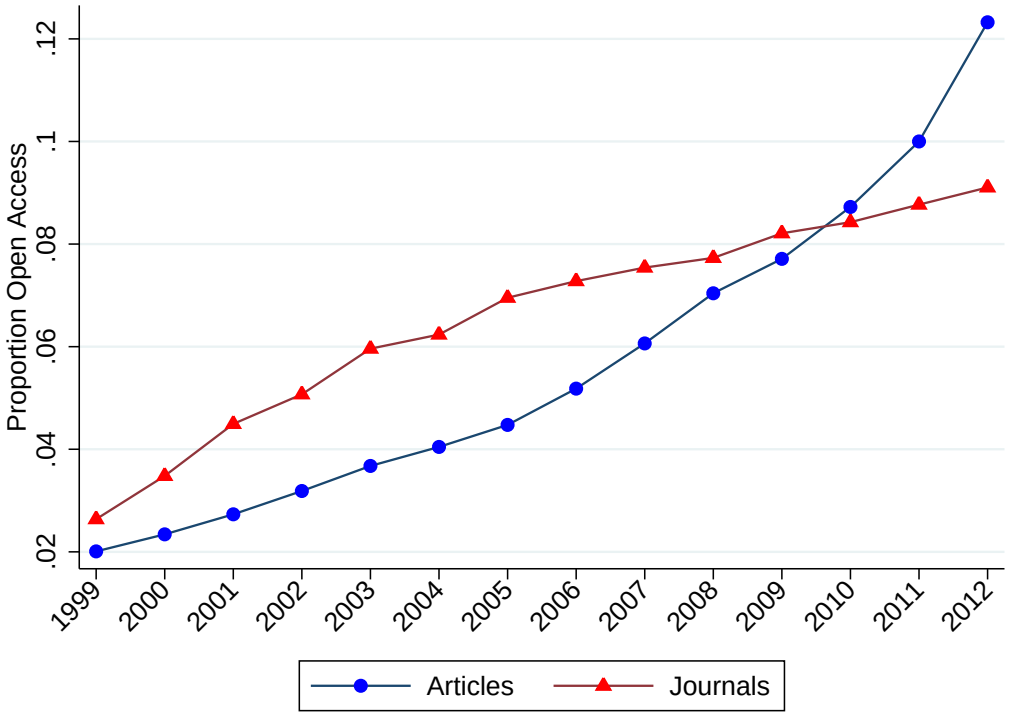
Notes–The data were obtained from *Nature's* blog: http://blogs.nature.com/news/2013/07/nih-sees-surge-in-open-access-manuscripts.html. Manuscript submission counts are monthly, and run from May 2005 to May 2013. The vertical red line indicates the month, April 2008, that the NIH Public Access Policy went into effect.

Figure 3: Total number of articles and journals.



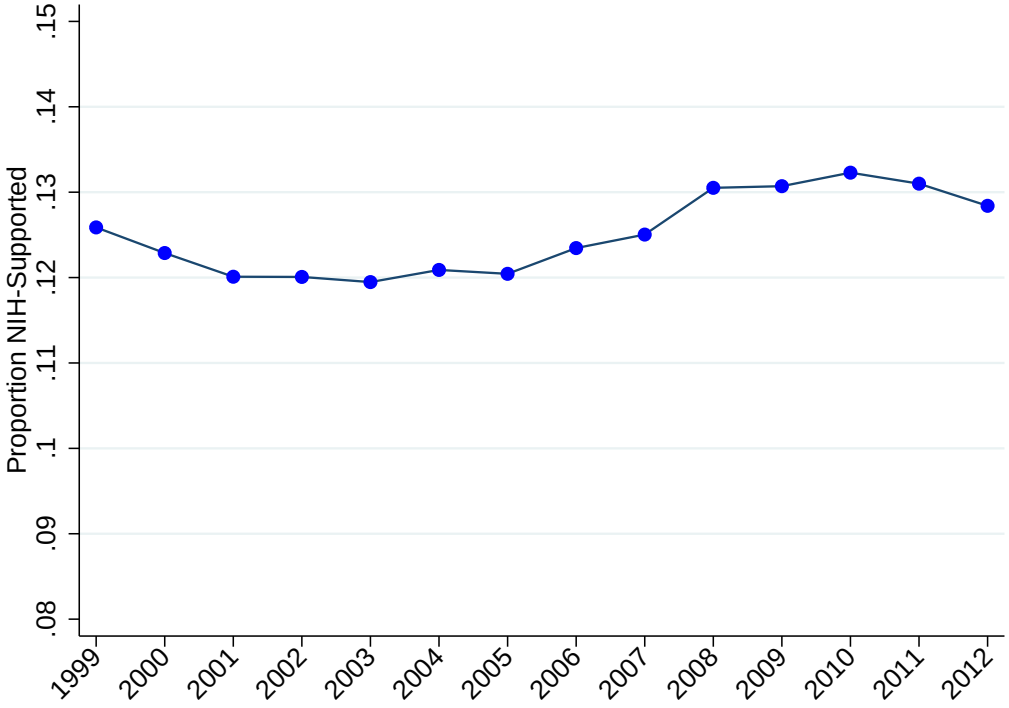
Notes–The data were obtained from the 2014 MEDLINE baseline files.

Figure 4: Proportion of articles and journals that are open access.



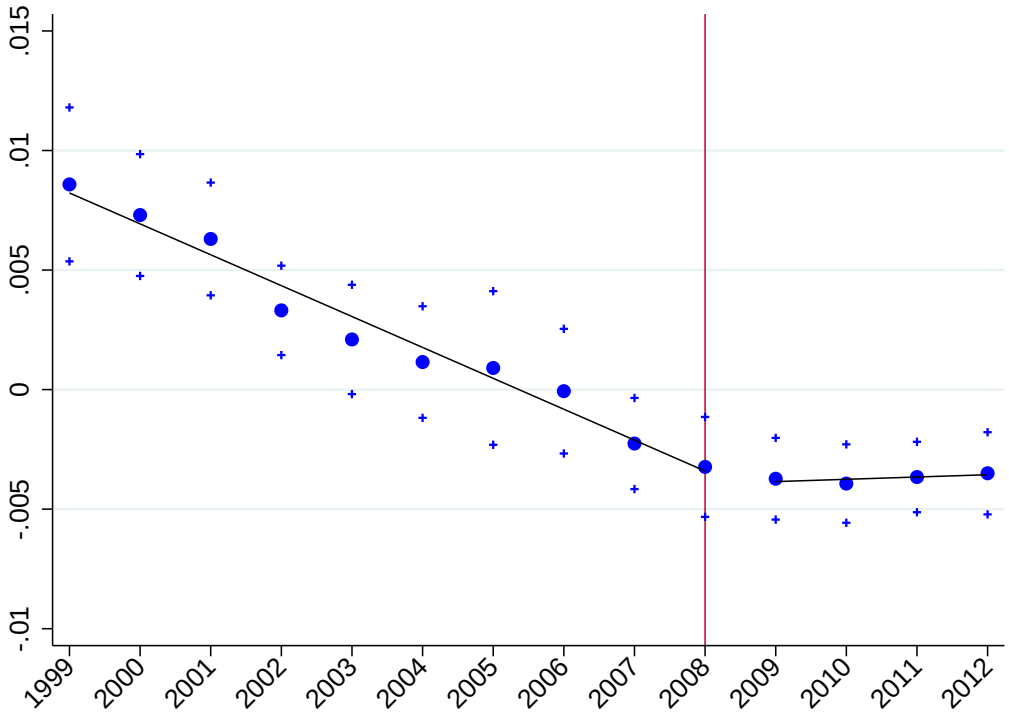
Notes–The data were obtained from the 2014 MEDLINE baseline files and the Directory of Open Access Journals (DOAJ). DOAJ was accessed on July 5, 2014 at 8:38 AM.

Figure 5: Proportion of articles that are NIH-supported.



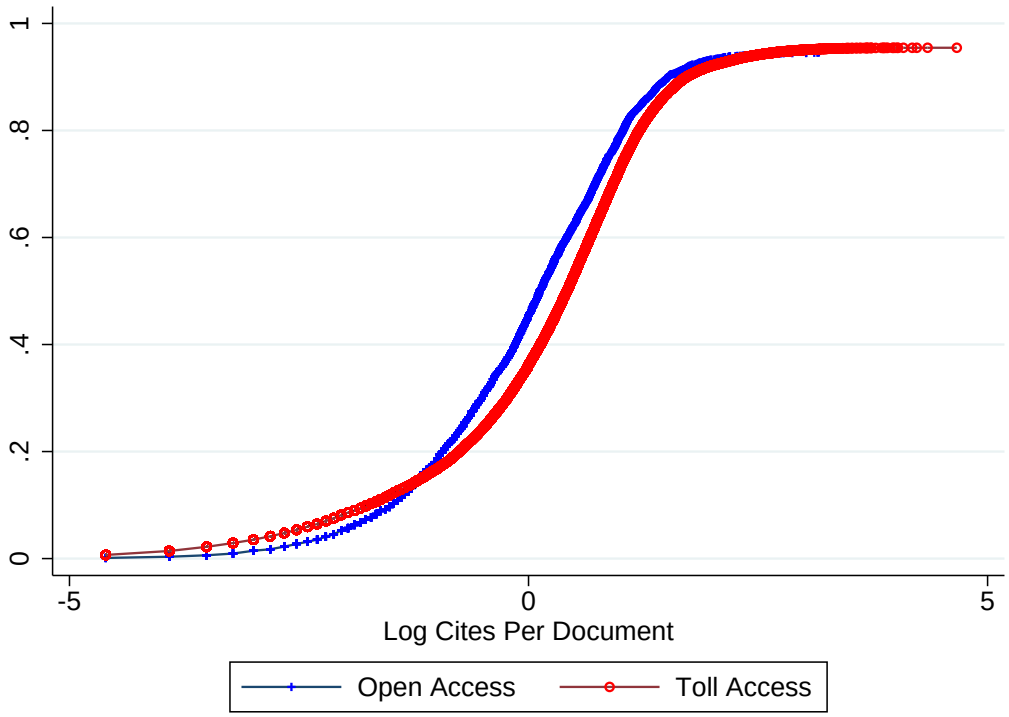
Notes–The data were obtained from the 2014 MEDLINE baseline files.

Figure 6: Estimated trends, before and after the NIH's 2008 open access mandate, of the impact of NIH support on the probability of publishing in an open access journal.



Notes–Each circle denotes the estimated impact, for a given year, of NIH support on the probability of an article being published in an open access journal. A FGLS estimator that assumes within-journal equicorrelated errors is used to obtain these estimates. The 95% confidence interval for a given estimated impact is denoted by plus signs on each side of a cirlce. These confidence intervals are computed using standard errors clustered at the journal level. The solid black line segments denote the fitted values from the regression, on a linear time trend allowed to differ before and after the 2008 mandate, of the estimated impacts of NIH support on the probability of publishing in an open access journal. The vertical red line indicates the year, 2008, that the NIH Public Access Policy went into effect.

Figure 7: Empirical cumulative distribution function (CDF) of the natural log of citations per document to a journal's articles during the previous two years.



Notes–The data were obtained from Scimago, which is based on Elsevier's Scopus database. 2,720 out of 59,068 journal-years received zero citations over the previous two years. We assign a value of 0.001 to these observations before logging them.

Table 1: Impact of sample restrictions on the number of articles and journals in the estimating sample.

	Articles Left	Articles Lost	Journals Left	Journals Lost
Starting Sample	7,952,412	-	8,258	-
Missing MeSH type	7,950,614	1,798	8,258	0
Missing article cites	7,952,405	7	8,258	0
Missing journal cites	7,409,427	542,985	5,787	2,471
Missing publisher type	7,586,833	365,579	7,315	943
Estimation Sample	7,271,545	-	5,613	-

Notes–The first column, titled "Articles Left" is the number of articles in the sample after a given sample restriction. The second column, titled, "Articles Lost", is the number of articles that are lost from the sample restriction. The third and fourth columns, "Journals Left" and "Journals Lost" are the analogous numbers for journals.

Table 2: The estimated impact, by year, of NIH support on the probability of an article being published in an open access journal.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1999	0.00858***	0.00858***	0.00858***	0.00856***	0.00856***	0.00855***	0.00865***	0.00865***	0.00864***	0.00865***	0.00865***	0.00864***
	(0.00164)	(0.00164)	(0.00164)	(0.00164)	(0.00164)	(0.00164)	(0.00165)	(0.00165)	(0.00165)	(0.00165)	(0.00165)	(0.00165)
2000	0.00730***	0.00730***	0.00729***	0.00728***	0.00728***	0.00727***	0.00734***	0.00734***	0.00733***	0.00734***	0.00734***	0.00733***
	(0.00130)	(0.00130)	(0.00130)	(0.00130)	(0.00130)	(0.00130)	(0.00131)	(0.00131)	(0.00131)	(0.00131)	(0.00131)	(0.00131)
2001	0.00630***	0.00630***	0.00630***	0.00629***	0.00628***	0.00628***	0.00641***	0.00641***	0.00641***	0.00641***	0.00641***	0.00641***
	(0.00120)	(0.00120)	(0.00121)	(0.00120)	(0.00120)	(0.00121)	(0.00123)	(0.00123)	(0.00123)	(0.00123)	(0.00123)	(0.00123)
2002	0.00331***	0.00331***	0.00331***	0.00329***	0.00329***	0.00329***	0.00337***	0.00336***	0.00337***	0.00337***	0.00336***	0.00337***
	(0.000954)	(0.000957)	(0.000962)	(0.000951)	(0.000954)	(0.000959)	(0.000967)	(0.000970)	(0.000975)	(0.000967)	(0.000970)	(0.000975)
2003	0.00210*	0.00209*	0.00210*	0.00208*	0.00207*	0.00208*	0.00214*	0.00214*	0.00214*	0.00214*	0.00214*	0.00214*
	(0.00117)	(0.00117)	(0.00118)	(0.00116)	(0.00117)	(0.00118)	(0.00117)	(0.00118)	(0.00119)	(0.00117)	(0.00118)	(0.00119)
2004	0.00115	0.00115	0.00116	0.00114	0.00113	0.00114	0.00120	0.00120	0.00121	0.00120	0.00120	0.00121
	(0.00119)	(0.00120)	(0.00121)	(0.00119)	(0.00119)	(0.00121)	(0.00120)	(0.00120)	(0.00122)	(0.00120)	(0.00120)	(0.00122)
2005	0.000905	0.000898	0.000913	0.000886	0.000880	0.000894	0.000925	0.000919	0.000933	0.000925	0.000919	0.000933
	(0.00164)	(0.00165)	(0.00166)	(0.00163)	(0.00164)	(0.00166)	(0.00164)	(0.00165)	(0.00166)	(0.00164)	(0.00165)	(0.00166)
2006	-6.59e-05	-6.95e-05	-5.38e-05	-8.58e-05	-8.94e-05	-7.44e-05	-6.29e-05	-6.64e-05	-5.21e-05	-6.28e-05	-6.64e-05	-5.21e-05
	(0.00133)	(0.00134)	(0.00135)	(0.00132)	(0.00133)	(0.00135)	(0.00133)	(0.00134)	(0.00135)	(0.00133)	(0.00134)	(0.00135)
2007	-0.00225**	-0.00226**	-0.00224**	-0.00226**	-0.00226**	-0.00224**	-0.00229**	-0.00230**	-0.00228**	-0.00229**	-0.00230**	-0.00228**
	(0.000972)	(0.000963)	(0.000945)	(0.000972)	(0.000963)	(0.000945)	(0.000981)	(0.000972)	(0.000954)	(0.000981)	(0.000972)	(0.000954)
2008	-0.00323***	-0.00324***	-0.00322***	-0.00325***	-0.00326***	-0.00324***	-0.00333***	-0.00333***	-0.00331***	-0.00333***	-0.00333***	-0.00331***
	(0.00107)	(0.00106)	(0.00104)	(0.00107)	(0.00106)	(0.00105)	(0.00109)	(0.00108)	(0.00106)	(0.00109)	(0.00108)	(0.00106)
2009	-0.00373***	-0.00374***	-0.00371***	-0.00375***	-0.00375***	-0.00373***	-0.00380***	-0.00381***	-0.00379***	-0.00380***	-0.00381***	-0.00379***
	(0.000872)	(0.000866)	(0.000852)	(0.000876)	(0.000870)	(0.000857)	(0.000889)	(0.000883)	(0.000869)	(0.000889)	(0.000883)	(0.000869)
2010	-0.00393***	-0.00394***	-0.00392***	-0.00395***	-0.00395***	-0.00394***	-0.00394***	-0.00395***	-0.00393***	-0.00394***	-0.00395***	-0.00393***
	(0.000836)	(0.000833)	(0.000820)	(0.000841)	(0.000837)	(0.000824)	(0.000841)	(0.000838)	(0.000825)	(0.000841)	(0.000838)	(0.000825)
2011	-0.00366***	-0.00367***	-0.00364***	-0.00366***	-0.00367***	-0.00365***	-0.00364***	-0.00365***	-0.00363***	-0.00364***	-0.00365***	-0.00363***
	(0.000751)	(0.000747)	(0.000735)	(0.000753)	(0.000749)	(0.000737)	(0.000750)	(0.000746)	(0.000734)	(0.000750)	(0.000746)	(0.000734)
2012	-0.00350***	-0.00351***	-0.00349***	-0.00351***	-0.00352***	-0.00351***	-0.00348***	-0.00349***	-0.00347***	-0.00348***	-0.00349***	-0.00347***
	(0.000877)	(0.000873)	(0.000860)	(0.000880)	(0.000876)	(0.000863)	(0.000875)	(0.000871)	(0.000858)	(0.000875)	(0.000871)	(0.000858)
Top-Level MeSH		×			×			×			×	
Second-Level MesH			×			×			×			×
Article Standardized Cites				×	×	×	×	×	×	×	×	×
Journal Cites Per Document							×	×	×	×	×	×
Publisher Type										×	×	×

Notes–A FGLS estimator that assumes equicorrelation of the within-journal error terms was used to obtain these estimates. The standard errors are in parentheses and are clustered at the journal level. * indicates statistical significance at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level. Our estimating sample contained 7,271,545 articles published in 5,613 journals between the years 1999 and 2012.

Table 3: Estimated break years.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Break Year	2008	2008	2008	2008	2008	2008	2008	2008	2008	2008	2008	2008
Bootstrap Proportion	0.9940	0.9440	0.9940	0.9970	0.9970	0.9970	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Top-Level MeSH		×			×			×			×	
Second-Level MesH			×			×			×			×
Article Standardized Cites				×	×	×	×	×	×	×	×	×
Journal Cites Per Document							×	×	×	×	×	×
Publisher Type										×	×	×

Notes–The row titled "Break Year" contains, for each specification, the year ($\bar{t}$) that minimizes equation (3) in section 5.2. The row titled "Bootstrap Proportion" contains, for each specification, the proportion of the 1,000 bootstrap samples that yield an estimated break year of 2008.

Table 4: The estimated impact of the NIH mandate on the probability of an article being published in an open access journal.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	-3.804*	-3.806*	-3.806*	-3.808*	-3.810*	-3.810*	-3.912*	-3.914*	-3.914*	-3.842*	-3.845*	-3.845*
	(2.025)	(2.025)	(2.025)	(2.025)	(2.025)	(2.025)	(2.027)	(2.028)	(2.027)	(2.027)	(2.027)	(2.027)
NIH Grant	2.563***	2.563***	2.557***	2.561***	2.561***	2.555***	2.597***	2.597***	2.591***	2.597***	2.597***	2.591***
	(0.454)	(0.454)	(0.453)	(0.453)	(0.454)	(0.453)	(0.462)	(0.462)	(0.461)	(0.462)	(0.462)	(0.461)
Post 2008	3.348	3.352	3.350	3.353	3.356	3.355	3.276	3.280	3.278	3.275	3.280	3.278
	(2.066)	(2.067)	(2.067)	(2.066)	(2.067)	(2.067)	(2.065)	(2.066)	(2.066)	(2.065)	(2.066)	(2.066)
NIH Grant × Post 2008	-2.764***	-2.761***	-2.754***	-2.766***	-2.764***	-2.756***	-2.859***	-2.856***	-2.849***	-2.859***	-2.856***	-2.849***
	(0.570)	(0.569)	(0.568)	(0.570)	(0.569)	(0.568)	(0.589)	(0.588)	(0.587)	(0.589)	(0.588)	(0.587)
Year	0.00194*	0.00194*	0.00194*	0.00194*	0.00194*	0.00194*	0.00199**	0.00199**	0.00199**	0.00199**	0.00199**	0.00199**
	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)	(0.00101)
NIH Grant × Year	-0.00128***	-0.00128***	-0.00128***	-0.00128***	-0.00128***	-0.00127***	-0.00130***	-0.00130***	-0.00129***	-0.00130***	-0.00130***	-0.00129***
	(0.000226)	(0.000226)	(0.000226)	(0.000226)	(0.000226)	(0.000226)	(0.000230)	(0.000230)	(0.000230)	(0.000230)	(0.000230)	(0.000230)
Post 2008 × Year	-0.00167	-0.00167	-0.00167	-0.00167	-0.00167	-0.00167	-0.00163	-0.00163	-0.00163	-0.00163	-0.00163	-0.00163
	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)
NIH Grant × Post 2008 × Year	0.00138***	0.00137***	0.00137***	0.00138***	0.00138***	0.00137***	0.00142***	0.00142***	0.00142***	0.00142***	0.00142***	0.00142***
	(0.000283)	(0.000283)	(0.000283)	(0.000284)	(0.000283)	(0.000283)	(0.000293)	(0.000292)	(0.000292)	(0.000293)	(0.000292)	(0.000292)
Top-Level MeSH		×			×			×			×	
Second-Level MesH			×			×			×			×
Article Standardized Cites				×	×	×	×	×	×	×	×	×
Journal Cites Per Document							×	×	×	×	×	×
Publisher Type										×	×	×

Notes–A FGLS estimator that assumes equicorrelation of the within-journal error terms was used to obtain these estimates. The standard errors are in parentheses and are clustered at the journal level. * indicates statistical significance at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level. Our estimating sample contained 7,271,545 articles published in 5,613 journals between the years 1999 and 2012.

Table 5: Illustrating the magnitude of the NIH mandate.

		(2008)	(2009)	(2010)	(2011)	(2012)	Total
(1)	Actual proportion of NIH-supported articles published in OA journals.	0.0469	0.0563	0.0690	0.0841	0.1116	-
(2)	Predicted proportion of NIH-supported articles published in OA journals.	0.0469	0.0455	0.0441	0.0427	0.0413	-
(3)	Actual number of NIH-supported articles published in any journal.	76,721	79,551	85,843	88,329	82,867	413,311
(4)	Actual number of NIH-supported articles published in OA journals.	3,601	4,478	5,919	7,425	9,250	30,673
(5)	Predicted number of NIH-supported articles published in OA journals.	3,601	3,620	3,786	3,772	3,422	18,198
(6)	Difference between (4) and (5).	0	858	2,133	3,653	5,828	12,475
(7)	Ratio of (6) and (4).	0.0000	0.1916	0.3604	0.4920	0.6301	0.4067

Notes–Row (1) is the *actual* proportion of NIH-supported articles published in open access journals for each year. Row (2) is computed as $0.0469 - 0.0014 * (t - 2008)$ for $t = 2008, ..., 2012$. Row (3) is the actual number of of NIH-supported articles published in any journal type for each year. Row (5) is computed as the second row times the third row. Row (6) is computed as the difference between row (4) and row (5). Row (7) is computed as row (6) divided by row (4).

Table 6: The estimated impact of the NIH mandate on the probability of an article funded by non-NIH grants being published in an open access journal (i.e., the "placebo estimates").

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
U.S. Government Grant (Non-NIH)	0.000186	0.000191	0.000196	0.000186	0.000192	0.000197	0.000153	0.000159	0.000164	0.000153	0.000159	0.000164
× Post 2008 × Year	(0.000615)	(0.000615)	(0.000616)	(0.000615)	(0.000615)	(0.000616)	(0.000615)	(0.000614)	(0.000615)	(0.000615)	(0.000614)	(0.000615)
HSS Grant (non-NIH)	0.000170	0.000176	0.000181	0.000170	0.000176	0.000182	0.000137	0.000143	0.000148	0.000137	0.000143	0.000148
× Post 2008 × Year	(0.000617)	(0.000617)	(0.000618)	(0.000617)	(0.000617)	(0.000618)	(0.000617)	(0.000616)	(0.000617)	(0.000617)	(0.000616)	(0.000617)

Notes–A FGLS estimator that assumes equicorrelation of the within-journal error terms was used to obtain these estimates. The standard errors are in parentheses and are clustered at the journal level. * indicates statistical significance at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level. Our estimating sample contained 7,271,545 articles published in 5,613 journals between the years 1999 and 2012. "HHS" stands for the Department of Health and Human Services.

Table 7: Using different samples to estimate the impact of the NIH mandate on the probability of an article being published in an open access journal.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	-3.208*	-3.214*	-3.213*	-3.213*	-3.218*	-3.218*	-3.706*	-3.710*	-3.714*	-3.837*	-3.846*	-3.846*
	(1.758)	(1.759)	(1.759)	(1.759)	(1.759)	(1.759)	(2.006)	(2.006)	(2.008)	(2.026)	(2.028)	(2.028)
NIH Grant	2.511***	2.511***	2.506***	2.509***	2.509***	2.504***	2.739***	2.738***	2.732***	2.598***	2.597***	2.591***
	(0.418)	(0.418)	(0.417)	(0.418)	(0.418)	(0.417)	(0.463)	(0.463)	(0.462)	(0.462)	(0.462)	(0.461)
Post 2008	2.877	2.883	2.882	2.883	2.887	2.887	3.358	3.362	3.366	3.271	3.282	3.280
	(1.816)	(1.817)	(1.816)	(1.816)	(1.817)	(1.817)	(2.056)	(2.056)	(2.059)	(2.063)	(2.067)	(2.067)
NIH Grant × Post 2008	-2.685***	-2.681***	-2.675***	-2.688***	-2.685***	-2.678***	-3.068***	-3.065***	-3.058***	-2.860***	-2.856***	-2.849***
	(0.541)	(0.540)	(0.539)	(0.541)	(0.540)	(0.540)	(0.598)	(0.597)	(0.596)	(0.589)	(0.588)	(0.587)
Year	0.00163*	0.00163*	0.00163*	0.00163*	0.00164*	0.00164*	0.00189*	0.00189*	0.00189*	0.00199**	0.00199**	0.00199**
	(0.000879)	(0.000880)	(0.000879)	(0.000880)	(0.000880)	(0.000880)	(0.00100)	(0.00100)	(0.00100)	(0.00101)	(0.00101)	(0.00101)
NIH Grant × Year	-0.00125***	-0.00125***	-0.00125***	-0.00125***	-0.00125***	-0.00125***	-0.00137***	-0.00137***	-0.00136***	-0.00130***	-0.00130***	-0.00129***
	(0.000208)	(0.000208)	(0.000208)	(0.000208)	(0.000208)	(0.000208)	(0.000231)	(0.000231)	(0.000230)	(0.000230)	(0.000230)	(0.000230)
Post 2008 × Year	-0.00143	-0.00144	-0.00144	-0.00144	-0.00144	-0.00144	-0.00167	-0.00168	-0.00168	-0.00163	-0.00164	-0.00163
	(0.000908)	(0.000908)	(0.000908)	(0.000908)	(0.000908)	(0.000908)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)	(0.00103)
NIH Grant × Post 2008 × Year	0.00134***	0.00133***	0.00133***	0.00134***	0.00134***	0.00133***	0.00153***	0.00153***	0.00152***	0.00142***	0.00142***	0.00142***
	(0.000269)	(0.000269)	(0.000268)	(0.000269)	(0.000269)	(0.000269)	(0.000297)	(0.000297)	(0.000297)	(0.000293)	(0.000292)	(0.000292)
Number of Articles	7,952,412	7,950,614	7,950,614	7,952,405	7,950,607	7,950,607	7,409,425	7,407,769	7,407,769	7,273,155	7,271,545	7,271,545
Number of Journals	8,258	8,258	8,258	8,258	8,258	8,258	5,787	5,787	5,787	5,613	5,613	5,613
Top-Level MeSH		×			×			×			×	
Second-Level MesH			×			×			×			×
Article Standardized Cites				×	×	×	×	×	×	×	×	×
Journal Cites Per Document							×	×	×	×	×	×
Publisher Type										×	×	×

Notes–A FGLS estimator that assumes equicorrelation of the within-journal error terms was used to obtain these estimates. The standard errors are in parentheses and are clustered at the journal level. * indicates statistical significance at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level.

Table 8: The estimated impact of the NIH mandate on the average number of yearly citations to NIH-supported articles in open access and toll access journals.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Impact of mandate on quality of articles published in:									
Panel A: OLS									
Open access journals	-0.880*	-0.754*	-0.751*	-0.453	-0.411	-0.408	-0.446	-0.402	-0.400
	(0.451)	(0.440)	(0.419)	(0.406)	(0.407)	(0.398)	(0.405)	(0.406)	(0.397)
Toll access journals	0.550***	0.648***	0.615***	0.594***	0.590***	0.574***	0.592***	0.588***	0.572***
	(0.206)	(0.207)	(0.204)	(0.137)	(0.135)	(0.134)	(0.134)	(0.133)	(0.132)
Panel B: Journal FE									
Open access journals	-0.136	-0.158	-0.149	-0.280	-0.302	-0.293	-0.280	-0.302	-0.293
	(0.267)	(0.265)	(0.263)	(0.283)	(0.281)	(0.280)	(0.283)	(0.281)	(0.280)
Toll access journals	0.379***	0.373***	0.367***	0.404***	0.399***	0.392***	0.404***	0.399***	0.392***
	(0.103)	(0.103)	(0.103)	(0.0987)	(0.0982)	(0.0983)	(0.0987)	(0.0982)	(0.0983)
Panel C: Tobit									
Open access journals	-0.433	-0.316	-0.303	-0.107	-0.0565	-0.0477	-0.0326	0.0243	0.0321
	(0.355)	(0.341)	(0.321)	(0.296)	(0.296)	(0.285)	(0.281)	(0.281)	(0.270)
Toll access journals	0.519***	0.617***	0.599***	0.540***	0.564***	0.560***	0.519***	0.546***	0.541***
	(0.133)	(0.134)	(0.132)	(0.0821)	(0.0805)	(0.0801)	(0.0783)	(0.0772)	(0.0768)

Notes–The standard errors are in parentheses and are clustered at the journal level. * indicates statistical significance at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level. Our estimating sample contained 7,271,545 articles published in 5,613 journals between the years 1999 and 2012. The Tobit estimates in Panel C are the estimates of the average partial effect of the mandate on the *unconditional* expectation of average yearly citations.

Table 9: The estimated impact of the NIH mandate on the percentile rank of journals in which NIH-supported research is published.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
OLS Estimates	-2.401***	-1.164**	-1.276**	-2.275***	-1.140**	-1.225**	-2.401***	-2.401***	-2.401***
	(0.618)	(0.563)	(0.543)	(0.545)	(0.496)	(0.482)	(0.618)	(0.618)	(0.618)

Notes–The standard errors are in parentheses and are clustered at the journal level. * indicates statistical significance at the 10 percent level, ** at the 5 percent level, and *** at the 1 percent level. Our estimating sample contained 7,271,545 articles published in 5,613 journals between the years 1999 and 2012.

A Data Appendix

As noted in section 4, we use two main sources of data for the analysis presented in this paper: MEDLINE and the Directory of Open Access Journals (DOAJ). However, we also use a variety of other sources to obtain information on the articles and journals in our sample. These include Thomson Reuters Web of Science (WOS), Scimago, UlrichsWeb Serials Directory, the List of Serials Indexed for Online Users (LSIOU), the NLM Catalog, as well as Google Searches and e-mail correspondence. We first discuss article-level variables and then move on to journal-level variables.

A.1 Article-Level Variables

A.1.1 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.[36] We use the 2014 baseline files.[37] These are distributed by the NLM as 746 compressed Extensible Markup Language (XML) files.[38]

We use a series of Perl scripts to extract data from the XML files and place it into tab-delimited text files.[39] The article-level elements that we extract are:[40] Status attribute, PMID (and the Version attribute), PubDate, ArticleDate, MedlineDate, PublicationTypeList, GrantList, and MeshHeadingList.

The top-level element for each record (article) in the MEDLINE XML files is MedlineCitation. This element has four attributes, but we are only interested in the Status attribute. This attribute indicates how thoroughly the record's information has been vetted. We only use records with the Status "MEDLINE" as these have undergone the most rigorous quality review and are the only true MEDLINE records. It is worth noting that when we include other Status types, none of our results change.

The PMID, or PubMed ID, is a unique identifier for every record in the MEDLINE database. 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.[41] The combination of the PMID and Version are crucial for linking various types of article-level data, both from MEDLINE and from Thomson Reuters Web of Science (see below).

MEDLINE has three date types that can be used 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 attribute. When dates do not follow this standard format, they are found in the element MedlineDate. For these non-standard dates, we manually coded the year. In some cases, there was a year range instead of a single year.

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

[37] http://www.nlm.nih.gov/bsd/licensee/2014_stats/baseline_doc.html

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

[39] We use the XML::Simple module from the Comprehensive Perl Archive Network (CPAN).

[40] For a description of all elements in MEDLINE, see: http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html#nl

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

For these cases, we took the first year in the range as the publication year. The element ArticleDate contains the date that a publisher first published an electronic version of an article. ArticleDate always follows a standard dating format, making it easy to identify the Year attribute. Often, the date information in the PubDate and 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. Since the NIH mandate requires all NIH-supported articles to be available in PubMed Central within 12 months of publication (regardless of format), we 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 elements.

Each MEDLINE record contains a GrantList element, which contains an element called Agency, which contains information on the organizations that funded the research contained in the record (article). We used this Agency element to identify which records were supported by a NIH grant. We also used it to determine whether each record was supported by other grant types.[42]. In some cases, the GrantList element does not contain all grants listed on the article for which the record was made. This is indicated by the attribute CompleteYN. Fortunately, only 0.4 percent of grant lists are incomplete.

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 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" may qualify the descriptor. We use only descriptors that are major topics to classify records by field. MeSH terms are hirearchical and get be very detailed. Thus, we confine ourselves to top-level (16 terms) and second-level (117 terms) MeSH terms. We use MeSH terms to control for the fields in which the article was published. Specifically, we generate a set of dummy variables that indicate whether an article contained MeSH terms that were classified under one of the top-level or second-level MeSH terms. Note that it is perfectly possible for a single article to be classified under more than one top-level or second-level MeSH term.

A.1.2 Web of Science

Thomson Reuters Web of Science (WOS) is a citation indexing database. Indeed, it is the most widely used source of citation data.[43] For each article that is indexed in both MEDLINE and WOS, we obtained the total number of citations received over the lifetime of the article (through mid-2014). For instance, if an article was published in 2006, we observe the number of citations the article received between 2006 and 2014. Since we do not observe

[42]For a list of all funding agencies indexed in MEDLINE, see: http://www.nlm.nih.gov/bsd/grant$_{a}$*cronym.html*

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

citation counts by year (e.g. the number of citations received in 2006, the number received in 2007, and so on) we are unable to construct measures such as 2-year forward citations. Since newer articles will mechanically receive fewer citations than older articles, we need a way to normalize the total number of citations. We use two methods. First, we compute the average number of citations per year. For instance, if an article published in 2006 received 206 citations between 2006 and 2014, we compute that its average number of citations per year is 206/(2014-2006)=25.75. Second, we standardize total citations by subtracting, from each article's total number of citations, the mean number of total citations received by other articles sharing the same publication year and then dividing by the standard deviation of the total number of citations received by other articles with the same publication year. For instance, if an article published in 2006 received 206 citations between 2006 and 2014, and all of the articles published in 2006 had a mean of 22.2 and a standard deviation of 48.9 citations between 2006 and 2014, the standardized citations for the single article is (206-22.2)/48.9=3.76. Fortunately, WOS includes the PMID of articles that are also indexed in MEDLINE. This enables us to link the two data sources.

A.2 Journal-Level Variables

A.2.1 MEDLINE

We discussed the general characteristics of MEDLINE and the article-level elements that we extracted from MEDLINE in section A.1.1. As with the article-level elements, we use a series of Perl scripts to extract journal-level data from the XML files and place it into tab-delimited text files. The journal-level elements that we extract are: NlmUniqueID, ISSN, and ISSNLinking.

The NlmUniqueID is a unique identifier of each journal in MEDLINE. It is crucial for linking jounal-level information within MEDLINE and other NLM sources such as the NLM-Catalog and the List of Serials Indexed for Online Users (see below). Unfortunately, other sources of data, such as DOAJ, Scimago, and UlrichsWeb, do not use the NlmUniqueID. Instead, they use the International Stanard Serial Number (ISSN) to identify journals. Thus, to link journal-level information in MEDLINE to these other data sources, we need to use the ISSN. Note that, in addition to using the NlmUniqueID as a linking variable, we also use it to estimate journal fixed effects and to cluster the standard errors at the journal level.

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. This causes some complications because most data sources organize data by journal titles, and assign each journal title a single ISSN without indicating whether it is the print or electronic format. Fortunately, MEDLINE, the NLM Catalog, and LSIOU typically include all formats, which allows us 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 us to uniquely identify journals with multiple ISSNs.

A.2.2 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 an independent organization. The database can be downloaded by anyone, free of charge.[44] The database is updated daily, and past versions are not readily available. We downloaded the file on July 5, 2014, and will make it available upon request. The database is distributed as a CSV (comma-separated) file. We use journals' International Standard Serial Number (ISSN) to match DOAJ data to the MEDLINE data. Crucially, DOAJ also includes the year in which journals started publishing open access content.[45]

Two other sources of data for determining whether a journal is open access are UlrichsWeb Serials Directory (see below) and the PubMed Central Journal List.[46] However, these sources were not suitable for our analysis because they only indicate whether a journal is *currently* open access. They contain no information on the dates that journals became or ceased being open access.

A.2.3 Scimago

Scimago is "a research group from the Consejo Superior de Investigaciones Cientficas (CSIC), University of Granada, Extremadura, Carlos III (Madrid) and Alcal de Henares."[47] They use data from Elsevier's Scopus database to compute various measures of journal-level and country-level scientific prestige. Their signature measure of journal quality is the Scimago Journal Rank (SJR) indicator (Guerrero-Bote and Moya-Anegon, 2012), which is a free alternative to Thompson Reuters' Impact Factor. They also compute measures of journal quality such as citations per document over a two year period[48] and the Hirsch index[49]. We connect these measures of journal-quality to MEDLINE data using journals' ISSN.

A.2.4 Commercial and Non-Profit Journals

For each journal indexed in MEDLINE, we attempted to obtain publisher information from UlrichsWeb Serials Directory, which is maintained by ProQuest. The process was fairly tedious. We had to search for each journal individually, save it to a list in UlrichsWeb, and then download the list when it reached 500 journals.[50] Our first step was to translate non-English publisher names into English. This was especially important when the publisher was an academy or government department. Most of the translation was done using Google

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

[45] See http://doaj.org/faqsearchresults for the fields contained in the DOAJ data file.

[46] The data from PubMed Central can be viewed and downloaded here: http://www.ncbi.nlm.nih.gov/pmc/journals/.

[47] http://www.scimagojr.com/aboutus.php

[48] This is defined as "Average citations per document in a 2 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the two previous years, –i.e. citations received in year X to documents published in years X-1 and X-2."

[49] This is defined as "The h index expresses the journal's number of articles (h) that have received at least h citations"

[50] UlrichsWeb only allows 500 downloads at a time.

Translate. However, when a translation was ambiguous, we consulted with several colleagues who were fluent in Chinese, Japanese, Korean, Russian, Turkish, and Slovak.[51]

Our next step was to standardize the names of the publishers. Often publishers use slightly different versions of the same name. For instance, Lippincott Williams & Wilkins sometimes goes by the acronym LWW. We used mostly Google searches to standardize the names.[52] Finally, we manually determine whether each publisher is a commercial (for-profit) or a non-profit publisher. We assumed that all academic, academy, and government publishers were non-profit. We then exhuastively went through the list and detrmined whether each publisher was commercial or non-profit. This was done mostly using Google searches. However, when their was ambiguity, we e-mailed publishers. Overall the response rate to these e-mails was quite high. When a publisher was from a non-English speaking country, we sent both an English and native language version of the e-mail.

B Bootstrapping Break Years

To determine the precision of our break year estimator in equation (3) in section 5.2, we use a simple non-parametric bootstrap procedure. Specifically, we repeat the following three steps 1,000 times:

1. Draw a simple random sample (with replacement) from the estimation sample of 7,271,545 articles.

2. For each of the 12 specifications in table 2, estimate equation (1) in section 3.2 using the FGLS estimator discussed in section section 3.3:

$$oa_{ajt} = \beta_t nihgrant_{ajt} + \delta_t controls_{ajt} + \nu_j + \alpha_t + \varepsilon_{ajt}.$$

   Collect the $\hat{\beta}_t$'s (the estimated β_t's) for the years $t = 1999, ..., 2012$.

3. For each year $\bar{t} = 1999...2012$, estimate the regression equation (3) from section 5.2 using OLS:

$$\hat{\beta}_t = \theta_1 + \theta_2 I(t \leq \bar{t}) + \pi_1 t + \pi_2 t * I(t \leq \bar{t}) + \varepsilon_t.$$

   Define $\bar{t}^{break}$ as the $\bar{t}$ that minimizes the sum of squared residuals. Call this the "break year".

At the end of every bootstrap repetition, we save $\bar{t}^{break}$, giving us a total of 1,000 break years for each specification. We use these to estimate the precision of our break year estimator.

C Details on NIH Public Access Policy and NLM Journal Selection

[51] A list of translations will be made available upon request.

[52] A list of all changes to publisher names will be made available upon request.